Phylogenetic Algebraic Geometry

Milosz Szymanski
Supervisor Dimitra Kosta

September 2024

Abstract

This report focuses on describing the phylogenetic invariants of claw trees on arbitrary number of leaves. More precisely we present an inductive algorithm, a result of *Chifman* and *Petrović* [1], which returns the generating set of the toric ideal associated with the monomial map associated to the phylogenetic tree. We then present this algorithim in action by computing a basic example in Section 6. We end the report by briefly commenting on how the description of phylogenetic invariants for claw trees, combined with the results of *Sturmfels* and *Sullivant* [2], allows for the description of phylogenetic invariants for a wide class of general phylogenetic trees.

Contents

1 Acknowledgements

I would like to thank the School of Mathematics at the University of Edinburgh and the Edinburgh Mathematical Society for providing the funding without which I couldn't devote my summer to this project. I would also like to thank the summer project coordinator, Wei En Tan, who has helped me navigate numerous applications and who has encouraged me to pursue this fantastic opportunity. Lastly, I would like to express my great gratitude to my project supervisor, Dimitra Kosta, whose guidance and insight were invaluable tools of research, and for introducing me to the wonderful world of Phylogenetic Algebraic Geometry.

2 Introduction

Phylogenetics is the field of biology concerned with the study of evolutionary relationships between species. The goal of phylogenetics is to represent these relationships between species and their common ancestors as a tree. Algebraic Geometry is the field of mathematics concerned with the connection between algebraic objects (ideals) and geometric objects (varieties) and how techniques from one field can be used to solve problems in the other field. Phylogenetic Algebraic Geometry is the application of the tools of Algebraic Geometry to the study of phylogenetic trees. We can think of each tree as a hidden Markov model, with each node of the tree representing a random variable and each edge representing a chance for that random variable to change state. We can then write down probabilities of observing certain states at the leaves of the tree as a polynomial in terms of the model parameters. Doing this for each possible collection of states at the leaves we obtain a system of polynomial equations in terms of the model parameters, the goal of Phylogenetic Algebraic Geometry is to solve this system of polynomial equations by finding the generating set of an ideal that this system of polynomial equations define. This ideal corresponds to a variety in the space of model parameters. Each point on this variety represents a solution to the system of polynomial equations. If now we use the DNA sequence data of the species represented by the leaves of the tree, we can use maximum likelihood estimation to find the closest point on our variety that matches the observed data, therefore providing the optimal estimate for the model parameters.

In Section 2, we discuss the concepts in Algebraic Geometry that will be used throughout this paper, as well as give a brief overview of the notation we will use throughout this paper and the general approach to finding the phylogenetic invariants for trees with the Jukes-Cantor Model. In Sections 3 and 4, we present example calculations for trees $K_{1,3}$ and $K_{1,4}$ respectivly, calculating explicitly the phylogenetic invariants. In Section 5, we state the inductive algorithm used to calculate phylogenetic invariants for claw trees with a higher number of leaves, and apply the algorithim to show how to obtain the phylogenetic invariants for $K_{1,4}$ from the phylogenetic invariants for $K_{1,3}$. In Section 6, we explain how the results of Section 5 can be used to compute the phylogenetic invariants for any tree with a friendly labelling.

3 Preliminaries

The goal of this paper is not only to present a way of calculating phylogenetic invariants for claw trees with an arbitrary number of leaves, but also to do so in a way which is accessible to an undergraduate student with knowledge of algebra up to the content of the Honours Algebra course. Below we present definitions of concepts in Algebraic Geometry, as well as other related fields, which such a student must grasp and for the curious mind we provide references which they should use to read up on the topics discussed.

3.1 Algebraic Geometry and More

We begin with some definitions and theorems from Algebraic Geometry. All of the definitions and theorems below (unless otherwise stated) are taken from *Cox, Little, O'Shea* [3]. I heavily recommend taking your time going through these definitions and theorems in detail and going over relevant proofs and exercises in the book. We begin by defining the concept of an **affine variety** below.

Definition 1. Let k be a field, and let $f_1, \dots, f_s$ be polynomials in $k[x_1, \dots, x_n]$. Then we set

$$\mathbf{V}(f_1, \dots, f_s) = \{(a_1, \dots, a_n) \in k^n : f_i(a_1, \dots, a_n) = 0\}.$$

We call $\mathbf{V}(f_1, \dots, f_s)$ the **affine variety** defined by $f_1, \dots, f_s$.

We next refresh our knowledge of **ideals**:

Definition 2. A subset $I \subset k[x_1, \dots, x_n]$ is an **ideal** if it satisfies the following three conditions:
(i) $0 \in I$.
(ii) If $f, g \in I$, then $f + g \in I$.
(iii) If $f \in I$ and $h \in k[x_1, \dots, x_n]$, then $hf \in I$.

Given a finite set of polynomials, we define their **generating set** to be as follows:

Definition 3. Let $f_1, \dots, f_s$ be polynomial in $k[x_1, \dots, x_n]$. Then we set

$$\langle f_1, \dots, f_s \rangle = \left\{ \sum_{i=1}^{s} h_i f_i : h_1, \dots, h_s \in k[x_1, \dots, x_n] \right\}.$$

The key observation here is that the generating set of polynomials $f_1, \dots, f_s$ as defined above is an ideal (the reader should prove this result for themselves!).

Algebraic Geometry revolves around the connection between varieties and ideals. Below we define a way of obtaining an ideal from a variety.

Definition 4. Let $V \subset k^n$ be an affine variety. Then we set,

$$\mathbf{I}(V) = \{f \in k[x_1, \dots, x_n] : f(a_1, \dots, a_n) = 0 \text{ for all } (a_1, \dots, a_n) \in k^n\}.$$

We stress that $\mathbf{I}(V)$ is an ideal, and we call it the ideal of $\mathbf{V}$.

Definition 5. A **monomial ordering** $>$ on $k[x_1, \dots, x_n]$ is any relation $>$ on $\mathbb{Z}^n_{\geq 0}$, or equivalently, any relation on the set of polynomials $x^\alpha, \alpha \in \mathbb{Z}^n_{\geq 0}$, satisfying:
(i) $>$ is a total (or linear) ordering on $\mathbb{Z}^n_{\geq 0}$.
(ii) if $\alpha > \beta$ and $\gamma \in \mathbb{Z}^n_{\geq 0}$, then $\alpha + \gamma > \beta + \gamma$.
(iii) $>$ is a well-ordering on $\mathbb{Z}^n_{\geq 0}$. This means that every nonempty subset of $\mathbb{Z}^n_{\geq 0}$ has a smallest element under $>$.

Definition 6 (**Lexicographic Order**)**.** Let $\alpha = (\alpha_1, \dots, \alpha_n)$ and $\beta = (\beta_1, \dots, \beta_n) \in \mathbb{Z}^n_{\geq 0}$. We say $\alpha >_{lex} \beta$ if, in the vector difference $\alpha - \beta \in \mathbb{Z}^n_{\geq 0}$, the leftmost nonzero entry is positive. We will write $x^\alpha >_{lex} x^\beta$ if $\alpha >_{lex} \beta$.

Definition 7. Let $f = \sum_\alpha a_\alpha x^\alpha$ be a nonzero polynomial in $k[x_1, \dots, x_n]$ and let $>$ be a monomial order.
(i) The multidegree of f is

$$multideg(f) = max(\alpha \in \mathbb{Z}^n_{\geq 0} : a_\alpha \neq 0)$$

(the maximum is taken with respect to $>$).
(ii) The leading coefficient of f is

$$LC(f) = a_{multideg(f)} \in k.$$

(iii) The leading monomial of f is

$$LM(f) = x^{multideg(f)}$$

(with coefficient 1).
(iv) The leading term of f is

$$LT(f) = LC(f) \cdot LM(f).$$

Definition 8. Let $I \subset k[x_1, \dots, x_n]$ be an ideal other than $\{0\}$.
(i) We denote by $LT(I)$ the set of leading terms of elements of I. Thus,

$$LT(I) = \{cx^\alpha : \text{ there exists } f \in I \text{ with } LT(f) = cx^\alpha\}.$$

(ii) We denote by $\langle LT(I) \rangle$ the ideal generated by the elements of $LT(I)$.

Theorem 9 (**Hilbert's Basis Theorem**)**.** *Every ideal $I \in k[x_1, \dots, x_n]$ has a finite generating set. That is, $I = \langle g_1, \dots, g_t \rangle$ for some $g_1, \dots, g_t \in I$.*

Definition 10. Fix a monomial order. A finite subset $G = \{g_1, \dots, g_t\}$ of an ideal I is said to be a **Gröbner basis** (or **standard basis**) if

$$\langle LT(g_1), \dots, LT(g_t) \rangle = \langle LT(I) \rangle .$$

Definition 11. Let $I \subset k[x_1, \ldots, x_n]$ be an ideal. We will denote by $\mathbf{V}(I)$ the set

$$\mathbf{V}(I) = \{(a_1, \ldots, a_n) \in k^n : f(a_1, \ldots, a_n) = 0 \text{ for all } f \in I\}.$$

Definition 4 gives us a mapping from affine varieties to ideals, meanwhile Definition 11 gives us a mapping from ideals to affine varieties, this correspondence between ideals and varieties is the heat of Algebraic Geometry. This correspondence is captured in the celebrated Hilbert's Nullstellensatz, for which we need two more definitions about radical ideals.

Definition 12. An ideal I is radical if $f^m \in I$ for some integer $m \geq 1$ implies that $f \in I$.

Definition 13. Let $I \subset k[x_1, \ldots, x_n]$ be an ideal. The radical of I, denoted $\sqrt{I}$, is the set

$$\{f : f^m \in I \; for \; some \; integer \; m \geq 1\}.$$

Theorem 14 (**Hilbert's Nullstellensatz**). *Let k be an algebraically closed field. If I is a radical ideal in $k[x_1, \ldots, x_n]$, then*

$$\boldsymbol{I}(\,\boldsymbol{V}(I)) = \sqrt{I}.$$

The most important consequence of Hilbert's Nullstellensatz is that it allows us to construct a bridge between geometry (varieties) and algebra (ideals). This is captured in the following theorem:

Theorem 15 (**The Ideal-Variety Correspondence**). *Let k be an arbitrary field.*
(i) The maps

$$\textit{affine varieties} \xrightarrow{\boldsymbol{I}} \textit{ideals}$$

and

$$\textit{ideals} \xrightarrow{\boldsymbol{V}} \textit{affine varieties}$$

are inclusion-reversing, i.e., if $I_1 \subset I_2$ are ideals, then $\boldsymbol{V}(I_1) \supset \boldsymbol{V}(I_2)$ and, similarly, if $V_1 \subset V_2$ are varieties, then $\boldsymbol{I}(V_1) \supset \boldsymbol{I}(V_2)$. Furthermore, for any variety V, we have

$$\boldsymbol{V}(\boldsymbol{I}(V)) = V,$$

so that $\boldsymbol{I}$ is always one-to-one.
(ii) If k is algebraically closed, and if we restrict to radical ideals, then the maps

$$\textit{affine varieties} \xrightarrow{\boldsymbol{I}} \textit{ideals}$$

and

$$\textit{ideals} \xrightarrow{\boldsymbol{V}} \textit{affine varieties}$$

are inclusion-reversing bijections which are inverses of each other.

The last bit of Algebraic Geometry we need comes in the form of the following four definitions:

Definition 16. The Zariski closure of a subset of affine space is the smallest affine algebraic variety containing the set. If $S \subset k^n$, the Zariski closure of S is denoted by $\bar{S}$ and is equal to $\mathbf{V}(\mathbf{I}(S))$.

Definition 17. An affine variety $V \subset k^n$ is irreducible if whenever V is written in the form $V = V_1 \cup V_2$, where V_1 and V_2 are affine varieties, then either $V_1 = V$ or $V_2 = V$.

Definition 18. An ideal $I \subset k[x_1, \ldots, x_n]$ is prime whenever $f, g \in k[x_1, \ldots, x_n]$ and $fg \in I$, then either $f \in I$ or $g \in I$.

The next definition comes from *Kosta*, *Thoma* and *Vladoiu* [4]:

Definition 19. Let k be a field and $A = [\mathbf{a_1}, \ldots, a_n] \in \mathbb{Z}^{m \times n}$be an integer matrix with the set of column vectors $\{\mathbf{a_1}, \ldots, a_n\}$ such that $Ker_{\mathbb{Z}}(A) \cap \mathbb{N}^n = \{\mathbf{0}\}$. The toric ideal of A is the ideal $I_A \subset k[x_1, \ldots, x_n]$ generated by the binomials $x^{\mathbf{u}^+} - x^{\mathbf{u}^-}$ where $\mathbf{u} \in Ker_{\mathbb{Z}}(A)$, and $\mathbf{u} = \mathbf{u}^+ - \mathbf{u}^-$ is the expression of $\mathbf{u}$ as a difference of two non-negative vectors with disjoint support.

Now we provide some definitions which whilst not directly linked with the field of Algebraic Geometry are nonetheless important for the discussion of Phylogenetic Algebraic Geometry.

This definition is taken from *Dieck* [3]:

Definition 20. The **n-dimensional standard simplex** is

$$\Delta^n = \Delta[n] = \{(t_0, \ldots, t_n) \in \mathbb{R}^{n+1} | \sum_{i=0}^{n} t_i = 1, t_i \geq 0\} \subset \mathbb{R}^{n+1}.$$

To give a bit of intuition on this, a 0-dimensional standard simplex is a point, a 1-dimensional standard simplex is a line, a 2-dimensional standard simplex is a equilateral triangle, and a 3-dimensional standard simplex is a tetrahedron.

Next we need some definitions from the field of combinatorics:

Definition 21. Let T be a graph with n vertices. T is a tree if T has no cycles and has $n - 1$ edges.

Definition 22 (Complete Bipartite graph). Let G be a graph,
(i) if the set of vertices V of G can be split into two disjoint set A and B so that $A \cup B = V$, and such that each edge of G joins a vertex of A and a vertex of B, then G is a bipartite graph.
(ii) G is a complete bipartite graph if G is a bipartite graph such that every vertex of A is joined t each vertex of B by exactly one edge. If G is a complete bipartite graph with $|A| = r$ and $|B| = s$, then we denote G by $K_{r,s}$

Definition 23. Let G be the complete bipartite graph $K_{1,m}$. Then we call G a star graph, or alternatively in the phylogenetics community a claw tree.

Definition 24. Let $T_1, \ldots, T_m$ be a collection of trees, we refer to the whole collection of trees $T_1, \ldots, T_m$ as a forest.

Finally, we end this subsection with a quick definition of character groups:

Definition 25. Let G be an abelian group. A function $f : G \to \mathbb{C}\backslash\{0\}$ which is a group homomorphism from the group G to the group of complex units $\mathbb{C}\backslash\{0\}$, i.e such that $\forall g_1, g_2 \in G$, $f(g_1 g_2) = f(g_1)f(g_2)$, is called a character of G. The group of all characters forms an abelian group under pointwise multiplication, that is if f_i and f_j are characters of G then $\forall g \in G$, $(f_i f_j)(g) = f_i(g)f_j(g)$. We denote the group of all characters by $G^{\times}$ and call it the character group of G.

3.2 Notation and Technical Definitions

Here we define the notation that we will use throughout the paper, and delve into the theory of computing phylogenetic invariants for claw trees.

Let T be a claw tree with m leaves. We denote the set of vertices of the tree by $\mathcal{V}$ and we label each leaf N_λ for $\lambda = 1, \dots, m$ such that the leftmost leaf is labeled N_1 and the rightmost leaf is labeled N_m. We call the top vertex of the tree the root and label it by N_r.
We label the edges of our tree by e^i for $\lambda = 1, \dots, m$ such that edge e^λ joins N_r and N_λ. Since T should be thought of as a hidden Markov model, to each edge e^λ of T we associate a transition matrix M_λ which encodes the probabilities of transitioning from one state to another as we travel from N_r to N_λ. In this spirit each vertex becomes a random variable which can take some number of discrete states. The number of states each N_λ can take is denoted by k, for the rest of this paper, unless otherwise stated, we assume $k = 2$, i.e for all $\lambda = 1, \dots, m, r$, $N_\lambda \in \{0, 1\}$. We represent the probability distribution at the root of the tree by $\pi = (\pi_0, \pi_1)$, where $\pi_0 = \mathbb{P}(N_r = 0)$ and $\pi_1 = \mathbb{P}(N_r = 1)$. In this paper we consider group-based phylogenetic models, which means that we fix an abelian group G such that $k = |G|$ and associate each element of G with a particular state our random variables can take. More rigorously, this means that we assume that our rate matrices are invariant under the actions of the group G, where we note that a transition matrix M and the related rate matrix Q have the relation $M = exp(Q)$. For more detail on this refer to page 4 of [5].

For the general model, where we don't assume any relationship between the entries of our transition matrices, the matrix M_λ would be given by:

$$M_\lambda = \begin{pmatrix} e^\lambda_{00} & e^\lambda_{01} \\ e^\lambda_{10} & e^\lambda_{11} \end{pmatrix},$$

where e^λ_{ij} represents the probability of transitioning from state i to state j as we travel along the edge e^λ. For this paper, and unless otherwise stated, we will consider the Jukes-Cantor evolution model, which lets us assume that the transition matrices have the following simplified structure:

$$M_\lambda = \begin{pmatrix} e^\lambda_0 & e^\lambda_1 \\ e^\lambda_1 & e^\lambda_0 \end{pmatrix}.$$

With this we can now write down the probabilities that we observe a certain distribution of bases at the leaves of our tree. For example, let $m = 3$ and let p_{000} represent the probability that all leaves of the tree will be the base 0, then our probability polynomial is given by:

$$p_{000} = \pi_0 e^1_0 e^2_0 e^3_0 + \pi_1 e^1_1 e^2_1 e^3_1,$$

and for general number of leaves m we have,

$$p_{x_1 \cdots x_m} = \sum_{i=0}^{1} \pi_i \prod_{j=1}^{m} e^j_{x_j},$$

where $x_i \in \{0, 1\}$ for $i = 1, \ldots, m$. We call the polynomials $p_{x_1 \cdots x_m}$ leaf probabilities.

We can write these polynomials $p_{x_1 \cdots x_m}$ in terms of the model parameters for all $k^m = 2^m$ possible states our leaves could be in. The solutions to this system of polynomial equations define a variety which contains all possible solutions in terms of our model's parameters. The phylogenetic invariants of the model which we seek is the set of polynomials in terms of the leaf probabilities which vanish for any choice of model parameters. These vanishing polynomials form a prime ideal in the polynomial ring over the leaf probabilities, and this prime ideal defines the variety mentioned above. For more detail refer to the introduction of *Toric Ideals of Phylogenetic Invariants* [2].
The remarkable result of *Evans* and *Speed* [6] is that there exists a linear transformation which translates the leaf probabilities into the transformed probabilities which are monomials, whose ideal generate a toric variety. This linear transformation is the Discrete Fourier Transform and it is given in [2] by the following equation:

$$\hat{f}(\chi) = \sum_{g \in G} \chi(g) f(g), \tag{1}$$

where $\chi \in G^{\times}$ and $f : G \to \mathbb{C}$. These functions f correspond to the model parameters, by which we mean that each edge e^i of T and the root N_r has an associated function $f^(i) : G \to \mathbb{C}$ defined below:

$$f^i(g) = \begin{cases} e^i_0 & \text{if } g = 0 \\ e^i_1 & \text{otherwise} \end{cases} \qquad\qquad \pi(g) = \begin{cases} \pi_0 & \text{if } g = 0 \\ \pi_1 & \text{otherwise} \end{cases}.$$

Since we will be considering phylogenetic models with two states, we need to choose an Abelian group of order two, $G = \mathbb{Z}_2$ is our sole choice. Below we give the table detailing the elements of the dual group $\mathbb{Z}_2$ which we will refer to throughout the paper:

	0	1
χ^0	1	1
χ^1	1	-1

which should be read as: applying the homomorphism $\chi \in \hat{G}$ in row i to the element $g \in G$ in column j returns the value in the i, j-th position, e.g. $\chi^1(1) = -1$.

Therefore our aim is to calculate the Gröbner basis for this ideal of transformed probabilities and describe a linear map which maps the transformed probabilities back to leaf probabilities.

4 Example 1: $K_{1,3}$

For this example we will consider the following claw tree T with 3 leaves:

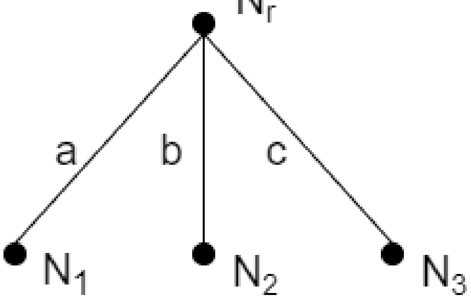

Figure 1: A phylogenetic tree with 3 leaves and 1 root

We make the following assumptions about the model: in this model we consider only two states $\{0, 1\}$, we keep the root distribution general, and we consider the Jukes-Cantor model.

The Jukes-Cantor DNA model of evolution assumption gives us the following nice structure on the transition matrices:

$$M_a = \begin{pmatrix} a_0 & a_1 \\ a_1 & a_0 \end{pmatrix}, \qquad M_b = \begin{pmatrix} b_0 & b_1 \\ b_1 & b_0 \end{pmatrix}, \qquad M_c = \begin{pmatrix} c_0 & c_1 \\ c_1 & c_0 \end{pmatrix}.$$

Since these are transition matrices, we can deduce the following relationships between our parameters:

$$a_0 = 1 - a_1, \qquad b_0 = 1 - b_1, \qquad c_0 = 1 - c_1.$$

We keep the root distribution general, i.e. $\pi = (\pi_0, \pi_1)$, and since these probabilities must sum to 1 we note that $\pi_0 = 1 - \pi_1$. Therefore the total number of parameters in this model is 4.
Since our nodes can take two possible values we need to choose an abelian group G of order 2, therefore we choose $G = \mathbb{Z}_2$, and we note that the state 0 corresponds to the identity of G, also labeled 0, and the state 1 corresponds to the non-identity element of G, also labeled 1.

Now, using the independence assumption, we can write down the probabilities of observing any combination of states at the leaves of the tree T. Below we give the probabilities of observing any combination of states at the leaves:

$$\begin{aligned}
p_{000} &= \pi_0 a_0 b_0 c_0 + \pi_1 a_1 b_1 c_1, & p_{001} &= \pi_0 a_0 b_0 c_1 + \pi_1 a_1 b_1 c_0,\\
p_{010} &= \pi_0 a_0 b_1 c_0 + \pi_1 a_1 b_0 c_1, & p_{011} &= \pi_0 a_0 b_1 c_1 + \pi_1 a_1 b_0 c_0,\\
p_{100} &= \pi_0 a_1 b_0 c_0 + \pi_1 a_0 b_1 c_1, & p_{101} &= \pi_0 a_1 b_0 c_1 + \pi_1 a_0 b_1 c_0,\\
p_{110} &= \pi_0 a_1 b_1 c_0 + \pi_1 a_0 b_0 c_1, & p_{111} &= \pi_0 a_1 b_1 c_1 + \pi_1 a_0 b_0 c_0.
\end{aligned}$$

In total we have eight polynomials p_{ijk} as for each of the three leaves we have two choices of base. Since these polynomials all represent probabilities, they will sum to 1, i.e. $\sum_{i \in G} \sum_{j \in G} \sum_{k \in G} p_{ijk} = 1$. We now wish to find the transformed probabilities q_{ijk} and find their Gröbner basis. To do this, we use the Discrete Fourier Transform using Equation 1 to find the transformed parameters in terms of the model parameters and substittue them into the leaf probabilities p_{ijk}.

In our example, we have four functions $f : G \to \mathbb{C}^{\times}$, one for each edge in our tree, $f^{(N_1)}, f^{(N_2)}, f^{(N_3)}$ and one for our root distribution, π, explicitly defined as follows for $g \in G$:

$$f^{(N_1)}(g) = \begin{cases} a_0 & \text{if } g = 0 \\ a_1 & \text{otherwise,} \end{cases} \qquad f^{(N_2)}(g) = \begin{cases} b_0 & \text{if } g = 0 \\ b_1 & \text{otherwise,} \end{cases}$$

$$f^{(N_3)}(g) = \begin{cases} c_0 & \text{if } g = 0 \\ c_1 & \text{otherwise,} \end{cases} \qquad \pi(g) = \begin{cases} \pi_0 & \text{if } g = 0 \\ \pi_1 & \text{otherwise.} \end{cases}$$

We refer to our dual group table 3.2 and apply the Discrete Fourier Transform equation 1 to compute the new parameters. An example calculation for edge a is given below, calculations for all other edges are analogous:

$$\begin{aligned}
\alpha_0 = f^{(\hat{N}_1)}(\chi^0) &= \sum_{g \in G} \chi^0(g) f^{(N_1)}(g)\\
&= \chi^0(0) f^{(N_1)}(0) + \chi^0(1) f^{(N_1)}(1)\\
&= a_0 + a_1,
\end{aligned}$$

$$\begin{aligned}
\alpha_1 = f^{(\hat{N}_1)}(\chi^1) &= \sum_{g \in G} \chi^1(g) f^{(N_1)}(g)\\
&= \chi^1(0) f^{(N_1)}(0) + \chi^1(1) f^{(N_1)}(1)\\
&= a_0 - a_1.
\end{aligned}$$

The complete set of transformed parameters is given below:

$$\begin{aligned}
r_0 &= \pi_0 + \pi_1, & \alpha_0 &= a_0 + a_1, & \beta_0 &= b_0 + b_1, & \gamma_0 &= c_0 + c_1,\\
r_1 &= \pi_0 - \pi_1, & \alpha_1 &= a_0 - a_1, & \beta_1 &= b_0 - b_1, & \gamma_1 &= c_0 - c_1.
\end{aligned}$$

The same information can be elegantly encoded in matrix form as $\epsilon = H\mathbf{e}$, where ϵ is the new set of parameters (here a 2×1 vector), H is the 2×2 Hadamard matrix encoding the information of the Fourier Transform, and $\mathbf{e}$ is the 2×1 vector of old parameters. An example for the parameters associated with edge a is given below:

$$\begin{pmatrix} \alpha_0 \\ \alpha_1 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix}. \tag{2}$$

We now refer to the result of *Theorem 6* from *Sturmfels* and *Sullivant* [2].

Theorem 26. *(**Sturmfels** and **Sullivant** [2]) Let $p(g_1, \ldots, g_m)$ be the joint distribution of a group based model for the phylogenetic tree T. Then the Fourier transform of p has the form*

$$q(\chi_1, \ldots, \chi_m) = \hat{\pi}(\chi_1 \ldots \chi_m) \cdot \prod_{v \in \mathcal{V}(T) \backslash \{r\}} \widehat{f^{(v)}} \left(\prod_{l \in \Lambda(v)} \chi_l \right) \tag{3}$$

where $\Lambda(v)$ is the set of leaves which have v as a common ancestor.

For claw trees, Equation 3 becomes:

$$q(\chi_1, \ldots, \chi_m) = \hat{\pi}(\chi_1 \ldots \chi_m) \cdot \prod_{i=1}^{m} \widehat{f^{(N_i)}}(\chi_i). \tag{4}$$

For our phylogenetic tree T we are considering the Fourier Transform of coordinates p_{ijk} which have the form:

$$q_{ijk} = q(\chi_1, \chi_2, \chi_3) = \hat{\pi}(\chi_1 \chi_2 \chi_3) \cdot \prod_{i=1}^{3} \widehat{f^{(N_i)}}(\chi_i).$$

Below we present a computation of q_{000}:

$$\begin{aligned} q_{000} = q(\chi_0, \chi_0, \chi_0) &= \hat{\pi}(\chi_0 \chi_0 \chi_0) \cdot \prod_{i=1}^{m} \widehat{f^{(v)}}(\chi_i) \\ &= \hat{\pi}(\chi_0) \cdot \left(\widehat{f^{(N_1)}}(\chi_0) \widehat{f^{(N_2)}}(\chi_0) \widehat{f^{(N_3)}}(\chi_0) \right) \\ &= r_0 \cdot (\alpha_0 \beta_0 \gamma_0) \\ &= r_0 \alpha_0 \beta_0 \gamma_0 \end{aligned}$$

and all other q_{ijk} are analogously calculated. We give the full list of transformed coordinates below:

$$q_{000} = r_0\alpha_0\beta_0\gamma_0, \quad q_{001} = r_1\alpha_0\beta_0\gamma_1, \quad q_{010} = r_1\alpha_0\beta_1\gamma_0, \quad q_{011} = r_0\alpha_0\beta_1\gamma_1,$$
$$q_{100} = r_1\alpha_1\beta_0\gamma_0, \quad q_{101} = r_0\alpha_1\beta_0\gamma_1, \quad q_{110} = r_0\alpha_1\beta_1\gamma_0, \quad q_{111} = r_1\alpha_1\beta_1\gamma_1,$$

and using the following Macaulay2 code, found in the Appendix here B, we compute the Gröbner basis of the ideal of our toric variety.
From the output $o3$ we can see that the Phylogenetic Invariant in transformed coordinates is given by the following set of three equations:

$$q_{000}q_{111} - q_{001}q_{110} = 0, \quad q_{000}q_{111} - q_{010}q_{101} = 0, \quad q_{000}q_{111} - q_{011}q_{100} = 0.$$

To write the Phylogenetic Invariant in terms of the original coordinates q_{ijk}, we utilise the relationship between the p_{ijk} and q_{ijk} coordinates given by the matrix equation $\mathbf{q} = \hat{H}\mathbf{p}$, where here $\mathbf{q}$ is the 8×1 vector of q_{ijk} coordinates, $\mathbf{p}$ is the 8×1 vector of p_{ijk}, and $\hat{H}$ is the 8×8 Hadamard matrix built out of the 2×2 Hadamard matrix H we encountered in equation 7. $\hat{H}$ is built recursively as follows:

$$H_1 = \begin{pmatrix} H & H \\ H & -H \end{pmatrix}, \qquad\qquad \hat{H} = \begin{pmatrix} H_1 & H_1 \\ H_1 & -H_1 \end{pmatrix}.$$

We can implement this in Macaulay2 with the following code found in the Appendix B, where we note the following labelling we have used,

$$x_0 = p_{000}, \qquad x_1 = p_{001}, \qquad x_2 = p_{010}, \qquad x_3 = p_{011},$$
$$x_4 = p_{100}, \qquad x_5 = p_{101}, \qquad x_6 = p_{110}, \qquad x_0 = p_{111}.$$

Now we can simply read of the Phylogenetic Invariant of this tree as the following set of three quadratic equations:

$$-p_{001}p_{010} + p_{000}p_{011} - p_{001}p_{100} + p_{000}p_{101} + p_{011}p_{110} + p_{101}p_{110} - p_{010}p_{111} - p_{100}p_{111} = 0$$
$$-p_{001}p_{010} + p_{000}p_{011} - p_{010}p_{100} + p_{011}p_{101} + p_{000}p_{110} + p_{101}p_{110} - p_{001}p_{111} - p_{100}p_{111} = 0$$
$$-p_{001}p_{100} - p_{010}p_{100} + p_{000}p_{101} + p_{011}p_{101} + p_{000}p_{110} + p_{011}p_{110} - p_{001}p_{111} - p_{010}p_{111} = 0.$$

5 Example 2: $K_{1,4}$

For this example we will consider the following claw tree T with 4 leaves:

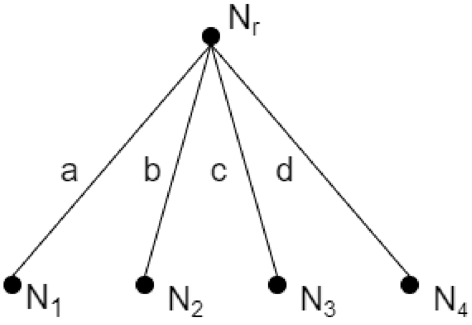

Figure 2: A phylogenetic tree with 4 leaves and 1 root

We make the following assumptions about the model: in this model we consider only two states $\{0, 1\}$, we keep the root distribution general, and we consider the Jukes-Cantor model.

The Jukes-Cantor DNA model of evolution assumption gives us the following nice structure on the transition matrices:

$$M_a = \begin{pmatrix} a_0 & a_1 \\ a_1 & a_0 \end{pmatrix}, \quad M_b = \begin{pmatrix} b_0 & b_1 \\ b_1 & b_0 \end{pmatrix}, \quad M_c = \begin{pmatrix} c_0 & c_1 \\ c_1 & c_0 \end{pmatrix}, \quad M_d = \begin{pmatrix} d_0 & d_1 \\ d_1 & d_0 \end{pmatrix}.$$

Since these are transition matrices, we can deduce the following relationships between our parameters:

$$a_0 = 1 - a_1, \qquad b_0 = 1 - b_1, \qquad c_0 = 1 - c_1, \qquad d_0 = 1 - d_1.$$

We keep the root distribution general, i.e. $\pi = (\pi_0, \pi_1)$, and since these probabilities must some to 1 we note that $\pi_0 = 1 - \pi_1$. Therefore the total number of parameters in this model is 5.

As discussed previously, since the nodes of T can take one of a possible two states, we choose $G = \mathbb{Z}_2$ as the abelian group we base our model on.

Now, using the independence assumption, we can write down the probabilities of observing any combination of states at the leaves of the tree T. Below we give the probabilities of observing any combination of states at the leaves:

$$\begin{aligned}
p_{0000} &= \pi_0 a_0 b_0 c_0 d_0 + \pi_1 a_1 b_1 c_1 d_1, & p_{0001} &= \pi_0 a_0 b_0 c_0 d_1 + \pi_1 a_1 b_1 c_1 d_0,\\
p_{0010} &= \pi_0 a_0 b_0 c_1 d_0 + \pi_1 a_1 b_1 c_0 d_1, & p_{0011} &= \pi_0 a_0 b_0 c_1 d_1 + \pi_1 a_1 b_1 c_0 d_0,\\
p_{0100} &= \pi_0 a_0 b_1 c_0 d_0 + \pi_1 a_1 b_0 c_1 d_1, & p_{0101} &= \pi_0 a_0 b_1 c_0 d_1 + \pi_1 a_1 b_0 c_1 d_0,\\
p_{0110} &= \pi_0 a_0 b_1 c_1 d_0 + \pi_1 a_1 b_0 c_0 d_1, & p_{0111} &= \pi_0 a_0 b_1 c_1 d_1 + \pi_1 a_1 b_0 c_0 d_0,\\
p_{1000} &= \pi_0 a_1 b_0 c_0 d_0 + \pi_1 a_0 b_1 c_1 d_1, & p_{1001} &= \pi_0 a_1 b_0 c_0 d_1 + \pi_1 a_0 b_1 c_1 d_0,\\
p_{1010} &= \pi_0 a_1 b_0 c_1 d_0 + \pi_1 a_0 b_1 c_0 d_1, & p_{1011} &= \pi_0 a_1 b_0 c_1 d_1 + \pi_1 a_0 b_1 c_0 d_0,\\
p_{1100} &= \pi_0 a_1 b_1 c_0 d_0 + \pi_1 a_0 b_0 c_1 d_1, & p_{1101} &= \pi_0 a_1 b_1 c_0 d_1 + \pi_1 a_0 b_0 c_1 d_0,\\
p_{1110} &= \pi_0 a_1 b_1 c_1 c_0 + \pi_1 a_0 b_0 c_0 d_1, & p_{1111} &= \pi_0 a_1 b_1 c_1 d_1 + \pi_1 a_0 b_0 c_0 d_0.
\end{aligned}$$

In total we have 16 polynomials p_{ijkl} as for each of the four leaves we have two choices of base. Since these polynomials all represent probabilities, they will sum to 1, i.e. $\sum_{i\in G}\sum_{j\in G}\sum_{k\in G}\sum_{l\in G} p_{ijkl} = 1$. We now wish to find the transformed probabilities q_{ijk} and find their Gröbner basis. To do this we use the Discrete Fourier Transform using Equation 1 to find the transformed parameters in terms of the model parameters and substittue them into the leaf probabilities p_{ijk}.

In our example, we have five such functions $f : G \to \mathbb{C}^{\times}$, one for each edge in our tree, $f^{(N_1)}, f^{(N_2)}, f^{(N_3)}, f^{(N_4)}$ and one for our root distribution, π, explicitly defined as follows for $g \in G$:

$$f^{(N_1)}(g) = \begin{cases} a_0 & \text{if } g = 0 \\ a_1 & \text{otherwise} \end{cases} \qquad f^{(N_2)}(g) = \begin{cases} b_0 & \text{if } g = 0 \\ b_1 & \text{otherwise} \end{cases}$$

$$f^{(N_3)}(g) = \begin{cases} c_0 & \text{if } g = 0 \\ c_1 & \text{otherwise} \end{cases} \qquad f^{(N_4)}(g) = \begin{cases} d_0 & \text{if } g = 0 \\ d_1 & \text{otherwise} \end{cases}$$

$$\pi(g) = \begin{cases} \pi_0 & \text{if } g = 0 \\ \pi_1 & \text{otherwise} \end{cases}$$

We now refer to our dual group table 3.2 and apply the Discrete Fourier Transform equation 1 to compute the new parameters. This calculation is the same as for Example 2, and as such we omit it here.
The complete set of transformed parameters is given below:

$$\begin{aligned}
r_0 &= \pi_0 + \pi_1, & \alpha_0 &= a_0 + a_1, & \beta_0 &= b_0 + b_1, & \gamma_0 &= c_0 + c_1, & \delta_0 &= d_0 + d_1\\
r_1 &= \pi_0 - \pi_1, & \alpha_1 &= a_0 - a_1, & \beta_1 &= b_0 - b_1, & \gamma_1 &= c_0 - c_1, & \delta_1 &= d_0 - d_1.
\end{aligned}$$

The same information can be elegantly encoded in matrix form as $\epsilon = H\mathbf{e}$, where ϵ is the new set of parameters (here a (2×1) vector), H is the (2×2) Hadamard matrix encoding the information of the Fourier Transform, and $\mathbf{e}$ is the (2×1) vector of old parameters. An example for the parameters associated with edge a is given below:

$$\begin{pmatrix} \alpha_0 \\ \alpha_1 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} \tag{5}$$

We now refer to the result of *Theorem 6* from *Toric Ideals of Phylogenetic Invariants* [2]. For our phylogenetic tree T we are considering, the Fourier Transform of coordinates p_{ijkl} has the form:

$$q_{ijkl} = q(\chi_1, \chi_2, \chi_3, \chi_4) = \hat{\pi}(\chi_1\chi_2\chi_3\chi_4) \cdot \prod_{u=1}^{4} \widehat{f^{(N_u)}}(\chi_u).$$

Below we present a computation of q_{0000}:

$$\begin{aligned} q_{0000} = q(\chi_0, \chi_0, \chi_0, \chi_0) &= \hat{\pi}(\chi_0\chi_0\chi_0\chi_0) \cdot \prod_{u=1}^{4} \widehat{f^{(N_u)}}(\chi_u) \\ &= \hat{\pi}(\chi_0) \cdot \left(\widehat{f^{(N_1)}}(\chi_0)\widehat{f^{(N_2)}}(\chi_0)\widehat{f^{(N_3)}}(\chi_0)\widehat{f^{(N_4)}}(\chi_0) \right) \\ &= r_0 \cdot (\alpha_0\beta_0\gamma_0\delta_0) \\ &= r_0\alpha_0\beta_0\gamma_0\delta_0 \end{aligned}$$

and all other q_{ijkl} are analogously calculated, we give the full list of transformed coordinates below:

$$\begin{aligned} &q_{0000} = r_0\alpha_0\beta_0\gamma_0\delta_0, \quad &&q_{0001} = r_1\alpha_0\beta_0\gamma_0\delta_1, \quad &&q_{0010} = r_1\alpha_0\beta_0\gamma_1\delta_0, \quad &&q_{0011} = r_0\alpha_0\beta_0\gamma_1\delta_1, \\ &q_{0100} = r_1\alpha_0\beta_1\gamma_0\delta_0, &&q_{0101} = r_0\alpha_0\beta_1\gamma_0\delta_1, &&q_{0110} = r_0\alpha_0\beta_1\gamma_1\delta_0, &&q_{0111} = r_1\alpha_0\beta_1\gamma_1\delta_1, \\ &q_{1000} = r_1\alpha_1\beta_0\gamma_0\delta_0, &&q_{1001} = r_0\alpha_1\beta_0\gamma_0\delta_1, &&q_{1010} = r_0\alpha_1\beta_0\gamma_1\delta_0, &&q_{1011} = r_1\alpha_1\beta_0\gamma_1\delta_1, \\ &q_{1100} = r_0\alpha_1\beta_1\gamma_0\delta_0, &&q_{1101} = r_1\alpha_1\beta_1\gamma_0\delta_1, &&q_{1110} = r_1\alpha_1\beta_1\gamma_1\delta_0, &&q_{1111} = r_0\alpha_1\beta_1\gamma_1\delta_1. \end{aligned}$$

Using the following Macaulay2 code, found in the Appendix here B, we compute the toric variety.

All that is left to do is to apply the inverse Fourier Transform, i.e. use the relationship $\mathbf{q} = \hat{H}\mathbf{p}$, where $\mathbf{q}$ is the (16x1) vector of q_{ijkl} coordinates, $\mathbf{p}$ is the (16x1) vector of p_{ijk}, and $\hat{H}$ is the (16x16) Hadamard matrix constructed analogously as in Example 2 but with an extra iteration.

We can implement this in Macaulay2 with the following code, found here in the Appendix B.

Reading the output of the code one can see that the toric variety in terms of the p_{ijkl} coordinates is given by 30 quadratic equations each with 16 terms. We don't include these here due to the length of these equations, but encourage the reader to run the code in the Appendix to see the equations explicitly.

6 Main Result: $K_{1,m}$

In this section we aim to generalise our past examples and comment on the claw tree with an arbitrary number of leaves. Let T be the following claw tree with m leaves:

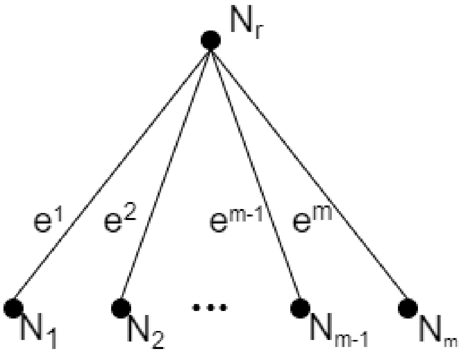

Figure 3: A phylogenetic tree with m leaves and 1 root

Like before, we make the following assumptions about our model: we consider two states $\{0,1\}$, with a general root distribution and we assume the Jukes-Cantor model of evolution.

The Jukes-Cantor model of evolution assumption gives us the following structure on the transisition matrices:

$$M_1 = \begin{pmatrix} e_0^1 & e_1^1 \\ e_1^1 & e_0^1 \end{pmatrix}, \quad \ldots, \quad M_i = \begin{pmatrix} e_0^i & e_1^i \\ e_1^i & e_0^i \end{pmatrix}, \quad \ldots, \quad M_m = \begin{pmatrix} e_0^m & e_1^m \\ e_1^m & e_0^m \end{pmatrix},$$

for all $i \in \{1,\ldots,m\}$. Since these are transition matrices, we can deduce the following relationships between our parameters:

$$e_0^1 = 1 - e_1^1, \qquad \ldots, \qquad e_0^i = 1 - e_1^i, \qquad \ldots, \qquad e_0^m = 1 - e_1^m,$$

for all $i \in \{1,\ldots,m\}$. We keep the root distribution general, i.e. $\pi = (\pi_0, \pi_1)$, and since these probabilities must some to 1 we note that $\pi_0 = 1 - \pi_1$. Therefore the total number of parameters in this model is $m+1$.
As before, our only choice of abelian group of order two is $\mathbb{Z}_2$, therefore we set $G = \mathbb{Z}_2$.

Now, using the independence assumption, we can write down the probabilities of observing any combination of states at the leaves of the tree T. Below we explicitly give the probability polynomial of observing the state 0 at all leaves N_i:

$$p_{00\dots 0} = \pi_0 e_0^1 e_0^2 \dots e_0^m + \pi_1 e_1^1 e_1^2 \dots e_1^m,$$

and in general, for any binary string $x_1 x_2 \dots x_m$ of length m, where here $\forall k$, $x_k \in \{0,1\}$, which represents the probability of the observation, $\mathbb{P}(N_1 = x_1, N_2 = x_2, \dots, N_m = x_m) = p_{x_1 x_2 \dots x_m}$ is given by the probability polynomial:

$$p_{x_1 x_2 \dots x_m} = \pi_0 \sum_{i=1}^{m} e^i_{x_i} + \pi_1 \sum_{i=1}^{m} e^i_{x_i+1},$$

where we note that for the subscripts we are working in $\mathbb{Z}_2$, therefore $\forall k$, $x_k+1 \in \{0,1\}$.
In total we have 2^m polynomials $p_{x_1 x_2 \dots x_m}$ as for each of the m leaves we have two choices of base. Since these polynomials all represent probabilities, they will sum to 1, i.e. $\sum_{(x_1,x_2,\dots,x_m)\in G^m} p_{x_1 x_2 \dots x_m} = 1$.

Now we must deviate from our usual strategy, which whilst computationally valid for small m, for large m it performs rather poorly. We turn instead to the theory presented in Section 6.2 of *Chifman* and *Petrović* [1]. We will present relevant definitions and propositions which allow us to construct an algorithm to inductively calculate the Gröbner basis for the set of transformed coordinates. We omit the proofs here and instead urge the reader to consult *Chifman* and *Petrović* [1] for full mathematical rigour.

Definition 27. (**Chifman** and **Petrović** [1]) Let $\pi_i(q)$ be the projection of q that eliminates the i^{th} index of each variable in q.

For example, we have $\pi_2(q_{1010}q_{0001} - q_{1011}q_{0000}) = q_{110}q_{001} - q_{111}q_{000}$.

Definition 28. (**Chifman** and **Petrović** [1]) Let $\mathcal{G}_n$ be the set of quadratic binomials $q \in I_n$ that can be written as

$$q = q^+ - q^- = q_{g_1{}^{(1)}\dots g_1{}^{(n)}} q_{g_2{}^{(1)}\dots g_2{}^{(n)}} - q_{h_1{}^{(1)}\dots h_1{}^{(n)}} q_{h_2{}^{(1)}\dots h_2{}^{(n)}}$$

such that one of the following two properties is satisfied:
Property (i): For some $1 \leq i \leq n, j \in \mathbb{Z}_2$,

$$g_1^{(i)} = g_2^{(i)} = j = h_1^{(i)} = h_2^{(i)}$$

and

$$\pi_i(q) \in I_{n-1}.$$

Property (ii): For each $1 \leq k \leq n$,

$$g_1^{(k)} + g_2^{(k)} = 1 = h_1^{(k)} + h_2^{(k)}$$

and

$$\pi_i(q) \in I_{n-1}.$$

Proposition 29. *(**Chifman** and **Petrović** [1]) The set of binomials $\mathcal{G}_n$ generates the ideal I_n. That is,*

$$I_n = (q : q^+ - q^- \in \mathcal{G}_n)$$

and in fact we can go one step further,

Proposition 30. *(**Chifman** and **Petrović** [1]) The set $\mathcal{G}_n$ is a lexicographic Gröbner basis of I_n, for any n.*

We now present the inductive algorithm which given $\mathcal{G}_{n-1}$ will generate $\mathcal{G}_n$.

Algorithm to construct $\mathcal{G}_n$:

1. First we generate all memberes of $\mathcal{G}_n$ which satisfy Property (i) of Definition 28.
Fix $q \in \mathcal{G}_{n-1}$ and write it as

$$q = q^+ - q^- = q_{g_1{}^{(1)} \ldots g_1{}^{(n-1)}} q_{g_2{}^{(1)} \ldots g_2{}^{(n-1)}} - q_{h_1{}^{(1)} \ldots h_1{}^{(n-1)}} q_{h_2{}^{(1)} \ldots h_2{}^{(n-1)}}.$$

We want to find the preimages of q under the projection maps π_i, where $1 \leq i \leq n$. This is quite straightforward, as it simply involves inserting a 0 or 1 in the appropriate index for each of the transformed coordinates $q_{g_1{}^{(1)} \ldots g_1{}^{(n-1)}}$, $q_{g_2{}^{(1)} \ldots g_2{}^{(n-1)}}$, $q_{h_1{}^{(1)} \ldots h_1{}^{(n-1)}}$, $q_{h_2{}^{(1)} \ldots h_2{}^{(n-1)}}$, as if we put a mix of 0 and 1 in the appropriate index of each transformed coordinate then we would violate Property (i). In general, let $\pi_i^{-1}(q)$ denote the preimage of q under the map π_i, which satisfies Property (i). Then explicitly $\pi_i^{-1}(q)$ is given by:

$$\begin{aligned} \pi_i^{-1}(q) = \{ & q_{g_1{}^{(1)} \ldots g_1{}^{(i-1)} 0 g_1{}^{(i)} \ldots g_1{}^{(n-1)}} q_{g_2{}^{(1)} \ldots g_2{}^{(i-1)} 0 g_2{}^{(i)} \ldots g_2{}^{(n-1)}} \\ & - q_{h_1{}^{(1)} \ldots h_1{}^{(i-1)} 0 h_1{}^{(i)} \ldots h_1{}^{(n-1)}} q_{h_2{}^{(1)} \ldots h_2{}^{(i-1)} 0 h_2{}^{(i)} \ldots h_2{}^{(n-1)}}, \\ & q_{g_1{}^{(1)} \ldots g_1{}^{(i-1)} 1 g_1{}^{(i)} \ldots g_1{}^{(n-1)}} q_{g_2{}^{(1)} \ldots g_2{}^{(i-1)} 1 g_2{}^{(i)} \ldots g_2{}^{(n-1)}} \\ & - q_{h_1{}^{(1)} \ldots h_1{}^{(i-1)} 1 h_1{}^{(i)} \ldots h_1{}^{(n-1)}} q_{h_2{}^{(1)} \ldots h_2{}^{(i-1)} 1 h_2{}^{(i)} \ldots h_2{}^{(n-1)}} \}, \end{aligned}$$

we'll see this more clearly in the example that follows. And by definition 28 we have that for all $1 \leq i \leq n$, for every $q \in \pi_i^{-1}(q)$, we have $q \in \mathcal{G}_n$. The binomials generated in this way need not be all unique, and in fact it's quite likely that repeats will occur!

2. Second we generate all members of $\mathcal{G}_n$ which satisfy Property (ii) of Definition 28. The idea will be to fix a good choice of q^- and find appropriate candidates for q^+ such that $q^+ - q^-$ satisfies Property (ii).
We first consider the case when n is odd. We fix $q^- = q_{01\ldots1} q_{10\ldots0}$ where each of the subscripts is a binary number of length n, and $1 \ldots 1$ implies that all digits between the second and last index are 1, and likewise for $0 \ldots 0$. Since the first subscript has an even number of 1's, our assumption that n is odd implies that $n - 1$ is even, and the second subscript has

and odd number of 1's, we can deduce that $r_0r_1|q^-$, i.e. r_0r_1 divides q^-. Therefore we need q^+ to be such that $r_0r_1|q^+$, as otherwise $q^+ - q^- \neq 0$. To satisfy Property (ii) we need to choose pairs of binary numbers of length n whose digits are complementary to each other to build q^+ out of. There are $2^{n-1} - 1$ such pairs, which are obtained by listing the smallest $2^{n-1} - 1$ binary numbers of length n, and then pairing them with the largest $2^{n-1} - 1$ binary numbers of length n in reverse order.

Next we consider the case when n is even. We have two choices to fix q^-; we can either fix $q^- = q_{01\ldots1}q_{10\ldots0}$ or $q^- = q_{01\ldots10}q_{10\ldots01}$, where for the first choice we have that $r_1^2|q^-$ and for the second choice we have that $r_0^2|q^-$. Like in the previous case, we consider pairs of binary numbers of length n constructed by pairing the smallest $2^{n-1} - 1$ binary numbers of length n with the largest $2^{n-1} - 1$ binary numbers of length n in reverse order. Then for the first choice of q^-, q^+ is constructed from the pairs whose binary numbers have an odd number of ones, and for the second choice of q^-, q^+ is constructed from the pairs whose binary numbers have an even number of ones. Again this should become much clearer with an example, which we promptly give below.

In the previous two Examples, 4 and 5, we have computed the phylogenetic invariants in the transformed coordinates for claw trees with $n = 3$ and $n = 4$ leaves. We will now use the algorithm above to construct the set of phylogenetic invariants for the claw tree on 4 leaves, $\mathcal{G}_4$, from the set of phylogenetic invariants for the claw tree on 3 leaves, $\mathcal{G}_3$.

From Example 4 we know that,

$$\mathcal{G}_3 = \{q_{000}q_{111} - q_{001}q_{110}, q_{000}q_{111} - q_{010}q_{101}, q_{000}q_{111} - q_{011}q_{100}\}.$$

For each $q \in \mathcal{G}_3$ we find the preimages $\pi_i(q)$ for $1 \leq i \leq 4$, which gives us the following 24 phylogenetic invariants.

- For $q = q_{000}q_{111} - q_{001}q_{110}$:

$$\begin{aligned}
\pi_1(q) &= \{q_{0000}q_{0111} - q_{0001}q_{0110}, q_{1000}q_{1111} - q_{1001}q_{1110}\}\\
\pi_2(q) &= \{q_{0000}q_{1011} - q_{0001}q_{1010}, q_{0100}q_{1111} - q_{0101}q_{1110}\}\\
\pi_3(q) &= \{q_{0000}q_{1101} - q_{0001}q_{1100}, q_{0010}q_{1111} - q_{0011}q_{1110}\}\\
\pi_4(q) &= \{q_{0000}q_{1110} - q_{0010}q_{1100}, q_{0001}q_{1111} - q_{0011}q_{1101}\}.
\end{aligned}$$

- For $q = q_{000}q_{111} - q_{010}q_{101}$:

$$\begin{aligned}
\pi_1(q) &= \{q_{0000}q_{0111} - q_{0010}q_{0101}, q_{1000}q_{1111} - q_{1010}q_{1101}\}\\
\pi_2(q) &= \{q_{0000}q_{1011} - q_{0010}q_{1001}, q_{0100}q_{1111} - q_{0110}q_{1101}\}\\
\pi_3(q) &= \{q_{0000}q_{1101} - q_{0100}q_{1001}, q_{0010}q_{1111} - q_{0110}q_{1011}\}\\
\pi_4(q) &= \{q_{0000}q_{1110} - q_{0100}q_{1010}, q_{0001}q_{1111} - q_{0101}q_{1011}\}.
\end{aligned}$$

- For $q = q_{000}q_{111} - q_{011}q_{100}$:

$$\begin{aligned}
\pi_1(q) &= \{q_{0000}q_{0111} - q_{0011}q_{0100}, q_{1000}q_{1111} - q_{1011}q_{1100}\}\\
\pi_2(q) &= \{q_{0000}q_{1011} - q_{0011}q_{1000}, q_{0100}q_{1111} - q_{0111}q_{1100}\}\\
\pi_3(q) &= \{q_{0000}q_{1101} - q_{0101}q_{1000}, q_{0010}q_{1111} - q_{0111}q_{1010}\}\\
\pi_4(q) &= \{q_{0000}q_{1110} - q_{0110}q_{1000}, q_{0001}q_{1111} - q_{0111}q_{1001}\}.
\end{aligned}$$

This gives us all elements of $\mathcal{G}_4$ which satisfy Property (i) of Definition 28. Now, since $n = 4$ is even, we consider two candidates for q^-, namely $q^- = q_{0111}q_{1000}$ divisible by r_1^2, and $q^- = q_{0110}q_{1001}$ divisible by r_0^2. We list the smallest $2^{4-1} - 1 = 7$ binary numbers of length 4 and the largest 7 binary numbers of length 4 in reverse order below:

0000	1111
0001	1110
0010	1101
0011	1100
0100	1011
0101	1010
0110	1001

However note that the pairing $(0110, 1001)$ corresponds to $q^+ = q_{0110}q_{1001} = q^-$, which trivially gives us $q = q^+ - q^- = 0$, and so we disregard this pairing.

The pairings whose constituents have an even number of ones are,

$$(0000, 1111) \qquad (0011, 1100) \qquad (0101, 1010),$$

therefore for $q^- = q_{0111}q_{1000}$ the possible phylogenetic invariants are,

$$q_{0000}q_{1111} - q_{0111}q_{1000} \qquad q_{0011}q_{1100} - q_{0111}q_{1000} \qquad q_{0101}q_{1010} - q_{0111}q_{1000}.$$

The pairings whose constituents have an odd number of ones are,

$$(0001, 1110) \qquad (0010, 1101) \qquad (0100, 1011),$$

therefore for $q^- = q_{0110}q_{1001}$ the possible phylogenetic invariants are,

$$q_{0001}q_{1110} - q_{0110}q_{1001} \qquad q_{0010}q_{1101} - q_{0110}q_{1001} \qquad q_{0100}q_{1011} - q_{0110}q_{1001}.$$

This results in an additional 6 elements of $\mathcal{G}_4$ which satisfy Property (ii) of Definition 28. In total we have computed 30 phylogenetic invariants which generate the toric ideal for the claw tree witj 4 leaves, which agrees with the Macaulay 2 output found in the Appendix B, and the table found on page 17 in *Sturmfels* and *Sullivant* [2].

We have a way of computing the Gröbner basis of the toric ideal of transformed coordinates for a claw tree with an arbitrary number of leaves! All that is left to do is to apply the inverse of the Discrete Fourier Transform to write each binomial of the Gröbner basis in terms of the leaf probabilities. Below we present our result for the strucutre of Hadamard matrices used in the inverse of the Discrete Fourier Transform. For a different presentation of the Discrete Foruier Transform and for further reading refer to Chapter 15 of *Pachter* and *Sturmfels* [7].

Theorem 31. *Let $m \in \mathbb{N}$ such that $m \geq 3$, and let $\mathbf{q}$ be the vector of 2^m Fourier coordinates, $\mathbf{p}$ be the vector of 2^m original coordinates and H_m be the ($2^m \times 2^m$) Hadamard matrix defined recursively by $H_1 = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$ and $H_m = \begin{pmatrix} H_{m-1} & H_{m-1} \\ H_{m-1} & -H_{m-1} \end{pmatrix}$. The inverse of the Discrete Fourier Transform, with respect to the Fourier coordinates of the claw tree with m leaves as described in this section, is given by:*

$$\mathbf{q} = H_m \mathbf{p}.$$

Proof. We begin by noting that proving the matrix equation in question is equivalent to proving the following equation holds for all $m \geq 3$,

$$q_{x_1 \dots x_m} = \sum_{(y_1,\dots,y_m) \in G^m} (-1)^{\sum_{i=1}^m x_i y_i} p_{y_1 \dots y_m}, \tag{6}$$

where $G = \mathbb{Z}_2$. This is equivalent since the equation above gives the rows of the product $H_m \mathbf{p}$. We proceed by induction.

Base Case ($m = 3$):

This is a matter of direct calculation which we present an example off (q_{000}) and leave the other calculations (which are completely analogous) to the interested reader:

$$\begin{aligned}
\sum_{(y_1,y_2,y_3)\in G^3} (-1)^{\sum_{i=1}^{3} 0 y_i} p_{y_1y_2y_3} &= \sum_{(y_1,y_2,y_3)\in G^3} p_{y_1y_2y_3} \\
&= p_{000} + \cdots + p_{111} \\
&= \pi_0 a_0 b_0 c_0 + \pi_1 a_1 b_1 c_1 + \cdots + \pi_0 a_1 b_1 c_1 + \pi_1 a_0 b_0 c_0 \\
&= \pi_0 (a_0 b_0 c_0 + \cdots + a_1 b_1 c_1) + \pi_1 (a_1 b_1 c_1 + \cdots + a_0 b_0 c_0) \\
&= (\pi_0 + \pi_1)(a_0 b_0 c_0 + \cdots + a_1 b_1 c_1) \\
&= (\pi_0 + \pi_1)(a_0 (b_0 c_0 + \cdots + b_1 c_1) + a_1 (b_0 c_0 + \cdots + b_1 c_1)) \\
&= (\pi_0 + \pi_1)(a_0 + a_1)(b_0 c_0 + \cdots + b_1 c_1) \\
&= (\pi_0 + \pi_1)(a_0 + a_1)(b_0 (c_0 + c_1) + b_1 (c_0 + c_1)) \\
&= (\pi_0 + \pi_1)(a_0 + a_1)(b_0 + b_1)(c_0 + c_1) \\
&= r_0 \alpha_0 \beta_0 \gamma_0 \\
&= q_{000}
\end{aligned}$$

Inductive Step ($m = k$):

We assume the claim is true for $m = k$ and investigate the case $m = k + 1$. We fix the index x_1 and look at cases when $x_1 = 0$ and $x_1 = 1$.

Firstly let's fix $x_1 = 0$, the right hand side of Equation 6 becomes:

$$RHS = \sum_{(y_1,\ldots,y_k,y_{k+1})\in G^{k+1}} (-1)^{0\cdot y_1 + \sum_{i=2}^{k+1} x_i y_i} p_{y_1\ldots y_k, y_{k+1}}$$

Now we sum over the y_1 index to get two sums,

$$\begin{aligned}
RHS = &\sum_{(y_2,\ldots,y_k,y_{k+1})\in G^{k}} (-1)^{0\cdot 0 + \sum_{i=2}^{k+1} x_i y_i} p_{0\ldots y_k, y_{k+1}} \\
+ &\sum_{(y_2,\ldots,y_k,y_{k+1})\in G^{k}} (-1)^{0\cdot 1 + \sum_{i=2}^{k+1} x_i y_i} p_{1\ldots y_k, y_{k+1}} \\
= &\sum_{(y_2,\ldots,y_k,y_{k+1})\in G^{k}} (-1)^{\sum_{i=2}^{k+1} x_i y_i} p_{0\ldots y_k, y_{k+1}} \\
+ &\sum_{(y_2,\ldots,y_k,y_{k+1})\in G^{k}} (-1)^{\sum_{i=2}^{k+1} x_i y_i} p_{1\ldots y_k, y_{k+1}}
\end{aligned}$$

Now we wish to factor out the first parameters a_0, a_1 out of the p coordinates. We note that for the coordinates $p_{0\ldots y_k, y_{k+1}}$, every term with a π_0 parameter will have an a_0 parameter, and every term with a π_1 parameter will have a a_1 parameter. Similarly for the $p_{1\ldots y_k, y_{k+1}}$ coordinates, every term with a π_0 parameter will have an a_1 parameter, and every term with a π_1 parameter will have a a_0 parameter. What this means it that when we factor out the a_0, a_1 terms we will end up with a sum of two terms A and B where A is a sum of terms each of which have a π_0 parameter and B is a sum of terms each of which have a π_1 parameter. In equations this looks as follows:

$$RHS = \sum_{(y_2,\ldots,y_k,y_{k+1})\in G^k} a_0 A + a_1 B + \sum_{(y_2,\ldots,y_k,y_{k+1})\in G^k} a_0 B + a_1 A,$$

where,

$$A = (-1)^{\sum_{i=2}^{k+1} x_i y_i} \cdot \pi_0 \cdot \Pi_{j=2}^{k+1} f^{(j)}(\chi^{y_j}), \quad B = (-1)^{\sum_{i=2}^{k+1} x_i y_i} \cdot \pi_1 \cdot \Pi_{j=2}^{k+1} f^{(j)}(\chi^{y_j}).$$

Continuing, we set:

$$\begin{aligned} RHS &= \sum_{(y_2,\ldots,y_k,y_{k+1})\in G^k} a_0 A + a_1 B + a_0 B + a_1 A \\ &= \sum_{(y_2,\ldots,y_k,y_{k+1})\in G^k} (a_0 + a_1) A + (a_0 + a_1) B \\ &= \sum_{(y_2,\ldots,y_k,y_{k+1})\in G^k} (a_0 + a_1)(A + B) \\ &= (a_0 + a_1) \sum_{(y_2,\ldots,y_k,y_{k+1})\in G^k} (A + B) \\ &= (a_0 + a_1) \sum_{(y_2,\ldots,y_k,y_{k+1})\in G^k} (-1)^{\sum_{i=2}^{k+1} x_i y_i} p_{y_2 \ldots y_k, y_{k+1}} \end{aligned}$$

Note that the sum we have now corresponds with the case $m = k$ (where we can relabel the y indices by $z_i = y_{i+1}$ for $i = 1, \ldots, k$.) and so we can apply the inductive hypothesis to get,

$$\begin{aligned} RHS &= (a_0 + a_1) q_{x_2,\ldots,x_k,x_{k+1}} \\ &= \alpha_0 q_{x_2,\ldots,x_k,x_{k+1}} \\ &= q_{0,x_2,\ldots,x_k,x_{k+1}}. \end{aligned}$$

Secondly we fix $x_1 = 1$, then we have:

$$RHS = \sum_{(y_1,\ldots,y_k,y_{k+1})\in G^{k+1}} (-1)^{1\cdot y_1 + \sum_{i=2}^{k+1} x_i y_i} p_{y_1 \ldots y_k, y_{k+1}}$$

Now we sum over the y_1 index to get two sums,

$$\begin{aligned}
RHS = &\sum_{(y_2,\dots,y_k,y_{k+1})\in G^k} (-1)^{1\cdot 0+\sum_{i=2}^{k+1} x_i y_i} p_{0\dots y_k,y_{k+1}} \\
+ &\sum_{(y_2,\dots,y_k,y_{k+1})\in G^k} (-1)^{1\cdot 1+\sum_{i=2}^{k+1} x_i y_i} p_{1\dots y_k,y_{k+1}} \\
= &\sum_{(y_2,\dots,y_k,y_{k+1})\in G^k} (-1)^{\sum_{i=2}^{k+1} x_i y_i} p_{0\dots y_k,y_{k+1}} \\
- &\sum_{(y_2,\dots,y_k,y_{k+1})\in G^k} (-1)^{\sum_{i=2}^{k+1} x_i y_i} p_{1\dots y_k,y_{k+1}}
\end{aligned}$$

Now we wish to factor out the first parameters a_0, a_1 out of the p coordinates. We note that for the coordinates $p_{0\dots y_k,y_{k+1}}$, every term with a π_0 parameter will have an a_0 parameter, and every term with a π_1 parameter will have a a_1 parameter. Similarly for the $p_{1\dots y_k,y_{k+1}}$ coordinates, every term with a π_0 parameter will have an a_1 parameter, and every term with a π_1 parameter will have a a_0 parameter. What this means it that when we factor out the a_0, a_1 terms we will end up with a sum of two terms A and B where A is a sum of terms each of which have a π_0 parameter and B is a sum of terms each of which have a π_1 parameter. In equations this looks as follows:

$$RHS = \sum_{(y_2,\dots,y_k,y_{k+1})\in G^k} a_0 A + a_1 B - \sum_{(y_2,\dots,y_k,y_{k+1})\in G^k} a_0 B + a_1 A,$$

where,

$$A = (-1)^{\sum_{i=2}^{k+1} x_i y_i} \cdot \pi_0 \cdot \Pi_{j=2}^{k+1} f^{(j)}(\chi^{y_j}), \quad B = (-1)^{\sum_{i=2}^{k+1} x_i y_i} \cdot \pi_1 \cdot \Pi_{j=2}^{k+1} f^{(j)}(\chi^{y_j}).$$

Continuing, then we have:

$$\begin{aligned}
RHS = &\sum_{(y_2,\dots,y_k,y_{k+1})\in G^k} a_0 A + a_1 B - a_0 B - a_1 A \\
= &\sum_{(y_2,\dots,y_k,y_{k+1})\in G^k} (a_0 - a_1)A - (a_0 - a_1)B \\
= &\sum_{(y_2,\dots,y_k,y_{k+1})\in G^k} (a_0 + a_1)(A - B) \\
= &(a_0 - a_1) \sum_{(y_2,\dots,y_k,y_{k+1})\in G^k} (A - B)
\end{aligned}$$

Here we take a quick detour, before we can proceed. We can rewrite every $q_{x_2,\dots,x_{k+1}} = r_i \varepsilon^2_{x_2} \cdots \varepsilon^m_{x_m}$ coordinate such that the r_i parameter is expressed

in terms of the original parameters π_0, π_1, therefore we can rewrite our q coordinates as either $(\pi_0 + \pi_1)\varepsilon^2_{x_2} \cdots \varepsilon^m_{x_m}$ or $(\pi_0 - \pi_1)\varepsilon^2_{x_2} \cdots \varepsilon^m_{x_m}$, where the first expression corresponds to the case where r_0 is present and the second expression corresponds to the case where r_1 is present. In either case we can expand the product and get the expression $A + B$, where A and B are as defined above. Now, if we switch the sign in $A + B$ so that we get $A - B$, we see that if we were to factor the π_0, π_1 parameters out again then in the first case we would end up with a $(\pi_0 - \pi_1) = r_1$ term, and in the second case we would end up with the a $(\pi_0 + \pi_1) = r_0$ term. Therefore the effect of changing the sign in $A + B$ to $A - B$ was to change our q coordinate from $q_{x_2,\ldots,x_{k+1}} = r_i\varepsilon^2_{x_2} \cdots \varepsilon^m_{x_m}$ to $r_{i+1}\varepsilon^2_{x_2} \cdots \varepsilon^m_{x_m}$.
Now we apply the inductive hypothesis and move on,

$$\begin{aligned} RHS &= (a_0 - a_1) \sum_{(y_2,\ldots,y_k,y_{k+1}) \in G^k} (A - B) \\ &= \alpha_1 r_{i+1}\varepsilon^2_{x_2} \cdots \varepsilon^m_{x_m} \\ &= q_{1,x_2,\ldots,x_k,x_{k+1}}. \end{aligned}$$

In either case for x_1, we see that the claim holds, therefore we have validated the inductive step.
With the base case and the inductive step shown, the proof is now complete by mathematical induction.

□

Therefore for a claw tree T with an arbitrary number of leaves, we can find its Phylogenetic Invariants by firstly computing the Gröbner basis of the toric ideal generated by the Fourier coordinates using the algorithm 6 and secondly writing the binomials generating the toric ideal in terms of leaf probabilities by applying the inverse Discrete Fourier Transform 31.

7 Conclusion

In this paper we have presented a method of obtaining phylogenetic invariants of claw trees on m leaves, where we have also assumed the Jukes-Cantor model of evolution, and we assumed that our nodes can take $k = 2$ states.

Theorem 24 of *Sturmfels* and *Sullivant* [2], reduces the computation of phylogenetic invariants for more general trees to calculating the phylogenetic invariants of the associated claw trees. Therefore combining Sturmfels and Sullivant's results with the findings presented in this paper, the phylogenetic invariants for a wide variety of trees can be calculated.

A Example - Calculations for a 3 leaf tree with 2 hidden nodes

Below we present a calculation of a tree with two hidden nodes, one root and one internal node, for the interested reader. This is an explicit calculation of the tree considered in *Eriksson*, *Ranestad*, *Sturmfels*, and *Sullivant* [8].

Let T be the following tree with 3 leaves:

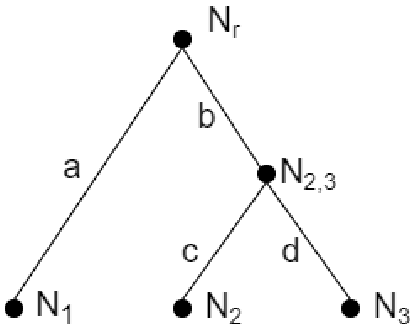

Figure 4: A phylogenetic tree with 3 leaves, 1 hidden node and 1 root

We make the following assumptions about the model: in this model we consider four states $\{A, T, C, G\}$, we assume that the root distribution is uniform, and we consider the Jukes-Cantor model.

The Jukes-Cantor DNA model of evolution assumption gives us the following structure on the transition matrices:

$$M_a = \begin{pmatrix} a_0 & a_1 & a_1 & a_1 \\ a_1 & a_0 & a_1 & a_1 \\ a_1 & a_1 & a_0 & a_1 \\ a_1 & a_1 & a_1 & a_0 \end{pmatrix}, \qquad M_b = \begin{pmatrix} b_0 & b_1 & b_1 & b_1 \\ b_1 & b_0 & b_1 & b_1 \\ b_1 & b_1 & b_0 & b_1 \\ b_1 & b_1 & b_1 & b_0 \end{pmatrix},$$

$$M_c = \begin{pmatrix} c_0 & c_1 & c_1 & c_1 \\ c_1 & c_0 & c_1 & c_1 \\ c_1 & c_1 & c_0 & c_1 \\ c_1 & c_1 & c_1 & c_0 \end{pmatrix}, \qquad M_d = \begin{pmatrix} d_0 & d_1 & d_1 & d_1 \\ d_1 & d_0 & d_1 & d_1 \\ d_1 & d_1 & d_0 & d_1 \\ d_1 & d_1 & d_1 & d_0 \end{pmatrix}.$$

Since these are transition matrices, we can deduce the following relationships between our parameters:

$$a_0 = 1 - 3a_1, \qquad b_0 = 1 - 3b_1, \qquad c_0 = 1 - 3c_1, \qquad d_0 = 1 - 3d_1.$$

We that the root distribution $\pi = (\pi_A, \pi_T, \pi_C, \pi_G)$ is uniform, i.e. $\pi = \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right)$. Therefore our total number of parameters of this model is 4. The model under study is group-based, therefore we need to chose an abelian group G of order 4 which we'll use to label our 4 states with. Here we will use $G = \mathbb{Z}_2 \times \mathbb{Z}_2$. In practice this means that we relabel each element of G with one of the four bases A, T, C, G, therefore $A = (0,0)$, $T = (1,0)$, $C = (0,1)$, $G = (1,1)$. This group identification will be useful later when we wish to change parameters.

To write down the probability polynomials we can condition on the probability of the root being in one of the four states A, T, C, G and use the independence assumption to write the probability of observing any particular state as a product of our parameters. For example, the probability that we observe the state AAA is given by the polynomial:

$$\begin{aligned}
p_{AAA} &= \mathbb{P}(X_1 = A, X_2 = A, X_3 = A | X_r = A) \\
&+ \mathbb{P}(X_1 = A, X_2 = A, X_3 = A | X_r = T) \\
&+ \mathbb{P}(X_1 = A, X_2 = A, X_3 = A | X_r = C) \\
&+ \mathbb{P}(X_1 = A, X_2 = A, X_3 = A | X_r = G) \\
&= \pi_A(a_0 b_0 c_0 d_0 + 3a_0 b_1 c_1 d_1) \\
&+ \pi_T(a_1 b_0 c_1 d_1 + a_1 b_1 c_0 d_0 + 2a_1 b_1 c_1 d_1) \\
&+ \pi_C(a_1 b_0 c_1 d_1 + a_1 b_1 c_0 d_0 + 2a_1 b_1 c_1 d_1) \\
&+ \pi_G(a_1 b_0 c_1 d_1 + a_1 b_1 c_0 d_0 + 2a_1 b_1 c_1 d_1) \\
&= \frac{1}{4}(a_0 b_0 c_0 d_0 + 3a_0 b_1 c_1 d_1) + \frac{3}{4}(a_1 b_0 c_1 d_1 + a_1 b_1 c_0 d_0 + 2a_1 b_1 c_1 d_1) \\
&= \frac{1}{4}(a_0 b_0 c_0 d_0 + 3a_0 b_1 c_1 d_1 + 3a_1 b_0 c_1 d_1 + 3a_1 b_1 c_0 d_0 + 6a_1 b_1 c_1 d_1)
\end{aligned}$$

In total we will have 64 such polynomials p_{ijk} as for each of the three leaves we have 4 choices of base. Since these polynomials all represent probabilities, they will sum to 1, i.e. $\sum_{i\in G}\sum_{j\in G}\sum_{k\in G} p_{ijk} = 1$. We now wish to find the transformed probabilities q_{ijk} and find their Gröbner basis. To do this we use the Discrete Fourier Transform using Equation 1 to find the transformed parameters in terms of the model parameters and substitute them into the leaf probabilities p_{ijk}.
In our example we have four functions $f : G\mathbb{C}^{\times}$, one for each edge in our tree, $f^{(N_1)}, f^{(N_{2,3})}, f^{(N_2)}, f^{(N_3)}$ explicitly defined as follows for $g, h \in G$:

$$f^{(N_1)}(g) = \begin{cases} a_0 & \text{if } g = A \\ a_1 & \text{otherwise} \end{cases}, \qquad f^{(N_{2,3})}(g) = \begin{cases} b_0 & \text{if } g = A \\ b_1 & \text{otherwise} \end{cases},$$

$$f^{(N_2)}(g) = \begin{cases} c_0 & \text{if } g = A \\ c_1 & \text{otherwise} \end{cases}, \qquad f^{(N_3)}(g) = \begin{cases} d_0 & \text{if } g = A \\ d_1 & \text{otherwise.} \end{cases}$$

Below we present the dual group table for $G = \mathbb{Z}_2 \times \mathbb{Z}_2$:

	A—(0,0)	T—(1,0)	C—(0,1)	G—(1,1)
χ_0	1	1	1	1
χ_1	1	-1	1	-1
χ_2	1	1	-1	-1
χ_3	1	-1	-1	1

Now we can use the Discrete Fourier Transform equation to compute the new parameters. We go through the calculation for edge a down below, calculations for all other edges are analogous:

$$\begin{aligned}
\alpha_0 &= f^{(\hat{N_1})}(\chi_0) \\
&= \sum_{g \in G} \chi_0(g) f^{(N_1)}(g) \\
&= \chi_0(A) f^{(N_1)}(A) + \chi_0(T) f^{(N_1)}(T) + \chi_0(C) f^{(N_1)}(C) + \chi_0(G) f^{(N_1)}(G) \\
&= a_0 + a_1 + a_1 + a_1 \\
&= a_0 + 3a_1
\end{aligned}$$

$$\begin{aligned}
\alpha_1 = f^{(\hat{N_1})}(\chi_1) &= \sum_{g \in G} \chi_1(g) f^{(N_1)}(g) \\
&= \chi_1(A) f^{(N_1)}(A) + \chi_1(T) f^{(N_1)}(T) + \chi_1(C) f^{(N_1)}(C) + \chi_1(G) f^{(N_1)}(G) \\
&= a_0 - a_1 + a_1 - a_1 \\
&= a_0 - a_1
\end{aligned}$$

One can check that $\alpha_1 = f^{(\hat{N_1})}(\chi_2) = f^{(\hat{N_1})}(\chi_3)$. Recalling the definition of the Discrete Fourier Transform, our function f can be any function which maps from G to $\mathbb{C}$, therefore we can also apply the Fourier Transform to our root distribution by defining a function $\pi : G \to \mathbb{C}$ such that $\pi(A) = \pi(T) = \pi(C) = \pi(G) = \frac{1}{4}$ and using Equation 1 to get:

$$\begin{aligned}
r_0 = \hat{\pi}(\chi_0) &= \sum_{g \in G} \chi_0(g) \pi(g) \\
&= \chi_0(A)\pi(A) + \chi_0(T)\pi(T) + \chi_0(C)\pi(C) + \chi_0(G)\pi(G) \\
&= \frac{1}{4} + \frac{1}{4} + \frac{1}{4} + \frac{1}{4} \\
&= 1
\end{aligned}$$

$$
\begin{aligned}
r_1 = \hat{\pi}(\chi_1) &= \sum_{g \in G} \chi_1(g)\pi(g) \\
&= \chi_1(A)\pi(A) + \chi_1(T)\pi(T) + \chi_1(C)\pi(C) + \chi_1(G)\pi(G) \\
&= \frac{1}{4} - \frac{1}{4} + \frac{1}{4} - \frac{1}{4} \\
&= 0
\end{aligned}
$$

Again, one can check that $r_1 = \hat{\pi}(\chi_2) = \hat{\pi}(\chi_3)$. The complete set of transformed parameters is given below:

$$
\begin{aligned}
r_0 &= 1, & \alpha_0 &= a_0 + 3a_1, & \beta_0 &= b_0 + 3b_1, & \gamma_0 &= c_0 + 3c_1, & \delta_0 &= d_0 + 3d_1, \\
r_1 &= 0, & \alpha_1 &= a_0 - a_1, & \beta_1 &= b_0 - b_1, & \gamma_1 &= c_0 - c_1, & \delta_1 &= d_0 - d_1.
\end{aligned}
$$

The same information can be elegantly encoded in matrix form as $\epsilon = H\mathbf{e}$, where ϵ is the new set of parameters (represented as a vector), H is the Hadamard matrix encoding the information on the Fourier Transform, and $\mathbf{e}$ is the old set of parameters. An example for the parameters associated with edge a is given below:

$$
\begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_1 \\ \alpha_1 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_1 \\ a_1 \end{pmatrix}. \tag{7}
$$

We now refer to the result of *Theorem 6* from *Toric Ideals of Phylogenetic Invariants* [2]. For our phylogenetic tree T that we are considering the Fourier Transform of coordinates p_{ijk} has the form:

$$
q_{ijk} = q(\chi_1, \chi_2, \chi_3) = \hat{\pi}(\chi_1\chi_2\chi_3) \cdot \prod_{v \in \mathcal{V}(T) \setminus \{r\}} \widehat{f^{(v)}}\Big(\prod_{l \in \Lambda(v)} \chi_l\Big)
$$

where $\mathcal{V}(T)$ is the set of vertices of our tree, and $\Lambda(v)$ is the set of leaves which share the vertex v as a common ancestor, e.g. $\Lambda(N_1) = \{N_1\}$, $\Lambda(N_3) = \{N_3\}$, $\Lambda(N_{2,3}) = \{N_2, N_3\}$. The products of dual group elements inside the functions $\hat{\pi}$ and $\widehat{f^{(v)}}$ is nothing new, it simply means that we should multiply the group elements using the group operation until we get a single element $\chi \in \hat{G}$ at which point we refer to our definitions. The best way to see how these definitions come together is with an example, so below we present a computation of q_{AAA}:

$$\begin{aligned} q_{AAA} = q(\chi_0,\chi_0,\chi_0) &= \hat{\pi}(\chi_0\chi_0\chi_0)\cdot \prod_{v\in\mathcal{V}(T)\setminus\{r\}} \widehat{f^{(v)}}(\prod_{l\in\Lambda(v)}\chi_l) \\ &= \hat{\pi}(\chi_0)\cdot\left(\widehat{f^{(N_1)}}(\chi_0)\widehat{f^{(N_{2,3})}}(\chi_0\chi_0)\widehat{f^{(N_2)}}(\chi_0)\widehat{f^{(N_3)}}(\chi_0)\right) \\ &= 1\cdot\left(\alpha_0\widehat{f^{(N_{2,3})}}(\chi_0)\gamma_0\delta_0\right) \\ &= \alpha_0\beta_0\gamma_0\delta_0 \end{aligned}$$

and all other q_{ijk} are analogously calculated.
There exists a lot of symmetry in our set of 64 original coordinates p_{ijk}, in fact there are only 5 distinct polynomials which we write below:

$$\begin{aligned} p_{123} &= p_{AAA} = p_{TTT} = p_{CCC} = p_{GGG} \\ p_{12} &= p_{AAT} = p_{AAC} = \cdots = p_{GGC} \\ p_{13} &= p_{ATA} = p_{ACA} = \cdots = p_{GCG} \\ p_{23} &= p_{TAA} = p_{CAA} = \cdots = p_{CGG} \\ p_{dis} &= p_{ATC} = p_{ATG} = \cdots = p_{GCT} \end{aligned}$$

where we have indexed the unique coordinates by the same convention as in (Erikson et al) [8]. Here p_{123} is the probability of observing the same base at all three leaves, pij is the probability of observing the same base at leaves i and j and p_{dis} is the probability of observing different bases in all three leaves.

Similarly for our transformed coordinates q_{ijk} we see a high degree of symmetry; most of the original coordinates p_{ijk} map to 0 under the Fourier Transform whenever we choose $i,j,k \in G$ such that the composition $\chi_1\chi_2\chi_3 \neq \chi_0$ as then $\hat{\pi}(\chi_1\chi_2\chi_3) = r_1 = 0$ and out transformed coordinate $q_{ijk} = 0$. It turns out that there are only 5 distinct non-zero transformed coordinates, which we write below:

$$\begin{aligned} &q_{0000} = \alpha_0\beta_0\gamma_0\delta_0 = q_{AAA} \\ &q_{0011} = \alpha_0\beta_0\gamma_1\delta_1 = q_{ATT} = q_{ACC} = q_{AGG} \\ &q_{1101} = \alpha_1\beta_1\gamma_0\delta_1 = q_{TAT} = q_{CAC} = q_{GAG} \\ &q_{1110} = \alpha_1\beta_1\gamma_1\delta_0 = q_{TTA} = q_{CCA} = q_{GGA} \\ &q_{1111} = \alpha_1\beta_1\gamma_1\delta_1 = q_{TCG} = q_{TGC} = q_{CTG} = q_{CGT} = q_{GTC} = q_{GCT} \end{aligned}$$

where we have indexed the unique coordinates by the same convention as in *(Erikson et al)* [8], that is, by the subforest of our tree, where a subforest is defined as any subgraph of the tree T which is a collection of trees (a forest) whose leaves are leaves of the original tree. In general q_{abcd} represents the graph which contains those edges a, b, c, d of the original tree which have a 1 in the subscript in there position. For example q_{0000} corresponds to the empty subtree, and q_{1101} corresponds to the subtree which contains edges a, b, d and whose leaves are N_1 and N_3. We see here that the more conventional labelling used in *Toric Ideals of Phylogenetic Invariants* [2] and the labelling convention used in *(Erikson et al)* [8] are equivalent and are a matter of preference.

With our set of transformed coordinates we can now find the Phylogenetic Invariant as a toric variety! To do so we utilise the following Macaulay2 code which can be found here in the Appendix B.
The output $o3$ gives us the equation which defines our Phylogenetic Invariant in transformed coordinates:

$$q_{0000}q_{1111}^2 - q_{0011}q_{1101}q_{1110} = 0 \tag{8}$$

Now all we have to do is change back into our original coordinates p_{ijk} and express Equation 8 in terms of p_{ijk} to obtain the Phylogenetic Invariant of our model. We utilise the fact that the change between p_{ijk} and q_{ijk} is given by the matrix equation $\mathbf{q} = H_4\mathbf{p}$, where $\mathbf{q}$ is the 64×1 vector of q_{ijk} coordinates, $\mathbf{p}$ is the 64×1 vector of p_{ijk}, and $\hat{H}$ is the 64×64 Hadamard matrix given by matrix built out of the 4×4 Hadamard matrix H we encountered in Equation 7. Built recursively as follows:

$$H_1 = \begin{pmatrix} H & H \\ H & -H \end{pmatrix}, \qquad H_2 = \begin{pmatrix} H_1 & H_1 \\ H_1 & -H_1 \end{pmatrix},$$
$$H_3 = \begin{pmatrix} H_2 & H_2 \\ H_2 & -H_2 \end{pmatrix}, \qquad H_4 = \begin{pmatrix} H_3 & H_3 \\ H_3 & -H_3 \end{pmatrix}.$$

We can implement this in Macaulay2 with the following code which can be found in the Appendix B.
Now noting the following correspondence, $x_0 = p_{123},\ x_1 = p_{12},\ x_2 = p_{13},\ x_3 = p_{23},\ x_4 = p_{dis}$, which follows from the way we have constructed our vector p in Macaulay2. With this identification we can now write the Phylogenetic Invariant for this example in terms coordinates p_{ijk}:

$$\begin{aligned} 0 = &-p_{123}p_{12}p_{13} + p_{12}^2p_{13} + p_{12}p_{13}^2 - p_{123}p_{12}p_{23} + p_{12}^2p_{23} - p_{123}p_{13}p_{23} \\ &- p_{12}p_{13}p_{23} + p_{13}^2p_{23} + p_{12}p_{23}^2 + p_{13}p_{23}^2 + p_{123}^2p_{dis} - p_{12}^2p_{dis} - p_{13}^2p_{dis} \\ &- p_{23}^2p_{dis} + p_{123}p_{dis}^2 - p_{12}p_{dis}^2 - p_{13}p_{dis}^2 - p_{23}p_{dis}^2 + 2p_{dis}^3 \end{aligned}$$

B Macaulay2 code for all examples

```
i1 : loadPackage(FourTiTwo)
i2 : A = matrix"1,0,0,1,0,1,1,0;
               0,1,1,0,1,0,0,1;
               1,1,1,1,0,0,0,0;
               0,0,0,0,1,1,1,1;
               1,1,0,0,1,1,0,0;
               0,0,1,1,0,0,1,1;
               1,0,1,0,1,0,1,0;
               0,1,0,1,0,1,0,1"
i3 : toricMarkov(A)
o3 = | 1 -1 0 0 0 0 -1 1 |
     | 1 0 -1 0 0 -1 0 1 |
     | 1 0 0 -1 -1 0 0 1 |
o3 : Matrix ZZ^3  <- ZZ^8

i1 : H = matrix"1,1;1,-1"
i2 : for i from 0 to 1 do H = matrix‘‘H,H;H,-H"
i3 : R = ZZ[x\_0..x\_7]
i4 : n = \{x\_0,x\_1,x\_2,x\_3,x\_4,x\_5,x\_6,x\_7\}
i5 : p = vector(n)
i6 : q = H*p
i7 : (q000, q001, q010, q011, q100, q101, q110, q111) =
     (q_0, q_1, q_2, q_3, q_4, q_5, q_6, q_7)
i8 : q000*q111 - q001*q110
o8 = -4x_1x_2 + 4x_0x_3 -4x_1x_4 + 4x_0x_5
     + 4x_3x_6 + 4x_5x_6 - 4x_2x_7 - 4x_4x_7
o8 : R
i9 : q000*q111 - q010*q101
o9 = - 4x_1x_2 + 4x_0x_3 - 4x_2x_4 + 4x_3x_5
     + 4x_0x_6 + 4x_5x_6 - 4x_1x_7 - 4x_4x_7
o9 : R
i10 : q000*q111 - q100*q011
o10 = - 4x_1x_4 - 4x_2x_4 + 4x_0x_5 + 4x_3x_5
      + 4x_0x_6 + 4x_3x_6 - 4x_1x_7 - 4x_2x_7
o10 : R
```

```
i1  : loadPackage(FourTiTwo)
i2  : A = matrix"1,0,0,1,0,1,1,0,0,1,1,0,1,0,0,1;
                 0,1,1,0,1,0,0,1,1,0,0,1,0,1,1,0;
                 1,1,1,1,1,1,1,1,0,0,0,0,0,0,0,0;
                 0,0,0,0,0,0,0,0,1,1,1,1,1,1,1,1;
                 1,1,1,1,0,0,0,0,1,1,1,1,0,0,0,0;
                 0,0,0,0,1,1,1,1,0,0,0,0,1,1,1,1;
                 1,1,0,0,1,1,0,0,1,1,0,0,1,1,0,0;
                 0,0,1,1,0,0,1,1,0,0,1,1,0,0,1,1;
                 1,0,1,0,1,0,1,0,1,0,1,0,1,0,1,0;
                 0,1,0,1,0,1,0,1,0,1,0,1,0,1,0,1”
i3  : toricMarkov(A)
o3 = | 0 0 0 0 0 0 0 0 0 0 1 -1 -1 1 0 0 |
     | 0 0 0 0 0 0 0 0 0 0 1 -1 0 0 -1 1 0 |
     | 0 0 0 0 0 0 0 0 0 1 0 -1 0 0 -1 0 1 |
                        ...
     | 1 0 0 0 0 -1 0 0 0 0 -1 0 0 0 0 1 |
     | 1 0 0 0 0 0 -1 0 0 0  0 0 0 0 1 0 |
     | 1 0 0 0 0 0 -1 0 0 -1 0 0 0 0 0 1 |
o3  : Matrix ZZ^{30}  <- ZZ^{16}}
```

```
i1  : H = matrix"1,1;1,-1”
i2  : for i from 0 to 2 do H = matrix‘‘H,H;H,-H”
i3  : R = ZZ[x_0..x_15]
i4  : n = {x_0,x_1,x_2,x_3,x_4,x_5,x_6,x_7,
           x_8,x_9,x_10,x_11,x_12,x_13,x_14,x_15}
i5  : p = vector(n)
i6  : q = H*p
```

```
i1  : loadPackage(FourTiTwo)
i2  : A = matrix ”1,1,0,0,0;
                  0,0,1,1,1;
                  1,1,0,0,0;
                  0,0,1,1,1;
                  1,0,1,0,0;
                  0,1,0,1,1;
                  1,0,0,1,0;
                  0,1,1,0,1”
i3  : toricMarkov(A)
o3 = | 1 -1 -1 -1 2 |
o3  : Matrix ZZ^1  <— ZZ^5

i5  : for i from 0 to 3 do H = matrix”H,H;H,-H”
i6  : R = ZZ[x_0..x_4]
i7  : l = (x_0,x_1,x_1,x_1,x_2,x_3,x_4,x_4,
           x_2,x_4,x_3,x_4,x_2,x_4,x_4,x_3,
           x_3,x_2,x_4,x_4,x_1,x_0,x_1,x_1,
           x_4,x_2,x_3,x_4,x_4,x_2,x_4,x_3)
i8  : r = reverse l}
i9  : n = splice(l,r)
i10 : n = toList n
i11 : p = vector(n)
    o11 = | x_0 |
          | x_1 |
          | x_1 |
          | x_1 |
            ...
          | x_1 |
          | x_1 |
          | x_1 |
          | x_0 |
i12 : q = H*p
i13 : (q0000, q0011, q1101, q1110, q1111) = (q_0, q_1, q_2, q_3, q_4)
i14 : PI = q0000*q1111^2 - q0011*q1101*q1110
o14 = -1024x_0x_1x_2 + 1024x_1^2x_2 + 1024x_1x_2^2 - 1024x_0x_1x_3
      + 1024x_1^2x_3 - 1024x_0x_2x_3  - 1024x_1x_2x_3 + 1024x_2^2x_3
      + 1024x_1x_3^2 + 1024x_2x_3^2 + 1024x_0^2x_4 - 1024x_1^2x_4
      -1024x_2^2x_4 - 1024x_3^2x_4  + 1024x_0x_4^2 - 1024x_1x_4^2
      - 1024x_2x_4^2 - 1024x_3x_4^2  + 2048x_4^3
o14 : R
```

References

[1] Julia Chifman; Sonja Petrović. 'Toric ideals of phylogenetic invariants for the general group-based model on claw trees $K_{1,n}$'. In: *Algebraic Biology*. Feb. 2007. URL: https://api.semanticscholar.org/CorpusID:1513816 (visited on 15/08/2024).

[2] Bernd Sturmfels; Seth Sullivant. 'Toric ideals of phylogenetic invariants'. In: *Journal of Computational Biology: a Journal of Computational Molecular Cell Biology* 12 (Feb. 2004), pp. 204–228. URL: https://api.semanticscholar.org/CorpusID:5539512 (visited on 05/07/2024).

[3] David Cox; John Little; Donal O'Shea. *Ideals, Varieties, and Algorithims*. Springer, Nov. 2015. URL: https://link.springer.com/book/10.1007/978-3-319-16721-3 (visited on 15/08/2024).

[4] Dimitra Kosta; Apostolos Thoma; Marius Vladoiu. 'On the strongly robustness property of toric ideals'. In: *Journal of Algebra* 616 (June 2022), pp. 1–25. URL: https://doi.org/10.1016/j.jalgebra.2022.11.002 (visited on 15/08/2024).

[5] Dimitra Kosta; Kaie Kubjas. 'Geometry of Group Based Models'. In: *arXiv* (Aug. 2017). URL: https://arxiv.org/abs/1705.09228 (visited on 01/08/2024).

[6] Steven Evans; Terrance Speed. 'Invariants of Some Probability Models Used In Phylogenetic Inference'. In: *The Annals of Statistics* 21 (Dec. 1993), pp. 355–377. URL: https://www.jstor.org/stable/i312925 (visited on 16/09/2024).

[7] Lior Pachter; Bernd Sturmfels. *Algebraic Statistics for Computational Biology*. Cambridge University Press, Aug. 2005. URL: https://www-cambridge-org.eux.idm.oclc.org/core/books/algebraic-statistics-for-computational-biology/2E5CCE6BB6751EB7423EE3D2BF40EBFF (visited on 15/09/2024).

[8] Nicholas Eriksson; Kristian Ranestad; Bernd Sturmfels; Seth Sullivant. 'Phylogenetic Algebraic Geometry'. In: *arXiv: Algebraic Geometry* (July 2004). URL: https://arxiv.org/abs/math/0407033 (visited on 05/07/2024).