Bulletin of Mathematical Biology (2024) 86:90
https://doi.org/10.1007/s11538-024-01314-z

Society for Mathematical Biology

ORIGINAL ARTICLE



Dimensions of Level-1 Group-Based Phylogenetic Networks

Elizabeth Gross[1] · **Robert Krone**[2] · **Samuel Martin**[3]

Received: 15 November 2023 / Accepted: 15 May 2024 / Published online: 17 June 2024
© The Author(s) 2024

Abstract
Phylogenetic networks represent evolutionary histories of sets of taxa where horizontal evolution or hybridization has occurred. Placing a Markov model of evolution on a phylogenetic network gives a model that is particularly amenable to algebraic study by representing it as an algebraic variety. In this paper, we give a formula for the dimension of the variety corresponding to a triangle-free level-1 phylogenetic network under a group-based evolutionary model. On our way to this, we give a dimension formula for codimension zero toric fiber products. We conclude by illustrating applications to identifiability.

Keywords Phylogenetic networks · Markov models of evolution · Group-based models · Dimension

1 Introduction

In evolutionary biology, phylogenetic networks are graphs used to represent the evolutionary history of a set of taxa or species. In molecular phylogenetics, these graphs are usually paired with a statistical model where the graph is a combinatorial parameter of the model. In this work, we focus on *network-based Markov models*. In particular, fixing a directed graph $\mathcal{N}$ with n leaves, i.e. a network, the associated network-based Markov model is the image of a polynomial parameterization in the space of probability distributions over the sample space, which commonly in applications is $\{A, G, C, T\}^n$ where A, G, C, T are the four-nucleic bases.

When understanding such models, the overarching goal is to be able to infer phylogenetic networks from molecular sequence data. To be able to do this, we must

✉ Samuel Martin
samuel.martin99@gmail.com

Elizabeth Gross
egross@hawaii.edu

[1] Department of Mathematics, University of Hawai'i at Mānoa, Mānoa, HI, USA

[2] Department of Mathematics, UC Davis, Davis, CA, USA

[3] Earlham Institute, Norwich, UK

first determine whether the model is identifiable from the observed data. By representing phylogenetic models as geometric objects called varieties, such questions can be reframed in terms of geometry, that is, we would like to know whether varieties representing distinct phylogenetic network models are themselves distinct. One of the first geometric descriptions we can give of a variety is its dimension. In Theorems 1 and 2, we give dimension results for all level-1, triangle-free phylogenetic networks under a class of Markov models called group-based models, and in Section 6 we give some identifiability results that follow easily from our dimension formula.

As described above, we are interested in the geometry of network-based Markov models, in particular, their dimensions. Such work is along the lines of Sturmfels and Sullivant (2005), Eriksson et al (2005), Allman and Rhodes (2007), Allman and Rhodes (2008), Casanellas and Fernández-Sánchez (2008), Zwiernik and Smith (2011), Casanellas and Fernández-Sánchez (2011), Michałek (2011), Casanellas et al (2017), Michałek and Ventura (2019), and Casanellas et al (2021), which study the geometry of tree-based Markov models. Indeed, by moving to $\mathbb{C}$ and taking Zariski closures, images of the parameterization maps correspond to algebraic varieties whose study can aid in model selection (see Pachter and Strumfels (2005), Drton et al (2009), and Sullivant (2018) for discussions). Popular constraints on the parameter space, such as Jukes-Cantor (JC), Kimura 2-parameter (K2P), and Kimura 3-parameter (K3P) constraints, give rise to a class of models referred to as *group-based models*. Assuming group-based constraints, the varieties associated to tree-based Markov models are toric varieties after a transformation of coordinates first described in Evans and Speed (1993) and Székely et al (1993) (see Sturmfels and Sullivant (2005) for an overview). The dimensions of tree varieties can be understood using tools from toric geometry. While under this same transformation, group-based network varieties have a lower dimensional toric action on them, and thus are T-varieties (see (Cummings et al (2021), Remark 4.1)), these varieties are generally less well understood. *In this paper, we expand our understanding of these varieties by giving a formula for the dimension for all level-1 triangle-free group-based network varieties.*

As described in Sect. 2, a group-based model of evolution is defined with a finite abelian group G and a subgroup B of the automorphism group of G, denoted $\mathrm{Aut}(G)$. In a network-based Markov model, each edge of the network has a transition matrix associated to it, representing the probabilities of each type of nucleotide (usually A, G, C, or T) mutating to another over an evolutionary time interval. The parameters of the model are the entries of these transition matrices along with a mixing parameter for each cycle. In a group-based model, the dimension of the parameter space is cut significantly by placing constraints on the transition matrices. In particular, each nucleotide is identified with an element of G, and the transition probability of a mutation from a to b depends only on $b - a$, reducing the number of free parameters in each matrix to $|G| - 1$. The parameter space is reduced further by identifying the parameters for all elements of G that are in the same B-orbit. If $l + 1$ is the number of B-orbits in G, the number of free parameters for each edge is then l.

For a phylogenetic network $\mathcal{N}$ with m edges and c cycles, the *expected dimension* of the group-based network variety $\dim V_{\mathcal{N}}^{M}$ is $l(m-c)+1$, and Proposition 12 shows that it is indeed an upper bound. The main theorem of this paper shows that most level-1 group-based network varieties have the expected dimension.

Theorem 1 *Let $\mathcal{N}$ be a level-1 triangle-free phylogenetic network with n leaves, m edges, and c cycles. Let G be a finite abelian group of order at least 3 and B a subgroup of $\mathrm{Aut}(G)$. Let $l+1$ be the number of B-orbits in G. Then the group-based network variety $V_{\mathcal{N}}^{(G,B)}$ has dimension $l(m-c)+1$.*

When $G = \mathbb{Z}/2\mathbb{Z}$, certain small phylogenetic networks do not have the expected dimension. In this case, since $\mathrm{Aut}(G)$ is the trivial group, there is only a single group-based model. This is the Cavender-Farris-Neyman (CFN) model, and has biological relevance, so we give the result for this group separately. Note that here we are able to give a full result for level-1 phylogenetic networks.

Theorem 2 *Let $G = \mathbb{Z}/2\mathbb{Z}$ and let $\mathcal{N}$ be a level-1 phylogenetic network with n leaves, m edges, $c_{\geq 5}$ cycles of length at least 5, c_4 4-cycles, and c_3 3-cycles. Then the group-based network variety $V_{\mathcal{N}}^{G}$ has dimension $m-(c_{\geq 5}+2c_4+3c_3)+1$.*

Our main tool for proving these theorems is the toric fiber product. This is an operation on ideals that was first introduced in Sullivant (2006) and generalises the Segre product. One of the first applications was to phylogenetic trees under group-based models, where the ideals of the model are toric fiber products, and the operation corresponds to the graph operation of cutting a tree at an internal edge. To some extent this remains true for phylogenetic networks and allows us to focus our attention on a family of phylogenetic networks called sunlet networks (defined in Sect. 2.1). In Sect. 3 we give a general dimension formula for toric fiber products (Theorem 9) and apply this to phylogenetic trees and networks.

2 Preliminaries

In this section, we lay out the background needed for the paper. In particular, we review group-based models of sequence evolution where the combinatorial parameters are phylogenetic networks, as well as two tools that underlie the proof of our main theorems: tropical geometry for dimension analysis and toric fiber products. The main objects of biological relevance in this paper are phylogenetic networks, and, thus, that is where we begin.

2.1 Phylogenetic Networks

The following network notation and terminology is adapted from Francis et al (2018), Francis and Steel (2015), and Semple (2016).

Definition 1 A *(binary rooted) phylogenetic network* $\mathcal{N}$ on a set X is a rooted, acyclic, directed graph with no parallel edges that satisfies:

- The root vertex has outdegree 2.
- All vertices of outdegree 0 have indegree 1. These vertices are called *leaves* and are labelled by X.
- All other vertices have either indegree 1 and outdegree 2 (called *tree vertices*), or indegree 2 and outdegree 1 (called *reticulation vertices*). The incoming edges of a reticulation vertex are called *reticulation edges*.

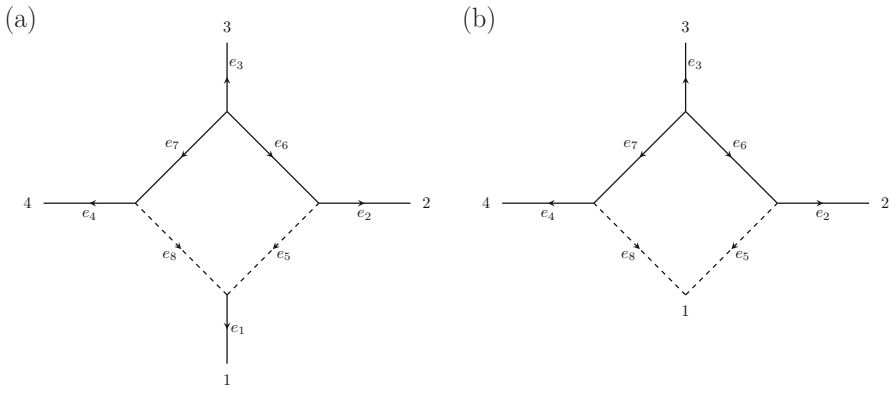


Fig. 1 **a** A leaf-labelled, directed 4-sunlet network, and **b** its corresponding contracted network (right)

A *level-1* phylogenetic network is a phylogenetic network where each cycle in the underlying undirected graph contains exactly one reticulation vertex. A *semi-directed phylogenetic network* is a mixed graph obtained from a phylogenetic network by suppressing the root node and undirecting all tree edges while the reticulation edges remain directed. In a semi-directed phylogenetic network, the reticulation vertices are the vertices of indegree two and level-1 is defined the same as for a rooted phylogenetic network. A triangle-free level-1 semi-directed phylogenetic network is a level-1 semi-directed phylogenetic network where every cycle in the unrooted skeleton has length greater than three. For our work, it will be helpful to reduce the number of edges in a semi-directed phylogenetic network that we consider. To this end we introduce *contracted semi-directed phylogenetic networks*. A contracted semi-directed phylogenetic network is a mixed graph obtained from a semi-directed phylogenetic network by contracting the non-reticulation edge of each reticulation vertex (see for example, Fig. 1). Note that since level-1 networks are tree-child networks, in a contracted level-1 semi-directed phylogenetic network, two distinct reticulation vertices are never identified, and thus each non leaf-adjacent reticulation vertex has indegree 2 and outdegree 2, and each leaf-adjacent reticulation vertex has indegree 2 and outdegree 0. Furthermore, the level-1 condition in a contracted level-1 semi-directed phylogenetic network means that at least one of the outgoing edges of a reticulation vertex is a non-reticulation edge.

Finally, an *n-sunlet network* is the semi-directed phylogenetic network topology with n leaves and a single cycle of length n, where each vertex in the cycle is adjacent to a leaf vertex and one vertex in the cycle is a reticulation vertex. The 4-sunlet network is depicted in Fig. 1. Since an arbitrary level-1 network can be decomposed into a collection of trees and sunlet networks, sunlet networks will play a key role in our study.

2.2 Group-Based Models of Evolution

Fix an abelian group G and a subgroup $B \subset \mathrm{Aut}(G)$. Denote by $B \cdot G$ the set of B-orbits in G and let $|B \cdot G| = l + 1$. For a phylogenetic tree or network $\mathcal{N}$, such a

choice of G and B defines a model of evolution on $\mathcal{N}$. From this model one can derive an algebraic variety, which we will denote $V_{\mathcal{N}}^{(G,B)}$. These varieties are our primary objects of study.

First, let us set up the notation and preliminaries for phylogenetic trees, i.e. phylogenetic networks with no reticulation vertices. For more details on group-based models on trees, see (Sullivant 2018, Section 15.3) and Sturmfels and Sullivant (2005). Let $\mathcal{T}$ be an n-leaf phylogenetic tree, with vertex set, edge set, and leaf set denoted by $\mathcal{V}(\mathcal{T})$, $\mathcal{E}(\mathcal{T})$, and $\mathcal{L}(\mathcal{T})$ respectively. Let $m = |\mathcal{E}(\mathcal{T})|$ be the number of edges in $\mathcal{T}$. A *consistent leaf G-labelling* of $\mathcal{T}$ is a function $\xi : \mathcal{L}(\mathcal{T}) \longrightarrow G$ that satisfies

$$\sum_{v \in \mathcal{L}(\mathcal{T})} \xi(v) = 0.$$

Note that the set of consistent leaf G-labellings depends only on n, and not on the edges of $\mathcal{T}$, so all n-leaf phylogenetic trees share the same set of consistent leaf G-labellings, which has size $|G|^{n-1}$. When G is clear, we will call ξ a consistent leaf labelling.

For a phylogenetic tree $\mathcal{T}$, each edge $e \in \mathcal{E}(\mathcal{T})$ is oriented away from the root vertex. Let $\mathcal{L}(e) \subset \mathcal{L}(\mathcal{T})$ be the set of leaves on the arrow side of e. A consistent leaf labelling ξ of $\mathcal{T}$ induces a consistent edge labelling of $\mathcal{T}$ (also denoted ξ), which is a map $\xi : \mathcal{E}(\mathcal{T}) \longrightarrow G$ given by

$$\xi(e) = \sum_{v \in \mathcal{L}(e)} \xi(v).$$

To each edge e in a phylogenetic tree or network we associate $l + 1$ parameters, denoted a_e^g, where g is a representative of the B-orbit $[g]$. For a tree $\mathcal{T}$ with n leaves, the parameterization in Fourier coordinates (see Sturmfels and Sullivant 2005) of the group-based model on $\mathcal{T}$ is

$$q_{g_1 g_2 \cdots g_n} = \prod_{e \in \mathcal{E}(\mathcal{T})} a_e^{\xi(e)}, \tag{1}$$

where ξ is given by the consistent leaf labelling $g_1, \ldots, g_n$. Index the standard basis of $\mathbb{C}^{m(l+1)}$ with upper indices g for some representatives of the orbits in $B \cdot G$, and lower indices by the edges $e \in \mathcal{E}(\mathcal{T})$, and index the standard basis of $\mathbb{C}^{|G|^{n-1}}$ by the consistent leaf-labellings. The parameterization map is the map

$$\phi_{\mathcal{T}} : \mathbb{C}^{m(l+1)} \to \mathbb{C}^{|G|^{n-1}}$$

where

$$(\phi_{\mathcal{T}}(w))_{g_1 \cdots g_n} = \prod_{e \in \mathcal{E}(\mathcal{T})} w_e^{\xi(e)},$$

for $w \in \mathbb{C}^{m(l+1)}$ and consistent leaf labellings ξ with leaf labels $g_1, \ldots, g_n$. The Zariski closure of the image of this map is called the *phylogenetic variety of* $\mathcal{T}$ *and* (G, B) and is denoted by $V_{\mathcal{T}}^{(G,B)}$.

Now, denote by R the $\mathbb{C}$-algebra $\mathbb{C}[q_{g_1 \cdots g_n} \mid g_1 + \cdots + g_n = 0]$ and by $S_{\mathcal{T}}$ the $\mathbb{C}$-algebra $\mathbb{C}[a_e^g \mid [g] \in B \cdot G,\ e \in \mathcal{E}(\mathcal{T})]$. The parameterization map $\phi_{\mathcal{T}}$ is a morphism of affine varieties, with comorphism given by the $\mathbb{C}$-algebra homomorphism $\psi_{\mathcal{T}} : R \to S_{\mathcal{T}}$ which acts on generators as

$$\psi_{\mathcal{T}}(q_{g_1 \cdots g_n}) = \prod_{e \in \mathcal{E}(\mathcal{T})} a_e^{\xi(e)}.$$

It follows that the vanishing ideal of $V_{\mathcal{T}}^{(G,B)}$, denoted $I_{\mathcal{T}}^{(G,B)}$, is the kernel of $\psi_{\mathcal{T}}$.

We now move from trees to networks. Let $\mathcal{N}$ be a level-1 phylogenetic network with n leaves, m edges, and k reticulation vertices. Since $\mathcal{N}$ is a phylogenetic network, if we remove one of the two reticulation edges for each reticulation vertex, we obtain a phylogenetic tree. We encode a choice of reticulation edge for each reticulation vertex with a vector $\sigma \in \{0, 1\}^k$, and denote the resulting n-leaf phylogenetic tree by $\mathcal{T}_\sigma$. Then the parameterization of our group based model on $\mathcal{N}$ is the map $\phi_{\mathcal{N}} : \mathbb{C}^{m(l+1)} \to \mathbb{C}^{|G|^{n-1}}$ given by

$$\phi_{\mathcal{N}} = \sum_{\sigma \in \{0,1\}^k} \phi_{\mathcal{T}_\sigma} \tag{2}$$

As above, we call the Zariski closure of the image of this map the *phylogenetic variety of* $\mathcal{N}$ *and* (G, B), and denote it $V_{\mathcal{N}}^{(G,B)}$. The vanishing ideal $I_{\mathcal{N}}^{(G,B)}$ of $V_{\mathcal{N}}^{(G,B)}$ is the kernel of the $\mathbb{C}$-algebra homomorphism $\psi_{\mathcal{N}}$ given by

$$\begin{aligned} \psi_{\mathcal{N}} : R &\to S_{\mathcal{N}} \\ q_{g_1 \cdots g_n} &\mapsto \sum_{\sigma \in \{0,1\}^k} \psi_{\mathcal{T}_\sigma}(q_{g_1 \cdots g_n}). \end{aligned} \tag{3}$$

where $S_{\mathcal{N}} = \mathbb{C}[a_e^g \mid [g] \in B \cdot G,\ e \in \mathcal{E}(\mathcal{N})]$ and we identify $S_{\mathcal{T}_\sigma}$ as a subalgebra of $S_{\mathcal{N}}$ in the obvious way.

When $B = \{\mathrm{id}\}$, we call the probabilistic model associated to (G, B) the *general group-based model* for the group G and denote the corresponding variety as $V_{\mathcal{N}}^{G} = V_{\mathcal{N}}^{(G,B)}$. The K3P model is the general group-based model for the Klein-4 group, and the CFN model is the general group-based model for the group $\mathbb{Z}/2\mathbb{Z}$. The pairs (G, B) corresponding to JC and K2P are $(\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}, \mathfrak{S}_3)$ and $(\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}, \mathfrak{S}_2)$, respectively, where $\mathrm{Aut}(\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z})$ is identified with the permutation group $\mathfrak{S}_3$.

In this paper, we are concerned with $\dim V_{\mathcal{N}}^{(G,B)}$ for a general $\mathcal{N}$ and (G, B). In previous work, it is shown that under the CFN, JC, K2P, and K3P models, two phylogenetic network varieties are the same if the two networks have the same underlying semi-directed phylogenetic network (Gross and Long 2018, Gross et al 2021). Here,

we extend these results to all group-based models. This allows us to focus our attention on semi-directed phylogenetic networks.

Lemma 3 *Let G be a finite abelian group and let B be a subgroup of* $\mathrm{Aut}(G)$. *If $\mathcal{N}_1$ and $\mathcal{N}_2$ are two phylogenetic networks with the same underlying semi-directed phylogenetic network, then $V_{\mathcal{N}_1}^{(G,B)} = V_{\mathcal{N}_2}^{(G,B)}$, where equality here means equality as sets.*

Proof Since phylogenetic networks have no parallel edges, it is clear that two phylogenetic networks $\mathcal{N}_1$ and $\mathcal{N}_2$ have the same semi-directed phylogenetic network if and only if their corresponding unrooted networks differ only by the directions of their non-reticulation edges. Therefore it is sufficient to show that suppressing vertices of degree 2 and changing the orientation of a single non-reticulation edge do not affect the model.

First notice that if we take a phylogenetic network $\mathcal{N}_1$ and reorient any collection of the non-reticulation edges to form a new network $\mathcal{N}_2$ (not necessarily a phylogenetic network), the maps described above are still well-defined and so too are the corresponding varieties. Thus, for this proof, we will relax the definition of phylogenetic networks to include such networks, as it will allow us to consider redirecting one edge at a time. Let $\mathcal{N}_1$ and $\mathcal{N}_2$ be two phylogenetic networks that are equal except for the direction of a single non-reticulation edge e, and let ξ_1 and ξ_2 be consistent leaf labellings on $\mathcal{N}_1$ and $\mathcal{N}_2$ respectively, with the property that $\xi_1(v) = \xi_2(v)$ for each leaf v in the skeleton of $\mathcal{N}_1$ and $\mathcal{N}_2$. Then it is clear that $\xi_1(e) = -\xi_2(e)$. This means that $\phi_{\mathcal{N}_2} = \phi_{\mathcal{N}_1} \circ \theta$, where θ is the automorphism of $\mathbb{C}^{m(l+1)}$ given by swapping the coefficients corresponding to a_e^g and a_e^{-g} whenever $g \neq -g$. Since θ is bijective, composing θ with $\phi_{\mathcal{N}_1}$ does not affect the image of $\phi_{\mathcal{N}_1}$, so it follows that $V_{\mathcal{N}_1}^{(G,B)} = V_{\mathcal{N}_2}^{(G,B)}$.

Next let $\mathcal{N}$ be a phylogenetic network with a vertex v of order 2 that has incident edges e_1 and e_2. Let ξ be a consistent leaf labelling of $\mathcal{N}$ with $\xi(e_1) = g$ so that either $\xi(e_2) = g$ or $\xi(e_2) = -g$. Let us suppose that $\xi(e_2) = g$ and note that the proof in the other case is similar. Let $\mathcal{N}'$ be the phylogenetic network got from $\mathcal{N}$ by suppressing v. Denote the new edge of $\mathcal{N}'$ by e' and, without loss of generality, give e' the same orientation as e_1 so that $\xi(e') = g$. Let $\theta : \mathbb{C}^{(m-1)(l+1)} \to \mathbb{C}^{m(l+1)}$ be the map from the parameter space of $\mathcal{N}'$ to the parameter space of $\mathcal{N}$ that is constant on all parameters for edges shared by $\mathcal{N}$ and $\mathcal{N}'$, takes all parameters for edge e' to the corresponding parameter for edge e_1, and sets all parameters corresponding to edge e_2 to 1. Then it is clear that we have

$$\phi_{\mathcal{N}'} = \phi_{\mathcal{N}} \circ \theta,$$

and thus $\mathrm{Im}\,\phi_{\mathcal{N}'} \subseteq \mathrm{Im}\,\phi_{\mathcal{N}}$. On the other hand, consider $\phi_{\mathcal{N}}(w)$ for $w \in \mathbb{C}^{m(l+1)}$. Taking $u \in \mathbb{C}^{(m-1)(l+1)}$ such that $u_{e'}^g = w_{e_1}^g w_{e_2}^g$ for all $[g] \in B \cdot G$ and $u_d^g = w_d^g$ for all edges $d \neq e'$ and all $[g] \in B \cdot G$, we see that $\phi_{\mathcal{N}'}(u) = \phi_{\mathcal{N}}(w)$. It follows that $\mathrm{Im}\,\phi_{\mathcal{N}'} = \mathrm{Im}\,\phi_{\mathcal{N}}$ and thus $V_{\mathcal{N}'}^{(G,B)} = V_{\mathcal{N}}^{(G,B)}$.

□

Note that since the orientation of the non-reticulation edges does not affect the variety, we may choose any orientation for non-reticulation edges, even if this is not consistent with any placement of a root vertex. Thus when considering a phylogenetic network $\mathcal{N}$, we may take the corresponding semi-directed phylogenetic network and arbitrarily assign orientations to each non-reticulation edge to obtain a parameterization of the model.

Example 1 Let G be a finite abelian group and $B = \{\mathrm{id}\}$. Let $\mathcal{N}$ be a 4-sunlet network with leaf labellings, edge orientations, and edge labellings as in Fig. 1. The map $\psi_{\mathcal{N}}$ is given by

$$q_{g_1g_2g_3g_4} \longmapsto a_1^{g_1}a_2^{g_2}a_3^{g_3}a_4^{g_4}a_5^{g_1}a_6^{g_1+g_2}a_7^{g_4} + a_1^{g_1}a_2^{g_2}a_3^{g_3}a_4^{g_4}a_6^{g_2}a_7^{g_1+g_4}a_8^{g_1},$$

where to simplify notation we write a_i^g for $a_{e_i}^g$. Here, the first monomial corresponds to the tree obtained by removing the edge e_8, and the second monomial corresponds to the tree obtained by removing the edge e_5.

To end this section, we show that sunlet networks and contracted sunlet networks have the same corresponding varieties.

Lemma 4 *Let G be a finite abelian group and B a subgroup of $Aut(G)$. Let $\mathcal{N}$ be a sunlet network and let $\mathcal{N}'$ be its contraction. Then $V_{\mathcal{N}}^{(G,B)} = V_{\mathcal{N}'}^{(G,B)}$.*

Proof Let $\mathcal{N}$ be an n -sunlet, so that $\mathcal{N}$ has $2n$ edges, denoted $e_1,\ldots,e_{2n}$, with e_1 the leaf edge adjacent to the reticulation vertex, and e_{n+1}, e_{2n} the two reticulation edges, as in Fig. 1. Let $\phi_{\mathcal{N}}$ denote the parameterization map for $\mathcal{N}$, and let $\phi_{\mathcal{N}'}$ denote the parameterization map for its contraction $\mathcal{N}'$. As before, index the parameter spaces $\mathbb{C}^{2n(l+1)}$ and $\mathbb{C}^{(2n-1)(l+1)}$ by the B-orbits in G and the edges of $\mathcal{N}$ and $\mathcal{N}'$ respectively.

It is clear that $\phi_{\mathcal{N}'} = \phi_{\mathcal{N}} \circ \iota$, where $\iota : \mathbb{C}^{(2n-1)(l+1)} \to \mathbb{C}^{2n(l+1)}$ is given by $\iota(w)_e^g = w_e^g$ for all $g \in B \cdot G$ and $e \in \mathcal{E}(\mathcal{N}')$ (i.e. $e \neq e_1$), and $\iota(w)_{e_1}^g = 1$ for all $g \in B \cdot G$. It follows that $\mathrm{Im}\,\phi_{\mathcal{N}'} \subseteq \mathrm{Im}\,\phi_{\mathcal{N}}$.

On the other hand, let $y = \phi_{\mathcal{N}}(w)$. Let $u \in \mathbb{C}^{(2n-1)(l+1)}$ be given by $u_e^g = w_e^g$ for $e \neq e_{n+1}, e_{2n}$, and $u_{e_{n+1}}^g = w_{e_1}^g w_{e_{n+1}}^g$, and $u_{e_{2n}}^g = w_{e_1}^g w_{e_{2n}}^g$ for all $g \in B \cdot G$. It follows that $\phi_{\mathcal{N}'}(u) = y = \phi_{\mathcal{N}}(w)$, so $\mathrm{Im}\,\phi_{\mathcal{N}} \subseteq \mathrm{Im}\,\phi_{\mathcal{N}'}$. □

In fact, by absorbing the parameters associated to the contracted edge of each reticulation vertex into the corresponding parameters of both reticulation edges, the above lemma can be extended for any level-1 phylogenetic network.

Lemma 5 *Let G be a finite abelian group and B a subgroup of $Aut(G)$. Let $\mathcal{N}$ be a level-1 semi-directed phylogenetic network and let $\mathcal{N}'$ be its contraction. Then $V_{\mathcal{N}}^{(G,B)} = V_{\mathcal{N}'}^{(G,B)}$.* □

2.3 Tropical Geometry

In Sect. 4 we give a lower bound on the dimension of the sunlet varieties by using the tropical geometry results of Draisma (2008). Here we will present the result tailored to our needs.

Let C be a Zariski-closed cone in a complex vector space V of dimension n, and let W be a complex vector space of dimension m such that we have a polynomial map $f : W \to V$ mapping W dominantly to C. A classic result is that the rank of the Jacobian matrix of f at any point in $x \in W$ gives a lower bound on $\dim(C)$ (and equality holds when x is generic). We similarly obtain a bound from the tropicalization of f.

Fix bases of V and W so that we may write $f = (f_b)_{b=1}^n$, where $f_b \in \mathbb{C}[x_1, \ldots, x_m]$ for $b = 1, \ldots n$. Write

$$f_b = \sum_{\alpha \in M_b} c_\alpha x^\alpha$$

for the finite subset $M_b \subset \mathbb{Z}^m_{\geq 0}$ consisting of those α for which $c_\alpha \neq 0$. The tropicalization of f_b is defined as the piece-wise linear function $\mathrm{Trop}(f_b) : \mathbb{R}^m \to \mathbb{R}$ given by

$$\mathrm{Trop}(f_b)(\lambda) := \min_{\alpha \in M_b} \langle \lambda, \alpha \rangle,$$

for $\lambda \in \mathbb{R}^m$. Then $\mathrm{Trop}(f) : \mathbb{R}^m \to \mathbb{R}^n$ is defined as $(\mathrm{Trop}(f_b))_{b=1}^n$.

We will not define here the tropical variety $\mathrm{Trop}(C) \subseteq \mathbb{R}^n$, but we note two relevant facts (see e.g. Maclagan and Sturmfels (2021)):

- $\mathrm{Trop}(C)$ is a polyhedral complex with dimension bounded by $\dim(C)$,
- and $\mathrm{Im}(\mathrm{Trop}(f)) \subseteq \mathrm{Trop}(C)$.

Therefore the Jacobian of $\mathrm{Trop}(f)$ at a point $\lambda \in \mathbb{R}^m$ where the map is differentiable gives a lower bound on $\dim(C)$ (although it is no longer true that equality necessarily holds when λ is generic).

Fix λ such that $\mathrm{Trop}(f)$ is differentiable at λ, meaning that $\mathrm{Trop}(f)$ is linear in an open neighborhood U of λ. Specifically $\mathrm{Trop}(f_b)(\mu) = \langle \mu, \alpha'_b \rangle$ for all $\mu \in U$ where α'_b is the unique vector in M_b that minimizes $\langle \lambda, \alpha'_b \rangle$. Then $\mathrm{Trop}(f)(\mu) = A_\lambda^T \mu$ where A_λ is the $m \times n$ matrix with columns $\alpha'_1, \ldots, \alpha'_n$. (Note that A_λ^T is also the Jacobian matrix of $\mathrm{Trop}(f)$ at λ.) The lemma below follows.

Lemma 6 *(Draisma (2008), Corollary 2.3) Let the notation be as above. Then*

$$\dim C \geq \max_{\lambda \in \mathbb{R}^m} \mathrm{rank}_{\mathbb{R}} A_\lambda.$$

For our purposes, f will be given by the polynomial parameterization map $\phi_{\mathcal{N}}$. Since the variety $V_{\mathcal{N}}^{(G,B)}$ is equal to the Zariski closure of $\phi_{\mathcal{N}}(\mathbb{C}^{|G|^{n-1}})$, and since each polynomial in the parameterization is homogeneous, $V_{\mathcal{N}}^{(G,B)}$ is a closed cone.

2.4 Toric Fiber Products

The toric fiber product is an algebraic operation that takes two homogeneous ideals with compatible multigradings and produces a new homogeneous ideal. It was introduced

in Sullivant (2006) in order to generalise the gluing operation for toric ideals that appear in tree-based models of evolution and elsewhere in algebraic statistics, and further studied in Engström et al (2014) and Kahle and Rauh (2014). More recently toric fiber products were introduced into the geometric modelling setting in Duarte et al (2023). Here, we will introduce the basic objects and recommend that the reader consult (Sullivant 2006) for further details.

Let $r \in \mathbb{N}$ and $s, t \in \mathbb{N}^r$, and let $\mathcal{A} = \{a_1, \ldots, a_r\} \subset \mathbb{Z}^D$ be a linearly independent set for some $D > 0$. Denote the affine semigroup generated by $\mathcal{A}$ by $\mathbb{N}\mathcal{A}$. Let $\mathbb{K}$ be an algebraically closed field and let

$$\mathbb{K}[x] = \mathbb{K}[x_j^i \mid i \in [r],\ j \in [s_i]],$$

and

$$\mathbb{K}[y] = \mathbb{K}[y_k^i \mid i \in [r], k \in [t_i]],$$

be multigraded polynomial rings with multidegree given by $\deg(x_j^i) = \deg(y_k^i) = a_i$ for all $i = 1, \ldots, r$, $j = 1, \ldots, s_i$, and $k = 1, \ldots, t_i$. Note that since the a_i are linearly independent, ideals in $\mathbb{K}[x]$ and $\mathbb{K}[y]$ that are homogeneous with respect to the multigrading are also homogeneous with respect to the total degree. For homogeneous ideals $I \subset \mathbb{K}[x]$ and $J \subset \mathbb{K}[y]$, let $R = \mathbb{K}[x]/I$ and $S = \mathbb{K}[y]/J$ be the corresponding quotient rings, which inherit the multigrading from $\mathbb{K}[x]$ and $\mathbb{K}[y]$ respectively. Let

$$\mathbb{K}[z] = \mathbb{K}[z_{jk}^i \mid i \in [r],\ j \in [s_i], k \in [t_i]]$$

be the polynomial ring with the analogous multigrading (i.e. $\deg z_{jk}^i = a_i$). Let ϕ_{IJ} be the ring homomorphism given by

$$\begin{aligned} \phi_{IJ} : \mathbb{K}[z] &\longrightarrow R \otimes_{\mathbb{K}} S \\ z_{jk}^i &\longmapsto x_j^i \otimes y_k^i. \end{aligned}$$

Definition 2 With notation as above, the *toric fiber product* of I and J is

$$I \times_{\mathcal{A}} J := \ker \phi_{IJ}.$$

Note that when I and J are prime ideals, since the toric fiber product $I \times_{\mathcal{A}} J$ is the kernel of a ring homomorphism into an integral domain, it is a prime ideal. It will be helpful for us to also consider the monomial homomorphism

$$\begin{aligned} \phi_B : \mathbb{K}[z] &\longrightarrow \mathbb{K}[x, y] \\ z_{jk}^i &\longmapsto x_j^i y_k^i, \end{aligned}$$

where we think of B as being the integral matrix of exponent vectors of this map. The ideal $I_B = \ker \phi_B$ is a toric ideal and is given by

$$I_B = \langle \mathrm{Quad}_B \rangle,$$

where

$$\text{Quad}_B = \{z^i_{j_1k_2}z^i_{j_2k_1} - z^i_{j_1k_1}z^i_{j_2k_2} \mid i \in [r],\ 1 \le j_1 < j_2 \le s_i,\ 1 \le k_1 < k_2 \le t_i\},$$

and Quad_B is a Gröbner basis for I_B with respect to any term order that selects the first term (as written above) as the initial term for each quadric (Sullivant 2006, Proposition 10).

We may also define $I \times_{\mathcal{A}} J$ as $\phi_B^{-1}(I + J)$, where we consider I and J as being their natural extensions in $\mathbb{K}[x, y]$. If ω_1 and ω_2 are weight vectors on $\mathbb{K}[x]$ and $\mathbb{K}[y]$ respectively, then we have a natural weight vector (ω_1, ω_2) on $\mathbb{K}[x, y]$, and the pullback $\phi_B^*(\omega_1, \omega_2)$ is a weight vector on $\mathbb{K}[z]$. These weight vectors have the property that for all monomials $z^a \in \mathbb{K}[z]$ we have $\text{wt}_{\phi_B^*(\omega_1,\omega_2)}(z^a) = \text{wt}_{(\omega_1,\omega_2)}(\phi_B(z^a))$.

Let $f \in \mathbb{K}[x]$ be a homogeneous polynomial with respect to the multigrading $\mathbb{N}\mathcal{A}$ and total degree d, so that we may write

$$f = \sum_{u=1}^{v} c_u x^{i_1}_{j_1^u} \cdots x^{i_d}_{j_d^u},$$

with each $j_l^u \in [s_{i_l}]$ and $c_u \in \mathbb{K}$. The upper indices $i_1, \ldots, i_d$ can be written independent of u since f is homogeneous with respect to $\mathbb{N}\mathcal{A}$ and $\mathcal{A}$ is linearly independent. For any $k = (k_1, \ldots, k_d)$ with $k_l \in [t_{i_l}]$ define the *lift* of f by k, denoted $f_k \in \mathbb{K}[z]$ by

$$f_k = \sum_{u=1}^{v} c_u z^{i_1}_{j_1^u k_1} \cdots z^{i_d}_{j_d^u k_d}. \tag{4}$$

For a set $F \subset \mathbb{K}[x]$ define Lift F to be the subset of $\mathbb{K}[z]$ consisting of all possible f_k with $f \in F$. We define Lift G for $G \subset \mathbb{K}[y]$ analogously. Observe that we have

$$\phi_B(f_k) = \sum_{u=1}^{v} c_u x^{i_1}_{j_1^u} \cdots x^{i_d}_{j_d^u} y^{i_1}_{k_1} \cdots y^{i_d}_{k_d} = (y^{i_1}_{k_1} \cdots y^{i_d}_{k_d}) f.$$

Since the weight of each monomial in f_k with respect to $\phi_B^*(\omega_1, \omega_2)$ is equal to the weight with respect to (ω_1, ω_2) of the image of that monomial under ϕ_B, and this is in turn given by the sum of the weight with respect to ω_1 of the corresponding monomial in f and the weight with respect to ω_2 of $y^{i_1}_{k_1} \cdots y^{i_d}_{k_d}$, we have that $\text{in}_{\phi_B^*(\omega_1.\omega_2)}(f_k) = (\text{in}_{\omega_1}(f))_k$. It follows that $\text{in}_{\phi_B^*(\omega_1.\omega_2)}(\langle \text{Lift } F\rangle) = \text{Lift}(\text{in}_{\omega_1}(\langle F\rangle))$, and by symmetry $\text{in}_{\phi_B^*(\omega_1.\omega_2)}(\langle \text{Lift } G\rangle) = \text{Lift}(\text{in}_{\omega_2}(\langle G\rangle))$.

One of the key results on toric fiber products is the following.

Theorem 7 *(Sullivant (2006), Theorem 13) Let F be a homogeneous Gröbner basis for I with respect to a weight vector ω_1, let G be a homogeneous Gröbner basis for J with respect to a weight vector ω_2, and let ω_q be a weight vector such that Quad_B is a Gröbner basis for I_B. Then*

$$Lift\ (F) \cup Lift\ (G) \cup Quad_B$$

is a Gröbner basis for $I \times_{\mathcal{A}} J$ with respect to the weight vector $\phi_B^*(\omega_1, \omega_2) + \varepsilon\omega_q$ for sufficiently small $\epsilon > 0$.

Note that if ε is chosen small enough, then $\text{in}_{\phi_B^*(\omega_1,\omega_2)+\varepsilon\omega_q}(f_k) = \text{in}_{\phi_B^*(\omega_1,\omega_2)}(f_k)$ for all $f_k \in \langle \text{Lift } F, \text{Lift } G \rangle$.

Remark 1 Since $\text{Quad}_B \subset \ker \phi_B$ we have that $\text{in}_{\phi_B^*(\omega_1,\omega_2)}(f) = f$ for all $f \in \text{Quad}_B$. Now $\text{in}_{\phi_B^*(\omega_1,\omega_2)+\varepsilon\omega_q}(f) = \text{in}_{\omega_q}(\text{in}_{\phi_B^*(\omega_1,\omega_2)}(f)) = \text{in}_{\omega_q}(f)$, so it follows that on Quad_B, the weight vector $\phi_B^*(\omega_1, \omega_2) + \varepsilon\omega_q$ chooses the same leading term as the weight vector ω_q.

3 Dimension of Toric Fiber Products

In this section we give a dimension formula for the toric fiber product of two prime ideals when the set $\mathcal{A}$ is linearly independent, and then apply this to level-1 phylogenetic networks.

Recall the following definitions, from e.g. Becker and Weispfenning (1993). Let I be an ideal in the polynomial ring $\mathbb{K}[x_1, \ldots, x_n]$. We say that a set $U \subseteq \{x_1, \ldots, x_n\}$ is *independent modulo I* if $I \cap \mathbb{K}[U] = \{0\}$. We say that U is *maximally independent modulo I* if it is independent modulo I and there exists no other set $U' \subseteq \{x_1, \ldots, x_n\}$ such that $U \subseteq U'$ and U' is independent modulo I. The *dimension* of I, denoted $\dim I$, is given by $\max\{|U| \mid U \subseteq \{x_1, \ldots, x_n\}$ is independent modulo $I\}$. If I is a prime ideal then for all sets $U \subseteq \{x_1, \ldots, x_n\}$ that are maximally independent modulo I we have $\dim I = |U|$. We begin with the following lemma.

Lemma 8 *Let $M \subset \mathbb{K}[x]$ be a set of monomials, and let $U \subseteq \{x^i_{j_i} \mid i \in [r],\ j_i \in [s_i]\}$ be maximally independent modulo $\langle M \rangle$, given by*

$$U = \{x^i_{j_i^h} \mid i \in \mathcal{I},\ h = 1, \ldots, n_i\},$$

where $\mathcal{I} \subset [r]$ and for each $i \in \mathcal{I}$ we have $j_i^h \in [s_i]$ for $h = 1, \ldots, n_i$. Then the set

$$\text{Lift}\, U = \{z^i_{j_i^h k} \mid i \in \mathcal{I},\ h = 1, \ldots, n_i,\ k \in [t_i]\} \subseteq \mathbb{K}[z]$$

is maximally independent modulo $\langle \text{Lift}\, M \rangle$.

Proof First observe that $\langle \text{Lift}\, M \rangle$ is a monomial ideal generated by monomials of the form m_k as in equation (4) for $m \in M$. Thus, in order to show independence, it is sufficient to only consider monomials m_k. Now if $m_k \in \langle \text{Lift}\, M \rangle \cap \mathbb{K}[z^i_{j_i^h k} \mid i \in \mathcal{I},\ h = 1, \ldots, n_i,\ k \in [t_i]]$, then $m \in M \cap \mathbb{K}[x^i_{j_i} \mid i \in \mathcal{I}, j_i \in [s_i]] = \{0\}$, and thus $\text{Lift}\, U$ is independent modulo $\langle \text{Lift}\, M \rangle$. Furthermore, if $\text{Lift}\, U$ is not maximal then there exists some $i' \in [r]$, $j' \in [s_{i'}]$, and $k' \in [t_{i'}]$ such that $\text{Lift}\, U \cup \{z^{i'}_{j'k'}\}$ is independent modulo $\langle \text{Lift}\, M \rangle$. But then $\text{Lift}\, U \cup \{z^{i'}_{j'k} \mid k \in [t_{i'}]\} = \text{Lift}\,(U \cup \{x^{i'}_{j'}\})$ is also

independent modulo $\langle \text{Lift}\, M\rangle$, so $U \cup \{x^{i'}_{j'}\}$ is independent modulo $\langle M\rangle$, contradicting the maximality of U. □

Note that we have the analogous result for a set of monomials $M \subset \mathbb{K}[y]$ and $U \subseteq \{y^i_{k_i} \mid i \in [r], k_i \in [t_i]\}$.

Theorem 9 *Let I and J be homogeneous ideals in $\mathbb{K}[x]$ and $\mathbb{K}[y]$ respectively, let ω_1 be a weight vector for $\mathbb{K}[x]$, and let ω_2 be a weight vector for $\mathbb{K}[y]$. Let the set $\{x^i_{j_i^h} \mid i \in \mathcal{I}_1,\ j_i^h \in [s_i],\ h = 1, \ldots, n_i\}$ be maximally independent modulo $\text{in}_{\omega_1}(I)$ for some $\mathcal{I}_1 \subseteq [r]$, and let the set $\{y^i_{k_i^g} \mid i \in \mathcal{I}_2,\ k_i^g \in [t_i],\ g = 1, \ldots, m_i\}$ be maximally independent modulo $\text{in}_{\omega_2}(J)$ with $\mathcal{I}_2 \subseteq [r]$. If the set $\mathcal{A}$ is linearly independent, then*

$$\dim I \times_{\mathcal{A}} J \geq \sum_{i \in \mathcal{I}_1 \cap \mathcal{I}_2} (n_i + m_i - 1). \tag{5}$$

Furthermore, if I and J are prime and we have $\mathcal{I}_1 = \mathcal{I}_2 = [r]$ then $\dim I \times_{\mathcal{A}} J = \dim I + \dim J - |\mathcal{A}|$.

Proof Let F and G be Gröbner bases of I and J with respect to the weight vectors ω_1 and ω_2 respectively, and let ω_q be a weight vector on $\mathbb{K}[z]$ that for all i chooses $z^i_{j_1k_2} z^i_{j_2k_1}$ as the initial term for each polynomial in Quad_B, where $1 \leq j_1 < j_2 \leq s_i$ and $1 \leq k_1 < k_2 \leq t_i$. By Theorem 7, we have that for the weight vector $\omega = \phi_B^*(\omega_1, \omega_2) + \varepsilon\omega_q$ and sufficiently small $\varepsilon > 0$, the set $\text{Lift}\,(F) \cup \text{Lift}\,(G) \cup \text{Quad}_B$ is a Gröbner basis of $I \times_{\mathcal{A}} J$. To prove inequality (5) it is sufficient to find a set of generators $z^{i_l}_{j_l k_l}$ that are maximally independent modulo $\text{in}_\omega(I \times_{\mathcal{A}} J)$ and that has size $\sum_{i \in \mathcal{I}_1 \cap \mathcal{I}_2} (n_i + m_i - 1)$.

As in the statement of the theorem, let the set $\{x^i_{j_i^h} \mid i \in \mathcal{I}_1,\ j_i^h \in [s_i],\ h = 1, \ldots, n_i\}$ be maximally independent modulo $\text{in}_{\omega_1}(I) = \text{in}_{\omega_1}(\langle F\rangle)$, and for each $i \in \mathcal{I}_1$ arrange the j_i^h so that $j_i^1 < j_i^2 < \cdots < j_i^{n_i}$. By Lemma 8, and since $\text{in}_\omega(\langle \text{Lift} F\rangle) = \text{Lift}(\text{in}_{\omega_1}\langle F\rangle)$, we have that the set

$$\{z^i_{j_i^h k} \mid i \in \mathcal{I}_1,\ h = 1, \ldots, n_i,\ k = 1, \ldots, t_i\} \subset \mathbb{K}[z]$$

is maximally independent modulo $\text{in}_\omega(\langle \text{Lift}\, F\rangle)$. Similarly, since the set $\{y^i_{k_i^g} \mid i \in \mathcal{I}_2,\ k_i^g \in [t_i],\ g = 1, \ldots, m_i\}$ is maximally independent modulo $\text{in}_{\omega_2}(J)$, we have that

$$\{z^i_{jk_i^g} \mid i \in \mathcal{I}_2,\ g = 1, \ldots, m_i,\ j = 1, \ldots, s_i\} \subset \mathbb{K}[z]$$

is maximally independent modulo $\text{in}_\omega(\langle \text{Lift}\, G\rangle)$. Again, for each $i \in \mathcal{I}_2$ arrange the k_i^h so that $k_i^1 < k_i^2 < \cdots < k_i^{m_i}$. We now have

$$\text{in}_\omega(\langle \text{Lift}\, F \cup \text{Lift}\, G\rangle) \cap \mathbb{K}[z^i_{j_i^h k_i^g} \mid i \in \mathcal{I}_1 \cap \mathcal{I}_2,\ h = 1, \ldots, n_i,\ g = 1, \ldots, m_i] = \{0\},$$

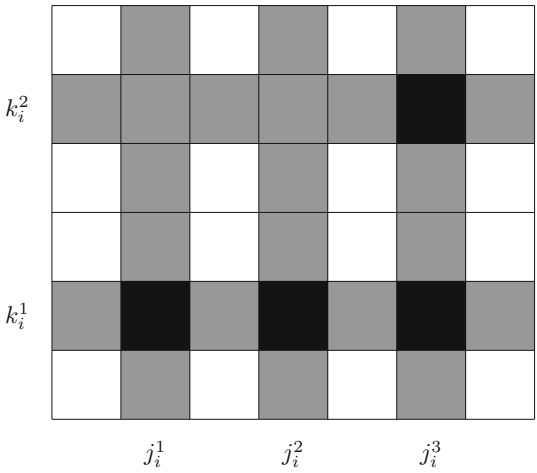


Fig. 2 Grid representing the generators z^i_{jk} for a fixed $i \in [r]$. Columns shaded grey give monomials coming from the lift of the maximally independent set modulo $\text{in}_{\omega_1}(I)$, and rows shaded grey give monomials coming from the lift of the maximally independent set modulo $\text{in}_{\omega_2}(J)$. Cells shaded black represent the elements of the set Z of degree a_i

and that the set $\{z^i_{j_i^h k_i^g} \mid i \in \mathcal{I}_1 \cap \mathcal{I}_2,\ h = 1, \ldots, n_i,\ g = 1, \ldots, m_i\}$ is maximal with respect to this condition. We claim that the set

$$Z = \{z^i_{j_i^h k_i^1} \mid i \in \mathcal{I}_1 \cap \mathcal{I}_2,\ h = 1, \ldots, n_i\} \cup \{z^i_{j_i^{n_i} k_i^g} \mid i \in \mathcal{I}_1 \cap \mathcal{I}_2,\ g = 1, \ldots, m_i\}$$

is maximally independent modulo $\text{in}_\omega(\langle \text{Lift } F \cup \text{Lift } G \cup \text{Quad}_B \rangle) = \text{in}_\omega(I \times_{\mathcal{A}} J)$ (see Fig. 2).

First we show that $\text{in}_\omega(\langle \text{Lift } F \cup \text{Lift } G \cup \text{Quad}_B \rangle) \cap \mathbb{K}[Z] = \{0\}$. Observe that since $\text{Lift } F \cup \text{Lift } G \cup \text{Quad}_B$ is a Gröbner basis, we have $\text{in}_\omega(\langle \text{Lift } F \cup \text{Lift } G \cup \text{Quad}_B \rangle) = \text{in}_\omega(\text{Lift } F) \cup \text{in}_\omega(\text{Lift } G) \cup \text{in}_\omega(\text{Quad}_B)$. Since $\text{in}_\omega(\langle \text{Lift } F \cup \text{Lift } G \rangle) \cap \mathbb{K}[Z] = \{0\}$, it is sufficient to show that for each $i \in \mathcal{I}_1 \cap \mathcal{I}_2$ the elements of degree a_i in Z do not appear together as a quadratic monomial in $\text{in}_\omega(\langle \text{Quad}_B \rangle)$. By Remark 1 and our choice of ω we have that

$$\text{in}_\omega(\text{Quad}_B) = \{z^i_{j_1 k_2} z^i_{j_2 k_1} \mid i \in [r],\ 1 \le j_1 < j_2 \le s_i,\ 1 \le k_1 < k_2 \le t_i\}.$$

Fix $i \in \mathcal{I}_1 \cap \mathcal{I}_2$ and observe that for any two elements of Z of degree a_i, say z^i_{jk} and $z^i_{j'k'}$ with $j \le j'$, we either have $j = j'$, $k = k'$, or $k < k'$. In all cases $z^i_{jk} z^i_{j'k'} \notin \text{in}_\omega(\text{Quad}_B)$.

Next we show that Z is maximal. By the maximal independence of $\mathbb{K}[z^i_{j_i^h k_i^g} \mid i \in \mathcal{I}_1 \cap \mathcal{I}_2,\ h = 1, \ldots, n_i,\ g = 1, \ldots, m_i]$ modulo $\text{in}_\omega(\langle \text{Lift } F \cup \text{Lift } G \rangle)$, we need only consider those $z^i_{jk} \notin Z$ with $i \in \mathcal{I}_1 \cap \mathcal{I}_2$, $j = j_i^h$ for some $h = 1, \ldots, n_i$, and $k = k_i^g$ for some $g = 1, \ldots, m_i$. But it is clear that for any such z^i_{jk}, we can find $z^i_{j_0 k_0} \in Z$ such that $z^i_{j_0 k_0} z^i_{j_h k_l} \in \text{in}_\omega(\text{Quad}_B)$. It follows that Z is maximally independent modulo $\text{in}_\omega(\langle \text{Lift } F \cup \text{Lift } G \cup \text{Quad}_B \rangle)$. Since

$$|Z| = \sum_{i \in \mathcal{I}_1 \cap \mathcal{I}_2} (n_i + m_i - 1),$$

inequality (5) is proved. For the final statement, observe that since I is prime, we have $\dim I = \sum_{i \in \mathcal{I}_1} n_i$, and since J is prime, we have $\dim J = \sum_{i \in \mathcal{I}_2} m_i$. If $\mathcal{I}_1 = \mathcal{I}_2 = [r]$ then we get

$$|Z| = \sum_{i=1}^{r} (n_i + m_i - 1) = \sum_{i=1}^{r} n_i + \sum_{i=1}^{r} m_i - r = \dim I + \dim J - |\mathcal{A}|.$$

Now since $I \times_{\mathcal{A}} J$ is prime, its dimension is equal to the size of any subset that is maximally independent modulo $\text{in}_\omega(\langle \text{Lift } F \cup \text{Lift } G \cup \text{Quad}_B \rangle)$. □

Remark 2 Observe that if both I and J are prime ideals and there exist maximally independent sets with $\mathcal{I}_1 = \mathcal{I}_2 = [r]$, then it is clear from the proof that $I \times_{\mathcal{A}} J$ is also a prime ideal and there exists a maximally independent set modulo $\text{in}_\omega(I \times_{\mathcal{A}} J)$ in $\{z^i_{jk} \mid i \in [r],\ j \in [s_i],\ k \in [t_i]\}$ with at least one element for each upper index $i \in [r]$.

For the remainder of this section we will apply our results on toric fiber products to level-1 phylogenetic networks. Fix a group-based model (G, B), and let $\mathcal{N}$ be a level-1 phylogenetic network with a (directed) cut edge e. Then the operation of cutting $\mathcal{N}$ at e results in two smaller level-1 networks, that we denote $\mathcal{N}_+$ and $\mathcal{N}_-$. We denote by e the new edge in both $\mathcal{N}_+$ and $\mathcal{N}_-$, and this edge inherits the direction from $\mathcal{N}$. We assume that the network $\mathcal{N}_+$ contains the leaves labelled $1, \ldots, n'$ for some $n' < n$, which are also leaves of $\mathcal{N}$, and the new leaf, which we denote by n_e. Then $\mathcal{N}_-$ contains the leaves labelled $n'+1, \ldots, n$, and the new leaf which we also denote by n_e.

The vanishing ideal $I_{\mathcal{N}_+}$ is contained in the polynomial ring $R_+ = \mathbb{C}[q^+_{g_1 \cdots g_{n'} g_{n_e}} \mid g_1 + \cdots + g_{n'} + g_{n_e} = 0]$, and $I_{\mathcal{N}_-}$ is contained in $R_- = \mathbb{C}[q^-_{g_{n_e} g_{n'+1} \cdots g_n} \mid g_{n_e} + g_{n'+1} + \cdots + g_n = 0]$. We give each polynomial ring the grading induced by $\deg(q_{g_1 \cdots g_N}) = E_{[\xi(e)]} \in \mathbb{Z}^{|B \cdot G|}_{\geq 0}$, where ξ is the consistent edge labelling induced by the consistent leaf labelling $g_1, \ldots, g_N$, and $\{E_{[g]} \mid [g] \in B \cdot G\}$ is the standard basis of $\mathbb{Z}^{|B \cdot G|}_{\geq 0}$. Note that the set $\mathcal{A}$ consisting of the image under deg of the generators of R_+ and R_- is given by the linearly independent set $\{E_{[g]} \mid [g] \in B \cdot G\}$, and each element of this set is the image under deg of a generator of both R_+ and R_-. We assume that the edge e in $\mathcal{N}$ is directed towards $\mathcal{N}_+$ so that $\xi(e) = g_1 + \cdots + g_{n'}$ and therefore $\deg(q_{g_1 \cdots g_n}) = E_{[g_1 + \cdots + g_{n'}]}$.

We have a natural $\mathbb{C}$-algebra homomorphism

$$\begin{aligned} R &\to R_+ \otimes_{\mathbb{C}} R_- \\ q_{g_1 \cdots g_n} &\mapsto q^+_{g_1 \cdots g_{n'} g_+} \otimes q^-_{g_- g_{n'+1} \cdots g_n}, \end{aligned} \tag{6}$$

where $g_+ = -(g_1 + \cdots + g_{n'})$ and $g_- = -(g_{n'+1} + \cdots + g_n)$. Note that $\deg(q^+_{g_1 \cdots g_{n'} g_+}) = E_{[g_1 + \cdots + g_{n'}]}$ and $\deg(q^-_{g_- g_{n'+1} \cdots g_n}) = E_{[-(g_{n'+1} + \cdots + g_n)]} =$

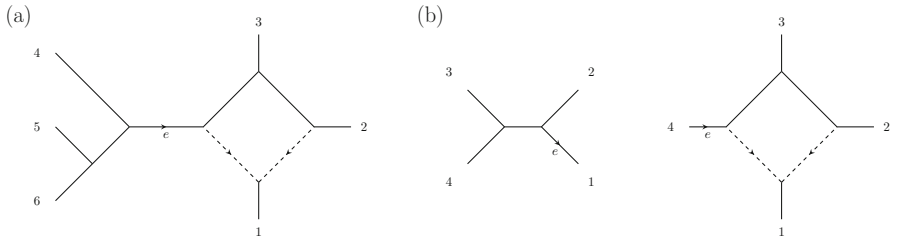


Fig. 3 **a** A level-1 phylogenetic network $\mathcal{N}$. **b** The phylogenetic networks $\mathcal{N}_-$ and $\mathcal{N}_+$ obtained by cutting $\mathcal{N}$ at e (right). The toric fiber product of $I_{\mathcal{N}_-}$ and $I_{\mathcal{N}_+}$ corresponds to gluing $\mathcal{N}_-$ and $\mathcal{N}_+$ along the edge labelled e

$E_{[g_1+\cdots+g_{n'}]}$. As in the proof of (Cummings et al (2021), Proposition 3.2), the network parameterisation map $\phi_{\mathcal{N}}$ factors through (6), so $I_{\mathcal{N}}$ is the toric fiber product $I_{\mathcal{N}_+} \times_{\mathcal{A}} I_{\mathcal{N}_-}$.

Example 2 We will consider the 2-state Cavender-Farris-Neyman model, for which $G = \mathbb{Z}/2\mathbb{Z}$, on the phylogenetic network $\mathcal{N}$ depicted in Figure 3. The corresponding ideal $I_{\mathcal{N}}$ is contained in the polynomial ring $R = \mathbb{C}[q_{g_1g_2g_3g_4g_5g_6} \mid g_1 + g_2 + g_3 + g_4 + g_5 + g_6 = 0]$. Cutting at the non-trivial cut edge e results in a 4 sunlet and a 4-leaf tree. Let $\mathcal{N}_+$ be the 4-sunlet, and $\mathcal{N}_-$ be the 4-leaf tree. Let

$$I_{\mathcal{N}_+} \subset R_+ = \mathbb{C}[q^+_{g_1g_2g_3g_4} \mid g_1 + g_2 + g_3 + g_4 = 0]$$

and let

$$I_{\mathcal{N}_-} \subset R_- = \mathbb{C}[q^-_{g_1g_2g_3g_4} \mid g_1 + g_2 + g_3 + g_4 = 0]$$

be the corresponding ideals. We give R_+ the grading induced by $\deg(q^+_{g_1g_2g_3g_4}) = E_{g_4} \in \mathbb{Z}^2 = \mathbb{Z}E_0 + \mathbb{Z}E_1$, so that the degree E_0 generators are $q^+_{0000}, q^+_{1100}, q^+_{1010}$, and q^+_{0110}, and the degree E_1 generators are $q^+_{0011}, q^+_{0101}, q^+_{1001}$, and q^+_{1111}. We give R_- the grading induced by $\deg(q^-_{g_1g_2g_3g_4}) = E_{g_1}$. In this case the degree E_0 generators are $q^-_{0000}, q^-_{0011}, q^-_{0101}$, and q^-_{0110}, and the degree E_1 generators are $q^-_{1001}, q^-_{1010}, q^-_{1100}$, and q^-_{1111}.

The multigrading in R is given by $\deg(q_{g_1g_2g_3g_4g_5g_6}) = E_{g_1+g_2+g_3}$ and the map ϕ_B is given by

$$\phi_B : R \to \mathbb{C}[q^+_{g_1g_2g_3g_4}, q^-_{h_1h_2h_3h_4} \mid g_1 + g_2 + g_3 + g_4 = h_1 + h_2 + h_3 + h_4 = 0]$$
$$\phi_B : q_{g_1g_2g_3g_4g_5g_6} \mapsto q^+_{g_1g_2g_3g_+} q^-_{g_-g_4g_5g_6},$$

where $g_+ = g_- = g_1 + g_2 + g_3$ and $\deg(q_{g_1g_2g_3g_+}) = \deg(q_{g_-g_4g_5g_6}) = E_{g_1+g_2+g_3}$. As described above, $I_{\mathcal{N}}$ is given by the toric fiber product $I_{\mathcal{N}_+} \times_{\mathcal{A}} I_{\mathcal{N}_-}$, where $\mathcal{A}$ is given by $\{E_0, E_1\}$.

First we describe Quad_B.Using that $\langle \text{Quad}_B \rangle = \ker \phi_B$ we see that Quad_B consists of all elements of the form

$$q_{g_1g_2g_3g_4g_5g_6}q_{h_1h_2h_3h_4h_5h_6} - q_{g_1g_2g_3h_4h_5h_6}q_{h_1h_2h_3g_4g_5g_6},$$

for generators $q_{g_1g_2g_3g_4g_5g_6}$ and $q_{h_1h_2h_3h_4h_5h_6}$ satisfying $\deg(q_{g_1g_2g_3g_4g_5g_6}) = \deg(q_{h_1h_2h_3h_4h_5h_6})$. Note that this condition guarantees that $(g_1, g_2, g_3, h_4, h_5, h_6)$ and $(h_1, h_2, h_3, g_4, g_5, g_6)$ are consistent leaf-labellings.

Next we consider lifts. For the generator q^+_{0000} we lift by elements of degree E_0 in R_-. Thus we have

$$\text{Lift}\, q^+_{0000} = \{q_{000000}, q_{0000011}, q_{000101}, q_{000110}\}.$$

Similarly, for the generator q^-_{0000} we have

$$\text{Lift}\, q^-_{0000} = \{q_{000000}, q_{011000}, q_{101000}, q_{110000}\}.$$

Observe that, for example, $\phi_B(q_{0000011}) = q^+_{0000}q^-_{0011}$ and $\phi_B(q_{101000}) = q^+_{1010}q^-_{0000}$. Note also that the generator q_{000000} can be obtained from lifting both q^+_{0000} and q^-_{0000}.

The ideal $I_{\mathcal{N}_+}$ is generated by the set F (consisting of a single quadratic), and the ideal $I_{\mathcal{N}_-}$ is generated by the set G. We give R_+, and R_- the monomial ordering $q_{g_1g_2g_3g_4} < q_{h_1h_2h_3h_4}$ if and only if $(g_1, g_2, g_3.g_4) < (h_1, h_2, h_3, , h_4)$ with respect to lexicographic ordering and with $0 < 1$ in $\mathbb{Z}/2\mathbb{Z}$. Then, with respect to lexicographic ordering on R_+ and R_-, F and G are Gröbner bases.

$$\begin{aligned}
F &= \{f = q^+_{0000}q^+_{1111} - q^+_{1100}q^+_{0011} + q^+_{1010}q^+_{0101} - q^+_{0110}q^+_{1001}\},\\
G &= \{g_1 = q^-_{1010}q^-_{1100} - q^-_{1001}q^-_{1111},\ g_2 = q^-_{0110}q^-_{1100} - q^-_{0101}q^-_{1010},\\
&\quad g_3 = q^-_{0011}q^-_{1100} - q^-_{0000}q^-_{1111},\ g_4 = q^-_{0110}q^-_{1001} - q^-_{0101}q^-_{1010},\\
&\quad g_5 = q^-_{0011}q^-_{1001} - q^-_{0000}q^-_{1010},\ g_6 = q^-_{0011}q^-_{0101} - q^-_{0000}q^-_{0110}\}.
\end{aligned}$$

Observe that in all cases, each polynomial is homogeneous in the grading. We give an example lift for f and for g_1. The degree of f is $E_0 + E_1$. For k corresponding to the pair (q^-_{0011}, q^-_{1010}) we have

$$f_k = q_{000011}q_{111010} - q_{110011}q_{001010} + q_{101011}q_{010010} - q_{011011}q_{100010}.$$

The polynomial g_1 has degree $2E_1$. For k corresponding to the pair (q^+_{0011}, q^+_{1001}) we have

$$(g_1)_k = q_{001010}q_{100100} - q_{001001}q_{100111}.$$

Finally, Lift F is given by all possible lifts f_k, and Lift G is given by all possible lifts $(g_i)_k$ for $i = 1, \ldots, 6$. Then the ideal $I_{\mathcal{N}}$ is generated by the elements of Lift F, Lift G, and Quad_B.

Corollary 10 *Fix a group-based model* (G, B). *Let* $\mathcal{N}$ *be a level-1 phylogenetic network with a cut edge* e, *and let* $\mathcal{N}_+$ *and* $\mathcal{N}_-$ *be the networks obtained by cutting* $\mathcal{N}$ *at* e. *Then* $\dim V_{\mathcal{N}}^{(G,B)} = \dim V_{\mathcal{N}_+}^{(G,B)} + \dim V_{\mathcal{N}_-}^{(G,B)} - |B \cdot G|$.

Proof As described above, the ideal $I_{\mathcal{N}}$ is the toric fiber product $I_{\mathcal{N}_+} \times_{\mathcal{A}} I_{\mathcal{N}_-}$, so we apply Theorem 9. Both $I_{\mathcal{N}_+}$ and $I_{\mathcal{N}_-}$ are prime ideals, so to prove the result, it is sufficient to show that for a phylogenetic network ideal I there exists a weight vector ω and a set $U \subset \{q_{g_1 \cdots g_n} \mid g_1 + \cdots + g_n = 0\}$ that is independent modulo $\text{in}_\omega(I)$ and that contains at least one element of degree a_i for each $a_i \in \mathcal{A}$. From Remark 2, we need only consider phylogenetic networks that are either sunlet networks or trees. Furthermore, if $\mathcal{N}$ is a sunlet network and $\mathcal{T}$ is a tree obtained from $\mathcal{N}$ be removing a reticulation edge, then $I_{\mathcal{N}} \subset I_{\mathcal{T}}$. It follows that if U is independent modulo $\text{in}_\omega(I_{\mathcal{T}})$ then U is also independent modulo $\text{in}_\omega(I_{\mathcal{N}})$, so in fact it is sufficient to show the result for any phylogenetic tree $\mathcal{T}$.

To show the result for a tree $\mathcal{T}$, we make the further observation that if $\mathcal{T}$ has an internal edge e, then $\mathcal{T}$ is a toric fiber product of the two trees given by cutting $\mathcal{T}$ at e. Thus in view of Remark 2 again, we need only consider claw trees. Since we are only considering binary phylogenetic trees, we need only consider the 3-claw tree T_3.

Fix a set of representatives $\mathcal{G} \subset G$ of the B-orbits in G, let $\mathcal{T} = T_3$, let $I = I_{\mathcal{T}}$, and let the set of multidegrees be given by $\mathcal{A} = \{E_g \mid g \in \mathcal{G}\}$. Note that $0 \in \mathcal{G}$ and that $[0] = \{0\}$. We may assume, without loss of generality, that $\deg(q_{g'hk}) = E_g$, where $g' \in [g]$ for some $g \in \mathcal{G}$. Recall that I is given by the kernel of the map $\psi_{\mathcal{T}}$ where

$$\begin{aligned} \psi_{\mathcal{T}} : \mathbb{C}[q_{ghk} \mid g + h + k = 0] &\longrightarrow \mathbb{C}[a_i^{[g]} \mid g \in \mathcal{G},\ i = 1, 2, 3] \\ q_{ghk} &\longmapsto a_1^{[g]} a_2^{[h]} a_3^{[k]}. \end{aligned}$$

Let $U = \{q_{g0(-g)} \mid g \in \mathcal{G}\}$. It is clear that U has exactly one element of each multidegree. Next, choose a term order on $\mathbb{C}[q_{ghk} \mid g + h + k = 0]$ such that $q_{ghk} < q_{g'h'k'}$ whenever $g \in \mathcal{G}$ and $g' \notin \mathcal{G}$, and let ω be a a weight vector whose induced term order satisfies this. We claim that $\mathbb{C}[U] \cap \text{in}_\omega(I) = \{0\}$.

To prove the claim, we will show that for any element $f \in I$, we have that $\text{in}_\omega(f)$ does not consist of a product of elements of U. Since I is homogeneous and generated by binomials, we may assume that f is a homogeneous binomial. Let $\mathcal{G}' \subseteq \mathcal{G}$ with $|\mathcal{G}'| = n$ and suppose that we can write

$$f = \prod_{g \in \mathcal{G}'} q_{g0(-g)} - m$$

for some other monomial m of total degree n. Since $f \in \ker \psi_{\mathcal{T}}$ we must have that

$$\psi_{\mathcal{T}}(m) = \psi_{\mathcal{T}}\Big(\prod_{g \in \mathcal{G}'} q_{g0(-g)}\Big) = \Big(\prod_{g \in \mathcal{G}'} a_1^{[g]}\Big)\Big(\prod_{g \in \mathcal{G}'} a_3^{[-g]}\Big)(a_2^{[0]})^n.$$

Now if $q_{g'h'k'}$ is a factor of m then we must have $\psi_{\mathcal{T}}(q_{g'h'k'}) = a_1^{[g]} a_2^{[0]} a_3^{[k]}$ for some $g, k \in \mathcal{G}$. Thus, $h' \in [0]$ so $h' = 0$ and $g' \in [g]$, and since $g' + 0 + k' = 0$ we must

have $k' = -g'$. Now if $g' = g$ then $q_{g'h'k'} = q_{g0(-g)}$ appears as a factor in the first monomial of f. If this holds for all factors of m then we have $f = 0$. If not, then for some factor $q_{g'0(-g')}$ we must have $g' \notin \mathcal{G}$, so we have $\text{in}_\omega(f) = m$. □

Remark 3 Notice that in the proof of Corollary 10, we made no assumptions on the number of reticulation vertices of $\mathcal{N}$. Since a binary phylogenetic tree can be thought of as a phylogenetic network with no reticulation vertices, the result also holds for binary phylogenetic trees. Explicitly, we have that if $\mathcal{T}$ is a binary phylogenetic tree with an interior edge e, with trees $\mathcal{T}_+$ and T_- obtained by cutting e, then we have

$$\dim V_{\mathcal{T}}^{(G,B)} = \dim V_{\mathcal{T}_+}^{(G,B)} + \dim V_{\mathcal{T}_-}^{(G,B)} - |B \cdot G|.$$

4 Sunlet Networks and Trees

If $\mathcal{N}$ is a level-1 phylogenetic network, then $\mathcal{N}$ can be decomposed along cut edges into a series of phylogenetic trees and sunlet networks. As shown in the previous section, the ideal structure of the corresponding varieties is given by the toric fiber product. It therefore remains for us give dimension results for the varieties corresponding to trees and sunlet networks. For an unrooted phylogenetic tree $\mathcal{T}$, the dimension of the variety $V_{\mathcal{T}}^{(G,B)}$ is well known. We give a proof using the dimension result of the previous section.

Lemma 11 *If $\mathcal{T}$ is a binary phylogenetic tree with m edges and no degree-2 vertices under a group-based evolutionary model (G, B), then the affine dimension of $V_{\mathcal{T}}^{(G,B)}$ is given by*

$$\dim V_{\mathcal{T}}^{(G,B)} = lm + 1.$$

Proof Denote by t the number of interior edges of $\mathcal{T}$. If $t = 0$ then $\mathcal{T}$ is the 3-claw tree. This has dimension $3l+1$ by (Baños et al (2019), Proposition 5.2), so the proposition is true in this case. Now suppose $\mathcal{T}$ is a binary phylogenetic tree with m edges and $t > 0$ interior edges. Let e be an interior edge and let $\mathcal{T}_+$ and $\mathcal{T}_-$ be the trees obtained by cutting at e. If m_+ and m_- are the number of edges of $\mathcal{T}_+$ and $\mathcal{T}_-$ respectively, we have $m = m_+ + m_- - 1$. Furthermore, the number of interior edges of $\mathcal{T}_+$ and of $\mathcal{T}_-$ is less than t, so by induction we have $\dim V_{\mathcal{T}_+}^{(G,B)} = lm_+ + 1$ and $\dim V_{\mathcal{T}_-}^{(G,B)} = lm_- + 1$. It follows from Remark 3 that

$$\dim V_{\mathcal{T}}^{(G,B)} = (lm_+ + 1) + (lm_- + 1) - (l + 1) = lm + 1.$$

□

Observe that one could extend the above proof to give the analogous dimension result for any phylogenetic tree with no degree 2 vertices. To do so, the base-case for the induction must be extended to cover all claw trees T_n with $n \geq 3$. That is, one must show that for each T_n with corresponding ideal I_n, there exists a maximal

independent set modulo $\text{in}_\omega(I_n)$ that contains at least one element of multidegree a_i for each $a_i \in \mathcal{A}$, as in the proof of Corollary 10.

The remainder of this section is dedicated to giving the dimension of the varieties corresponding to sunlet networks. As we have already seen, the variety associated to a phylogenetic network $\mathcal{N}$ is equal to the variety associated to the corresponding contracted semi-directed phylogenetic network, so from this point onwards we will only consider contracted semi-directed phylogenetic networks. First we will give an upper bound on the dimension.

Proposition 12 *If $\mathcal{N}$ is a contracted semi-directed phylogenetic network with only disjoint cycles and with m edges then*

$$\dim V_{\mathcal{N}}^G \leq lm + 1.$$

Proof Let c denote the number of cycles in $\mathcal{N}$. The affine variety $V_{\mathcal{N}}^G$ is parameterized by $(l+1)m$ parameters, but the map is multihomogeneous. It is linear in the set of parameters for each non-reticulation edge, and in the union of the parameters for the two reticulation edges of each cycle. Thus we may think of the parameterization map as a projective map

$$\mathbb{P}^l \times \cdots \times \mathbb{P}^l \times \mathbb{P}^{2l+1} \times \cdots \times \mathbb{P}^{2l+1} \dashrightarrow \mathbb{P}^{(l+1)^{n-1}},$$

where $\mathbb{P}^l$ appears $m-2c$ times (once for each non-reticulation edge), and $\mathbb{P}^{2l+1}$ appears c times (once for each cycle). We use a dashed arrow to indicate that in order for the map to be well-defined we may need to take a subset of the domain. It follows that $V_{\mathcal{N}}^G$ has projective dimension at most $lm + c$, and thus its affine dimension is at most $lm + c + 1$.

Now consider $v = \phi_{\mathcal{N}}(w) \in \mathbb{C}^{|G|^{n-1}}$, where $w \in \mathbb{C}^{m(l+1)}$. For each pair of reticulation edges e_1, e_2, a consistent leaf labelling of $\mathcal{N}$ assigns both edges the same label. For each consistent leaf labelling of $\mathcal{N}$ in which they are labelled 0, the edges along the cycle all receive the same labels in both trees, so the coordinate of v corresponding to the consistent leaf labelling has a factor of $w_{e_1}^0 + w_{e_2}^0$. For every consistent leaf labelling in which they are not labelled 0, the coordinate does not depend on $w_{e_1}^0$ or $w_{e_2}^0$. Therefore the map depends only on the sum $w_{e_1}^0 + w_{e_2}^0$. This reduces the number of independent parameters by c, so the affine dimension of $V_{\mathcal{N}}^G$ is at most $lm + 1$. □

4.1 General Group-Based Models

First, we restrict our attention to sunlet networks under general group-based models of evolution, i.e., those where the group B consists only of the identity automorphism. We will deal with the case $G = \mathbb{Z}/2\mathbb{Z}$ separately.

Proposition 13 *Let $\mathcal{N}$ be the n-sunlet network with $n \geq 4$ and let G be an abelian group with $|G| = l + 1 > 2$. Then*

$$\dim V_{\mathcal{N}}^G = l(2n-1) + 1.$$

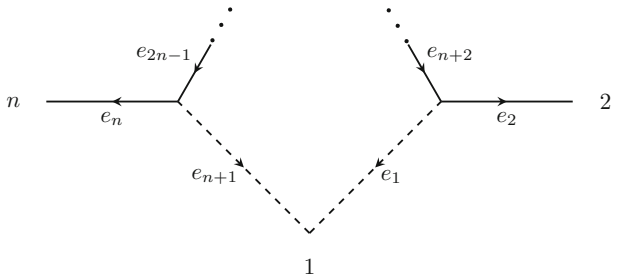


Fig. 4 A contracted n-sunlet network with $n \geq 4$. Arrows indicate the orientation used for assigning a consistent edge labelling from a consistent leaf labelling

Proof Using Lemma 4, we may replace $\mathcal{N}$ by its contraction. This network has $m = 2n - 1$ edges, so by Proposition 12 we have $\dim V_{\mathcal{N}}^G \leq l(2n-1)+1$.

Label the edges and vertices as in Fig. 4, and let $\mathcal{T}_1$ and $\mathcal{T}_2$ be the two trees got by removing the edges e_{n+1} and e_1 respectively. The parameterization map for $\mathcal{N}$ is given by

$$q_{g_1 g_2 \cdots g_n} = a_2^{g_2} \cdots a_n^{g_n} \left(a_1^{g_1} a_{n+2}^{g_1+g_2} \cdots a_{2n-1}^{g_n} + a_{n+1}^{g_1} a_{n+2}^{g_2} \cdots a_{2n-1}^{g_1+g_n} \right), \tag{7}$$

where $g_1, \ldots, g_n$ is a consistent leaf labelling, the first monomial corresponds to $\mathcal{T}_1$, and the second monomial corresponds to $\mathcal{T}_2$. With notation as in Sect. 2.3, our aim will be to find λ that maximises $\operatorname{rank}_{\mathbb{R}} A_\lambda$, which by Lemma 6 gives a lower bound on $\dim V_{\mathcal{N}}^G$.

Let $\{E_i^g \mid g \in G, i = 1, \ldots, 2n-1\}$ be the standard basis of $\mathbb{R}^{m(l+1)}$, indexed by the edges of $\mathcal{N}$ and elements of G, and consider the dual vector space $V = (\mathbb{R}^{m(l+1)})^*$ with the dual basis. Choose $\lambda \in V$ such that $\lambda_{n+2}^0 = -2$, $\lambda_{n+1}^g = 1$ for all $g \in G$, and all other entries are 0. Let $g_1, \ldots, g_n$ be a consistent leaf labelling of $\mathcal{N}$. Then the corresponding column of A_λ has the following properties:

- If $g_1 = 0$, then the monomial from $\mathcal{T}_1$ is chosen.
- If $g_1 \neq 0$ and $g_2 = 0$, then the monomial from $\mathcal{T}_2$ is chosen.
- In all other cases the monomial from $\mathcal{T}_1$ is chosen.

We will show that $\operatorname{rank} A_\lambda \geq l(2n-1)+1$ to give the lower bound.

Consider the submatrix given by consistent leaf labellings where $g_1 = 0$, so that each column is an exponent vector coming from a monomial in $\mathcal{T}_1$. Perform column operations on A_λ so that the first $(l+1)^{n-2}$ columns are given by this submatrix. Let $\mathcal{S}$ be the tree with $n-1$ leaves obtained from $\mathcal{N}$ by deleting the reticulation vertex. The consistent leaf labellings of $\mathcal{N}$ in which $g_1 = 0$ give all of the consistent leaf labellings of $\mathcal{S}$. Since $\mathcal{S}$ is a phylogenetic tree, its corresponding variety $V_{\mathcal{S}}^G$ is toric, and therefore the rank of its corresponding matrix A is equal to the dimension of the variety. By Lemma 11, the variety $V_{\mathcal{S}}^G$ has dimension $l(2n-5)+1$, so the submatrix of A_λ consisting only of the columns where $g_1 = 0$ has rank $l(2n-5)+1$ (note that since $\mathcal{S}$ has a monomial parameterization, for all choices of λ we have that this submatrix is the same).

We make the following observations about this submatrix. First, since the monomial from $\mathcal{T}_1$ is always chosen, the entries corresponding to the parameters a^g_{n+1} are 0 for all $g \in G$. Similarly, for the edge e_1, only the parameter a^0_1 appears in the parameterization of $q_{g_1 g_2 \cdots g_n}$, so the entries corresponding to the parameters a^g_1 are 0 for all $g \in G$ except $g = 0$. Next, observe that in this submatrix, the row corresponding to the parameter a^g_2 is equal to the row corresponding to the parameter a^g_{n+2} for all $g \in G$ and similarly the row corresponding to the parameter a^g_n is equal to the row corresponding to the parameter a^g_{2n-1} for all $g \in G$. This is because the label of e.g. the edge e_{n+2} is $g_1 + g_2 = 0 + g_2 = g_2$, which is also the label of the edge e_2. We perform row operations on A_λ so that for each $g \in G$, the first $(l+1)^{n-2}$ entries of the rows corresponding to the parameters a^g_2 and a^g_n are zero, by subtracting the rows a^g_{n+2} and a^g_{2n-1} respectively. Now we perform further row operations to swap rows and obtain a matrix of the following form, where the upper left block is a $(4l+3) \times (l+1)^{n-2}$ matrix consisting of zeros,

$$A_\lambda = \left[\begin{array}{c|c} 0 & B \\ \hline A'_\lambda & * \end{array}\right],$$

and rank $A'_\lambda = l(2n-5)+1$. It follows that rank $A_\lambda \geq l(2n-5)+1+\text{rank } B$, so it is sufficient to show that rank $B \geq 4l$.

The columns of B correspond to consistent leaf labellings $g_1, \ldots, g_n$ with $g_1 \neq 0$. Recall that λ was such that if $g_2 = 0$ then the monomial from $\mathcal{T}_2$ is chosen, and otherwise the monomial from $\mathcal{T}_1$ is chosen. The rows of B correspond to the parameters a^g_1 for $g \neq 0$, and a^g_{n+1}, a^g_2, and a^g_n for all $g \in G$. However, we performed row operations on the rows corresponding to a^g_2 and a^g_n, so for each column of B the coefficient of the standard basis vector E^g_2 is given by the exponent of a^g_2 minus the exponent of a^g_{n+2}, and the coefficient of E^g_n is given by the exponent of a^g_n minus the exponent of a^g_{2n-1} in the corresponding monomial from the parameterization (7). Thus the columns of B are given by

$$(E^{g_n}_n - E^{g_1+g_n}_n) + E^{g_1}_{n+1}, \tag{8}$$

if $g_2 = 0$ (so the monomial comes from $\mathcal{T}_2$), and

$$(E^{g_2}_2 - E^{g_1+g_2}_2) + E^{g_1}_1, \tag{9}$$

otherwise (so the monomial comes from $\mathcal{T}_1$), where $g_1, \ldots, g_n$ is a consistent leaf labelling with $g_1 \neq 0$. Note that since $n \geq 4$, we can find a consistent leaf labelling $g_1, \ldots, g_n$ for any choice of g_1, g_2, g_n. Denote by X_1 the vector space spanned by all the vectors of the form in equation (8). We have

$$\sum_{g \in G} \left((E^g_n - E^{g_1+g}_n) + E^{g_1}_{n+1}\right) = (l+1)E^{g_1}_{n+1} \in X_1,$$

so $E_{n+1}^{g_1} \in X_1$ for all $g_1 \neq 0$. It follows immediately that for a fixed $g_n \in G$, we have $E_n^{g_n} - E_n^{g_1+g_n} \in X_1$ for all $g_1 \neq 0$, and thus $\dim X_1 \geq 2l$.

Next denote by X_2 the vector space spanned by all the vectors of the form in equation (9). Using that $\sum_{g\in G} E_2^g - E_2^{g_1+g} = 0$, we see that

$$\sum_{g\in G\setminus\{0\}} \left((E_2^g - E_2^{g_1+g}) + E_1^{g_1}\right) = lE_1^{g_1} + E_2^{g_1} - E_2^0 \in X_2,$$

for each $g_1 \neq 0$. Now fix $g \in G \setminus \{0\}$ and let $g_2 = g$, and $g_1 = -g$, so that $E_1^{-g} + E_2^g - E_2^0 \in X_2$. Then we have

$$(lE_1^g + E_2^g - E_2^0) - (E_1^{-g} + E_2^g - E_2^0) = lE_1^g - E_1^{-g} \in X_2.$$

Now if $g = -g$ then we have $E_1^g \in X_2$. If not, by swapping g_1 and g_2 we have $lE_1^{-g} - E_1^g \in X_2$, so $(l-1)E_1^g - (l-1)E_1^{-g} \in X_2$. Then $lE_1^g - lE_1^{-g} \in X_2$, and subtracting $lE_1^{-g} - E_1^g$ gives $E_1^g \in X_2$, for all $g \in G\setminus\{0\}$. As before, it follows that for a fixed g_2 we have $E_2^{g_2} - E_2^{g_1+g_2} \in X_2$ for all $g_1 \in G\setminus\{0\}$, so $\dim X_2 \geq 2l$. It follows that rank $B \geq 4l$. □

We expect the result to hold for the case $n = 3$ once the size of G is large enough, and this is explored in a forthcoming paper. Here, the proof of Proposition 13 breaks down in this case because, when finding the rank of B, we we have only $l+1$ columns when $g_2 = 0$, since in this case $g_1 = -g_n$. Thus the dimension of X_1 is strictly less than $2l$.

Next we deal with the case $G = \mathbb{Z}/2\mathbb{Z}$. The expected dimension for n-sunlets here is $2n$. However, if $n = 3$ then we only have $4 < 2n$ consistent leaf labellings of $\mathcal{N}$, so in this case the expected dimension cannot be reached. When $n = 4$ we have $8 = 2n$ consistent leaf labellings, however, in this case $\dim V_{\mathcal{N}}^G = 7$. This can be shown by direct computation.

Proposition 14 *Let $\mathcal{N}$ be the n-sunlet network with $n \geq 5$ and let $G = \mathbb{Z}/2\mathbb{Z}$. Then*

$$\dim V_{\mathcal{N}}^G = 2n.$$

Proof As before, Proposition 12 gives the upper bound. Label the edges and vertices as in Fig. 5, and let $\mathcal{T}_1$ and $\mathcal{T}_2$ be the two trees got by removing the edges e_{n+1} and e_1 respectively. Observe that we have at least one edge on the cycle, e.g. e_{n+3}, that is not adjacent to either reticulation edge. We proceed as in Proposition 13, this time choosing $\lambda \in \mathbb{R}^{2m}$ such that $\lambda^0_{n+1} = \lambda^1_{n+1} = 1$, $\lambda^0_{n+3} = 2$. and all other entries are 0. Let $g_1, \ldots, g_n$ be a consistent leaf labelling of $\mathcal{N}$. Then the corresponding column of A_λ has the following properties:

- If $g_1 = 0$, then the monomial from $\mathcal{T}_1$ is chosen.
- If $g_1 = 1$ and $g_2 + g_3 = 1$, then the monomial from $\mathcal{T}_2$ is chosen.
- If $g_1 = 1$ and $g_2 + g_3 = 0$, then monomial from $\mathcal{T}_1$ is chosen.

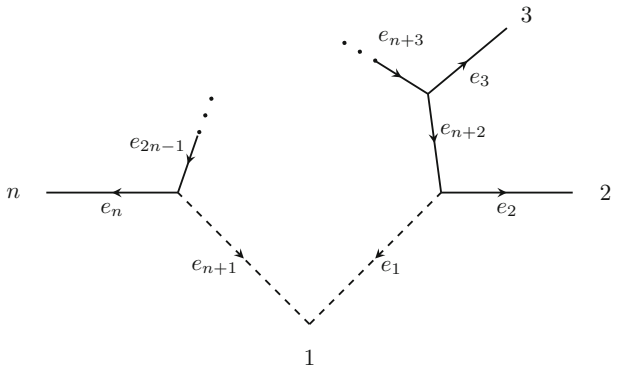


Fig. 5 A contracted n-sunlet network with $n \geq 5$. Arrows indicate the orientation used for assigning a consistent edge labelling from a consistent leaf labelling

As in the proof of Proposition 13, we perform column operations so that the first 2^{n-2} columns are indexed by consistent leaf labellings where $g_1 = 0$, and each column is an exponent vector coming from the corresponding $\mathcal{T}_1$ monomial. The submatrix consisting of these columns has rank $2n-4$, and we perform the same row operations as before to give the block triangular matrix

$$A_\lambda = \left[\begin{array}{c|c} 0 & B \\ \hline A'_\lambda & * \end{array}\right],$$

where the submatrix B is given by rows corresponding to the parameters a_1^1, a_2^0, a_2^1, a_n^0, a_n^1, a_{n+1}^0, and a_{n+1}^1. However, we performed row operations on the rows corresponding to a_2^g and a_n^g, so for each column of B the coefficient of E_2^g is given be the exponent of a_2^g minus the exponent of a_{n+2}^g, and the coefficient of E_n^g is given by the exponent of a_n^g minus the exponent of a_{2n-1}^g for $g = 0, 1$. Consider the following columns of B. For a consistent leaf labelling with $g_1 = g_n = 1$ and $g_2 = g_3 = 0$, the monomial from $\mathcal{T}_1$ is chosen, so the labels assigned to a_n and a_{2n-1} are equal, and the labels assigned to a_2 and a_{n+2} are not equal. Thus the column is given by

$$E_1^1 + E_2^0 - E_2^1.$$

Next for a consistent leaf labelling with $g_1 = g_2 = g_3 = g_n = 1$, the monomial from $\mathcal{T}_1$ is chosen so the column is given by

$$E_1^1 - E_2^0 + E_2^1.$$

For $g_1 = g_3 = 1$ and $g_2 = g_n = 0$, the monomial from $\mathcal{T}_2$ is chosen so the column is given by

$$E_{n+1}^1 + E_n^0 - E_n^1.$$

Finally, for a consistent leaf labelling with $g_1 = g_3 = g_n = 1$ and $g_2 = 0$, the monomial from $\mathcal{T}_2$ is chosen so the column is given by

$$E^1_{n+1} - E^0_n + E^1_n.$$

These vectors are linearly independent, so rank $B \geq 4$ and the result follows. □

Observe that in the case $n = 4$, the vector $E^1_{n+1} - E^0_n + E^1_n$, is not obtained, since there is no consistent leaf-labelling with $g_1 = g_2 = g_4 = 1$ and $g_2 = 0$. In this case B has four columns, corresponding to the consistent leaf labellings when $g_1 = 1$. The two columns assigned to $\mathcal{T}_2$ correspond to $(1, 0, 1, 0)$ and $(1, 1, 0, 0)$. The reader can check that in both cases the column vector is E^1_5. The two columns assigned to $\mathcal{T}_1$ remain linearly independent, so in this case rank $B = 3$.

4.2 Group-Based Models

In this section, we use our results on general group-based models to obtain the result for all group-based models, following the method of (Baños et al (2019), Lemma 4.2). Throughout, let $\mathcal{N}$ be the contracted n-sunlet network, so the number of edges m is equal to $2n - 1$. Let G be a finite abelian group, and let B be a subgroup of the automorphism group Aut(G) with $|B \cdot G| = l + 1$. Let $(\mathbb{R}^{|G|m})^*$ have standard basis elements ε^g_e where $g \in G$ and $e \in \mathcal{E}(\mathcal{N})$. Next, pick representatives $g_0 = 0, g_1, \ldots, g_l$ in G for each B-orbit, and let $(\mathbb{R}^{(l+1)m})^*$ have standard basis elements $\varepsilon^{[g_i]}_e$ for $i = 0, \ldots, l$ and $e \in \mathcal{E}(\mathcal{N})$.

Let $p : (\mathbb{R}^{|G|m})^* \longrightarrow (\mathbb{R}^{(l+1)m})^*$ be the map that sums coefficients of the unit vectors for each orbit, i.e.

$$\sum_{e\in\mathcal{E}(\mathcal{N})} \sum_{g\in G} c^g_e \varepsilon^g_e \longmapsto \sum_{e\in\mathcal{E}(\mathcal{N})} \sum_{i=0}^{l} \Big(\sum_{g\in[g_i]} c^g_e \Big) \varepsilon^{[g_i]}_e,$$

where $c^g_e \in \mathbb{R}$. It is clear that p is a surjective, linear map, so $\dim \ker p = (|G|-l-1)m$. Now consider the parameterizations of $V^G_{\mathcal{N}}$ and $V^{(G,B)}_{\mathcal{N}}$. For a fixed consistent leaf labelling ξ, let α_1 and α_2 be the exponent vectors of the monomials corresponding to $\mathcal{T}_1$ and $\mathcal{T}_2$ respectively, in the parameterization of $V^G_{\mathcal{N}}$. Similarly let α'_1 and α'_2 be the corresponding monomials in the parameterization of $V^{(G,B)}_{\mathcal{N}}$. Then $p(\alpha_i) = \alpha'_i$ for $i = 1, 2$. Furthermore, observe that if $\lambda \in \mathbb{R}^{|G|m}$ is such that $\lambda^g_e = \lambda^h_e$ whenever g and h are in the same B orbit for all edges $e \in \mathcal{E}(\mathcal{N})$, then there exists $\lambda' \in \mathbb{R}^{(l+1)m}$ satisfying $\lambda = \lambda' \circ p$ (where we are considering λ' as an element of the dual space of $(\mathbb{R}^{(l+1)m})^*$).

Proposition 15 *With the notation as above, let $\lambda \in \mathbb{R}^{|G|m}$ be such that $\lambda^g_e = \lambda^h_e$ whenever g and h are in the same B-orbit, for all $e \in \mathcal{E}_{\mathcal{N}}$. Then there exists $\lambda' \in \mathbb{R}^{(l+1)m}$ such that*

$$p \circ A_\lambda = A_{\lambda'}.$$

Proof First observe that for any $\alpha \in (\mathbb{R}^{|G|m})^*$ we have

$$\langle \lambda, \alpha \rangle = \lambda(\alpha) = \lambda' \circ p(\alpha) = \langle \lambda', p(\alpha) \rangle.$$

Now consider the polynomials of the parameterizations of $V_{\mathcal{N}}^{G}$ and $V_{\mathcal{N}}^{(G,B)}$, for a consistent leaf labelling ξ. Let α_1 and α_2 be as above, and suppose that $\langle \lambda, \alpha_1 \rangle < \langle \lambda, \alpha_2 \rangle$. Then $\langle \lambda', p(\alpha_1) \rangle < \langle \lambda', p(\alpha_2) \rangle$, so both λ and λ' pick the monomial corresponding to $\mathcal{T}_1$. Since $\alpha_1' = p(\alpha_1)$, the result follows. □

Note that Proposition 15 is easily generalizable to level-1 phylogenetic networks.

Corollary 16 *Let $\mathcal{N}$ be the n-sunlet network with $n \geq 4$, let G be a finite abelian group, and let B be a non-trivial subgroup of the automorphism group* Aut(G)*, with $|B \cdot G| = l + 1$. Then*

$$\dim V_{\mathcal{N}}^{(G,B)} = l(2n-1) + 1.$$

Proof As in the case for general group-based models, the upper bound is given by Proposition 12. For the lower bound, first observe that since B is a non-trivial subgroup, we must have $|G| > 2$. Next observe that the vector λ chosen in the proof of Proposition 13 satisfies the condition in Proposition 15, so using Proposition 15 (and Lemma 6) there exists some λ' such that

$$\dim V_{\mathcal{N}}^{(G,B)} \geq \operatorname{rank}_{\mathbb{R}} A_{\lambda'} = \operatorname{rank}_{\mathbb{R}}(p \circ A_{\lambda'}).$$

Finally, since p is a surjective linear map with kernel of dimension $(|G| - l - 1)m$, we have

$$\operatorname{rank}_{\mathbb{R}} A_{\lambda'} \geq (|G| - 1)m + 1 - (|G| - l - 1)m = lm + 1.$$

□

We summarise our results on sunlet networks in a single theorem. Note that the final two cases are given by direct computation.

Theorem 17 *Let $\mathcal{N}$ be a sunlet network with n leaves. Let G be a finite abelian group and let B be a subgroup of* Aut(G)*. Denote by $l + 1$ the number of B-orbits in G. Then* $\dim V_{\mathcal{N}}^{(G,B)}$ *is given in the following cases.*

- *If $n \geq 4$ and $|G| > 2$ then* $\dim V_{\mathcal{N}}^{(G,B)} = l(2n-1) + 1$.
- *If $n \geq 5$ and $G = \mathbb{Z}/2\mathbb{Z}$ so that $B = \{\mathrm{id}\}$ then* $\dim V_{\mathcal{N}}^{\mathbb{Z}/2\mathbb{Z}} = 2n$.
- *If $n = 4$ then* $\dim V_{\mathcal{N}}^{\mathbb{Z}/2\mathbb{Z}} = 7$.
- *If $n = 3$ then* $\dim V_{\mathcal{N}}^{\mathbb{Z}/2\mathbb{Z}} = 4$.

□

Table 1 Values for the deficiency of dim $V_{\mathcal{N}}^{(G,B)}$, where $\mathcal{N}$ is an n-sunlet

n	$\mathbb{Z}/2\mathbb{Z}$	$\mathbb{Z}/3\mathbb{Z}$	JC	K2P	$(\mathbb{Z}/2\mathbb{Z})^2$	$\mathbb{Z}/4\mathbb{Z}$	$\mathbb{Z}/5\mathbb{Z}$	$\mathbb{Z}/6\mathbb{Z}$	$\mathbb{Z}/7\mathbb{Z}$
3	**2**	**2**	**1**	**1**	1	0	0	0	0
4	1	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0

5 Proof of Theorems 1 and 2

We are now able to give simple inductive proofs of Theorems 1 and 2. Below we give only the proof of Theorem 1. The proof of Theorem 2 is almost identical, and is left to the reader with the aid of Table 1.

Proof of Theorem 1 We will prove the result using induction on the number of non-trivial cut edges of a level-1, triangle-free phylogenetic network $\mathcal{N}$. For the case when there are no non-trivial cut edges, we must have that $\mathcal{N}$ is either the 3-claw tree, in which case the dimension of $V_{\mathcal{N}}^{(G,B)}$ is equal to $lm + 1$ by Lemma 11, or $\mathcal{N}$ is an n-sunlet network with $n \geq 4$, in which case the dimension is $l(2n - 1) + 1$ by Theorem 17. In both cases the result holds.

Now suppose that $\mathcal{N}$ is a level-1, triangle-free phylogenetic network with a non-trivial cut edge e, and m edges and c cycles. Let $\mathcal{N}_1$ and $\mathcal{N}_2$ be the networks obtained by cutting at e, and let m_i and c_i denote the number of edges and cycles in $\mathcal{N}_i$ respectively for $i = 1, 2$. Since the number of non-trivial cut edges in $\mathcal{N}_1$ and $\mathcal{N}_2$ must be fewer than the number of non-trivial cut edges in $\mathcal{N}$, by induction we have $\dim V_{\mathcal{N}_i}^{(G,B)} = l(m_i - c_i) + 1$ for $i = 1, 2$. By Corollary 10 we have

$$\begin{aligned}\dim V_{\mathcal{N}}^{(G,B)} &= \dim V_{\mathcal{N}_1}^{(G,B)} + \dim V_{\mathcal{N}_2}^{(G,B)} - (l+1)\\ &= l(m_1 + m_2 - c_1 - c_2) + 2 - (l+1)\\ &= l(m-c) + 1,\end{aligned}$$

where $m_1 + m_2 = m + 1$ and $c_1 + c_2 = c$. □

6 Application to Identifiability

In this section we apply Theorems 1 and 2 to give some immediate identifiability results. Throughout, fix an abelian group G and subgroup B of $\mathrm{Aut}(G)$, and let $l + 1$ be the number of orbits in $B \cdot G$. First, we extend the definition of *distinguishability* from Gross and Long (2018) to all group-based models of evolution

Definition 3 Let (G, B) be a group-based model of evolution. Two distinct n-leaf networks $\mathcal{N}_1$ and $\mathcal{N}_2$ are *distinguishable over* (G, B) if $V_{\mathcal{N}_1}^{(G,B)} \nsubseteq V_{\mathcal{N}_2}^{(G,B)}$ and $V_{\mathcal{N}_2}^{(G,B)} \nsubseteq V_{\mathcal{N}_1}^{(G,B)}$.

When G and B are clear, we will simply say that $\mathcal{N}_1$ and $\mathcal{N}_2$ are distinguishable. Observe that if $V_{\mathcal{N}_1}^{(G,B)}$ and $V_{\mathcal{N}_2}^{(G,B)}$ are irreducible varieties of equal dimension, then in order to determine whether $\mathcal{N}_1$ and $\mathcal{N}_2$ are distinguishable it is sufficient to show that either $V_{\mathcal{N}_1}^{(G,B)} \not\subseteq V_{\mathcal{N}_2}^{(G,B)}$ or $V_{\mathcal{N}_2}^{(G,B)} \not\subseteq V_{\mathcal{N}_1}^{(G,B)}$. One of the key results we will use to show identifiability is the following.

Lemma 18 *(Gross et al (2021) Lemma 3) Let $\mathcal{N}_1$ and $\mathcal{N}_2$ be n-leaf networks. If for some $A \subseteq [n]$, we have that $V_{\mathcal{N}_1|_A}^{(G,B)} \not\subseteq V_{\mathcal{N}_2|_A}^{(G,B)}$, then $V_{\mathcal{N}_1}^{(G,B)} \not\subseteq V_{\mathcal{N}_2}^{(G,B)}$.*

Corollary 19 *Let $\mathcal{N}_1$ and $\mathcal{N}_2$ be n-leaf networks with $\dim V_{\mathcal{N}_1}^{(G,B)} = \dim V_{\mathcal{N}_2}^{(G,B)}$. If for some $A \subseteq [n]$ we have $V_{\mathcal{N}_1|_A}^{(G,B)} \not\subseteq V_{\mathcal{N}_2|_A}^{(G,B)}$, then $\mathcal{N}_1$ and $\mathcal{N}_2$ are distinguishable over (G, B).*

Proof By Lemma 18, $V_{\mathcal{N}_1}^{(G,B)} \not\subseteq V_{\mathcal{N}_2}^{(G,B)}$. Since they are irreducible varieties of the same dimension, they are distinguishable. □

We will use Corollary 19 in conjunction with the following dimension results.

Lemma 20 *Let $\mathcal{N}_1$ and $\mathcal{N}_2$ be n-leaf, level-1 phylogenetic networks, both with exactly c cycles, where each cycle has length at least 4 when $|G| > 2$ and at least 5 when $G = \mathbb{Z}/2\mathbb{Z}$. Then $\dim V_{\mathcal{N}_1}^{(G,B)} = \dim V_{\mathcal{N}_2}^{(G,B)}$.*

Proof Observe that $\mathcal{N}_1$ and $\mathcal{N}_2$ have the same number of edges. To see this, suppose that $\mathcal{N}_1$ and $\mathcal{N}_2$ have m_1 and m_2 edges respectively. Then the corresponding contracted networks $\mathcal{N}_1'$ and $\mathcal{N}_2'$ have $m_1 - c$ and $m_2 - c$ edges, since for each reticulation vertex the outgoing edge is removed. Next for each of the c reticulation vertices $v_1, \ldots, v_c$ in $\mathcal{N}_1'$ arbitrarily pick a reticulation edge (u_i, v_i) and remove it. After removal, the vertex u_i has degree 2 and can be suppressed. The result is an unrooted binary phylogenetic tree on n leaves with $m_1 - 3c$ edges. Performing the same operations on $\mathcal{N}_2'$ we also obtain a (possibly different) unrooted binary phylogenetic tree on n leaves with $m_2 - 3c$ edges. Since all unrooted binary phylogenetic trees on n leaves have $2n - 3$ edges, we have that $m_1 = m_2$. Now since both $\mathcal{N}_1$ and $\mathcal{N}_2$ have exactly c cycles, the result follows from Theorems 1 and 2. □

Remark 4 From the proof of Lemma 20 it is clear that the number of edges of an unrooted level-1 phylogenetic network on n leaves with c cycles is $2n - 3 + 3c$

For the remaining results in this section we will need to use the fact that binary phylogenetic trees with group-based models of evolution are distinguishable. This result is well-known in the algebraic phylogenetics community, but we give a direct proof here for completeness.

Lemma 21 *Let (G, B) be a group-based model of evolution, and let $\mathcal{T}_1$ and $\mathcal{T}_2$ be two distinct n-leaf, unrooted, binary phylogenetic trees. Then $\mathcal{T}_1$ and $\mathcal{T}_2$ are distinguishable over (G, B).*

Proof First observe that since $\mathcal{T}_1$ and $\mathcal{T}_2$ are determined by their quartets, there exists a subset $A \subset [n]$ with $|A| = 4$ such that $\mathcal{T}_1$ restricted to A and $\mathcal{T}_2$ restricted to A are

distinct four-leaf, binary phylogenetic trees. By Corollary 19, it is sufficient to show that $V^{(G,B)}_{\mathcal{T}_1|_A} \not\subseteq V^{(G,B)}_{\mathcal{T}_2|_A}$. Since the dimensions of these varieties are equal (Lemma 11), this is equivalent to the restricted trees being distinguishable.

We will show that the four leaf binary phylogenetic trees are distinguishable. Let $\mathcal{T}$ be the four-leaf tree with split 12|34, and the corresponding interior edge denoted e_5. Pick $g, h \in G$ such that $h \notin [g]$ and consider the polynomial $f = q_{\mathbf{g}}q_{\mathbf{h}} - q_{\mathbf{g}'}q_{\mathbf{h}'}$ where $\mathbf{g} = (g, -g, g, -g)$, $\mathbf{h} = (h, -h, h, -h)$, $\mathbf{g}' = (g, -g, h, -h)$, and $\mathbf{h}' = (h, -h, g, -g)$. We have

$$\psi_{\mathcal{T}}(f) = a_1^g a_1^h a_2^{-g} a_2^{-h} a_3^g a_3^h a_4^{-g} a_4^{-h} a_5^0 a_5^0 - a_1^g a_1^h a_2^{-g} a_2^{-h} a_3^h a_3^g a_4^{-h} a_4^{-g} a_5^0 a_5^0 = 0,$$

so that $f \in \ker(\psi_{\mathcal{T}}) = I^{(G,B)}_{\mathcal{T}}$. On the other hand, by looking at the parameters corresponding to the interior edge, the reader can check that f does not belong to the ideals corresponding to the trees with splits 13|24 and 14|23 respectively.

In a similar manner one can find polynomials belonging only to the ideal of the tree with split 13|24 and only to the ideal of the tree with the split 14|23. It follows that the four leaf binary phylogenetic trees are distinguishable. □

Proposition 22 *Let $\mathcal{N}_1$ and $\mathcal{N}_2$ be two distinct n-sunlet networks with $n \geq 5$ and distinct leaves adjacent to the reticulation vertex. Then $\mathcal{N}_1$ and $\mathcal{N}_2$ are distinguishable over (G, B).*

Proof By Theorem 17 we have that $\dim V^{(G,B)}_{\mathcal{N}_1} = \dim V^{(G,B)}_{\mathcal{N}_2}$. Assume, without loss of generality, that for $\mathcal{N}_1$ the leaf adjacent to the reticulation vertex is leaf 1. Let $A = \{2, \ldots, n\}$, so that $\mathcal{N}_1|_A$ is a caterpillar tree on $n-1$ leaves and $\mathcal{N}_2|_A$ is an $(n-1)$-sunlet network. Then

$$\dim V^{(G,B)}_{\mathcal{N}_1|_A} = l(2n-5) + 1 < l(2n-3) + 1 = \dim V^{(G,B)}_{\mathcal{N}_2|_A},$$

and so $V^{(G,B)}_{\mathcal{N}_2|_A} \not\subseteq V^{(G,B)}_{\mathcal{N}_1|_A}$. By Corollary 19, $\mathcal{N}_1$ and $\mathcal{N}_2$ are distinguishable. □

Proposition 23 *Let $\mathcal{N}_1$ and $\mathcal{N}_2$ be two distinct n-sunlet networks with $n \geq 4$ such that the leaf adjacent to the reticulation vertex is the same for both networks, and the trees obtained from each network by removing the reticulation vertex and adjacent leaf are distinct. Then $\mathcal{N}_1$ and $\mathcal{N}_2$ are distinguishable over (G, B).*

Proof Assume that $\mathcal{N}_1$ and $\mathcal{N}_2$ both have leaf 1 adjacent to the reticulation vertex. Let $A = \{2, \ldots, n\}$ so that by assumption $\mathcal{N}_1|_A$ and $\mathcal{N}_2|_A$ are distinct caterpillar trees with $n-1$ leaves. Since these are distinguishable (Lemma 21), the result follows from Corollary 19. □

Observe that Propositions 22 and 23 are not sufficient to give identifiability for all sunlet networks. For example, take an n sunlet with leaves labelled in ascending order clockwise around the sunlet with 1 at the reticulation. Then obtain a distinct sunlet by swapping leaves 2 and 3. The caterpillar trees obtained from both of these sunlets by restricting to $\{2, \ldots, n\}$ are the same, so neither Proposition 22 nor Proposition 23 applies.

More generally we can give the following identifiability result for triangle-free, level-1 phylogenetic networks. The result relies on the existence of a subset A of the leaf set with particular properties.

Proposition 24 *Let $\mathcal{N}_1$ and $\mathcal{N}_2$ be two triangle-free, level-1 phylogenetic networks on n leaves and both with exactly c cycles, and let G be an abelian group with $|G| > 2$. If there exists a subset $A \subset [n]$ such that either*

1. *$\mathcal{N}_1|_A$ and $\mathcal{N}_2|_A$ are triangle-free level-1 phylogenetic networks with distinct number of cycles, or*
2. *$\mathcal{N}_1|_A$ is a tree and $\mathcal{N}_2|_A$ is a triangle-free level-1 phylogenetic network, or*
3. *$\mathcal{N}_1|_A$ and $\mathcal{N}_2|_A$ are distinct trees,*

then $\mathcal{N}_1$ and $\mathcal{N}_2$ are distinguishable over (G, B).

Proof First observe that $\dim V_{\mathcal{N}_1}^{(G,B)} = \dim V_{\mathcal{N}_2}^{(G,B)}$ by Lemma 20. Let $\mathcal{N}_1|_A$ and $\mathcal{N}_2|_A$ have m_1 and m_2 edges respectively, and c_1 and c_2 cycles respectively.

For case 1, assume without loss of generality that $c_1 < c_2$. Then by Remark 4 we have that $m_1 = 2|A|-1+3c_1$ and $m_2 = 2|A|-1+3c_2$. In particular, $m_1-c_1 < m_2-c_2$. Then by Theorem 1 we have that

$$\dim V_{\mathcal{N}_1|_A}^{(G,B)} = l(m_1 - c_1) + 1 < l(m_2 - c_2) + 1 = \dim V_{\mathcal{N}_2|_A}^{(G,B)}.$$

It follows that $V_{\mathcal{N}_2|_A}^{(G,B)} \not\subseteq V_{\mathcal{N}_1|_A}^{(G,B)}$. For case 2 let us assume that $\mathcal{N}_1|_A$ is a tree and $\mathcal{N}_2|_A$ is a triangle-free level-1 phylogenetic network. Then $\dim V_{\mathcal{N}_1|_A}^{(G,B)} < \dim V_{\mathcal{N}_2|_A}^{(G,B)}$ so as above $V_{\mathcal{N}_2|_A}^{(G,B)} \not\subseteq V_{\mathcal{N}_1|_A}^{(G,B)}$. For case 3 we have that $V_{\mathcal{N}_1|_A}^{(G,B)} \not\subseteq V_{\mathcal{N}_2|_A}^{(G,B)}$ and $V_{\mathcal{N}_2|_A}^{(G,B)} \not\subseteq V_{\mathcal{N}_1|_A}^{(G,B)}$ by Lemma 21. In all three cases the result now follows by Corollary 19. □

7 Discussion

In this paper we have given a dimension formula for all triangle-free, level-1 phylogenetic networks under a group-based model of evolution. Our main tool was the toric fiber product, for which we gave a dimension formula that we hope will be useful beyond this work.

Our results confirmed a conjecture of Gross and Long which states that under the JC model of evolution, the dimensions of large cycle networks (that is, level-1 phylogenetic networks with a single cycle of length at least 4) are equal (Gross and Long 2018, Conjecture 5.1). In fact, as we have shown, this is true for all group-based models and level-1 phylogenetic networks where the number of cycles is equal. We were also able to give partial identifiability results for sunlet networks and larger level-1 networks that followed immediately from our results on dimension.

We were unable to give a general dimension result for 3-sunlets. For this case, our upper bound (Proposition 12) still holds, but our proof for the lower bound does not work. This is because with the λ we have chosen, when $n = 3$ we have only l columns in the matrix A_λ coming from $\mathcal{T}_2$, whilst the rest come from $\mathcal{T}_1$. Thus the maximum

rank of A_λ is dim $V_{\mathcal{T}_1}^{(G,B)} + l = (2n-2)l + 1$, and this is too small. Nonetheless, we believe the result still holds, and we make the following conjecture.

Conjecture 25 *If $\mathcal{N}$ is the 3-sunlet network and $|G| > 4$ then*

$$\dim V_{\mathcal{N}}^{G} = lm + 1.$$

Our conjecture is backed up by calculations of the dimension $V_{\mathcal{N}}^{(G,B)}$ for small sunlet networks and small groups. The deficiencies (i.e., the number of dimensions less than the expected dimension $l(2n-1)+1$) are shown in Table 1.

Bold values in Table 1 indicate that the variety fills the whole space $\mathbb{C}^{(l+1)^{n-1}}$, and this has dimension less than the expected dimension. Note that for the JC and K2P models we have binomial linear invariants, and it is customary to identify these and reduce the dimension of the ambient space. From Table 1, it appears that we only have two cases where the dimension of $V_{\mathcal{N}}^{(G,B)}$ is less than expected for unknown reasons. These are when $G = \mathbb{Z}/2\mathbb{Z}$ and $n = 4$, and when $G = \mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}$ and $n = 3$. The latter case has implications for models of DNA sequence evolution, since the group $G = \mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}$ is usually identified with the four nucleic acids, and the corresponding general group-based model of evolution is the Kimura 3-parameter model (K3P). The 3-sunlet network models events such as hybridisation, so a good understanding of this case will be useful for models in molecular phylogenetics.

A full identifiability result, generalising (Gross et al 2021, Theorem 2), remains open. For the DNA group-based models (JC, K2P, and K3P), one of the key results is that the variety corresponding to the 3-sunlet has smaller dimension than expected. This result can be exploited to give identifiability results on level-1 phylogenetic networks with four leaves (e.g. (Gross and Long 2018, Corollary 4.8)), since for a fixed number of leaves a 3-cycle network will have a strictly lower dimension than a 4-cycle network. For general G however, this is not the case, as shown in Table 1, so an alternative approach will be necessary to show identifiability for general G.

As the authors note in Gross and Long (2018), this dimension deficiency is in contrast to group-based mixture models, where the number of leaves determines the dimension. Here, we have shown that the dimension of a triangle-free level-1 phylogenetic network variety is fully determined by the number of leaves and the number of cycles (see Theorem 1), and for large enough G we expect this to be true for all level-1 phylogenetic networks.

Acknowledgements Elizabeth Gross is supported by the National Science Foundation (NSF) under grant DMS-1945584. Samuel Martin is supported by the Biotechnology and Biological Sciences Research Council (BBSRC), part of UK Research and Innovation, through the Core Capability Grant BB/CCG1720/1 at the Earlham Institute, and is grateful for funding from EPSRC (grant number EP/W007134/1) and BBSRC (grant number BB/X005186/1).

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Allman ES, Rhodes JA (2007) Phylogenetic invariants. Reconstructing evolution: new mathematical and computational advances, pp 108–146

Allman ES, Rhodes JA (2008) Phylogenetic ideals and varieties for the general Markov model. Adv Appl Math 40(2):127–148

Baños H, Bushek N, Davidson R et al (2019) Dimensions of group-based phylogenetic mixtures. Bull Math Biol 81(2):316–336

Becker T, Weispfenning V (1993) Gröbner bases, a computational approach to commutative algebra. Springer-Verlag, New York

Casanellas M, Fernández-Sánchez J (2008) Geometry of the Kimura 3-parameter model. Adv Appl Math 41(3):265–292

Casanellas M, Fernández-Sánchez J (2011) Relevant phylogenetic invariants of evolutionary models. J Math Pures Appl 96(3):207–229

Casanellas M, Fernández-Sánchez J, Michałek M (2017) Local equations for equivariant evolutionary models. Adv Math 315:285–323

Casanellas M, Fernández-Sánchez J, Garrote-López M (2021) Distance to the stochastic part of phylogenetic varieties. J Symb Comput 104:653–682

Cummings J, Hollering B, Manon C (2021) Invariants for level-1 phylogenetic networks under the Cavendar-Farris-Neyman model. arXiv preprint arXiv:2102.03431

Draisma J (2008) A tropical approach to secant dimensions. J Pure Appl Algebra 212:349–363

Drton M, Sturmfels B, Sullivant S (2009) Lectures on algebraic statistics, Oberwolfach seminars, vol 39. Birkhäuser Basel

Duarte E, Hollering B, Wiedmann M (2023) Toric fiber products in geometric modeling. arXiv preprint arXiv:2303.08754

Engström A, Kahle T, Sullivant S (2014) Multigraded commutative algebra of graph decompositions. J Algebraic Combinatorics 39:335–372

Eriksson N, Ranestad K, Sturmfels B et al (2005) Phylogenetic algebraic geometry. Projective varieties with unexpected properties. 237:255

Evans SN, Speed TP (1993) Invariants of some probability models used in phylogenetic inference. Ann Stat 21(1):355–377

Francis A, Semple C, Steel M (2018) New characterisations of tree-based networks and proximity measures. Adv Appl Math 93:93–107

Francis AR, Steel M (2015) Which phylogenetic networks are merely trees with additional arcs? Syst Biol 64(5):768–777

Gross E, Long C (2018) Distinguishing phylogenetic networks. SIAM J Appl Algebra Geom 2(1):72–93

Gross E, van Iersel L, Janssen R et al (2021) Distinguishing level-1 phylogenetic networks on the basis of data generated by Markov processes. J Math Biol 83(32):1

Kahle T, Rauh J (2014) Toric fiber products versus Segre products. Abhandlungen aus dem Mathematischen Seminar der Universität Hamburg. 84:187–201

Maclagan D, Sturmfels B (2021) Introduction to tropical geometry, vol 161. American Mathematical Society

Michałek M (2011) Geometry of phylogenetic group-based models. J Algebra 339(1):339–356

Michałek M, Ventura E (2019) Phylogenetic complexity of the Kimura 3-parameter model. Adv Math 343:640–680

Pachter L, Strumfels B (2005) Algebraic statistics for computational biology. Cambridge University Press. https://doi.org/10.1017/CBO9780511610684

Semple C (2016) Phylogenetic networks with every embedded phylogenetic tree a base tree. Bull Math Biol 78(1):132–137

Sturmfels B, Sullivant S (2005) Toric ideals of phylogenetic invariants. J Comput Biol 12(4):457–481

Sullivant S (2006) Toric fiber products. J. Algebra 316:560–577

Sullivant S (2018) Algebraic statistics, vol 194. American Mathematical Soc

Székely LA, Steel MA, Erdős PL (1993) Fourier calculus on evolutionary trees. Adv Appl Math 14:200–216

Zwiernik P, Smith JQ (2011) Implicit inequality constraints in a binary tree model. Electron J Stat 5:1276–1312

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.