Identifiability in Phylogenetics Using Algebraic Matroids

Ben Hollering and Seth Sullivant

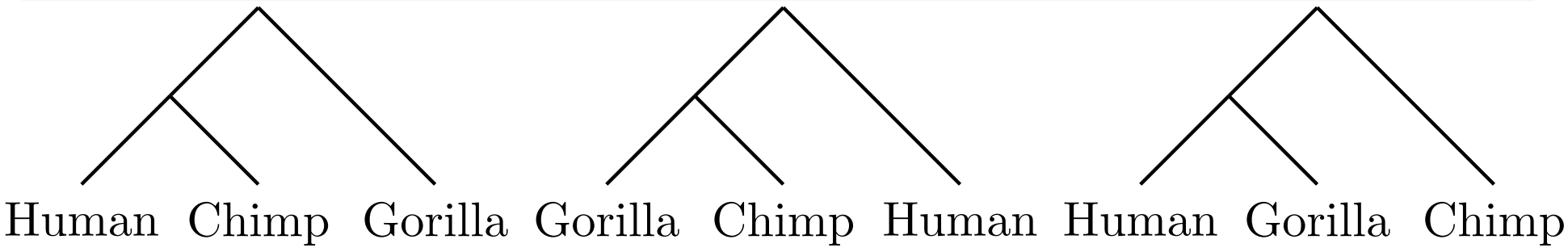
North Carolina State University

April 9, 2020

Phylogenetics

Problem

Given a collection of species, find the tree that explains their evolutionary history.

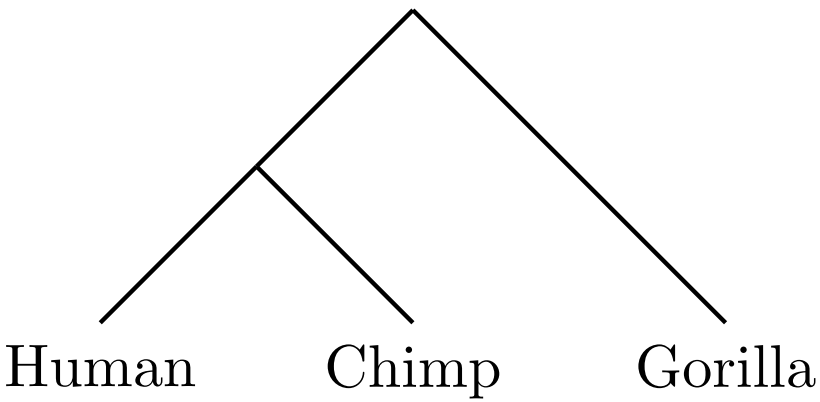


Building Trees with DNA Sequence Data

- DNA bases are A, T, G, C
- DNA sequences of related species all evolved from some common ancestor
- *Align* sequences for a gene that appears in all species

Human :	GATCTCAAGGAC
Chimp :	GGCCTCAAGGAT
Gorilla :	GATCTCCAGGCA

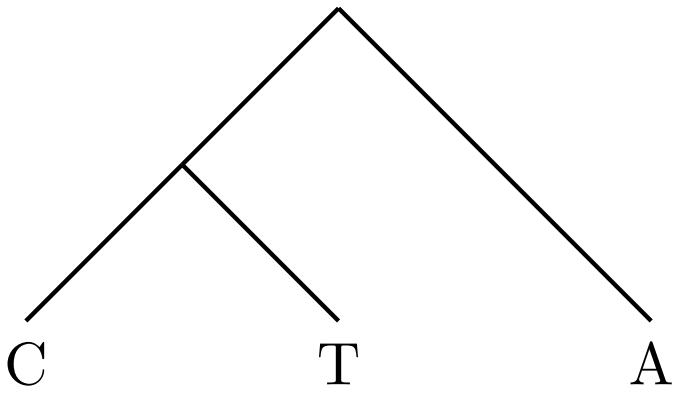
Phylogenetic Models



Human : GATCTCAAGGAC
Chimp : GGCCTCAAGGAT
Gorilla : GATCTCCAGGCA

- We label the leaves of the tree with the base that each species has at a fixed site in their DNA
- Each tree gives a family of distributions on columns in the alignment
- Maximum Likelihood Estimation can then be used to find the tree that maximizes the probability of the data

Phylogenetic Models



Human : AATGGGACATG**C**

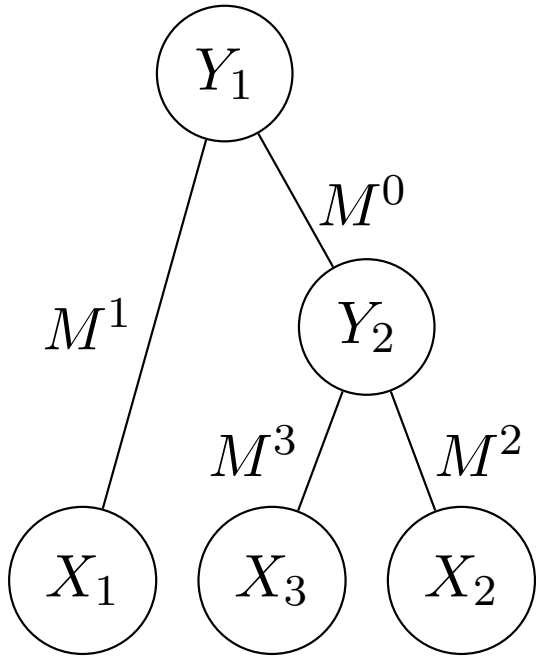
Chimp : AATGGCACATG**T**

Gorilla : AACGGGACATA**A**

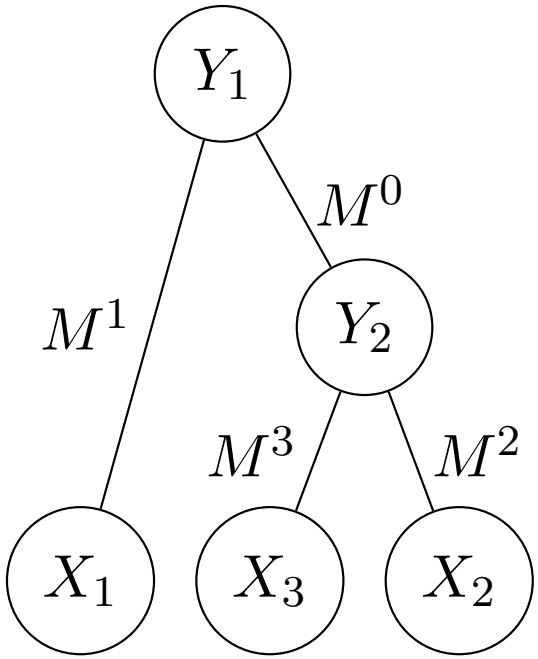
- We label the leaves of the tree with the base that each species has at a fixed site in their DNA
- Each tree gives a family of distributions on columns in the alignment
- Maximum Likelihood Estimation can then be used to find the tree that maximizes the probability of the data

Phylogenetic Models

- Assume each site evolves independently
- Phylogenetic models are hidden variable graphical models
- Each leaf v is an observed random variable $X_v \in \{\text{A, C, G, T}\}$
- Each internal node v is a hidden random variable Y_v
- Associate a transition matrix M^e to each edge $e = (u, v)$ and a distribution π to the root



Phylogenetic Models



- The probability of observing $(x_1, x_2, x_3) \in \{A, C, G, T\}^3$ is

$$P(x_1, x_2, x_3) = \sum_{y_1} \sum_{y_2} \pi_{y_1} M^0_{y_1,y_2} M^1_{y_1,x_1} M^2_{y_2,x_2} M^3_{y_2,x_3}$$

Types of Phylogenetic Models

- First require that $M^e = \exp(Q^e t)$ for a *rate matrix* Q^e and parameter t_e
- Further restrictions can be imposed on the rate matrices

$$\begin{bmatrix} * & \alpha \\ \alpha & * \end{bmatrix}$$

CFN

$$\begin{bmatrix} * & \beta & \alpha & \gamma \\ \beta & * & \gamma & \alpha \\ \alpha & \gamma & * & \beta \\ \gamma & \alpha & \beta & * \end{bmatrix}$$

K3P

$$\begin{bmatrix} * & \alpha & \alpha & \alpha \\ \alpha & * & \alpha & \alpha \\ \alpha & \alpha & * & \alpha \\ \alpha & \alpha & \alpha & * \end{bmatrix}$$

JC

$$\begin{bmatrix} * & \beta & \alpha & \beta \\ \beta & * & \beta & \alpha \\ \alpha & \beta & * & \beta \\ \beta & \alpha & \beta & * \end{bmatrix}$$

K2P

Algebraic Perspective on Phylogenetic Models

- Once we fix a tree T with n leaves we get a polynomial map in the entries of π and the M^e

$$\psi_T : \Theta_T \to \mathbb{R}^{4^n}$$

- The phylogenetic model associated to T is $M_T = \text{im}(\psi_T) \subseteq \mathbb{R}^{4^n}$
- $\Theta \subset \mathbb{R}^d$ is the space of numerical parameters (rate matrices Q^e and time parameters t^e)
- This gives a family of parametric algebraic statistical models indexed by the discrete parameter T
- Let V_T be the Zariski closure of the model

Phylogenetic Mixture Models

- Mixture models can be used to model more complicated evolutionary events such as horizontal gene transfer or hybridization
- The 2-tree mixture model for trees T_1 and T_2 is parameterized by

$$\psi_{T_1,T_2} : \Theta_{T_1} \times \Theta_{T_2} \times [0,1] \to \Delta_{4^n-1}$$

  defined by

$$\psi_{T_1,T_2}(\theta_1, \theta_2, \lambda) = \lambda\psi_{T_1}(\theta_1) + (1-\lambda)\psi_{T_2}(\theta_2)$$

- This gives a family of parametric algebraic statistical models indexed by multisets $\{T_1, T_2\}$
- The Zariski closure of the image is the *join variety* $V_{T_1} * V_{T_2}$

Identifiability

Definition

A parametric statistical model is *identifiable* if it gives a 1-1 map from parameters to probability distributions.

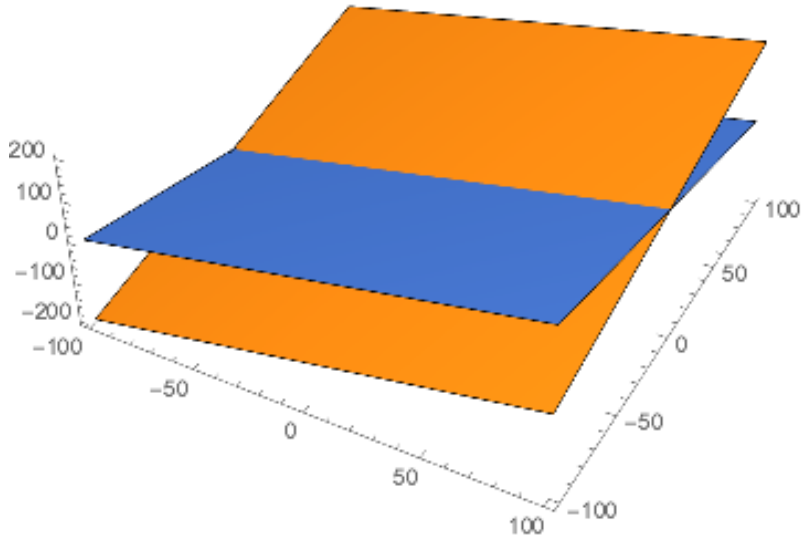
- Identifiability is needed for consistency of inference
- In phylogenetics, the identifiability of the tree parameter is particularly important
- Can T or $\{T_1, T_2\}$ be recovered from DNA sequence data?

Generic Identifiability of Discrete Parameters

Definition

Let $\{M_s\}_{s=1}^k$ be a collection of algebraic models that sit inside the probability simplex Δ_r, then the discrete parameter s is *generically identifiable* if for each 2-subset $\{s_1, s_2\} \subset [k]$

$$\dim(M_{s_1} \cap M_{s_2}) < \min(\dim(M_{s_1}), \dim(M_{s_2}))$$



Algebraic Tools for Testing Generic Identifiability

- Let $k[p] = k[p_1, p_2, \ldots p_r]$ denote the polynomial ring in indeterminates $p_1, p_2, \ldots p_r$

Definition

Let $S \subseteq k^r$. The *vanishing ideal* of S, denoted $\mathcal{I}(S)$ is

$$\mathcal{I}(S) = \{f \in k[p] : f(a) = 0 \text{ for all } a \in S\} \subseteq k[p]$$

- The ideal $I_T = \mathcal{I}(M_T)$ is called the ideal of *phylogenetic invariants* of T

Algebraic Tools for Testing Generic Identifiability

Proposition

Let M_1 and M_2 be two irreducible algebraic models which sit inside the probability simplex Δ_r. If there exists polynomials f_1 and f_2 such that

$$f_1 \in \mathcal{I}(M_1) \setminus \mathcal{I}(M_2) \text{ and } f_2 \in \mathcal{I}(M_2) \setminus \mathcal{I}(M_1)$$

then $\dim(M_1 \cap M_2) < \min(\dim(M_1), \dim(M_2))$.

- Since the models are irreducible, the ideals $\mathcal{I}(M_s)$ are prime
- If the models are the same dimension, then it suffices to show $\mathcal{I}(M_1) \neq \mathcal{I}(M_2)$
- Finding polynomials f_1 and f_2 can be quite difficult

Generic Identifiability of Tree Parameters

- The tree parameter is identifiable of the JC, CFN, K2P, and K3P models are generically identifiable
- The tree parameters of the 2-tree JC and K2P mixture models are generically identifiable (Allman-Petrovic-Rhodes-Sullivant 2009)
- The tree parameters of the 3-tree JC mixture model are generically identifiable (Long - Sullivant 2015)

Matroids

- A matroid is a combinatorial object used to axiomatize independence
- Characterized by a ground set E and independent sets $I \subseteq E$

Definition

A *matroid* is a pair $(E, \mathcal{I})$, where $I \subseteq 2^E$ that satisfies

1. $\emptyset \in \mathcal{I}$
2. If $S \subseteq T$ and $T \in I$, then $S \in I$
3. If $S, T \in \mathcal{I}$ and $\#S < \#T$, then there exists $e \in T \setminus S$ such that $S \cup \{e\} \in \mathcal{I}$

Linear Matroids

Definition

A *linear matroid* is one where $E \subset k^n$ is a finite subset, and $S \in \mathcal{I}$ if and only if S is linearly independent over k

Example (Linear Matroid)

$$A = \begin{bmatrix} 1 & 1 & -1 & -2 \\ 3 & 1 & 2 & 4 \\ 0 & -1 & 1 & 2 \end{bmatrix}$$

- $E = [4]$
- The independent sets are

$$\{1\}, \{2\}, \{3\}, \{4\}, \{1,2\}, \{1,3\}, \{1,4\}, \{2,3\}, \{2,4\}, \{1,2,3\}, \{1,2,4\}.$$

Algebraic Matroids

- Since $\mathcal{I}(M_s)$ is a prime ideal it defines an *algebraic matroid* on the set of coordinates $E = \{p_i : i \in [r+1]\}$ with independent sets

$$\{S \subseteq E : \mathcal{I}(M_s) \cap \mathbb{C}[S] = \langle 0 \rangle\}$$

- Let $M_s = \text{im}(\phi)$ with $\phi(\theta_1, \ldots, \theta_d) = (\phi_1(\theta), \ldots, \phi_{r+1}(\theta))$ and let

$$J(\phi) = \left(\frac{\partial \phi_j}{\partial \theta_i} \right), 1 \leq i \leq d,\ 1 \leq j \leq r+1$$

- The matroid defined by the columns of $J(\phi)$ over the fraction field $\mathbb{C}(\theta)$ is the same matroid defined by $\mathcal{I}(M_s)$
- Let $\mathcal{M}(M_s)$ be the independence matroid of the model defined in either of these ways

Proving Identifiability with Algebraic Matroids

Proposition (H - Sullivant)

Let M_1 and M_2 be two irreducible algebraic models which sit inside the probability simplex Δ_r. Without loss of generality assume $\dim(M_1) \geq \dim(M_2)$. *If there exists a subset S of the coordinates such that*

$$S \in \mathcal{M}(M_2) \setminus \mathcal{M}(M_1)$$

then $\dim(M_1 \cap M_2) < \min(\dim(M_1), \dim(M_2))$.

- Allows us to prove identifiability results without computing $\mathcal{I}(M_s)$
- Still requires symbolic computation over $k(\theta)$

Specializing the Jacobian

Proposition

Let k be a field of characteristic zero and ϕ be a rational map. Then the matrix obtained by plugging generic parameter values into $J(\phi)$ gives a linear matroid over k which is the same as that defined by $J(\phi)$ with symbolic parameters over $k(\theta)$

- $\mathcal{M}(J(\phi), k(\theta))$ = independence matroid over $k(\theta)$
- $\mathcal{M}(J(\phi), k)$ = independence matroid over k obtained by plugging in random values for θ

Certifying Identifiability with Algebraic Matroids

Algorithm 1: matroidSeparate

Input : Two maps ϕ_1, ϕ_2 parameterizing models M_1 and M_2 in k^n with $\dim(M_1) \geq \dim(M_2)$, a number of trials t.

Output: A certificate S

for $i = 0$ **to** t **do**
  Randomly select $T \subseteq [n]$ such that $|T| \leq \dim(M_2)$;
  if $T \in \mathcal{M}(J(\phi_2), k) \setminus \mathcal{M}(J(\phi_1), k)$ **then**
    if $T \in \mathcal{M}(J(\phi_2), k(\theta)) \setminus \mathcal{M}(J(\phi_1), k(\theta))$ **then**
      S = T;
      Break;
return *S or report that no certificate was found.*

- Still requires symbolic computation over $k(\theta)$
- Embarrassingly parallel

The Schwartz-Zippel Lemma

Lemma (Schwartz-Zippel)

Let $f \in k[x_1, \ldots x_n]$ be a non-zero polynomial of total degree α. Let E be a finite subset of k and $r_1, \ldots r_n$ be selected at random independently and uniformly from E. Then

$$P(f(r_1, \ldots, r_n) = 0) \leq \frac{\alpha}{|E|}.$$

- $S \notin \mathcal{M}(J(\phi_1), k(\theta))$ if the corresponding minor of $J(\phi_1)$ vanishes
- Main algorithm can be modified to avoid symbolic computation and produce a certificate that holds with probability $1 - \varepsilon$ by using this lemma

Six-to-Infinity Theorem

Theorem (Six-To-Infinity Theorem (Matsen-Mossel-Steel 2008))

Suppose that the tree parameters T_1, T_2 are identifiable for a 2-tree mixture model for trees with six leaves. Then the tree parameters are identifiable for trees with n leaves for all $n \geq 6$.

- Only finitely many cases to check since it is enough to check for every pair of 2-multisets of 6 leaf trees

Identifiability for CFN and K3P

Theorem (H - Sullivant)

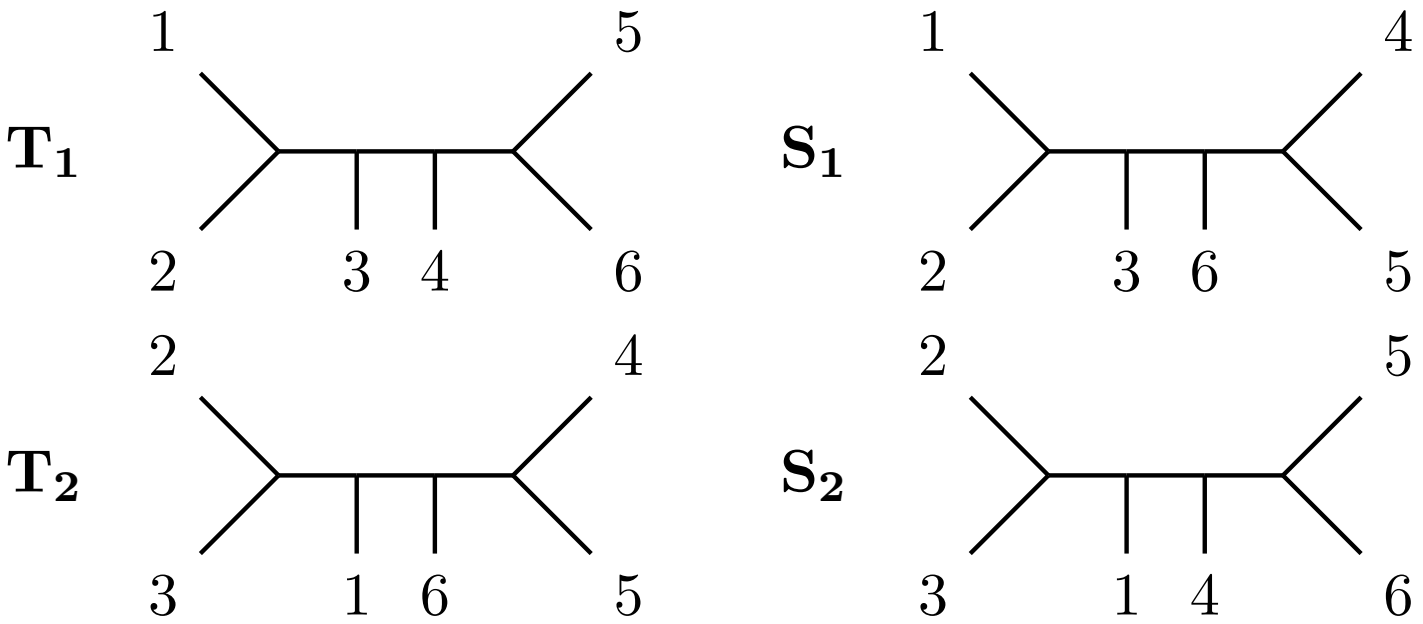
The tree parameters of the 2-tree CFN mixture model are generically identifiable for trees with at least six leaves and the tree parameters of the 2-tree K3P mixture model are generically identifiable for trees with at least four leaves.

Proof idea:

- By the Six-To-Infinity Theorem of Matsen, Mossel, and Steel (2008) its enough to prove identifiability for six leaf trees
- There are 22,773 cases to check up to symmetry
- Run the main algorithm for each case to find a certificate of identifiability
- In one case it failed but we were able to compute a degree-bounded Gröbner basis in this case

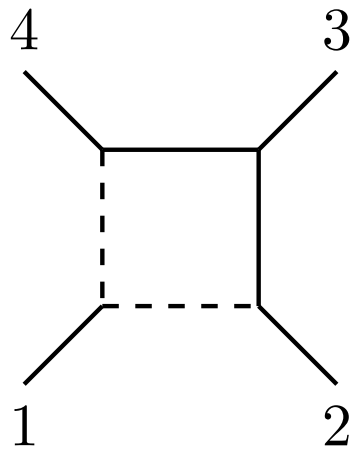
Why Did the Algorithm Fail?

- Different prime ideals can have the same matroid
- We conjecture that the ideals we get from the trees below have the same matroid despite having different ideals



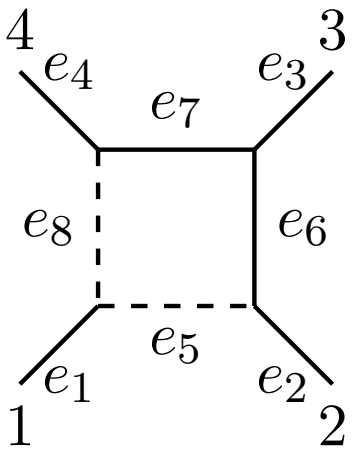
Phylogenetic Networks

- Recent tool that has emerged to model evolutionary phenomena that are non-treelike such as horizontal gene transfer
- Solid edges are called *tree edges*
- Dotted edges are *reticulation edges* which represent horizontal gene transfer
- Networks can be thought of as cycles connected by trees

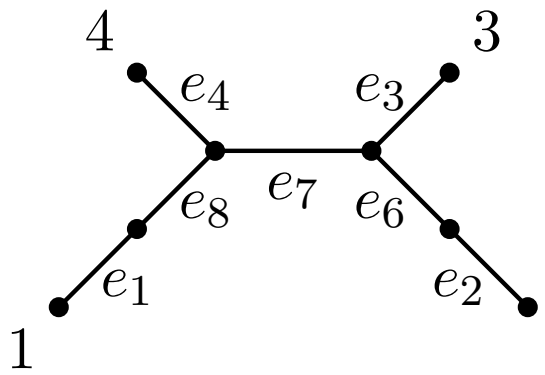


Phylogenetic Networks

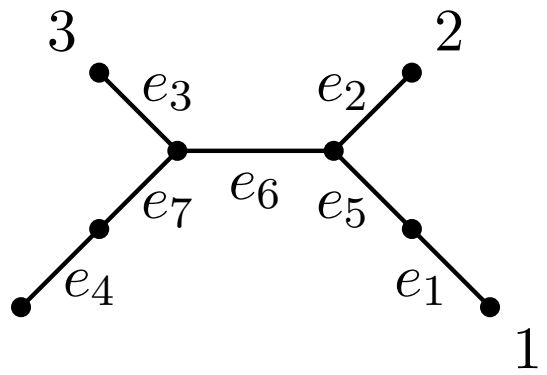
- As the number of cycles and number of allowable reticulation edges increases the model becomes increasingly complicated
- A good starting point is a single cycle with a single reticulation vertex, called a *cycle network*
- Deleting a reticulation edge e_i from the network N gives a tree T_i



(a) N



(b) T_1



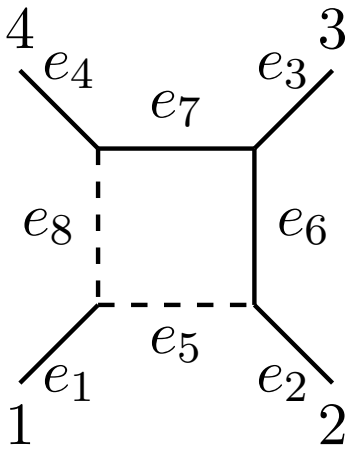
(c) T_2

Phylogenetic Network Models

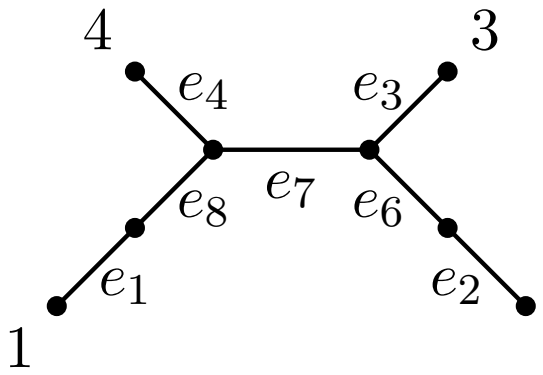
- A model for trees ψ_T gives us a model ψ_N for cycle networks where

$$\psi_N = \lambda\psi_{T_1} + (1-\lambda)\psi_{T_2}$$

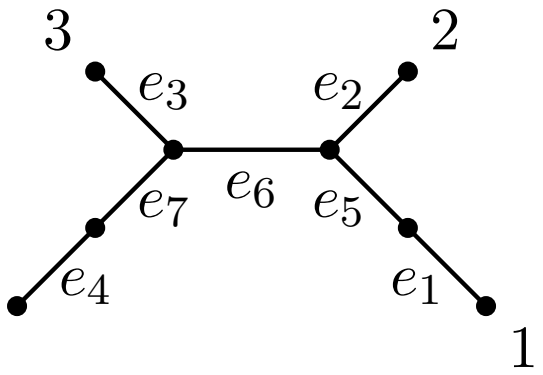
- This is not the same as mixture model since the parameters on each tree are not independent



(a) N



(b) T_1



(c) T_2

Identifiability for Phylogenetic Network Models

- If T is one of the trees obtained from a network N then $\text{im}(\psi_T) \subseteq \text{im}(\psi_N)$ so in general the cycle-network parameter is not identifiable
- Gross and Long suggested limiting the question to large cycle networks (cycle size $k \geq 4$)
- They proved that the network parameter is identifiable for large cycle networks under the JC model
- Similar to the tree case, they show that the question can be reduced to a finite number of cases and then computed ideals explicitly in these cases

Identifiability for Phylogenetic Network Models

Theorem (H - Sullivant)

The semi-directed network parameter of large-cycle K2P and K3P network models is generically identifiable.

Proof idea:

- Use results of Gross and Long to reduce to a finite number of cases
- Use our matroid algorithm to prove identifiability in each case

Summary

- Algebraic matroids can be used to show discrete parameters are generically identifiable
- Using matroids allows us to avoid computing $\mathcal{I}(M)$
- Using the Schwartz-Zippel Lemma we can completely avoid computing over $k(\theta)$ and give a certificate of generic identifiability with probability $1 - \epsilon$
- We used it to prove that the tree parameters of 2-tree CFN and K3P mixture models are generically identifiable
- We also used this method to prove that the network parameter in K2P and K3P large-cycle network models is generically identifiable

References

Elizabeth S Allman, Sonia Petrovic, John A Rhodes, and Seth Sullivant.
Identifiability of two-tree mixtures for group-based models.
IEEE/ACM transactions on computational biology and bioinformatics, 8(3):710–722, 2010.

Elizabeth Gross and Colby Long.
Distinguishing phylogenetic networks.
SIAM Journal on Applied Algebra and Geometry, 2(1):72–93, 2018.

Colby Long and Seth Sullivant.
Identifiability of 3-class Jukes-Cantor mixtures.
Adv. in Appl. Math., 64:89–110, 2015.

Frederick A. Matsen, Elchanan Mossel, and Mike Steel.
Mixed-up trees: the structure of phylogenetic mixtures.
Bull. Math. Biol., 70(4):1115–1139, 2008.

Zvi Rosen.
Computing algebraic matroids.
arXiv preprint arXiv:1403.8148, 2014.

Seth Sullivant.
Algebraic statistics, volume 194 of *Graduate Studies in Mathematics*.
American Mathematical Society, Providence, RI, 2018.