Identifiability in Phylogenetics using Algebraic Matroids

Benjamin Hollering and Seth Sullivant

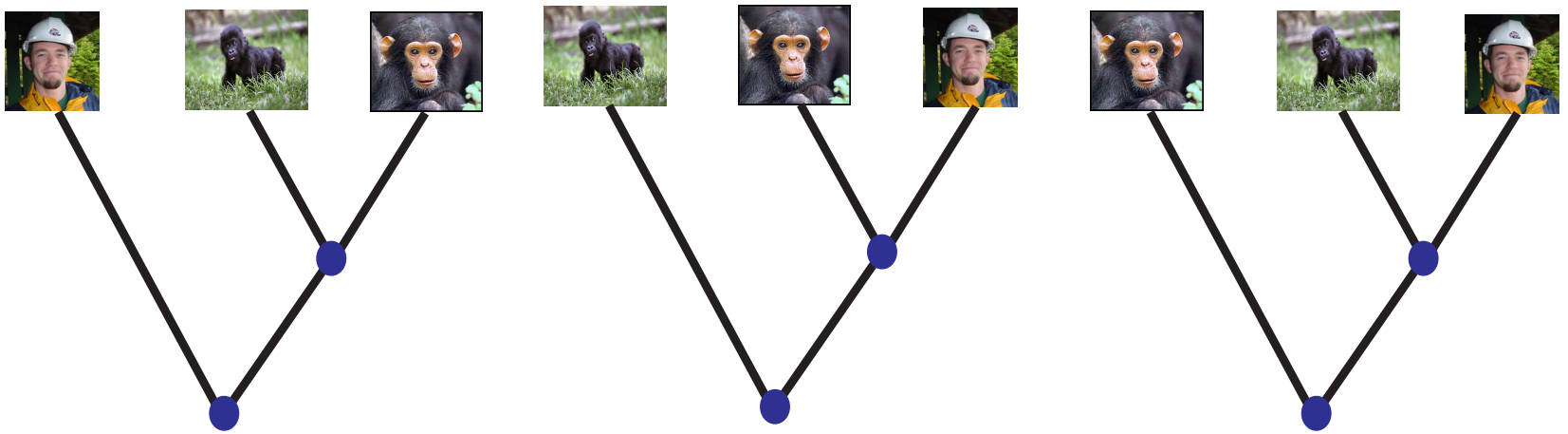
North Carolina State University

March 19, 2025

Phylogenetics

Problem

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



- Data consists of aligned DNA sequences from homologous genes

Human:	...ACCGTGCAACGTGAACGA...
Chimp:	...ACCTTGGAAGGTAAACGA...
Gorilla:	...ACCGTGCAACGTAAACTA...

Model-Based Phylogenetics

- Use a probabilistic model of mutations
- Parameters for the model are the combinatorial tree T, and rate parameters for mutations on each edge
- Models give a probability for observing a particular aligned collection of DNA sequences

Human:	ACCGTGCAACGTGAACGA
Chimp:	ACGTTGCAAGGTAAACGA
Gorilla:	ACCGTGCAACGTAAACTA

- Assuming site independence, data is summarized by empirical distribution of columns in the alignment.
- e.g. $\hat{p}(AAA) = \frac{6}{18}$, $\hat{p}(CGC) = \frac{2}{18}$, etc.
- Use empirical distribution and test statistic to find tree best explaining data
- This talk introduces some algebraic methods for proving identifiability of these models.

Identifiability of Statistical Models

- A parametric algebraic statistical model M for discrete random variables is the image of a rational map

$$\phi : \Theta \to \Delta_r = \left\{ p \in \mathbb{R}^{r+1} \ : \ \sum_{i=0}^{r} p_i = 1, \ p_i \geq 0 \text{ for all } i \right\}$$

Definition

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

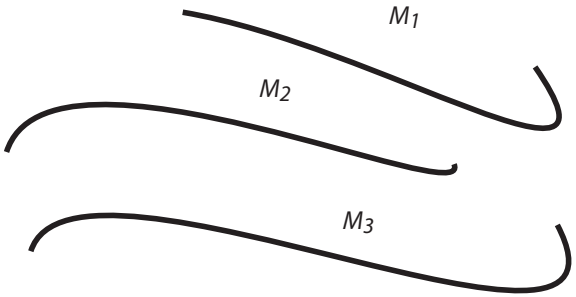
- Identifiability is needed for consistency of inference (e.g. consistency of ML)
- "Is it possible to infer model parameters, even with perfect data?"

Identifiability of Discrete Parameters

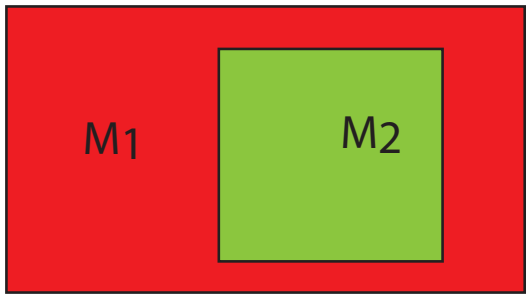
Definition

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 identifiable if

$$M_{s_1} \cap M_{s_2} = \emptyset \text{ for all } s_1 \neq s_2.$$



Identifiable



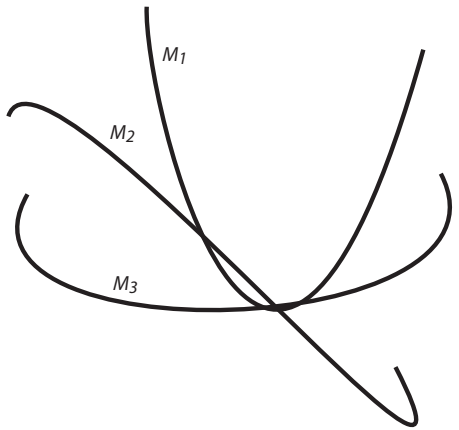
Not Identifiable

Generic Identifiability

Definition

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

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



Motivation: Phylogenetic Models

- For each tree T with leaf label set $[n] = \{1, 2, \ldots, n\}$, we get a phylogenetic model $M_T \subseteq \Delta_{4^n}$. Consists of all probability distributions that come from T.
- So our collection of models is $\{M_T : T \in BT(n)\}$, one model for each (binary) tree on n leaves.
- (Generic) Identifiability of the discrete parameter $\leftrightarrow$ Can the tree parameter be recovered from (perfect) data
- More complex phylogenetic settings might have a more complex discrete parameter:
  - Mixture models: (Multi)-Sets of trees
  - Networks

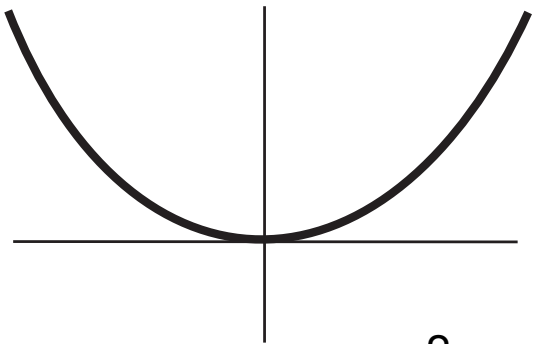
Proving Generic Identifiability with Algebra

- $\mathbb{R}[p] = \mathbb{R}[p_0, \ldots, p_r]$ denotes the polynomial ring in indeterminates $p_0, \ldots, p_r$ with coefficients in $\mathbb{R}$.

Definition

Let $S \subseteq \mathbb{R}^{r+1}$. The vanishing ideal of S, denoted $I(S)$ is the set

$$\mathcal{I}(S) = \{f \in \mathbb{R}[p] : f(\mathbf{a}) = 0 \text{ for all } \mathbf{a} \in S\}.$$



$$\mathcal{I}(S) = \langle y - x^2 \rangle$$

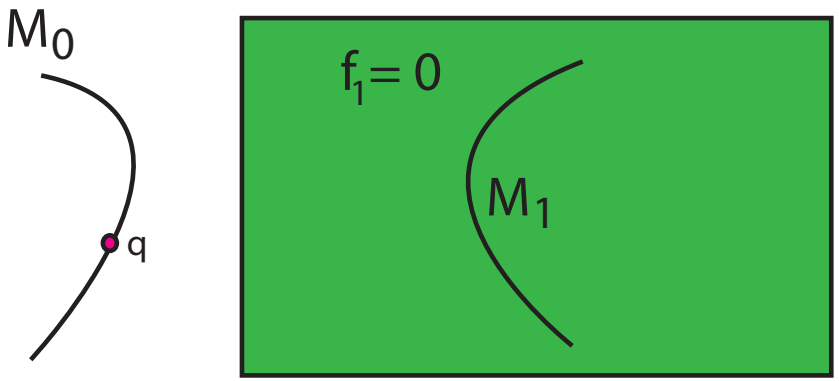
Proving Generic Identifiability with Algebra

Proposition

Let M_0 and M_1 be two irreducible algebraic models which sit inside the probability simplex Δ_r. If there exists polynomials f_0 and f_1 such that

$$f_0 \in \mathcal{I}(M_0) \setminus \mathcal{I}(M_1) \text{ and } f_1 \in \mathcal{I}(M_1) \setminus \mathcal{I}(M_0)$$

then $\dim(M_0 \cap M_1) < \min(\dim(M_0), \dim(M_1))$.



Applications to Phylogenetics

- The algebraic perspective has been very useful for proving generic identifiability in phylogenetic models.
  - Invariable Sites (Allman-Rhodes 2008)
  - Covarion models (Allman-Rhodes 2009)
  - Two-tree mixtures (Allman-Petrovic-Rhodes-Sullivant 2010)
  - Same tree mixtures (Stefankovic-Vigoda 2007, Rhodes-Sullivant 2012)
  - Three tree Jukes-Cantor mixtures (Long-Sullivant 2015)

- Requires knowing structural information about $\mathcal{I}(M_i)$, or
- Requires very time consuming Gröbner basis calculations.

Matroids

- A matroid is a combinatorial object used to axiomatize independence
- Characterized by a ground set E and independent sets $I \subseteq E$
- Axiom capture the "combinatorial essence" of independence structures

Definition

A matroid is a pair $(E, \mathcal{I})$, with $\mathcal{I} \subseteq 2^E$ satisfying

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

Example: Linear Matroids

Definition

A linear matroid is one where $E \subseteq \mathbb{K}^n$, and $S \in \mathcal{I}$ if and only if S is linearly independent.

Example (Linear Matroid)

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

- $E = [4] = \{1, 2, 3, 4\}$
- The independent sets are

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

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

Example: Algebraic Matroids

- Let $P \subseteq \mathbb{R}[p]$ be a prime ideal. It defines an algebraic matroid on the set of coordinates $E = \{p_i : i \in [r]\}$ with independent sets

$$\{S \subseteq E : P \cap \mathbb{C}[S] = \langle 0 \rangle\}$$

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

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

- 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 algebraic matroid of the model defined in either of these ways

Using Algebraic Matroids to Prove Identifiability

Proposition (Hollering-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 $k(\theta)$

Practical Aspects

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 $k(\theta)$

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

Practical Aspects

Algorithm (Hollering- Sullivant)

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 ensuring $\dim(M_1 \cap M_2) < \dim(M_2)$.
For $i = 1, 2, \ldots t$:

- Randomly select $T \subseteq [n]$ such that $|T| \leq \dim(M_2)$.
- If $T \in \mathcal{M}(J(\phi_2), k) \setminus \mathcal{M}(J(\phi_1), k)$
  - If $T \in \mathcal{M}(J(\phi_2), k(\theta)) \setminus \mathcal{M}(J(\phi_1), k(\theta))$
    - Then, $S = T$

Output: S or announce no certificate was found

- Can skip the symbolic verification step and use Schwarz-Zippel Lemma to prove identifiability with probabilistic guarantees.

Applications: Phylogenetic Mixture Models

- Given trees $T_1, \ldots, T_r$ on n leaves have models $\mathcal{M}_{T_1}, \ldots, \mathcal{M}_{T_r}$.
- Mixture model $\mathcal{M}_{T_1} * \mathcal{M}_{T_2} * \cdots * \mathcal{M}_{T_r}$ is

$$\{\sum_i \pi_i \mathbf{p}^i : \pi \in \Delta_{r-1}, \mathbf{p}^i \in \mathcal{M}_{T_i}, i = 1, \ldots, r\}$$

- Used to model heterogeneity in substitution processes.

Problem

For a given model structure and number of mixture classes prove that the multiset of tree parameters $\{T_1, \ldots, T_r\}$ is generically identifiable.

- Two-tree Jukes-Cantor and Kimura 2P mixtures (Allman-Petrovic-Rhodes-Sullivant 2010)
- Three tree Jukes-Cantor mixtures (Long-Sullivant 2015)

“Six to Infinity Theorem” for 2-Tree Mixtures

- To prove identifiability results computationally requires a combinatorial theorem to reduce to a finite number of cases.

Theorem

Every binary phylogenetic tree is uniquely determined by its restriction trees to four leaves (quartets).

Corollary

For ordinary phylogenetic models, identifiability of the tree parameter on 4 leaf trees implies identifiability for arbitrary number of leaves.

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

For 2 tree mixtures, it suffices to prove identifiability of the tree parameters on trees with 6 leaves.

Results: Two Tree Mixtures

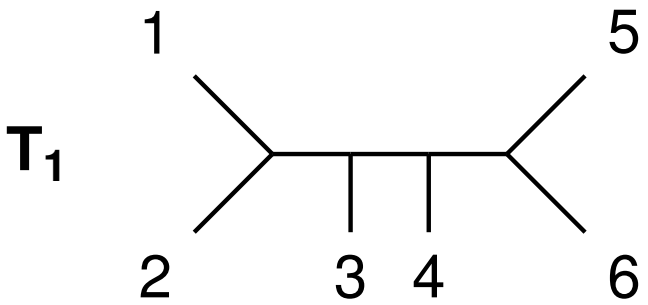
Theorem (Hollering - Sullivant)

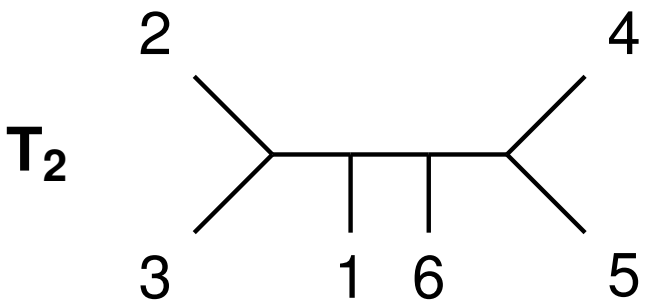
The tree parameters of the 2-tree CFN mixture model are generically identifiable for trees with at least six leaves.

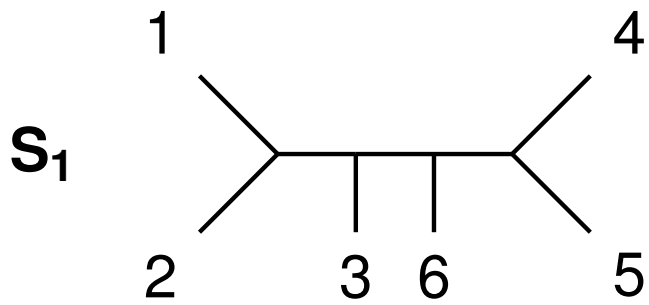
Proof idea:

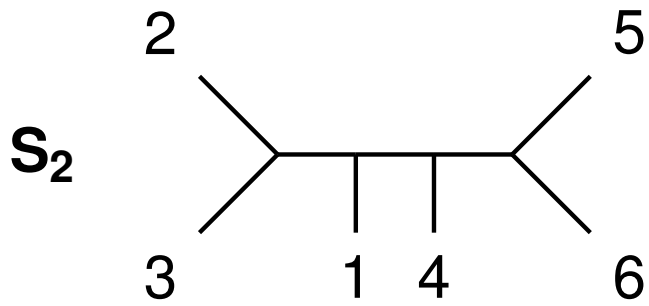
- By the Six-To-Infinity Theorem its enough to prove identifiability for six leaf trees
- There are 22,773 cases to check up to symmetry (need to check pairs of pairs of 6 leaf trees)
- Run the matroid algorithm for each case to find a certificate of identifiability
- In one case it failed but we were able to compute a degree-bounded Gröbner basis in this case

An Interesting Example







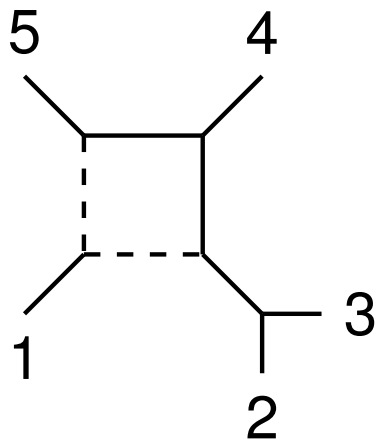


- Different prime ideals might have the same matroid.
- These two pairs of trees seem to have the same matroid (in our random sampling).
- We needed to compute a Gröbner basis to rule out this pair in the case of CFN model.

Level 2 Networks (On going project)

Theorem (Gross-van Iersel-Janssen-Jones-Long-Murakami)

The network parameter of a triangle-free level-1 semidirected network with a fixed number of reticulations is generically identifiable, under JC, K2P, K3P markov models ("displayed tree model", no coalescent).

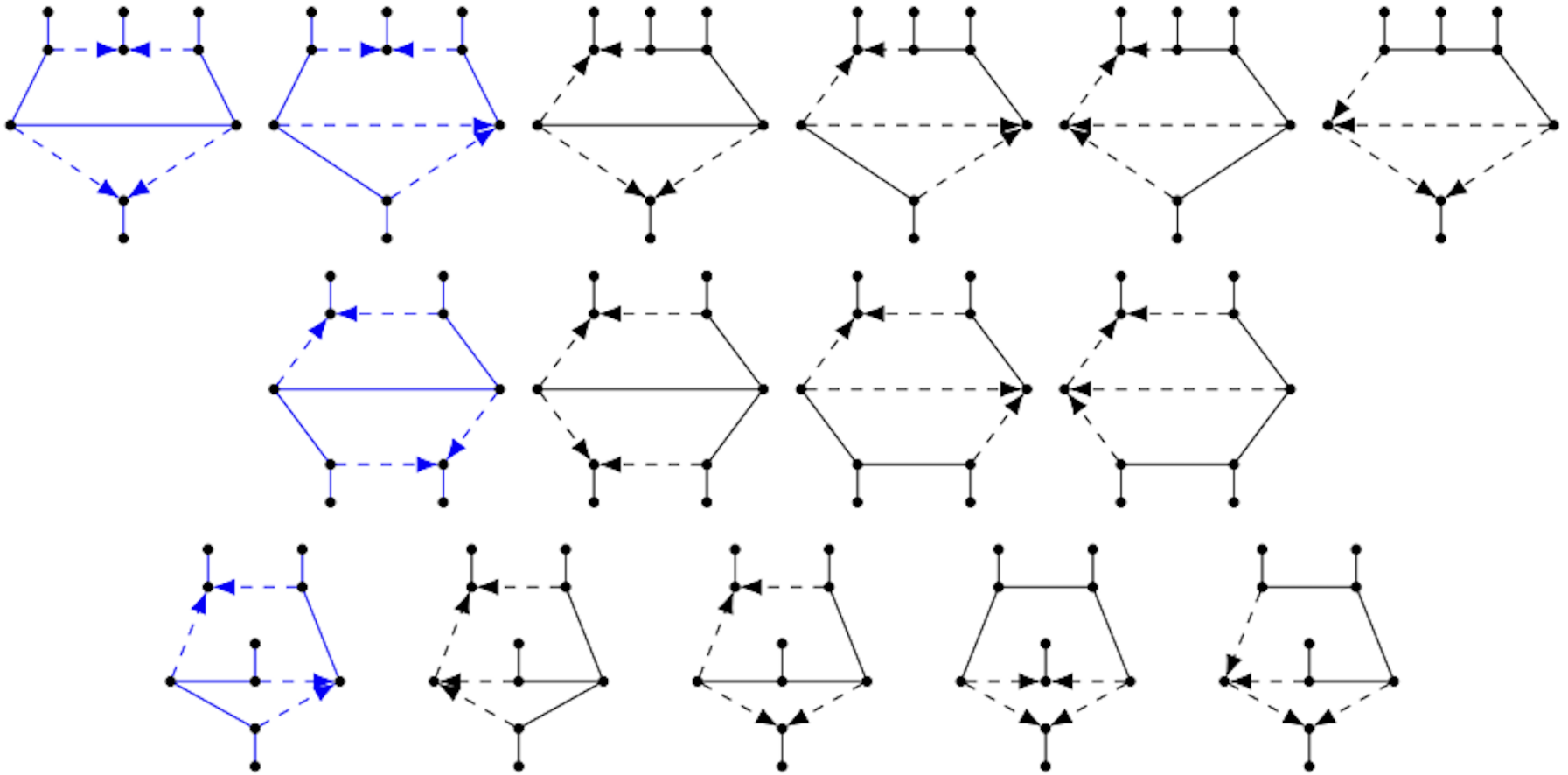


Question

Can we extend these identifability results to level 2 networks?

- with Englander, Frohn, Gross, Holtgrefe, van Iersel, Jones

Level 2 Networks



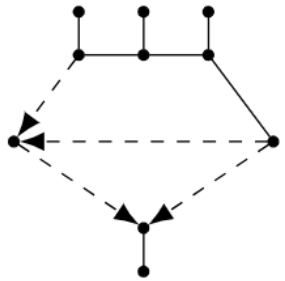
- There are 294 simple level-2 quarnet 4 blobs to distinguish.
- So we must check distinguishability for $\binom{294}{2} = 43071$ pairs.

Identifiability of Level-2 Networks

Theorem (EFGHvIJS)

The network parameter under the Jukes-Cantor model is generically identifiable when the network parameter is an n-leaf binary, triangle-free, strongly tree-child, level-2 semi-directed phylogenetic network.

- Proof uses a range of tools.
  - Matroids
  - Phylogenetic Invariants/Ideals
  - Inequalities
- Challenges: Stacked reticulations, triangles



Summary

- Tools from algebraic geometry can be useful for producing new identifiability results.
- Methods based on ideals are powerful, but can run out of steam if ideal generators are not easy to determine.
- Methods based on algebraic matroids can be more computationally tractable, including giving results with probabilistic guarantees.
- We applied methods to phylogenetic mixture models and network models.

References


E. Allman, S. Petrovic, J. Rhodes, S. Sullivant. Identifiability of two-tree mixtures for group-based models. *IEEE/ACM TCBB* **8** no. 3 (2011) 710-722. 0909.1854


ES Allman, JA Rhodes. Identifying evolutionary trees and substitution parameters for the general Markov model with invariable sites *Mathematical Biosciences* **211** (no. 1) (2008) 18-33


ES Allman, JA Rhodes. The identifiability of covarion models in phylogenetics, with E. S. Allman, *IEEE/ACM Transactions in Computational Biology and Bioinformatics*, **6** no. 1 (2009), 76–88.


E Gross, C Long. Distinguishing Phylogenetic Networks. *SIAM J. Appl. Algebra Geometry.* **2** (2018), no. 1, 72–93


B Hollering, S Sullivant. Identifiability in Phylogenetics using Algebraic Matroids. *J. Symbolic Computation***104** (2021),142-158. arXiv:1909.13754


C. Long, S. Sullivant. Identifiability of Jukes-Cantor 3-tree mixtures *Adv. in Appl. Math.* **64** (2015), 89–110. 1406.7256


F Matsen, E Mossel, and M Steel. Mixed-up trees: the structure of phylogenetic mixtures. *Bull. Math. Biol.*, **70** (4):1115–1139, 2008.


J. Rhodes, S. Sullivant. Identifiability of large phylogenetic mixture models. *Bull. Math. Biol.* **74** (2012), no. 1, 212–231. 1011.4134


D. Stefankovic and E. Vigoda. Pitfalls of Heterogeneous Processes for Phylogenetic Reconstruction *Systematic Biology* **56**(1): 113-124, 2007.