# COMPUTING IMPLICITIZATIONS OF MULTI-GRADED POLYNOMIAL MAPS

#### JOSEPH CUMMINGS AND BENJAMIN HOLLERING

ABSTRACT. In this paper, we focus on computing the kernel of a map of polynomial rings  $\varphi$ . This core problem in symbolic computation is known as implicitization. While there are extremely effective Gröbner basis methods used to solve this problem, these methods can become infeasible as the number of variables increases. In the case when the map  $\varphi$  is multigraded, we consider an alternative approach. We demonstrate how to quickly compute a matrix of maximal rank for which  $\varphi$  has a positive multigrading. Then in each graded component we compute the minimal generators of the kernel in that multidegree with linear algebra. We have implemented our techniques in Macaulay2 and show that our implementation can compute many generators of low degree in examples where Gröbner techniques have failed. This includes several examples coming from phylogenetics where even a complete list of quadrics and cubics were unknown. When the multigrading refines total degree, our algorithm is embarassingly parallel and a fully parallelized version of our algorithm will be forthcoming in OSCAR.

#### 1. Introduction

Implicitization is a core problem in symbolic computations with many applications in a variety of scientific fields. This problem is focused on computing the kernel of a ring homomorphism  $\varphi$ 

$$\varphi: R = \mathbb{C}[x_1, \dots, x_n] \to S = \mathbb{C}[t_1, \dots, t_m]$$
  
 $x_i \mapsto \varphi_i(x_i)$ 

This means one seeks to find a Gröbner basis or even just a generating set for the ideal  $\ker(\varphi)$ . Standard techniques for this typically rely on computing a Gröbner basis for the elimination ideal  $\langle x_i - \varphi_i(x_i) \rangle$  with respect to an elimination order for the variables  $t_j$  [8]. While modern Gröbner bases algorithms are extremely effective at solving a wide array problems, they still often become too expensive as the number of variables and the degree of the polynomials involved grows and are difficult to parallelize effectively [5, 14, 15].

In this paper we focus on computing the kernel of polynomial maps which arise in algebraic statistics though our techniques apply more broadly. Many problems in algebraic statistics are fundamentally implicitization problems; however, for many families of interesting statistical models the number of variables involved grows exponentially. For instance, to compute the ideal of phylogenetic invariants for an n-leaf phylogenetic tree or network, there are  $\approx 4^n$  variables involved [3]. This means that it is often impossible to compute polynomials in  $\ker(\varphi)$  for small trees or networks with a computer algebra system. In many algebraic statistics problems one may only need to find a single polynomial in  $\ker(\varphi)$  to prove identifiability results [1, 18, 24] or a collection of statistically meaningful polynomials which can be used for model selection [7, 12, 32]. Even in these cases where only some polynomials

in the  $\ker(\varphi)$  are needed, modern Gröbner bases algorithms which leverage homogeneity and degree-limiting may still fail to compute low-degree polynomials for small examples since there are so many variables involved [25]. For a thorough (and rather enjoyable) treatment of the Gröbner-based approach in the multi-graded setting, we refer the reader to [23].

In this paper we provide an alternative algorithm to the common Gröbner-based approach which exploits the fact that many polynomial maps in algebraic statistics are actually homogeneous in a  $\mathbb{Z}^k$ -multigrading. Our approach is inspired by the technique the authors used in [10] to compute the quadratic polynomials which vanish on certain phylogenetic network models as well as [9] where the authors study multigraded Macaulay dual spaces. In the following section we show how an essentially maximal multigrading in which  $\ker(\varphi)$  is homogeneous can be computed without computing  $\ker(\varphi)$ . We then describe how the generators in  $\ker(\varphi)$  which have a given multidegree  $\beta \in \mathbb{Z}^k$  can be computed by solving large linear systems. This means that if  $\varphi$  is also homogeneous in the usual sense of total degree, then one compute all generators of total degree d by computing the homogeneous component of  $\ker(\varphi)$  with multidegree  $\beta \in \mathbb{Z}^k$  for all  $\beta$  which are the multidegree of a monomial of total degree d. Moreover, this step is embarassingly parallel meaning that the computation of each homogeneous component corresponding to  $\beta \in \mathbb{Z}^k$  can be computed completely in parallel. This makes our algorithm extremely effective at computing all of the low-degree polynomials in the kernel of a polynomial map which is homogeneous in a large multigrading.

The basic idea behind this technique has been noted before in the case that  $\varphi$  is homogeneous with respect to the usual Z-grading given by total degree. However, for many large examples, this technique fails since the linear systems which one needs to solve grow exponentially in the number of variables, which in algebraic statistics often grows exponentially itself. By leveraging multigradings, we are able to instead solve many smaller systems completely in parallel. In our last section we showcase this technique on several examples from algebraic statistics and phylogenetics which Gröbner bases techniques or the previously known total-degree version of this algorithm are unable to solve. This includes finding all degree 2 and degree 3 phylogenetic invariants for 4 leaf networks under the Kimura 3-Parameter model. Recently, [25] attempted this same computation with degree-limited Gröbner bases and were unable to find these degree 3 generators even after 100 days of computation time. Another model of interest is the Timura-Nei model [22]. This model is more flexible than group-based models and is used more widely in practice. While the vanishing ideal for a generic tree is still currently unknown, the authors in [6] showed that on an open subset, the ideal for a 4 leaf tree is a complete intersection of dimension 16, and they explicitly produce the ideal. Using our methods, we were able to show that the full ideal is not a complete intersection by exhibiting that there are 375 minimal quadrics in the vanishing ideal.

All of our code along with detailed explanations can be found on our MathRepo page

https://mathrepo.mis.mpg.de/MultigradedImplicitization

or in the GitHub repository

https://github.com/bkholler/MultigradedImplicitization.

This includes a Macaulay2 [17] package with our main algorithm implemented. A fully parallelized version of our algorithm will also be available in OSCAR [28] as soon as this functionality is supported.

The remainder of this paper is organized as follows. In Section 2 we show how multigradings on polynomial maps can be computed and then describe our algorithm which computes  $\ker(\varphi)$  up to a given degree. In Section 3 we examine several different applications of our algorithm to open implicitization problems in phylogenetics which are known to be difficult and show that many low degree polynomials can be found with our algorithm.

## 2. The Main Algorithm

In this section we show how to compute the homogeneity space of the kernel of a polynomial map  $\varphi$  without actually computing  $\ker(\varphi)$ . The homogeneity space of any ideal in a polynomial ring induces a maximal  $\mathbb{Z}^k$ -multigrading in which the ideal is homogeneous. We then leverage this multigrading to give an *embarrassingly parallelizable* algorithm for computing homogeneous components of  $\ker(\varphi)$  which is powered by solving large linear systems. Our notation throughout this section is adapted from [21].

**Definition 2.1.** Let  $f = \sum_{\alpha} c_{\alpha} x^{\alpha} \in \mathbb{K}[x_1, \dots, x_n]$  and let  $\omega \in \mathbb{R}^n$ . Then the *initial form* of f with respect to  $\omega$  is

$$\operatorname{in}_{\omega}(f) = \sum_{\substack{c_{\alpha} \neq 0 \\ \omega \cdot \alpha \text{ is minimal}}} c_{\alpha} x^{\alpha}.$$

For an ideal  $I \subseteq \mathbb{K}[x_1,\ldots,x_n]$  and  $\omega \in \mathbb{R}^n$ , the *initial ideal* of I with respect to  $\omega$  is  $\operatorname{in}_{\omega}(I) = \langle \operatorname{in}_{\omega}(f) \mid f \in I \rangle$ .

**Definition 2.2.** Let  $I \subseteq \mathbb{K}[x_1, \dots, x_n]$  be an ideal. The *homogeneity space* of I is the linear space

$$C_0(I) = \{ \omega \in \mathbb{R}^n \mid \operatorname{in}_{\omega}(I) = I \}.$$

The homogeneity space  $C_0(I)$  consists of vectors which yield a grading in which I is homogeneous. Indeed if  $\omega \in C_0(I)$  and we set  $\deg(x_i) = \omega_i$ , then by definition  $\operatorname{in}_{\omega}(f)$  is homogeneous with respect to this grading. Since  $\operatorname{in}_{\omega}(I) = I$ , we have that I is generated by homogeneous polynomials, i.e. I has a (possibly non-standard)  $\mathbb{Z}$ -grading given by  $\omega$ . Now, let  $b_1, \ldots, b_r \in \mathbb{Z}^n$  be a basis for  $C_0(I)$  and consider the matrix  $A = (b_1|b_2|\ldots|b_r)^T \in \mathbb{Z}^{r \times n}$ . Then I is homogeneous in the  $\mathbb{Z}^r$ -multigrading given by  $\deg(x_j) = A_{x_j}$  where  $A_{x_j} \in \mathbb{Z}^r$  is the j-th column of A which naturally corresponds to  $x_j$ . Moreover, this multigrading is maximal in the sense of the following lemma.

**Lemma 2.3.** Let A be as above and suppose  $A' \in \mathbb{Z}^{k \times n}$  is another matrix for which I is homogeneous in the multigrading induced by A'. Then the row space of A' is contained in the row space of A. Note that k need not be equal to r.

*Proof.* Suppose A' is equal to  $(b'_1|b'_2|\dots|b'_k)^T \in \mathbb{Z}^{k\times n}$ . It is enough to show that  $b'_i$  is in  $C_0(I)$  for  $i=1,\dots,k$ . To this end, consider any homogeneous element  $f\in I$  of degree  $\beta\in\mathbb{Z}^k$ . Then f is of the form

$$f = \sum_{A'\alpha = \beta} c_{\alpha} x^{\alpha}.$$

Then for any i, we have that

$$\operatorname{in}_{b_i'}(f) = \sum_{\substack{A'\alpha = \beta \\ b_i' \cdot \alpha \text{ is minimal} \\ 3}} c_{\alpha} x^{\alpha}$$

As f is homogeneous,  $b'_i \cdot \alpha = \beta_i$  for all  $\alpha$  appearing in f, so it follows that  $\operatorname{in}_{b'_i}(f) = f$  for all such homogeneous polynomials. In particular, we conclude that  $\operatorname{in}_{b'_i}(I) = I$  and  $b'_i \in C_0(I)$  completing the proof.

We now focus on the problem of finding the homogeneity space of the kernel of a polynomial map. So consider ring homomorphism  $\varphi$  of the form

$$\varphi: R = \mathbb{K}[x_1, \dots, x_n] \to S = \mathbb{K}[t_1, \dots, t_m]$$
  
 $x_i \mapsto \varphi(x_i)$ 

The following theorem and lemma gives an immediate technique to partially compute the homogeneity space of  $\ker(\varphi)$ .

**Lemma 2.4.** Let  $\varphi : R \to S$  be as above, and let  $J = \langle x_i - \varphi(x_i) \mid i \in [n] \rangle$  be the elimination ideal. Then

$$C_0(J) = \{ \omega \in \mathbb{R}^n \mid \text{in}_{\omega}(x_i - \varphi(x_i)) = x_i - \varphi(x_i) \text{ for all } i \in [n] \}.$$

*Proof.* One inclusion is obvious. If  $\operatorname{in}_{\omega}(x_i - \varphi(x_i)) = x_i - \varphi(x_i)$  for all i, then  $\omega \in C_0(J)$ . Indeed, if this is the case, then we automatically have that  $\operatorname{in}_{\omega}(J) \supseteq J$ . On the other hand, if f is in J, then  $f = \sum_i h_i(x_i - \varphi(x_i))$  and  $\operatorname{in}_{\omega}(f) = \sum_i \operatorname{in}_{\omega}(h_i)(x_i - \varphi(x_i))$ , so  $\operatorname{in}_{\omega}(f) \in J$ . As  $\operatorname{in}_{\omega}(J)$  is generated by all such initial forms we have  $\operatorname{in}_{\omega}(J) = J$ .

Now, let  $\omega \in C_0(J)$ . Fix an  $i \in [n]$  and consider  $x_i - \varphi(x_i)$ . Since J is homogeneous with respect to  $\omega$ , we could rewrite  $x_i - \varphi(x_i)$  as  $\sum_{i=1}^N f_i$  where each  $f_i$  satisfies  $\operatorname{in}_{\omega}(f_i) = f_i$  and  $f_i \in J$  for all  $i \in [N]$ . We can assume that  $f_1$  has  $x_i$  as one of its terms, so  $f_1$  takes the form  $x_i - \sum_{\alpha} c_{\alpha} t^{\alpha}$ . The fact that  $f_1$  is in the ideal implies that  $\sum_{j=2}^N f_j = \varphi(x_i) - (f_1 - x_i)$  lies in  $J \cap S$ . Since  $x_1, \ldots, x_n$  are algebraically independent,  $J \cap S = \{0\}$ . Therfore we must have  $f_j = 0$  for each  $j \geq 2$ . It follows that  $\operatorname{in}_{\omega}(x_i - \varphi(x_i)) = x_i - \varphi(x_i)$  for all  $i \in [n]$  as claimed.

**Theorem 2.5.** Let  $\varphi: R \to S$  be a homomorphism of polynomial rings as above and  $J = \langle x_i - \varphi(x_i) \mid i \in [n] \rangle \subseteq \mathbb{K}[x_1, \dots, x_n, t_1, \dots t_m]$  be the associated elimination ideal. Let  $b_1, \dots, b_r \in \mathbb{Z}^{n+m}$  be a basis for the homogeneity space  $C_0(J)$  and write  $b_i = (b_i', b_i'') \in \mathbb{Z}^{n+m}$ . Then  $b_1', \dots, b_r' \in \mathbb{Z}^n$  are contained in the homogeneity space of  $\ker(\varphi)$ .

Proof. Let  $i \in [n]$ . We will show that  $b'_i \in C_0(\ker(\varphi))$ . By Lemma 2.4 the generators of J, i.e.  $x_j - \varphi(x_j)$ , are all homogeneous with respect to the  $\mathbb{Z}$ -grading given by  $b_i$ . Now fix a lexicographic term ordering where  $t_i \succ x_j$  for all  $(i,j) \in [m] \times [n]$ . In order to compute  $\ker(\varphi)$ , we need to compute a Gröbner basis with respect to  $\prec$ . Our key insight is that each step in Buchberger's algorithm always adds homogeneous (with respect to  $b_i$ ) polynomials to the generating set. This is because an S-pair of two homogeneous polynomials is also homogeneous and the reduction of an S-pair by homogeneous polynomials will also be homogeneous.

Now let  $\mathcal{G}$  be the resulting Gröbner basis and let  $\mathcal{G}' = \mathcal{G} \cap R$ . As discussed above, each element of  $\mathcal{G}'$  is homogeneous with respect to  $b_i$ , but as they involve no  $t_j$ 's, they are also homogeneous with respect to  $b'_i$ . It follows that  $\operatorname{in}_{b'_i}(\mathcal{G}') = \mathcal{G}'$  and that  $\operatorname{in}_{b'_i}(\ker(\varphi)) = \ker(\varphi)$ .

**Remark 2.6.** It is important to note that the elements in the homogeneity space of  $\ker(\varphi)$  obtained from Theorem 2.5 are generally not independent nor spanning. We will see in Example 2.8, that the rows of A are not independent. For an example where they aren't spanning, consider the following map.

$$\varphi : \mathbb{K}[x, y, z] \to \mathbb{K}[a, b]$$
$$x \mapsto (a + b)^{2}$$
$$y \mapsto a^{2} - b^{2}$$
$$z \mapsto (a - b)^{2}$$

The kernel of this map is the toric ideal  $\langle xz - y^2 \rangle$ ; however, using Theorem 2.5, we only detect a 1-dimensional subspace of the 2-dimensional homogeneity space.

Corollary 2.7. Let  $\varphi: R \to S$  be a homomorphism of polynomial rings, J be the associated elimination ideal, and  $b_i = (b'_i, b''_i) \in \mathbb{Z}^{n+m}$ ,  $i \in [k]$  be a basis for  $C_0(J)$ . Let  $A = (b_1|b_2|\dots|b_r)^T \in \mathbb{Z}^{r \times (n+m)}$ . Then  $\varphi$  is homogeneous in the multigrading given by  $\deg(t_j) = A_{t_j}$  and  $\deg(x_j) = A_{x_j}$ . Moreover,  $\ker(\varphi)$  is homogeneous in the induced multigrading which is  $\deg(x_j) = A_{x_j}$ .

Note that in the previous corollary we make the natural identification between the columns of the matrix A with the corresponding variables in the polynomial rings R and S for simplicity of notation. The following example illustrates this corollary.

**Example 2.8.** Consider the Plücker embedding of Gr(2,4). Set  $R = \mathbb{C}[p_{ij} \mid 1 \leq i < j \leq 4]$  and  $S = \mathbb{C}[x_{ij} \mid i \in [2], j \in [4]]$ .

$$\varphi: R \to S$$
$$p_{ij} \mapsto \det(M_{ij})$$

Here M is a  $2 \times 5$  matrix whose entries are the variables  $x_{ij}$  and  $M_{ij}$  corresponds to the square  $2 \times 2$  sub-matrix of M whose columns are the  $i^{\text{th}}$  and  $j^{\text{th}}$  columns of M. If we let J be the elimination ideal of  $\varphi$ , the homogeneity space  $C_0(J) \subset \mathbb{R}^5$  is the rowspan of the following matrix.

The lattice spanned by the first 6 columns of the matrix above has rank 4, so we deduce that there is an action of  $(\mathbb{C}^{\times})^4$  on the affine cone of Gr(2,4). The row of ones corresponds to the usual scaling action on  $\mathbb{C}^6$  used to construct  $\mathbb{P}^5 \cong \mathbb{C}^6/\mathbb{C}^{\times}$ , so there is a 3 dimensional torus  $(\mathbb{C}^{\times})^4/\mathbb{C}^{\times}$  acting on  $Gr(2,4) \subseteq \mathbb{P}^5$ .

So given any polynomial map  $\varphi$ , Corollary 2.7 allows one to inexpensively compute a multigrading in which  $\ker(\varphi)$  is homogeneous. We now focus on the task of computing

minimal generators of  $\ker(\varphi)$  of a fixed total degree d. One naive way of doing this is to consider an arbitrary element

$$f^{(d)} = \sum_{\substack{\alpha \\ \alpha : 1 = d}} c_{\alpha} x^{\alpha}$$

of total degree d. Then of course we know that  $f \in \ker(\varphi)$  if and only if  $\varphi(f) = 0$ . Observe that we can simply compute  $\varphi(f)$ , collect the coefficients of each monomial  $t^{\gamma}$ , and set each of these to 0. This gives us necessary and sufficient linear conditions on the coefficients  $c_{\alpha}$ . Computing a basis for the set of all such  $c_{\alpha}$  then gives us a set of minimal generators of degree d for  $\ker(\varphi)$ . This is demonstrated by the following example.

**Example 2.9.** We continue with Example 2.8 by finding the quadrics in  $I = \ker(\varphi)$ . Consider a generic quadric in R.

$$f^{(2)} = c_{(1,2),(1,2)}p_{1,2}^2 + c_{(1,2),(1,3)}p_{1,2}p_{1,3} + c_{(1,3),(1,3)}p_{1,3}^2 + c_{(1,2),(2,3)}p_{1,2}p_{2,3} + c_{(1,3),(2,3)}p_{1,3}p_{2,3} + c_{(2,3),(2,3)}p_{2,3}^2 + c_{(1,2),(1,4)}p_{1,2}p_{1,4} + c_{(1,3),(1,4)}p_{1,3}p_{1,4} + c_{(2,3),(1,4)}p_{2,3}p_{1,4} + c_{(1,4),(1,4)}p_{1,4}^2 + c_{(1,2),(2,4)}p_{1,2}p_{2,4} + c_{(1,3),(2,4)}p_{1,3}p_{2,4} + c_{(2,3),(2,4)}p_{2,3}p_{2,4} + c_{(1,4),(2,4)}p_{1,4}p_{2,4} + c_{(2,4),(2,4)}p_{2,4}^2 + c_{(1,2),(3,4)}p_{1,2}p_{3,4} + c_{(1,3),(3,4)}p_{1,3}p_{3,4} + c_{(2,3),(3,4)}p_{2,3}p_{3,4} + c_{(1,4),(3,4)}p_{1,4}p_{3,4} + c_{(2,4),(3,4)}p_{2,4}p_{3,4} + c_{(3,4),(3,4)}p_{3,4}^2$$

As stated above, we can apply  $\varphi$  to  $f^{(2)}$ , collect coefficients, and get necessary and sufficient linear conditions on the  $c_{(i,j),(k,\ell)}$ 's to find a basis for the kernel of  $\varphi$  in degree 2. This is more than a little cumbersome, so we will forego showing you this computation explicitly; however, we will describe how to implement this in your favorite computer algebra system.

There are  $\binom{6+2-1}{2} = 21$  monomials spanning  $R_2$  and the images of each monomial are supported on 72 monomials of degree 4 in S. In order to find a basis for  $\ker(\varphi)$  in degree 2, we need to find the *linear* relations among the polynomials  $\varphi(p_{ij}p_{k\ell})$ . This amounts to finding the kernel of a 72 × 21 matrix C. The columns of this matrix are indexed by the monomials spanning  $R_2$  and the rows are indexed by the monomials in  $S_4$  on which  $\varphi(p_{ij}p_{k\ell})$  are supported. The entry  $C_{x^{\alpha},p_{ij}p_{k\ell}}$  is the coefficient of  $x^{\alpha}$  in  $\varphi(p_{ij}p_{k\ell})$ . The kernel of C is generated by exactly one element and it corresponds to the Plücker relation  $p_{2,3}p_{1,4} - p_{1,3}p_{2,4} + p_{1,2}p_{3,4}$ .

While the previous approach can be used occasionally it is often not helpful since the generic polynomial f which one has to consider has  $\binom{n+d-1}{d}$  terms which grows exponentially in d. However, if we instead apply this technique to each homogeneous component of a finer multigrading, we can solve much smaller linear systems instead. So let  $A \in \mathbb{Z}^{r \times n}$  be a maximal rank multigrading on  $\ker(\varphi)$  and assume that the  $\mathbb{1} \in \operatorname{rowspan}(A)$  which guarantees that  $\ker(\varphi)$  is also homogeneous in the typical sense of total degree. This implies that  $\ker(\varphi)$  has a minimal generating set  $\{f_1, \ldots f_k\}$  consisting of polynomials where each minimal generator  $f_i$  is homogeneous in the multigrading determined by A. So if we wish to compute a set of minimal generators of  $\ker(\varphi)$  which are total degree d, then we can instead consider each multidegree  $A\alpha = \beta \in \mathbb{N}A$  such that  $\alpha \cdot \mathbb{1} = d$  separately. This means to compute all degree d minimal generators, we no longer consider a polynomial of the form

found in Eq. (2.1) but instead we consider a polynomial of the form

$$(2.2) f^{(\beta)} = \sum_{\substack{\alpha \\ A\alpha = \beta}} c_{\alpha} x^{\alpha}$$

for each homogeneous component of degree  $\beta \in \mathbb{N}A$  such that  $\beta = A\alpha$  and  $\alpha \cdot \mathbb{1} = d$ . Generally, as the multigrading A becomes finer, meaning rank(A) becomes larger, the size of the monomial basis  $M_{\beta} = \{\alpha \in \mathbb{N}^n \mid A\alpha = \beta\}$  for the homogeneous component  $R_{\beta} = \{f \in R \mid \deg(f) = \beta\}$  becomes smaller. This means instead of solving one extremely large linear system which corresponds to  $\varphi(f^{(d)}) = 0$  as we would get from Eq. (2.1), we can solve many small linear systems which come from settings  $\varphi(f^{(\beta)}) = 0$  for each  $\beta$ . The following example elucidates this.

**Example 2.10.** We continue with our running example of Gr(2,4). As we saw in Example 2.9. There was a single quadratic Plücker relation. However, we had to compute the kernel of a  $72 \times 21$  matrix. Using the multigrading from Example 2.8, we can greatly reduce the size of this computation. The total degree 2 component of R can be divided into 19 homogeneous components using the grading matrix from Example 2.8. Only one of these homogeneous components has a basis with more than a single element. This is  $R_{(2,1,1,1,-1)}$  and is spanned by  $M_{(2,1,1,1,-1)} = \{p_{1,2}p_{3,4}, p_{1,3}p_{2,4}, p_{2,3}p_{1,4}\}$ . There will be no relations supported on the other components since there are evidently no monomials in  $\ker(\varphi)$ . Consider a generic polynomial of degree (2,1,1,1,-1).

$$f^{(2,1,1,1,-1)} = c_1 p_{1,2} p_{3,4} + c_2 p_{1,3} p_{2,4} + c_3 p_{2,3} p_{1,4}.$$

Now, we can find the necessary and sufficient linear relations among the  $c_i$ 's to ensure  $f^{(2,1,1,1,-1)}$  is in  $\ker(\varphi)$ . This can be done as follows.

There are 6 monomials that appear when we apply  $\varphi$  to the monomial basis of  $R_{(2,1,1,1,-1)}$ , so we construct a  $6 \times 3$  matrix whose columns are indexed by the elements of  $M_{(2,1,1,1,-1)}$  and whose rows are indexed by these 6 monomials. The entry in row  $x^{\alpha}$  and column  $p_{i,j}p_{k,\ell}$  is the coefficient of  $x^{\alpha}$  in  $\varphi(p_{i,j}p_{k,\ell})$ .

$$\begin{array}{c} p_{1,2}p_{3,4} & p_{1,3}p_{2,4} & p_{2,3}p_{1,4} \\ x_{1,3}x_{1,4}x_{2,1}x_{2,2} & 0 & 1 & 1 \\ x_{1,2}x_{1,4}x_{2,1}x_{2,3} & 1 & 0 & -1 \\ x_{1,1}x_{1,4}x_{2,2}x_{2,3} & -1 & -1 & 0 \\ x_{1,2}x_{1,3}x_{2,1}x_{2,4} & -1 & -1 & 0 \\ x_{1,1}x_{1,3}x_{2,2}x_{2,4} & 1 & 0 & -1 \\ x_{1,1}x_{1,2}x_{2,3}x_{2,4} & 0 & 1 & 1 \end{array} \right)$$

The kernel of this matrix is spanned by  $(1, -1, 1)^T$  giving us the Plücker relation  $p_{1,2}p_{3,4} - p_{1,3}p_{2,4} + p_{2,3}p_{1,4}$ .

This idea gives an immediate algorithm for computing all of the minimal generators of  $\ker(\varphi)$  of degree at most d which can be found at the end of this section. First, we note that while building the set of minimal generators in degree d, one may further reduce the set  $M_{\beta}$  for each  $\beta$  such that  $\beta \cdot \mathbb{1} = d$  using the set of generators of degree strictly less than d. This idea is captured by the following proposition and example.

**Proposition 2.11.** Suppose H is a minimal homogeneous generating set for  $\ker(\varphi)$ . Let d be a positive integer, let  $G = \{g \in H \mid \deg(g) = \beta_g = A\alpha_g \text{ with } \alpha_g \cdot \mathbb{1} < d\}$  and let  $\beta = A\alpha$  where  $\alpha \cdot \mathbb{1} = d$ . Consider the vector space

$$\operatorname{Lift}(G) = \operatorname{span}_{\mathbb{K}} \{ x^{\gamma} g \mid \deg(x^{\gamma}) = \beta - \deg(g) \text{ and } g \in G \} \subseteq R_{\beta}.$$

We can write  $R_{\beta}$  as a direct sum  $\text{Lift}(G) \oplus V_{\beta}$ . Then the minimal generators of  $\ker(\varphi)$  of degree  $\beta$  can be chosen to be supported on  $V_{\beta}$ .

Proof. Suppose f is a minimal generator of degree  $\beta$ . This polynomial can be rewritten as  $f_G + f_{V_\beta}$  where  $f_G \in \text{Lift}(G)$  and  $f_{V_\beta} \in V_\beta$ . By definition,  $f_G$  is in the ideal generated by G; hence,  $f_G \in \text{ker}(\varphi)$ . It follows that  $f_{V_\beta} \in \text{ker}(\varphi)$ , and this polynomial can be chosen as a minimal generator instead of f.

**Example 2.12.** We illustrate Proposition 2.11 by continuing our running example, Gr(2, 4). We will compute minimal generators of  $ker(\varphi)$  of degree  $(3, 1, 1, 2, -1) = deg(p_{3,4}) + deg(f)$  where f is the quadratic Plücker relation in  $ker(\varphi)$  from Example 2.9 and Example 2.10. It is well known that the Plücker relation forms a universal Gröbner basis for this ideal; therefore, we should find that there are no minimal generators of degree (3, 1, 1, 2, -1).

If we continue as before, we would compute the kernel of the matrix below.

$$\begin{array}{c} p_{1,2}p_{3,4}^2 \\ p_{1,3}p_{2,4}p_{3,4} \\ p_{2,3}p_{1,4}p_{3,4} \end{array} \begin{pmatrix} 0 & -1 & 1 & 0 & 2 & -2 & 0 & -1 & 1 & 0 \\ -1 & 0 & 1 & 1 & 1 & -1 & -1 & -1 & 0 & 1 \\ -1 & 1 & 0 & 1 & -1 & 1 & -1 & 0 & -1 & 1 \end{pmatrix}^T$$

The rows are indexed by the 10 degree 6 monomials in S which appear after applying  $\varphi$  to the monomial basis of  $R_{(3,1,1,2,-1)}$ . The kernel is generated by  $(1,-1,1)^T$ , and it corresponds to  $p_{3,4}f$ . This is clearly not a minimal generator.

Instead of considering the monomial basis, we could have used the basis

$$\{p_{3,4}f\} \cup \{p_{1,3}p_{2,4}p_{3,4}, p_{2,3}p_{1,4}p_{3,4}\}$$

of  $R_{(3,1,1,2,-1)}$ . Note that  $\{p_{3,4}f\}$  is a basis for Lift(f) as in Proposition 2.11 and the second set of monomials is a basis for  $V_{(3,1,1,2,-1)}$ . Since it is already evident that  $p_{3,4}f \in \ker(\varphi)$ , we only need to search for linear relations among  $\varphi(p_{1,3}p_{2,4}p_{3,4})$  and  $\varphi(p_{2,3}p_{1,4}p_{3,4})$ . Of course, there are none since any such linear relation corresponds to an element of the kernel of the matrix above with the first column removed. Since this kernel is trivial, we see that there are no minimal generators of degree (3, 1, 1, 2, -1).

Lastly, we discuss an additional speed-up based on [20, 30] which uses the following proposition to throw out some homogeneous components that cannot have generators.

**Proposition 2.13.** [30] Let  $\varphi : R = \mathbb{K}[x_1, \dots, x_n] \to S = \mathbb{K}[t_1, \dots, t_m]$  be a ring homomorphism,  $J(\varphi)$  be the matrix  $J(\varphi)_{ij} = \left(\frac{\partial \varphi_j}{\partial t_i}\right)$ , and  $S \subset [n]$ . Then

$$\mathbb{K}[x_i \mid i \in S] \cap \ker(\varphi) = \langle 0 \rangle \iff \operatorname{rank}(J(\varphi)_S) = |S|$$

**Remark 2.14.** For those familiar with matroid theory, the previous proposition essentially states that the *algebraic matroid* defined by the prime ideal  $\ker(\varphi)$  is the same as the linear matroid defined by  $J(\varphi)$  over the fraction field  $\mathbb{K}(t_1, \dots t_m)$ . For a more detailed discussion of different cryptomorphic constructions of algebraic matroids we refer the reader to [20, 30].

Now suppose we want to compute the degree  $\beta$  homogeneous component of  $\ker(\varphi)$ . Let  $S \subseteq \{x_1, \ldots, x_n\}$  be the subset of the variables which  $M_{\beta}$  is supported on. Then by Proposition 2.13, if  $\operatorname{rank}(J(\phi)_S) = |S|$ , then there are no generators in  $\ker(\phi)$  whose support is S. This immediately implies the following corollary.

Corollary 2.15. Let  $\varphi: R = \mathbb{K}[x_1, \dots, x_n] \to S = \mathbb{K}[t_1, \dots, t_m]$  be a ring homomorphism,  $J(\varphi)$  be the matrix  $J(\varphi)_{ij} = \left(\frac{\partial \varphi_j}{\partial t_i}\right)$ . Let  $M_\beta$  be a monomial basis for the homogeneous component of degree  $\beta$  of  $\ker(\varphi)$  and  $S \subset [n]$  correspond to the subset of variables on which  $M_\beta$  is supported. If  $\operatorname{rank}(J(\varphi)_S) = |S|$  then there are no generators of degree  $\beta$  in  $\ker(\varphi)$ 

What makes the above corollary extremely powerful for the purpose of the task at hand is the observation from [30], that if one plugs in random values for the variables  $t_j$  into  $J(\varphi)$ , then Proposition 2.13 still holds with probability 1. This means that for the purposes of our algorithm, we can simply compute the matrix  $J(\varphi)$  and then substitute in random values for the parameters  $t_j$ . This allows us to skip over many components which can never yield generators by simply computing the rank of  $J(\varphi)_S$  which is a  $m \times |S|$  matrix with entries in  $\mathbb{K}$  which is extremely cheap compared to the time it takes to evaluate  $\varphi$  on  $M_\beta$ . Also, we note that when plugging in random values for the  $t_j$ , the rank of  $J(\varphi)_S$  can only drop. This means in our algorithm we would just unnecessarily compute the component of  $\ker(\varphi)$  of degree  $\beta$ . Thus even though we are leveraging some numerical speed-ups, the output is always still correct. In the next section we will show how effective this step can be at reducing the total computation time on several large examples.

We end the section with Algorithm 1 which naturally arises from the above discussion. Observe that one major advantage that this algorithm has over other approaches comes from parallelization. The inner loop in Algorithm 1 below which runs over all multidegrees  $\beta$  that correspond to degree d monomials is *embarassingly parallel*. This means that massive speedups can be achieved if the algorithm is run in parallel on a large cluster. In our last section we showcase how effective this algorithm can be on some difficult examples from phylogenetics.

Remark 2.16. Throughout the latter part of this section, we assumed that  $\mathbb{1}$  was in the row-space of A. If instead we just assumed that the grading were positive i.e. there is a vector  $\mathbf{a} = (a_1, \ldots, a_n) \in \mathbb{Z}_{>0}^n$  in the row-space of A, you can still construct an algorithm similar to Algorithm 1. The only difference would be to replace  $P_i = \{\beta \in \mathbb{N}A \mid \beta = A\alpha, \ \alpha \cdot \mathbb{1} = i\}$  with the set  $P'_i = \{\beta \in \mathbb{N}A \mid \beta = A\alpha, \ \alpha \cdot \mathbf{a} = i\}$ .

## 3. Applications to Algebraic Statistics and Phylogenetics

In this section we apply Algorithm 1 to find low-degree minimal generators for several examples in algebraic statistics which come from mathematical phylogenetics. These examples have been previously shown to be extremely difficult and Gröbner basis algorithms typically do not terminate when applied to them even when degree-limiting is utilized [25]. In all of the cases which we describe below, we have used the Macaulay2 implementation of our algorithm which is not parallelized since this functionality. This means that the main advantage of this technique is not being fully leveraged in the below examples. Despite this, the algorithm still performs extremely well. All of the code for constructing the polynomial maps below can be found at our MathRepo page [27].

## Algorithm 1: componentsOfKernel

```
Input: A polynomial map \varphi: R = \mathbb{K}[x_1, \dots, x_n] \to S = \mathbb{K}[t_1, \dots, t_m] and a total
                  degree d
    Output: All minimal generators of ker(\varphi) of total degree at most d
 1 Compute the homogeneity space of \langle x_i - \varphi(x_i) \rangle and set A \in \mathbb{Z}^{r \times n} to be the
      associated multigrading on \ker(\varphi)
 2 Set G := \{\} to be the set of computed minimal generators
 3 Set B := \{\} to be the list of monomial bases for each multidegree
 4 Set J := J(\varphi) and substitute random values for the t_i
 5 for i=1 to d do
         Set P_i := \{ \beta \in \mathbb{N}A \mid \beta = A\alpha, \ \alpha \cdot \mathbb{1} = i \}
 6
         for \beta \in P_i do
 7
              Compute the monomial basis M_{\beta} for R_{\beta} and set S to be the variable support
 8
              if \operatorname{rank}(J(\varphi)_S) = |S| then
 9
                   continue
10
              Compute a basis for V_{\beta} where R_{\beta} = \text{Lift}(G) \oplus V_{\beta} using Proposition 2.11
11
              Expand \varphi(f^{(\beta)}) and extract the linear system L_{\beta}c = 0 where c = (c_{\alpha}) which is
12
             obtained by setting \varphi(f^{(\beta)}) = 0.
Compute a basis v^{(1)}, v^{(2)}, \dots, v^{(\ell)} \in \mathbb{K}^{M_{\beta}} for \ker(L_{\beta})
13
              G = G \cup \{\sum_{\alpha \in M_\beta} v_\alpha^{(i)} x^\alpha \mid i \in [\ell]\}
14
15 return G
```

3.1. The General Markov Model on a Phylogenetic Tree. In this subsection we provide a very brief overview of phylogenetic Markov models and the general Markov model. Since our main purpose here is to simply showcase the effectiveness of this algorithm on some notoriously difficult polynomial maps, we do not provide significant detail or background on phylogenetics and describe the polynomial maps involved primarily from an algebraic perspective. For a more detailed discussion on phylogenetics we refer the reader to [31, 34].

A  $\kappa$ -state phylogenetic Markov model on a n-leaf, leaf-labelled rooted binary tree T is a directed acyclic graphical model in which all of the internal nodes are hidden. The model produces a joint distribution on all possible joint states  $(i_1,\ldots,i_n)\in [\kappa]^n$  which can be observed at the leaves of T. This distribution is determined by associating a  $\kappa$ -state random variable  $X_v$  to each internal vertex v of T and a  $\kappa\times\kappa$  transition matrix  $M^e$  to each directed edge e=(u,v) of T such that  $M^e_{i,j}=P(X_v=j|X_u=i)$ . A root distribution  $\pi$  for the root  $\rho$  of T is also needed. Then the probability of observing  $(i_1,\ldots i_n)\in [\kappa]^n$  of states at the leaves is

$$p_{i_1...i_n} = P(X_1 = i_1, ..., X_n = i_n) = \sum_{j \in [\kappa]^{Int(T)}} \pi_{j_\rho} \prod_{(u,v) \in E(T)} M_{j_u,j_v}^{(u,v)}.$$

which as we can see is a polynomial expression in the parameters  $M_{i,j}^e$  and  $\pi_k$ . This means the model can essentially be viewed as the image of a polynomial map, and thus the vanishing

ideal of the model is the kernel of the map below.

(3.1) 
$$\psi_{T}: \mathbb{K}[p_{i_{1}...i_{n}} \mid (i_{1},...,i_{n}) \in [\kappa]^{n}] \to \mathbb{K}[M_{i,j}^{e}, \pi_{k} \mid e \in E(T), i, j, k \in [\kappa]]$$
$$p_{i_{1}...i_{n}} \mapsto \sum_{j \in [\kappa]^{Int(T)}} \pi_{j_{\rho}} \prod_{(u,v) \in E(T)} M_{j_{u},j_{v}}^{(u,v)}$$

If no other restrictions are made on the transition matrices  $M^e$  and the root distribution  $\pi$ , then resulting phylogenetic model is called the *general Markov model* [3]. For any algebraic phylogenetic model,  $\psi_T$ , the kernel  $\psi_T$ , denoted  $I_T$ , is often called the *ideal of phylogenetic invariants* of the model. The number of variables involved here grows exponentially in the number of leaves n of the tree T. This means for large trees it is often impossible to compute the kernel of  $\psi_T$  with standard methods.

Finding a complete set of generators for  $\ker(\psi_T)$  when n=3 and  $\kappa=4$  is still an open question, though the *Salmon Conjecture* [2, 3] contains a conjectural set of generators which have been shown to define the model set theoretically [16]. Further numerical evidence has also been found in [4].

We tried to find all degree 5 polynomials in the kernel which are known with our Macaulay2 implementation of Algorithm 1. In this case there are  $\binom{64+5-1}{5} = 10424128$  monomials which yield a total of 175616 unique multidegrees. While our current Macaulay2 implementation was able to compute some components, our current estimate is that it would take approximately 130 hours to compute all components, but it typically runs out of RAM. Based on our current benchmarks, we expect these issues to be solved by our OSCAR implementation. We end this section with a short application of our algorithm to the easier problem of when  $\kappa = 3$ .

**Example 3.1.** When  $\kappa = 3$ , it is known that the  $\ker(\psi_T)$  is cut out by 27 quartics [29]. We were able to verify that there are indeed 27 minimal quartics using our unparallelized Macaulay2 implementation in 29.76 seconds. We also tried to verify this using Gröbner bases; however, we killed this computation after an 76 minutes.

3.2. The K3P Model on a Phylogenetic Network. Another well studied family of phylogenetic models are group-based models. These models have been studied extensively from an algebraic perspective [11, 13, 19, 26, 33] and many algebraic problems are well understood including a complete description of the Gröbner basis for the vanishing ideal of the model [33]. This is because these models allow for a linear change of coordinates [13, 19] in which the parameterization of the model becomes a monomial map and thus the vanishing ideal becomes toric [33]. While group-based models on trees are relatively well understood, more interest recently in phylogenetics has been focused on phylogenetic networks which will be our main focus in this subsection. We begin with a description of the monomial parameterization for trees since this will be used to define the network parameterization.

In a group-based model, the states of the random variables involved are identified with the elements of a finite abelian group G. This allows a simultaneous coordinate change on both the domain and codomain of  $\psi_T$  which essentially comes from applying to the discrete Fourier transform to the expression for the joint probabilities Equation 3.1. For a more detailed explanation of this coordinate change we refer the reader to [34, Chapter 15] and instead focus on defining the polynomial map in this new coordinate system which is what we will run our algorithm on.

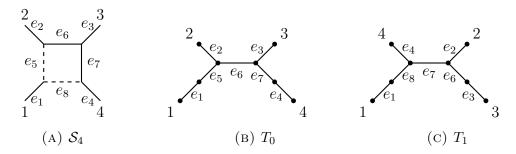


FIGURE 1. A 4 leaf sunlet network N and the two trees  $T_0$  and  $T_1$  that are obtained by deleting the reticulation edges  $e_8$  and  $e_5$  respectively.

The transformed coordinates of the domain of  $\psi_T$  are denoted by  $q_{g_1...g_n}$  and are typically called the *Fourier coordinates*. We then have new parameters  $a_g^e$  for each edge  $e \in E(T)$  and  $g \in G$ . Since T is a tree, removing any edge e of T naturally induces a partition of the leaf set into two connected components which is called a *split* of T and is denoted by  $A_e|B_e$ . The parameterization of the model in these coordinates is given by

(3.2) 
$$q_{g_1,\dots g_n} = \begin{cases} \prod_{e \in E(T)} a^e_{\sum_{i \in A_e} g_i} & \text{if } \sum_{i \in [n]} g_i = 0\\ 0 & \text{otherwise} \end{cases}$$

Many well known phylogenetic models are group-based such as the Cavendar-Farris-Neyman model, the Jukes-Cantor model, the Kimura 2-Parameter model, and the Kimura 3-Parameter (K3P) model which is typically the most difficult to compute and will be our main object of interest later in this subsection. As discussed previously, group-based models on trees are relatively well understood but many open questions remain. The simplest type of network from an algebraic perspective is called a *sunlet network* and was first introduced in [18] and further studied algebraically in [10].

**Definition 3.2.** A *n-sunlet network* is a semi-directed graph with a distinguished vertex called the *reticulation vertex* and whose underlying graph is obtained by adding a leaf to every vertex of a *n*-cycle and then directing the non-leaf edges which are adjacent to the reticulation vertex towards it.

The two directed edges which point into the reticulation vertex are often called reticulation edges and are drawn as dotted edges instead of directed edges since they are implicitly directed toward the vertex at which they meet. This is illustrated in Figure 1. Observe that deleting either of the reticulation edges from the sunlet network yields a tree. These underlying trees are used to construct the parameterization of the network model. For any phylogenetic model  $\psi_T$  which is defined for trees, it is naturally extended to a sunlet network N by defining

$$\psi_N = \lambda \psi_{T_0} + (1 - \lambda) \psi_{T_1}$$

We now focus on the concrete problem of computing the ideal of phylogenetic invariants for a 4-leaf sunlet network  $N_4$  under the K3P model. The K3P model is the generic group-based model for the group  $G = \mathbb{Z}_2 \times \mathbb{Z}_2$ . This means for each edge of the network  $N_4$  and

each  $g \in \mathbb{Z}_2 \times \mathbb{Z}_2$  we have a parameter  $a_q^e$ . The parameterization  $\psi_{N_4}$  is then given by

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

Since in this case  $G = \mathbb{Z}_2 \times \mathbb{Z}_2$ , there are a total of  $|G|^4 = 4^{4-1} = 64$  variables in the domain of  $\psi_{N_4}$  and  $4 \times 8 = 32$  parameters; however, by exploiting the fact the associated map of varieties is actually of the form  $\psi_T: \prod_{i=1}^8 \mathbb{P}^3 \to \mathbb{P}^{63}$ , one can naturally reparameterize with only  $8 \times (4-1) + 1 = 25$  parameters. This means that in total the elimination ideal will be in 89 variables. Recently, the authors of [25] attempted to find all generators up to total degree 3 in ker $(\psi_{N_4})$  using standard Gröbner basis algorithms in Macaulay2 with degree-limiting. They were able to find all degree 2 generators however after 100 days the Gröbner basis algorithm still did not terminate to provide all degree 3 generators.

We ran our Macaulay2 implementation of Algorithm 1 which has no parallelization features on a MacBook Pro with an Apple M2 chip and 16 GB of RAM. It takes slightly over 8 minutes for Algorithm 1 to produce all minimal generators of  $\ker(\psi_{N_4})$  of total degree at most 3. We also ran this computation without the speed-up from Corollary 2.15. For the degree 2 generators the computation time was quite similar however for the degree 3 generators the computation took approximately 30 minutes instead of 8. The final results are summarized in the following theorem.

**Theorem 3.3.** The ideal of phylogenetic invariants  $I_{N_4} = \ker(\psi_{N_4})$  for K3P model on a four leaf sunlet network has 12 minimal quadratic and 64 minimal cubics generators.

We were actually able to compute all minimal degree 2 generators for 5-leaf sunlets as well. In this case  $\psi_{N_5}$  maps from a ring in 256 variables into a ring with 31 variables so the elimination ideal is in 287 variables total. Despite this our algorithm is still able to compute all degree 2 generators in only 25 minutes and with parallelization could compute all degree 3 generators as well based on our current benchmarking. As one can see, Algorithm 1 can scale to extremely large numbers of variables provided that the generators of interest are of low total degree and the map is homogeneous in a reasonably fine multigrading. The results are summarized in the following theorem and broken down in Table 1 below.

**Theorem 3.4.** There are 648 minimal quadratic invariants of the K3P model on a 5-leaf sunlet network.

3.3. The TN93 Model on a 4-Leaf Tree. As discussed in the previous section, groupbased models for trees have many nice algebraic properties associated to them. In particular, there is a linear change of coordinates which realizes the associated varieties as toric varieties. In practice, these models may not be the most biologically relevant. For example, it might not be a reasonable assumption for the root distribution  $\pi$  to be uniform.

In this section, we consider the Timura-Nei (TN93) model [22] as studied in [6] and compute all of the quadratic invariants for a 4-leaf tree. This model is algebraic timereversible meaning that for each transition matrix M, we have that

$$\pi_i M_{i,j} = \pi_j M_{j,i}$$
13

Minimal Generators for 4 and 5 Leaf Sunlet Networks							
Leaves	Total Degree	Monomials	Grading	Multidegrees	Skipped	Min.	Time
			Rank		Components	Gens.	(sec)
4	2	2080	13	1720	1708	12	9.66
4	3	45,760	13	25,152	24,304	64	492.31
5	2	32,896	16	19,936	19,312	648	1504.03
5	3	2,829,056	16	637,440	_	_	-

Table 1. This table shows the number of monomials, distinct multidegrees, the number of minimal generators, and the time each computation took in each total degree for both 4 and 5 leaf sunlet networks. The column skipped components corresponds to the number of components which can be skipped completely using Corollary 2.15.

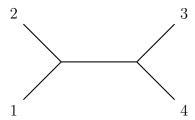


FIGURE 2. A 4-leaf tree with one non-trivial split 12|34.

and that the collection of transition matrices all commute with each other. These assumptions amount to saying that the transition matrices are simultaneously diagonalizable and that the root distribution  $\pi$  is an eigenvector of  $M^T$  with eigenvalue 1. The TN93 model enjoys much more flexibility than group-based models.

**Definition 3.5.** The TN93 model is a 4-state algebraic time-reversible model with transition matrices of the form

$$\begin{pmatrix} *_1 & \pi_2 c & \pi_3 b & \pi_4 b \\ \pi_1 c & *_2 & \pi_3 b & \pi_4 b \\ \pi_1 b & \pi_2 b & *_3 & \pi_4 d \\ \pi_1 b & \pi_2 b & \pi_3 d & *_4 \end{pmatrix}$$

where  $*_i$  is chosen so that each row sums to 1 where the root distribution is  $\pi = (\pi_1, \pi_2, \pi_3, \pi_4)$ .

We will focus on the quartet tree T which is pictured in Figure 2 under the TN93 model. Since the transition matrices are simultaneously diagonalizable, if we ignore the stochastic restrictions on these matrices, we see that the variety is parameterized by the  $5 \times 4 = 20$ eigenvalues of these matrices. We also assume that the root distribution is fixed and generic, so instead of working over  $\mathbb{C}$ , we work over the fraction field  $K = \mathbb{C}(\pi_1, \pi_2, \pi_3, \pi_4)$ . These observations along with the fact that this is a 4-state model means that the parametrization takes the following form.

$$\phi_T: K[p_{i_1,i_2,i_3,i_4} \mid (i_1,i_2,i_3,i_4) \in [4]^4] \to K[\lambda_{1,1},\ldots,\lambda_{5,4}]$$

In [6], the authors describe a linear change of coordinates from the probability coordinates to a new set of coordinates  $q_{i_1,i_2,i_3,i_4}$  which has two nice properties: (1) 176 of the  $q_{i_1,i_2,i_3,i_4}$ 's map to 0 and (2) 71 of the remaining non-zero coordinates are monomials in the eigenvalues of the transition matrices. We will refer to the set of indices of the 80 non-zero coordinates by  $\mathcal{N}\mathcal{Z}_T$ . In particular, we can greatly reduce the number of variables in the elimination ideal from 276 to just 100. The new parametrization takes the following form.

$$\varphi_T: K[q_{i_1,i_2,i_3,i_4} \mid (i_1,i_2,i_3,i_4) \in \mathcal{NZ}_T] \to K[\lambda_{1,1},\ldots,\lambda_{5,4}]$$

We let  $I_T$  denote the kernel of  $\varphi_T$ . The authors showed that on an open set of  $\mathcal{V}(I_T)$  the variety is a complete intersection and is cut out by 64 equations of degree at most 4 [6, Theorem 5.14].

While the number of parameters is greatly reduced from the general Markov model, computing a Gröbner basis for  $I_T$  is probably still out of reach. However, using Algorithm 1, we found all minimal quadrics in  $I_T$ . We see that there are many more minimal quadrics cutting out the full variety.

**Theorem 3.6.** There are 375 minimal quadratic invariants of T under the TN93 model.

#### ACKNOWLEDGEMENTS

Part of this research was performed while the authors were visiting the Institute for Mathematical and Statistical Innovation (IMSI), which is supported by the National Science Foundation (Grant No. DMS-1929348). Benjamin Hollering was supported by the Alexander von Humboldt Foundation. Joseph Cummings was supported by NSF CCF-1812746.

## References

- [1] Elizabeth S Allman, Sonia Petrovic, John A Rhodes, and Seth Sullivant. Identifiability of two-tree mixtures for group-based models. *IEEE/ACM transactions on computational biology and bioinformatics*, 8(3):710–722, 2010.
- [2] Elizabeth S. Allman and John A. Rhodes. Phylogenetic invariants for the general markov model of sequence mutation. *Mathematical Biosciences*, 186(2):113–144, 2003.
- [3] Elizabeth S. Allman and John A. Rhodes. Phylogenetic ideals and varieties for the general markov model. Advances in Applied Mathematics, 40(2):127–148, 2008.
- [4] Daniel J. Bates and Luke Oeding. Toward a salmon conjecture. Exp. Math., 20(3):358–370, 2011.
- [5] Jérémy Berthomieu, Christian Eder, and Mohab Safey El Din. Msolve: A library for solving polynomial systems. ISSAC '21, page 51–58, New York, NY, USA, 2021. Association for Computing Machinery.
- [6] Marta Casanellas, Roser Homs Pons, and Angélica Torres. A novel algebraic approach to time-reversible evolutionary models, 2023.
- [7] Julia Chifman and Laura Kubatko. Quartet Inference from SNP Data Under the Coalescent Model. *Bioinformatics*, 30(23):3317–3324, 08 2014.
- [8] David A. Cox, John Little, and Donal O'Shea. *Ideals, varieties, and algorithms*. Undergraduate Texts in Mathematics. Springer, Cham, fourth edition, 2015. An introduction to computational algebraic geometry and commutative algebra.
- [9] Joseph Cummings and Jonathan Hauenstein. Multi-graded macaulay dual spaces, 2023.
- [10] Joseph Cummings, Benjamin Hollering, and Christopher Manon. Invariants for level-1 phylogenetic networks under the cavendar-farris-nevman model, 2021.
- [11] Jan Draisma and Jochen Kuttler. On the ideals of equivariant tree models. *Math. Ann.*, 344(3):619–644, 2009.
- [12] Nicholas Eriksson. Tree construction using singular value decomposition. In *Algebraic statistics for computational biology*, pages 347–358. Cambridge Univ. Press, New York, 2005.

- [13] Steven N. Evans and T. P. Speed. Invariants of some probability models used in phylogenetic inference. Ann. Statist., 21(1):355–377, 1993.
- [14] Jean Charles Faugère. A new efficient algorithm for computing gröbner bases without reduction to zero (f5). In *Proceedings of the 2002 International Symposium on Symbolic and Algebraic Computation*, ISSAC '02, page 75–83, New York, NY, USA, 2002. Association for Computing Machinery.
- [15] Jean-Charles Faugère, Mohab Safey El Din, and Thibaut Verron. On the complexity of computing gröbner bases for weighted homogeneous systems. *Journal of Symbolic Computation*, 76:107–141, 2016.
- [16] Shmuel Friedland and Elizabeth Gross. A proof of the set-theoretic version of the salmon conjecture. J. Algebra, 356:374–379, 2012.
- [17] Daniel R. Grayson and Michael E. Stillman. Macaulay2, Version 1.20, 2022. http://www.math.uiuc.edu/Macaulay2/.
- [18] Elizabeth Gross and Colby Long. Distinguishing phylogenetic networks. SIAM Journal on Applied Algebra and Geometry, 2(1):72–93, 2018.
- [19] Michael D Hendy and David Penny. Complete families of linear invariants for some stochastic models of sequence evolution, with and without the molecular clock assumption. *Journal of Computational Biology*, 3(1):19–31, 1996.
- [20] Benjamin Hollering and Seth Sullivant. Identifiability in phylogenetics using algebraic matroids. *J. Symbolic Comput.*, 104:142–158, 2021.
- [21] Anders Nedergaard Jensen. Computing gröbner fans and tropical varieties in gfan. 2008.
- [22] Tamura K and Nei M. Estimation of the number of nucleotide substitutions in the control region of mitochondrial dna in humans and chimpanzees. *Mol Biol Evol.*, 10(3):512–26, 1993 May.
- [23] Martin Kreuzer and Lorenzo Robbiano. Computational commutative algebra. 2. Springer-Verlag, Berlin, 2005.
- [24] Colby Long and Seth Sullivant. Identifiability of 3-class Jukes-Cantor mixtures. Adv. in Appl. Math., 64:89–110, 2015.
- [25] Samuel Martin, Vincent Moulton, and Richard M. Leggett. Algebraic invariants for inferring 4-leaf semi-directed phylogenetic networks. bioRxiv, 2023.
- [26] Mateusz Michał ek. Geometry of phylogenetic group-based models. J. Algebra, 339:339–356, 2011.
- [27] MATHREPO Mathematical Data and Software. https://mathrepo.mis.mpg.de/MultigradedImplicitization, 2023. [Online; accessed 1 November 2023].
- [28] Oscar open source computer algebra research system, version 0.14.0-dev, 2023.
- [29] Lior Pachter and Bernd Sturmfels. Tropical geometry of statistical models. *Proc. Natl. Acad. Sci. USA*, 101(46):16132–16137, 2004.
- [30] Zvi Rosen. Computing algebraic matroids. arXiv preprint arXiv:1403.8148, 2014.
- [31] Mike Steel. Phylogeny: discrete and random processes in evolution. SIAM, 2016.
- [32] Nils Sturma, Mathias Drton, and Dennis Leung. Testing many constraints in possibly irregular models using incomplete u-statistics, 2023.
- [33] Bernd Sturmfels and Seth Sullivant. Toric ideals of phylogenetic invariants. *Journal of Computational Biology*, 12(2):204–228, 2005.
- [34] Seth Sullivant. Algebraic statistics, volume 194 of Graduate Studies in Mathematics. American Mathematical Society, Providence, RI, 2018.

University of Notre Dame

Email address: jcummin7@nd.edu

Technische Universität München, 85748 Garching B. München, Boltzmannstr. 3., Germany

Email address: benhollering@gmail.com