Recovering a tree from randomly sampled phylogenetic diversities

Steven N. Evans

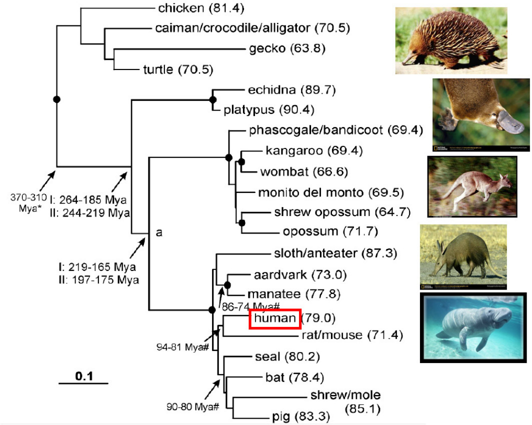
June, 2015

Department of Mathematics & Department of Statistics
Group in Computational and Genomic Biology
Group in Computational Science and Engineering
University of California at Berkeley

Collaborator

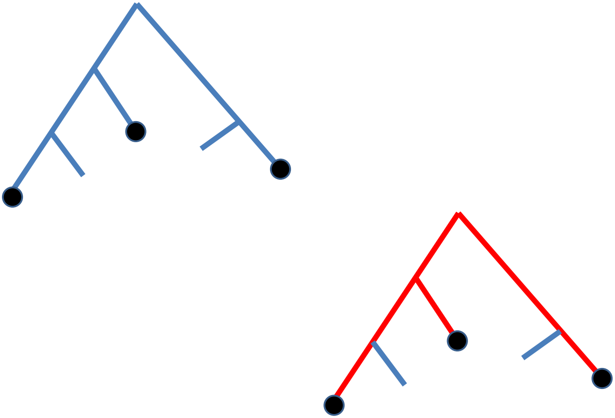
Daniel Lanoue, Berkeley
Pre-print at http://arxiv.org/abs/1506.01091.

Phylogenetic trees



A phylogenetic tree is just a tree with weights / lengths on the edges and labels on the leaves.

Phylogenetic diversity



- The phylogenetic diversity of a collection of taxa (= leaves) is the total length of the subtree they span.
- This quantity is important in ecology and conservation.

Some notation

- Consider a tree $\mathbf{T}$ with
  - vertex set $\mathbf{V}(\mathbf{T})$,
  - edge set $\mathbf{E}(\mathbf{T})$,
  - leaf set $\mathbf{L}(\mathbf{T})$,
  - edge-lengths (edge-weights) $\mathbf{W}_{\mathbf{T}} : \mathbf{E}(\mathbf{T}) \to \mathbb{R}_{++}$.
- For $x, y \in \mathbf{V}(\mathbf{T})$ let $r_{\mathbf{T}}(x, y) :=$ length of the the (unique) path between x and y ($=$ sum of the lengths of the edges on the path).
- Given $K \subseteq \mathbf{L}(\mathbf{T})$, write $\mathbf{W}_{\mathbf{T}}(K)$ for the length of the subtree spanned by K ($=$ the phylogenetic diversity of K).

Some notation

- Consider a tree $\mathbf{T}$ with
  - vertex set $\mathbf{V}(\mathbf{T})$,
  - edge set $\mathbf{E}(\mathbf{T})$,
  - leaf set $\mathbf{L}(\mathbf{T})$,
  - edge-lengths (edge-weights) $\mathbf{W}_{\mathbf{T}} : \mathbf{E}(\mathbf{T}) \to \mathbb{R}_{++}$.
- For $x, y \in \mathbf{V}(\mathbf{T})$ let $r_{\mathbf{T}}(x, y) :=$ length of the the (unique) path between x and y ($=$ sum of the lengths of the edges on the path).
- Given $K \subseteq \mathbf{L}(\mathbf{T})$, write $\mathbf{W}_{\mathbf{T}}(K)$ for the length of the subtree spanned by K ($=$ the phylogenetic diversity of K).

Some notation

- Consider a tree $\mathbf{T}$ with
  - vertex set $\mathbf{V}(\mathbf{T})$,
  - edge set $\mathbf{E}(\mathbf{T})$,
  - leaf set $\mathbf{L}(\mathbf{T})$,
  - edge-lengths (edge-weights) $\mathbf{W}_{\mathbf{T}} : \mathbf{E}(\mathbf{T}) \to \mathbb{R}_{++}$.
- For $x, y \in \mathbf{V}(\mathbf{T})$ let $r_{\mathbf{T}}(x, y) :=$ length of the the (unique) path between x and y ($=$ sum of the lengths of the edges on the path).
- Given $K \subseteq \mathbf{L}(\mathbf{T})$, write $\mathbf{W}_{\mathbf{T}}(K)$ for the length of the subtree spanned by K ($=$ the phylogenetic diversity of K).

How do we compute phylogenetic diversities easily?

- Write $d_{\mathbf{T}}(v)$ for the degree of $v \in \mathbf{V}(\mathbf{T})$.
- For distinct $x, y \in \mathbf{L}(\mathbf{T})$,
  - $I_{\mathbf{T}}(x,y) :=$ the set of interior vertices on the (unique) path in $\mathbf{T}$ between x and y,
  - $h_{\mathbf{T}}(x,y) := \prod_{v \in I_{\mathbf{T}}(x,y)} ((d_{\mathbf{T}}(v) - 1)!)^{-1}$,
  - $r_{\mathbf{T}}(x,y) :=$ length of the the path between x and y as above.
- Then (Semple & Steel '04 extending Pauplin '00), the total length of $\mathbf{T}$ is

$$\mathbf{W}_{\mathbf{T}}(\mathbf{L}(\mathbf{T})) = \sum_{\{x,y\} \subseteq \mathbf{L}(\mathbf{T}), x \neq y} h_{\mathbf{T}}(x,y) r_{\mathbf{T}}(x,y).$$

A similar formula holds for general $W_{\mathbf{T}}(K)$.

How do we compute phylogenetic diversities easily?

- Write $d_{\mathbf{T}}(v)$ for the degree of $v \in \mathbf{V}(\mathbf{T})$.
- For distinct $x, y \in \mathbf{L}(\mathbf{T})$,
  - $I_{\mathbf{T}}(x,y) :=$ the set of interior vertices on the (unique) path in $\mathbf{T}$ between x and y,
  - $h_{\mathbf{T}}(x,y) := \prod_{v \in I_{\mathbf{T}}(x,y)} ((d_{\mathbf{T}}(v) - 1)!)^{-1}$,
  - $r_{\mathbf{T}}(x,y) :=$ length of the the path between x and y as above.
- Then (Semple & Steel '04 extending Pauplin '00), the total length of $\mathbf{T}$ is

$$\mathbf{W}_{\mathbf{T}}(\mathbf{L}(\mathbf{T})) = \sum_{\{x,y\} \subseteq \mathbf{L}(\mathbf{T}), x \neq y} h_{\mathbf{T}}(x,y) r_{\mathbf{T}}(x,y).$$

A similar formula holds for general $W_{\mathbf{T}}(K)$.

How do we compute phylogenetic diversities easily?

- Write $d_{\mathbf{T}}(v)$ for the degree of $v \in \mathbf{V}(\mathbf{T})$.
- For distinct $x, y \in \mathbf{L}(\mathbf{T})$,
  - $I_{\mathbf{T}}(x,y) :=$ the set of interior vertices on the (unique) path in $\mathbf{T}$ between x and y,
  - $h_{\mathbf{T}}(x,y) := \prod_{v \in I_{\mathbf{T}}(x,y)} ((d_{\mathbf{T}}(v) - 1)!)^{-1}$,
  - $r_{\mathbf{T}}(x,y) :=$ length of the the path between x and y as above.
- Then (Semple & Steel '04 extending Pauplin '00), the total length of $\mathbf{T}$ is

$$\mathbf{W}_{\mathbf{T}}(\mathbf{L}(\mathbf{T})) = \sum_{\{x,y\} \subseteq \mathbf{L}(\mathbf{T}), x \neq y} h_{\mathbf{T}}(x,y) r_{\mathbf{T}}(x,y).$$

A similar formula holds for general $W_{\mathbf{T}}(K)$.

The main question

WHAT DO WE LEARN
ABOUT A TREE
FROM THE PHYLOGENETIC DIVERSITIES
OF RANDOMLY SAMPLED SUBSETS OF TAXA???

Recognizing/reconstructing trees in general

- In general, what information do we need to reconstruct an edge-weighted tree?
- What information do we need to determine whether two edge-weighted trees are the same?
- The answer depends on what we mean by the term tree.

Recognizing/reconstructing trees in general

- In general, what information do we need to reconstruct an edge-weighted tree?
- What information do we need to determine whether two edge-weighted trees are the same?
- The answer depends on what we mean by the term tree.

Recognizing/reconstructing trees in general

- In general, what information do we need to reconstruct an edge-weighted tree?
- What information do we need to determine whether two edge-weighted trees are the same?
- The answer depends on what we mean by the term tree.

Leaf-labeled trees: leaf-to-leaf distances

- A leaf-labeled, edge-weighted tree can be reconstructed from its matrix of leaf-to-leaf distances (Zaretskii ‘65, Simões Peraira ‘69, Buneman ‘71, Buneman ‘74).
- We will recall WHY on the next two slides.

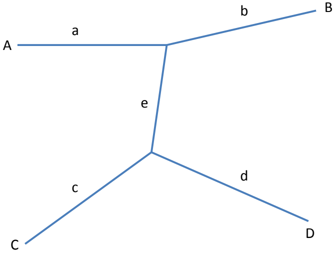
Leaf-labeled trees: leaf-to-leaf distances

- A leaf-labeled, edge-weighted tree can be reconstructed from its matrix of leaf-to-leaf distances (Zaretskii '65, Simões Peraira '69, Buneman '71, Buneman '74).
- We will recall WHY on the next two slides.

Leaf-labeled trees: why are leaf-to-leaf distances enough?

- In this four-taxon tree we can tell that A, B and C, D are siblings because

$$r_{\mathbf{T}}(A,B) + r_{\mathbf{T}}(C,D) \le r_{\mathbf{T}}(A,C) + r_{\mathbf{T}}(B,D) = r_{\mathbf{T}}(A,D) + r_{\mathbf{T}}(C,D).$$



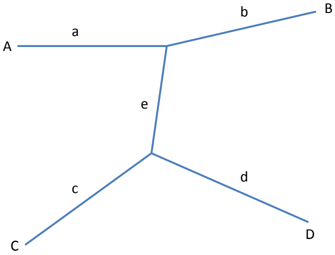
- We can recover the edge-lengths by solving six linear equations in five unknowns:

$$r_{\mathbf{T}}(A,B) = a+b,\ r_{\mathbf{T}}(A,C) = a+e+c,\ \ldots, r_{\mathbf{T}}(C,D) = c+d.$$

Leaf-labeled trees: why are leaf-to-leaf distances enough?

- In this four-taxon tree we can tell that A, B and C, D are siblings because

$$r_{\mathbf{T}}(A,B) + r_{\mathbf{T}}(C,D) \le r_{\mathbf{T}}(A,C) + r_{\mathbf{T}}(B,D) = r_{\mathbf{T}}(A,D) + r_{\mathbf{T}}(C,D).$$



- We can recover the edge-lengths by solving six linear equations in five unknowns:

$$r_{\mathbf{T}}(A,B) = a + b,\ r_{\mathbf{T}}(A,C) = a + e + c,\ \ldots,\ r_{\mathbf{T}}(C,D) = c + d.$$

Leaf-labeled trees: why are leaf-to-leaf distances enough?

- Lastly, knowing the subtree spanned by every four taxa (= quartet) suffices to determine the whole tree (“quartet puzzling”).

Leaf-labeled trees: lengths of subtrees spanned by m leaves

- A leaf-labeled, edge-weighted tree with n leaves can be reconstructed from the collection of total lengths of subtrees spanned by all subsets of m leaves provided $n \geq 2m - 1$ (Pachter & Speyer '04).

Unlabeled trees: leaf-to-leaf distances are NOT enough

- The multiset of leaf-to-leaf distances does not determine an unlabeled tree up to isomorphism. An example follows.

Caterpillars

- A tree is a caterpillar if the deletion of the leaves along with the edges adjacent to them results in a path.

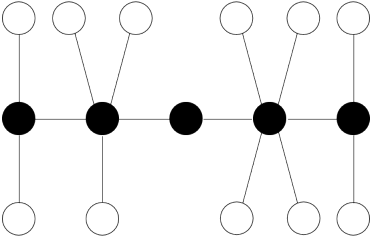


Figure: A caterpillar. Removing the leaves (white vertices) results in a path of length 5 (black vertices).

Unlabeled trees: leaf-to-leaf distance counterexample

- Consider the two caterpillars $\mathbf{T}'$ and $\mathbf{T}''$ with 25 leaves each, where
  - $\mathbf{T}'$ has 3 internal vertices in order along a path that are adjacent respectively to $2, 11, 12$ leaves,
  - $\mathbf{T}''$ has 3 internal vertices in order along a path that are adjacent respectively to $3, 14, 8$ leaves,
  - all edges have length 1.
- Taking the $\binom{25}{2}$ pairs of distinct leaves in each tree,
  - the distance 2 appears $\binom{2}{2} + \binom{11}{2} + \binom{12}{2} = 122$ times in $\mathbf{T}'$ and $\binom{3}{2} + \binom{14}{2} + \binom{8}{2} = 122$ times in $\mathbf{T}''$,
  - the distance 3 appears $2 \times 11 + 11 \times 12 = 154$ times in $\mathbf{T}'$ and $3 \times 14 + 14 \times 8 = 154$ times in $\mathbf{T}''$,
  - the distance 4 appears $2 \times 12 = 24$ times in $\mathbf{T}'$ and $3 \times 18 = 24$ times in $\mathbf{T}''$.
- Probabilistically, if we pick two distinct leaves uniformly at random from $\mathbf{T}'$ and $\mathbf{T}''$, then the two random leaf-to-leaf distances have the same probability distribution.

Unlabeled trees: leaf-to-leaf distance counterexample

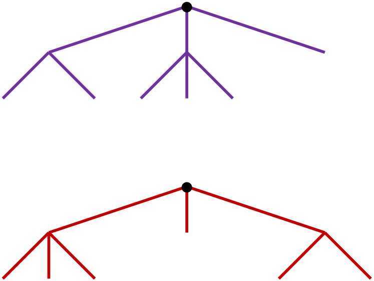
- Consider the two caterpillars $\mathbf{T}'$ and $\mathbf{T}''$ with 25 leaves each, where
  - $\mathbf{T}'$ has 3 internal vertices in order along a path that are adjacent respectively to $2, 11, 12$ leaves,
  - $\mathbf{T}''$ has 3 internal vertices in order along a path that are adjacent respectively to $3, 14, 8$ leaves,
  - all edges have length 1.
- Taking the $\binom{25}{2}$ pairs of distinct leaves in each tree,
  - the distance 2 appears $\binom{2}{2} + \binom{11}{2} + \binom{12}{2} = 122$ times in $\mathbf{T}'$ and $\binom{3}{2} + \binom{14}{2} + \binom{8}{2} = 122$ times in $\mathbf{T}''$,
  - the distance 3 appears $2 \times 11 + 11 \times 12 = 154$ times in $\mathbf{T}'$ and $3 \times 14 + 14 \times 8 = 154$ times in $\mathbf{T}''$,
  - the distance 4 appears $2 \times 12 = 24$ times in $\mathbf{T}'$ and $3 \times 18 = 24$ times in $\mathbf{T}''$.
  -
- Probabilistically, if we pick two distinct leaves uniformly at random from $\mathbf{T}'$ and $\mathbf{T}''$, then the two random leaf-to-leaf distances have the same probability distribution.

Unlabeled trees: leaf-to-leaf distance counterexample

- Consider the two caterpillars $\mathbf{T}'$ and $\mathbf{T}''$ with 25 leaves each, where
  - $\mathbf{T}'$ has 3 internal vertices in order along a path that are adjacent respectively to $2, 11, 12$ leaves,
  - $\mathbf{T}''$ has 3 internal vertices in order along a path that are adjacent respectively to $3, 14, 8$ leaves,
  - all edges have length 1.
- Taking the $\binom{25}{2}$ pairs of distinct leaves in each tree,
  - the distance 2 appears $\binom{2}{2} + \binom{11}{2} + \binom{12}{2} = 122$ times in $\mathbf{T}'$ and $\binom{3}{2} + \binom{14}{2} + \binom{8}{2} = 122$ times in $\mathbf{T}''$,
  - the distance 3 appears $2 \times 11 + 11 \times 12 = 154$ times in $\mathbf{T}'$ and $3 \times 14 + 14 \times 8 = 154$ times in $\mathbf{T}''$,
  - the distance 4 appears $2 \times 12 = 24$ times in $\mathbf{T}'$ and $3 \times 18 = 24$ times in $\mathbf{T}''$.
  -
- Probabilistically, if we pick two distinct leaves uniformly at random from $\mathbf{T}'$ and $\mathbf{T}''$, then the two random leaf-to-leaf distances have the same probability distribution.

Rooted unlabeled trees with equal edge lengths

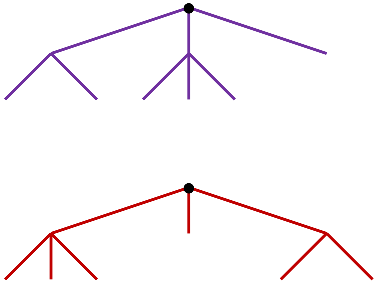
- Consider two rooted trees $\mathbf{T}'$ and $\mathbf{T}''$ with all edge lengths 1.



- There is an isomorphism that preserves roots if and only if
  - the two roots have the same number of children,
  - there is an ordering of these children for each tree such that the subtree below the i^{th} child of the root of $\mathbf{T}'$ is isomorphic (as a rooted tree) to the subtree below the i^{th} child of the root of $\mathbf{T}''$.
- This observation can be turned into a linear time algorithm for determining whether $\mathbf{T}'$ and $\mathbf{T}''$ are isomorphic.
- An extension of this algorithm works for general rooted trees.

Rooted unlabeled trees with equal edge lengths

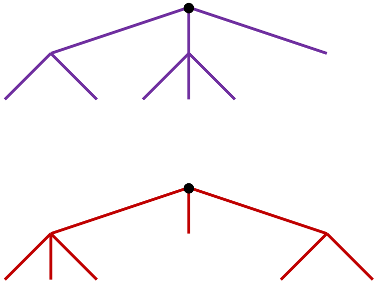
- Consider two rooted trees $\mathbf{T}'$ and $\mathbf{T}''$ with all edge lengths 1.



- There is an isomorphism that preserves roots if and only if
  - the two roots have the same number of children,
  - there is an ordering of these children for each tree such that the subtree below the i^{th} child of the root of $\mathbf{T}'$ is isomorphic (as a rooted tree) to the subtree below the i^{th} child of the root of $\mathbf{T}''$.
- This observation can be turned into a linear time algorithm for determining whether $\mathbf{T}'$ and $\mathbf{T}''$ are isomorphic.
- An extension of this algorithm works for general rooted trees.

Rooted unlabeled trees with equal edge lengths

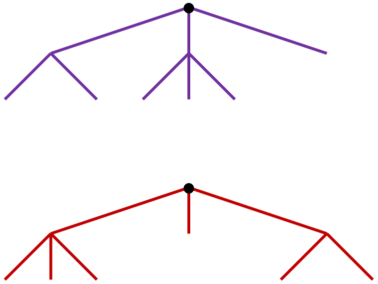
- Consider two rooted trees $\mathbf{T}'$ and $\mathbf{T}''$ with all edge lengths 1.



- There is an isomorphism that preserves roots if and only if
  - the two roots have the same number of children,
  - there is an ordering of these children for each tree such that the subtree below the i^{th} child of the root of $\mathbf{T}'$ is isomorphic (as a rooted tree) to the subtree below the i^{th} child of the root of $\mathbf{T}''$.
- This observation can be turned into a linear time algorithm for determining whether $\mathbf{T}'$ and $\mathbf{T}''$ are isomorphic.
- An extension of this algorithm works for general rooted trees.

Rooted unlabeled trees with equal edge lengths

- Consider two rooted trees $\mathbf{T}'$ and $\mathbf{T}''$ with all edge lengths 1.



- There is an isomorphism that preserves roots if and only if
  - the two roots have the same number of children,
  - there is an ordering of these children for each tree such that the subtree below the i^{th} child of the root of $\mathbf{T}'$ is isomorphic (as a rooted tree) to the subtree below the i^{th} child of the root of $\mathbf{T}''$.
- This observation can be turned into a linear time algorithm for determining whether $\mathbf{T}'$ and $\mathbf{T}''$ are isomorphic.
- An extension of this algorithm works for general rooted trees.

Unrooted unlabeled trees with equal edge lengths

- Two unrooted trees with equal edge lengths are isomorphic if and only if there is some choice of roots such the resulting rooted trees are isomorphic.
-
    - The center of a tree with equal edge lengths is a vertex with minimal greatest distance to a leaf.
    - A tree with equal edge lengths has either one or two centers (Jordan 1869).
    - Rooting each tree at one of its centers followed by a determination of whether the resulting two rooted trees are isomorphic requires linear time to detect isomorphism.

Unrooted unlabeled trees with equal edge lengths

- Two unrooted trees with equal edge lengths are isomorphic if and only if there is some choice of roots such the resulting rooted trees are isomorphic.
-
  - The center of a tree with equal edge lengths is a vertex with minimal greatest distance to a leaf.
  - A tree with equal edge lengths has either one or two centers (Jordan 1869).
  - Rooting each tree at one of its centers followed by a determination of whether the resulting two rooted trees are isomorphic requires linear time to detect isomorphism.

Summary statistics

- Are there “statistics” of a more numerical character that can be used to decide isomorphism of unlabeled trees?
- Such statistics may be constructed using a labeling of the tree, but they must be invariant under relabeling.

Summary statistics

- Are there "statistics" of a more numerical character that can be used to decide isomorphism of unlabeled trees?
- Such statistics may be constructed using a labeling of the tree, but they must be invariant under relabeling.

Spectral statistics

- The multiset of eigenvalues of the adjacency matrix fails in a very strong sense for trees with unit edge lengths (Schwenk '73, Botti & Merris '93, Steyaert & Flajolet '83, Flajolet, Gourdon & Martínez '97, Matsen & Evans '11, Bhamidi, Evans & Sen '12).
- The same is true for the eigenvalues of the matrix of leaf-to-leaf distances and the matrix of vertex-to-vertex distances.
- Indeed, the proportion of trees of various types with n leaves that share a spectrum with another tree of the same type converges to 1 as $n \to \infty$.

Spectral statistics

- The multiset of eigenvalues of the adjacency matrix fails in a very strong sense for trees with unit edge lengths (Schwenk '73, Botti & Merris '93, Steyaert & Flajolet '83, Flajolet, Gourdon & Martínez '97, Matsen & Evans '11, Bhamidi, Evans & Sen '12).
- The same is true for the eigenvalues of the matrix of leaf-to-leaf distances and the matrix of vertex-to-vertex distances.
- Indeed, the proportion of trees of various types with n leaves that share a spectrum with another tree of the same type converges to 1 as $n \to \infty$.

Spectral statistics

- The multiset of eigenvalues of the adjacency matrix fails in a very strong sense for trees with unit edge lengths (Schwenk '73, Botti & Merris '93, Steyaert & Flajolet '83, Flajolet, Gourdon & Martínez '97, Matsen & Evans '11, Bhamidi, Evans & Sen '12).
- The same is true for the eigenvalues of the matrix of leaf-to-leaf distances and the matrix of vertex-to-vertex distances.
- Indeed, the proportion of trees of various types with n leaves that share a spectrum with another tree of the same type converges to 1 as $n \to \infty$.

Symmetric polynomials

- Gordon, McDonnell, Orloff & Yung ‘95 conjectured that the greedoid Tutte polynomial determines the isomorphism type of a tree with unit edge lengths.
- Eisenstat & Gordon ‘06 produced an infinite family of counterexamples.
- Stanley ‘95 conjectured that the chromatic symmetric function determines the isomorphism type of a tree with unit edge lengths.
- There have been some positive results and no counterexamples.

Symmetric polynomials

- Gordon, McDonnell, Orloff & Yung ‘95 conjectured that the greedoid Tutte polynomial determines the isomorphism type of a tree with unit edge lengths.
- Eisenstat & Gordon ‘06 produced an infinite family of counterexamples.
- Stanley ‘95 conjectured that the chromatic symmetric function determines the isomorphism type of a tree with unit edge lengths.
- There have been some positive results and no counterexamples.

Symmetric polynomials

- Gordon, McDonnell, Orloff & Yung '95 conjectured that the greedoid Tutte polynomial determines the isomorphism type of a tree with unit edge lengths.
- Eisenstat & Gordon '06 produced an infinite family of counterexamples.
- Stanley '95 conjectured that the chromatic symmetric function determines the isomorphism type of a tree with unit edge lengths.
- There have been some positive results and no counterexamples.

Symmetric polynomials

- Gordon, McDonnell, Orloff & Yung ‘95 conjectured that the greedoid Tutte polynomial determines the isomorphism type of a tree with unit edge lengths.
- Eisenstat & Gordon ‘06 produced an infinite family of counterexamples.
- Stanley ‘95 conjectured that the chromatic symmetric function determines the isomorphism type of a tree with unit edge lengths.
- There have been some positive results and no counterexamples.

Random length sequences

- Suppose that $\#\mathbf{L}(\mathbf{T}) = n$ and $Y_1, \ldots, Y_n$ is the result of sampling the leaves of $\mathbf{T}$ uniformly at random without replacement.
- The random variable $W_k := \mathbf{W}_{\mathbf{T}}(\{Y_1, \ldots, Y_k\})$ is the length of the subtree spanned by the first k randomly chosen leaves.
- The $(n-1)$-dimensional random vector $\mathcal{W}_{\mathbf{T}} := (W_2, \ldots, W_n)$ is the random length sequence of $\mathbf{T}$.
- Is it possible to reconstruct the edge-weighted tree $\mathbf{T}$ up to isomorphism from the joint probability distribution of the random length sequence $\mathcal{W}_{\mathbf{T}}$?
- NB: Clearly, we must restrict to trees where no vertex has degree 2 ($=:$ simple trees).

Random length sequences

- Suppose that $\#\mathbf{L}(\mathbf{T}) = n$ and $Y_1, \ldots, Y_n$ is the result of sampling the leaves of $\mathbf{T}$ uniformly at random without replacement.
- The random variable $W_k := \mathbf{W}_\mathbf{T}(\{Y_1, \ldots, Y_k\})$ is the length of the subtree spanned by the first k randomly chosen leaves.
- The $(n-1)$-dimensional random vector $\mathcal{W}_\mathbf{T} := (W_2, \ldots, W_n)$ is the random length sequence of $\mathbf{T}$.
- Is it possible to reconstruct the edge-weighted tree $\mathbf{T}$ up to isomorphism from the joint probability distribution of the random length sequence $\mathcal{W}_\mathbf{T}$?
- NB: Clearly, we must restrict to trees where no vertex has degree 2 ($=:$ simple trees).

Random length sequences

- Suppose that $\#\mathbf{L}(\mathbf{T}) = n$ and $Y_1, \ldots, Y_n$ is the result of sampling the leaves of $\mathbf{T}$ uniformly at random without replacement.
- The random variable $W_k := \mathbf{W}_\mathbf{T}(\{Y_1, \ldots, Y_k\})$ is the length of the subtree spanned by the first k randomly chosen leaves.
- The $(n-1)$-dimensional random vector $\mathcal{W}_\mathbf{T} := (W_2, \ldots, W_n)$ is the random length sequence of $\mathbf{T}$.
- Is it possible to reconstruct the edge-weighted tree $\mathbf{T}$ up to isomorphism from the joint probability distribution of the random length sequence $\mathcal{W}_\mathbf{T}$?
- NB: Clearly, we must restrict to trees where no vertex has degree 2 ($=:$ simple trees).

Random length sequences

- Suppose that $\#\mathbf{L}(\mathbf{T}) = n$ and $Y_1, \ldots, Y_n$ is the result of sampling the leaves of $\mathbf{T}$ uniformly at random without replacement.
- The random variable $W_k := \mathbf{W}_{\mathbf{T}}(\{Y_1, \ldots, Y_k\})$ is the length of the subtree spanned by the first k randomly chosen leaves.
- The $(n-1)$-dimensional random vector $\mathcal{W}_{\mathbf{T}} := (W_2, \ldots, W_n)$ is the random length sequence of $\mathbf{T}$.
- Is it possible to reconstruct the edge-weighted tree $\mathbf{T}$ up to isomorphism from the joint probability distribution of the random length sequence $\mathcal{W}_{\mathbf{T}}$?
- NB: Clearly, we must restrict to trees where no vertex has degree 2 ($=:$ simple trees).

Random length sequences

- Suppose that $\#\mathbf{L}(\mathbf{T}) = n$ and $Y_1, \ldots, Y_n$ is the result of sampling the leaves of $\mathbf{T}$ uniformly at random without replacement.
- The random variable $W_k := \mathbf{W}_\mathbf{T}(\{Y_1, \ldots, Y_k\})$ is the length of the subtree spanned by the first k randomly chosen leaves.
- The $(n-1)$-dimensional random vector $\mathcal{W}_\mathbf{T} := (W_2, \ldots, W_n)$ is the random length sequence of $\mathbf{T}$.
- Is it possible to reconstruct the edge-weighted tree $\mathbf{T}$ up to isomorphism from the joint probability distribution of the random length sequence $\mathcal{W}_\mathbf{T}$?
- NB: Clearly, we must restrict to trees where no vertex has degree 2 ($=:$ simple trees).

Trees with four leaves can be reconstructed

Theorem 1

The isomorphism class of a simple, edge-weighted tree $\mathbf{T}$ *with* 4 *leaves is uniquely determined by the joint probability distribution of its random length sequence.*

Proof of the result for trees with four leaves

- The total length of $\mathbf{T}$ is W_4.
- The multiset of lengths of the pendent edges (= edges adjacent to leaves) can be determined from the distribution of $W_4 - W_3$; e.g. the pendent edges are a,a,b,c if and only if $W_4 - W_3$ takes the values a, b, c with probabilities $\frac{1}{2}, \frac{1}{4}, \frac{1}{4}$.
- If the lengths of the pendent edges sum to the total length of $\mathbf{T}$, then $\mathbf{T}$ is a star and its isomorphism class is determined.
- Otherwise, $\mathbf{T}$ has two degree 3 internal vertices and we can determine the length e of the single internal edge.

Proof of the result for trees with four leaves

- The total length of $\mathbf{T}$ is W_4.
- The multiset of lengths of the pendent edges ($=$ edges adjacent to leaves) can be determined from the distribution of $W_4 - W_3$; e.g. the pendent edges are a,a,b,c if and only if $W_4 - W_3$ takes the values a, b, c with probabilities $\frac{1}{2}, \frac{1}{4}, \frac{1}{4}$.
- If the lengths of the pendent edges sum to the total length of $\mathbf{T}$, then $\mathbf{T}$ is a star and its isomorphism class is determined.
- Otherwise, $\mathbf{T}$ has two degree 3 internal vertices and we can determine the length e of the single internal edge.

Proof of the result for trees with four leaves

- The total length of $\mathbf{T}$ is W_4.
- The multiset of lengths of the pendent edges (= edges adjacent to leaves) can be determined from the distribution of $W_4 - W_3$; e.g. the pendent edges are a,a,b,c if and only if $W_4 - W_3$ takes the values a, b, c with probabilities $\frac{1}{2}, \frac{1}{4}, \frac{1}{4}$.
- If the lengths of the pendent edges sum to the total length of $\mathbf{T}$, then $\mathbf{T}$ is a star and its isomorphism class is determined.
- Otherwise, $\mathbf{T}$ has two degree 3 internal vertices and we can determine the length e of the single internal edge.

Proof of the result for trees with four leaves

- The total length of $\mathbf{T}$ is W_4.
- The multiset of lengths of the pendent edges ($=$ edges adjacent to leaves) can be determined from the distribution of $W_4 - W_3$; e.g. the pendent edges are a,a,b,c if and only if $W_4 - W_3$ takes the values a, b, c with probabilities $\frac{1}{2}, \frac{1}{4}, \frac{1}{4}$.
- If the lengths of the pendent edges sum to the total length of $\mathbf{T}$, then $\mathbf{T}$ is a star and its isomorphism class is determined.
- Otherwise, $\mathbf{T}$ has two degree 3 internal vertices and we can determine the length e of the single internal edge.

Proof of the result for trees with four leaves – continued

- Suppose $\mathbf{T}$ has two degree 3 internal vertices and a single internal edge of length e.
- If the multiset of pendent edge lengths is of the form $\{a, a, a, a\}$ or $\{a, a, a, b\}$, then $\mathbf{T}$ is determined.
- Suppose the pendent edge lengths are of the form $\{a, a, b, b\}$.
  - If the possible values of W_2 are $(a + a), (b + b), (a + b + e)$ with probabilities $\frac{1}{6}, \frac{1}{6}, \frac{2}{3}$, then the leaves with pendent edges of length a (resp. b) are siblings.
  - If the possible values of W_2 are $(a + b), (a + b + e), (a + a + e), (b + b + e)$ with probabilities $\frac{1}{3}, \frac{1}{3}, \frac{1}{6}, \frac{1}{6}$, then the two pairs of sibling leaves each have one pendent edge of length a and one of length b.

Proof of the result for trees with four leaves – continued

- Suppose $\mathbf{T}$ has two degree 3 internal vertices and a single internal edge of length e.
- If the multiset of pendent edge lengths is of the form $\{a, a, a, a\}$ or $\{a, a, a, b\}$, then $\mathbf{T}$ is determined.
- Suppose the pendent edge lengths are of the form $\{a, a, b, b\}$.
  - If the possible values of W_2 are $(a + a), (b + b), (a + b + e)$ with probabilities $\frac{1}{6}, \frac{1}{6}, \frac{2}{3}$, then the leaves with pendent edges of length a (resp. b) are siblings.
  - If the possible values of W_2 are $(a + b), (a + b + e), (a + a + e), (b + b + e)$ with probabilities $\frac{1}{3}, \frac{1}{3}, \frac{1}{6}, \frac{1}{6}$, then the two pairs of sibling leaves each have one pendent edge of length a and one of length b.

Proof of the result for trees with four leaves – continued

- Suppose $\mathbf{T}$ has two degree 3 internal vertices and a single internal edge of length e.
- If the multiset of pendent edge lengths is of the form $\{a, a, a, a\}$ or $\{a, a, a, b\}$, then $\mathbf{T}$ is determined.
- Suppose the pendent edge lengths are of the form $\{a, a, b, b\}$.
    - If the possible values of W_2 are $(a + a), (b + b), (a + b + e)$ with probabilities $\frac{1}{6}, \frac{1}{6}, \frac{2}{3}$, then the leaves with pendent edges of length a (resp. b) are siblings.
    - If the possible values of W_2 are $(a + b), (a + b + e), (a + a + e), (b + b + e)$ with probabilities $\frac{1}{3}, \frac{1}{3}, \frac{1}{6}, \frac{1}{6}$, then the two pairs of sibling leaves each have one pendent edge of length a and one of length b.

Proof of the result for trees with four leaves – continued

- Suppose $\mathbf{T}$ has two degree 3 internal vertices and a single internal edge of length e, and the multiset of pendent edge lengths is of the form $\{a, a, b, c\}$.
- The leaves with pendent edge lengths a are siblings if and only if $(2a, 2a + e + b)$ occurs as a value of (W_2, W_3) with positive probability.

Proof of the result for trees with four leaves – continued

- Suppose $\mathbf{T}$ has two degree 3 internal vertices and a single internal edge of length e, and the multiset of pendent edge lengths is of the form $\{a, a, b, c\}$.
- The leaves with pendent edge lengths a are siblings if and only if $(2a, 2a + e + b)$ occurs as a value of (W_2, W_3) with positive probability.

Proof of the result for trees with four leaves – continued

- Suppose $\mathbf{T}$ has two degree 3 internal vertices and a single internal edge of length e, and the multiset of pendent edge lengths is of the form $\{a, b, c, d\}$ with $a < b < c < d$.
-
  - If the leaves with pendent edge lengths a and b are siblings, then the possible values of W_2 are $(a+b), (c+d), (a+c+e), (a+d+e), (b+c+e), (b+d+e)$ with equal probabilities $\frac{1}{6}$.
  - If the leaves with pendent edge lengths a and c are siblings, then the possible values of W_2 are $(a+c), (b+d), (a+b+e), (a+d+e), (b+c+e), (c+d+e)$ with equal probabilities $\frac{1}{6}$.
  - If the leaves with pendent edge lengths a and d are siblings, then the possible values of W_2 are $(a+d), (b+c), (a+c+e), (a+b+e), (c+d+e), (b+d+e)$ with equal probabilities $\frac{1}{6}$.
- It is possible to distinguish which alternative holds.

Proof of the result for trees with four leaves – continued

- Suppose $\mathbf{T}$ has two degree 3 internal vertices and a single internal edge of length e, and the multiset of pendent edge lengths is of the form $\{a, b, c, d\}$ with $a < b < c < d$.
-
    - If the leaves with pendent edge lengths a and b are siblings, then the possible values of W_2 are $(a+b), (c+d), (a+c+e), (a+d+e), (b+c+e), (b+d+e)$ with equal probabilities $\frac{1}{6}$.
    - If the leaves with pendent edge lengths a and c are siblings, then the possible values of W_2 are $(a+c), (b+d), (a+b+e), (a+d+e), (b+c+e), (c+d+e)$ with equal probabilities $\frac{1}{6}$.
    - If the leaves with pendent edge lengths a and d are siblings, then the possible values of W_2 are $(a+d), (b+c), (a+c+e), (a+b+e), (c+d+e), (b+d+e)$ with equal probabilities $\frac{1}{6}$.
- It is possible to distinguish which alternative holds.

Proof of the result for trees with four leaves – continued

- Suppose $\mathbf{T}$ has two degree 3 internal vertices and a single internal edge of length e, and the multiset of pendent edge lengths is of the form $\{a, b, c, d\}$ with $a < b < c < d$.
-
  - If the leaves with pendent edge lengths a and b are siblings, then the possible values of W_2 are $(a+b), (c+d), (a+c+e), (a+d+e), (b+c+e), (b+d+e)$ with equal probabilities $\frac{1}{6}$.
  - If the leaves with pendent edge lengths a and c are siblings, then the possible values of W_2 are $(a+c), (b+d), (a+b+e), (a+d+e), (b+c+e), (c+d+e)$ with equal probabilities $\frac{1}{6}$.
  - If the leaves with pendent edge lengths a and d are siblings, then the possible values of W_2 are $(a+d), (b+c), (a+c+e), (a+b+e), (c+d+e), (b+d+e)$ with equal probabilities $\frac{1}{6}$.
- It is possible to distinguish which alternative holds.

“Generic” trees can be reconstructed

- The edge weights of an edge-weighted tree $\mathbf{T}$ are in general position if the sums of the lengths of any two distinct subset of edges of $\mathbf{T}$ are not equal.

Theorem 2

The isomorphism class of a simple, edge-weighted tree with edge weights in general position is uniquely determined by the joint probability distribution of its random length sequence.

- The proof uses the result for trees with 4 leaves.

"Generic" trees can be reconstructed

- The edge weights of an edge-weighted tree $\mathbf{T}$ are in general position if the sums of the lengths of any two distinct subset of edges of $\mathbf{T}$ are not equal.

Theorem 2

The isomorphism class of a simple, edge-weighted tree with edge weights in general position is uniquely determined by the joint probability distribution of its random length sequence.

- The proof uses the result for trees with 4 leaves.

“Generic” trees can be reconstructed

- The edge weights of an edge-weighted tree $\mathbf{T}$ are in general position if the sums of the lengths of any two distinct subset of edges of $\mathbf{T}$ are not equal.

Theorem 2

The isomorphism class of a simple, edge-weighted tree with edge weights in general position is uniquely determined by the joint probability distribution of its random length sequence.

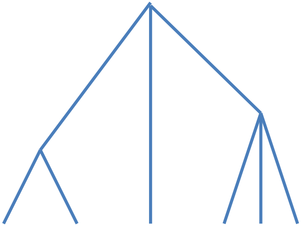
- The proof uses the result for trees with 4 leaves.

Ultrametric trees

- Recall that for $i, j \in \mathbf{L}(\mathbf{T})$, $r_{\mathbf{T}}(i,j)$ is the sum of the lengths of the edges on the path between i and j.
- An edge-weighted tree $\mathbf{T}$ is ultrametric if for any $i, j, k \in \mathbf{L}(\mathbf{T})$ we have

$$r_{\mathbf{T}}(i,k) \le r_{\mathbf{T}}(i,j) \vee r_{\mathbf{T}}(j,k),$$

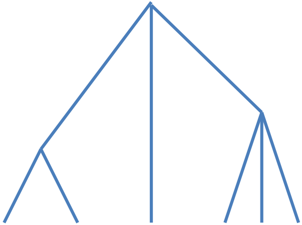
  from which it follows that at least two of $r_{\mathbf{T}}(i,j)$, $r_{\mathbf{T}}(i,k)$, and $r_{\mathbf{T}}(j,k)$ are equal while the third is no greater than that common value.



- Alternatively, $\mathbf{T}$ is ultrametric if, when it is thought of as a real tree (that is, a metric space where the edges are treated as real intervals), then there is a (unique) point ρ (which may be in the interior of an edge) such that the distance from ρ to a leaf is the same for all leaves.

Ultrametric trees

- Recall that for $i, j \in \mathbf{L}(\mathbf{T})$, $r_{\mathbf{T}}(i,j)$ is the sum of the lengths of the edges on the path between i and j.
- An edge-weighted tree $\mathbf{T}$ is ultrametric if for any $i, j, k \in \mathbf{L}(\mathbf{T})$ we have
$$r_{\mathbf{T}}(i,k) \le r_{\mathbf{T}}(i,j) \vee r_{\mathbf{T}}(j,k),$$
from which it follows that at least two of $r_{\mathbf{T}}(i,j)$, $r_{\mathbf{T}}(i,k)$, and $r_{\mathbf{T}}(j,k)$ are equal while the third is no greater than that common value.



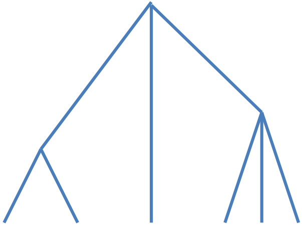
- Alternatively, $\mathbf{T}$ is ultrametric if, when it is thought of as a real tree (that is, a metric space where the edges are treated as real intervals), then there is a (unique) point ρ (which may be in the interior of an edge) such that the distance from ρ to a leaf is the same for all leaves.

Ultrametric trees

- Recall that for $i, j \in \mathbf{L}(\mathbf{T})$, $r_{\mathbf{T}}(i,j)$ is the sum of the lengths of the edges on the path between i and j.
- An edge-weighted tree $\mathbf{T}$ is ultrametric if for any $i, j, k \in \mathbf{L}(\mathbf{T})$ we have

  $$r_{\mathbf{T}}(i,k) \le r_{\mathbf{T}}(i,j) \vee r_{\mathbf{T}}(j,k),$$

  from which it follows that at least two of $r_{\mathbf{T}}(i,j)$, $r_{\mathbf{T}}(i,k)$, and $r_{\mathbf{T}}(j,k)$ are equal while the third is no greater than that common value.



- Alternatively, $\mathbf{T}$ is ultrametric if, when it is thought of as a real tree (that is, a metric space where the edges are treated as real intervals), then there is a (unique) point ρ (which may be in the interior of an edge) such that the distance from ρ to a leaf is the same for all leaves.

Ultrametric trees can be reconstructed

Theorem 3

The isomorphism class of an ultrametric, simple, edge-weighted tree is uniquely determined by the joint probability distribution of its random length sequence.

- Indeed, it suffices to know the $(n-1)$-tuple in the support of $(W_2, \ldots, W_n)$ that is minimal in the lexicographic order.

Ultrametric trees can be reconstructed

Theorem 3

The isomorphism class of an ultrametric, simple, edge-weighted tree is uniquely determined by the joint probability distribution of its random length sequence.

- Indeed, it suffices to know the $(n-1)$-tuple in the support of $(W_2, \ldots, W_n)$ that is minimal in the lexicographic order.

Caterpillars can be reconstructed

- Recall that a tree is a caterpillar if the deletion of the leaves along with the edges adjacent to them results in a path.

Theorem 4

The isomorphism class of a caterpillar with all edges of weight 1 is uniquely determined by the joint probability distribution of its random length sequence.

Caterpillars can be reconstructed

- Recall that a tree is a caterpillar if the deletion of the leaves along with the edges adjacent to them results in a path.

Theorem 4

The isomorphism class of a caterpillar with all edges of weight 1 is uniquely determined by the joint probability distribution of its random length sequence.

Caterpillars: what goes into the proof

- Suppose that the caterpillar has $\ell + 1$ internal vertices with respective numbers of leaves $n_0, \ldots, n_\ell$. Write $(W_2, \ldots, W_n)$ for the random subtree length sequences.
- Consider a box with n tickets. Each ticket has a label belonging to $\{0, 1, \ldots, \ell\}$ and there are n_i tickets with label i for $0 \le i \le \ell$.
- Let $X_1, X_2, \ldots, X_n$ be the result of drawing tickets uniformly at random from the box without replacement and noting their labels.
- Set

$$\begin{aligned} K_r &:= \max_{1 \le j \le r} X_j - \min_{1 \le j \le r} X_j \\ &= \text{difference between the largest and smallest labels seen in first } r \text{ draws.} \end{aligned}$$

- Note that $(W_2, W_3, \ldots, W_n)$ has the same distribution as $(K_2 + 3, K_3 + 3, \ldots, K_n + n)$. It suffices to show that it is possible to determine $\{(n_0, n_1, \ldots, n_{\ell-1}, n_\ell), (n_\ell, n_{\ell-1}, \ldots, n_1, n_0)\}$ from the distribution of $\mathcal{K} := (K_2, \ldots, K_n)$.
- The proof uses some Fourier analysis similar to that used in crystallography to recover molecular structures from distances between atoms and some commutative algebra.

Caterpillars: what goes into the proof

- Suppose that the caterpillar has $\ell + 1$ internal vertices with respective numbers of leaves $n_0, \ldots, n_\ell$. Write $(W_2, \ldots, W_n)$ for the random subtree length sequences.
- Consider a box with n tickets. Each ticket has a label belonging to $\{0, 1, \ldots, \ell\}$ and there are n_i tickets with label i for $0 \leq i \leq \ell$.
- Let $X_1, X_2, \ldots, X_n$ be the result of drawing tickets uniformly at random from the box without replacement and noting their labels.
- Set

$$\begin{aligned} K_r &:= \max_{1 \leq j \leq r} X_j - \min_{1 \leq j \leq r} X_j \\ &= \text{difference between the largest and smallest labels seen in first } r \text{ draws.} \end{aligned}$$

- Note that $(W_2, W_3, \ldots, W_n)$ has the same distribution as $(K_2 + 3, K_3 + 3, \ldots, K_n + n)$. It suffices to show that it is possible to determine $\{(n_0, n_1, \ldots, n_{\ell-1}, n_\ell), (n_\ell, n_{\ell-1}, \ldots, n_1, n_0)\}$ from the distribution of $\mathcal{K} := (K_2, \ldots, K_n)$.
- The proof uses some Fourier analysis similar to that used in crystallography to recover molecular structures from distances between atoms and some commutative algebra.

Caterpillars: what goes into the proof

- Suppose that the caterpillar has $\ell + 1$ internal vertices with respective numbers of leaves $n_0, \ldots, n_\ell$. Write $(W_2, \ldots, W_n)$ for the random subtree length sequences.
- Consider a box with n tickets. Each ticket has a label belonging to $\{0, 1, \ldots, \ell\}$ and there are n_i tickets with label i for $0 \le i \le \ell$.
- Let $X_1, X_2, \ldots, X_n$ be the result of drawing tickets uniformly at random from the box without replacement and noting their labels.
- Set

$$K_r := \max_{1 \le j \le r} X_j - \min_{1 \le j \le r} X_j$$

$$= \text{difference between the largest and smallest labels seen in first } r \text{ draws.}$$

- Note that $(W_2, W_3, \ldots, W_n)$ has the same distribution as $(K_2 + 3, K_3 + 3, \ldots, K_n + n)$. It suffices to show that it is possible to determine $\{(n_0, n_1, \ldots, n_{\ell-1}, n_\ell), (n_\ell, n_{\ell-1}, \ldots, n_1, n_0)\}$ from the distribution of $\mathcal{K} := (K_2, \ldots, K_n)$.
- The proof uses some Fourier analysis similar to that used in crystallography to recover molecular structures from distances between atoms and some commutative algebra.

Caterpillars: what goes into the proof

- Suppose that the caterpillar has $\ell + 1$ internal vertices with respective numbers of leaves $n_0, \ldots, n_\ell$. Write $(W_2, \ldots, W_n)$ for the random subtree length sequences.
- Consider a box with n tickets. Each ticket has a label belonging to $\{0, 1, \ldots, \ell\}$ and there are n_i tickets with label i for $0 \le i \le \ell$.
- Let $X_1, X_2, \ldots, X_n$ be the result of drawing tickets uniformly at random from the box without replacement and noting their labels.
- Set

$$K_r := \max_{1 \le j \le r} X_j - \min_{1 \le j \le r} X_j$$

$$= \text{difference between the largest and smallest labels seen in first } r \text{ draws.}$$

- Note that $(W_2, W_3, \ldots, W_n)$ has the same distribution as $(K_2 + 3, K_3 + 3, \ldots, K_n + n)$. It suffices to show that it is possible to determine $\{(n_0, n_1, \ldots, n_{\ell-1}, n_\ell), (n_\ell, n_{\ell-1}, \ldots, n_1, n_0)\}$ from the distribution of $\mathcal{K} := (K_2, \ldots, K_n)$.
- The proof uses some Fourier analysis similar to that used in crystallography to recover molecular structures from distances between atoms and some commutative algebra.

Caterpillars: what goes into the proof

- Suppose that the caterpillar has $\ell + 1$ internal vertices with respective numbers of leaves $n_0, \ldots, n_\ell$. Write $(W_2, \ldots, W_n)$ for the random subtree length sequences.
- Consider a box with n tickets. Each ticket has a label belonging to $\{0, 1, \ldots, \ell\}$ and there are n_i tickets with label i for $0 \leq i \leq \ell$.
- Let $X_1, X_2, \ldots, X_n$ be the result of drawing tickets uniformly at random from the box without replacement and noting their labels.
- Set

$$K_r := \max_{1 \leq j \leq r} X_j - \min_{1 \leq j \leq r} X_j$$

$$= \text{difference between the largest and smallest labels seen in first } r \text{ draws.}$$

- Note that $(W_2, W_3, \ldots, W_n)$ has the same distribution as $(K_2 + 3, K_3 + 3, \ldots, K_n + n)$. It suffices to show that it is possible to determine $\{(n_0, n_1, \ldots, n_{\ell-1}, n_\ell), (n_\ell, n_{\ell-1}, \ldots, n_1, n_0)\}$ from the distribution of $\mathcal{K} := (K_2, \ldots, K_n)$.
- The proof uses some Fourier analysis similar to that used in crystallography to recover molecular structures from distances between atoms and some commutative algebra.

Caterpillars: what goes into the proof

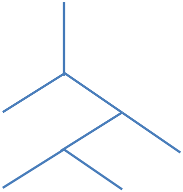
- Suppose that the caterpillar has $\ell + 1$ internal vertices with respective numbers of leaves $n_0, \ldots, n_\ell$. Write $(W_2, \ldots, W_n)$ for the random subtree length sequences.
- Consider a box with n tickets. Each ticket has a label belonging to $\{0, 1, \ldots, \ell\}$ and there are n_i tickets with label i for $0 \leq i \leq \ell$.
- Let $X_1, X_2, \ldots, X_n$ be the result of drawing tickets uniformly at random from the box without replacement and noting their labels.
- Set

$$\begin{aligned} K_r &:= \max_{1 \leq j \leq r} X_j - \min_{1 \leq j \leq r} X_j \\ &= \text{difference between the largest and smallest labels seen in first } r \text{ draws.} \end{aligned}$$

- Note that $(W_2, W_3, \ldots, W_n)$ has the same distribution as $(K_2 + 3, K_3 + 3, \ldots, K_n + n)$. It suffices to show that it is possible to determine $\{(n_0, n_1, \ldots, n_{\ell-1}, n_\ell), (n_\ell, n_{\ell-1}, \ldots, n_1, n_0)\}$ from the distribution of $\mathcal{K} := (K_2, \ldots, K_n)$.
- The proof uses some Fourier analysis similar to that used in crystallography to recover molecular structures from distances between atoms and some commutative algebra.

$k+1$-valent trees can be reconstructed

- For $k \geq 2$, a $(k+1)$-valent tree is a tree for which all internal vertices have degree $k+1$.

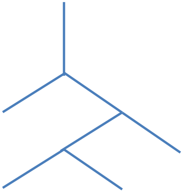


Theorem 5

The isomorphism class of a $(k+1)$-valent tree with all edges of length 1 is uniquely determined by the joint distribution of its random length sequence.

$k+1$-valent trees can be reconstructed

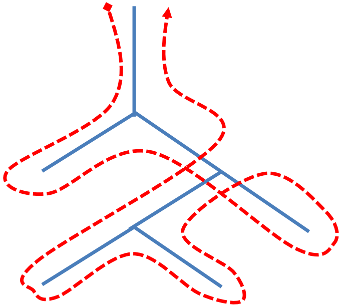
- For $k \geq 2$, a $(k+1)$-valent tree is a tree for which all internal vertices have degree $k+1$.



Theorem 5

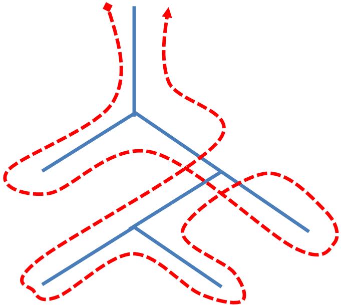
The isomorphism class of a $(k+1)$-valent tree with all edges of length 1 is uniquely determined by the joint distribution of its random length sequence.

$k + 1$-valent trees: idea of proof



- If we know the leaves are visited in a traversal, then the sequence of subtree lengths determines the tree.

$k+1$-valent trees: idea of proof



- If we know the leaves are visited in a traversal, then the sequence of subtree lengths determines the tree.

$k+1$-valent trees: more idea of proof

- A subtree $\mathbf{S}$ of a $(k+1)$-valent tree $\mathbf{T}$ has all vertices of degree $k+1$ or 1 except for a single vertex of degree k if and only if

$$\#\mathbf{E}(\mathbf{S}) = \frac{k}{k-1}(\#\mathbf{L}(\mathbf{S}) - 1).$$

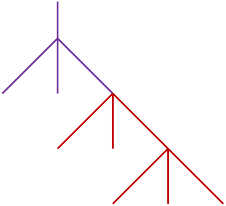


Figure: Here $k=3$. The red subtree $\mathbf{S}$ has $\#\mathbf{E}(\mathbf{S}) = 6$ and $\#\mathbf{L}(\mathbf{S}) = 5$. Note that $6 = \frac{3}{2}(5-1)$.

- There is a total order on the set of possible length sequences for a $(k+1)$-valent tree with unit edge lengths such that the minimal sequence is guaranteed to come from a traversal.

$k + 1$-valent trees: more idea of proof

- A subtree $\mathbf{S}$ of a $(k+1)$-valent tree $\mathbf{T}$ has all vertices of degree $k+1$ or 1 except for a single vertex of degree k if and only if

$$\#\mathbf{E}(\mathbf{S}) = \frac{k}{k-1}(\#\mathbf{L}(\mathbf{S}) - 1).$$

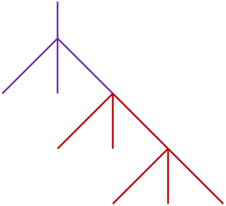


Figure: Here $k = 3$. The red subtree $\mathbf{S}$ has $\#\mathbf{E}(\mathbf{S}) = 6$ and $\#\mathbf{L}(\mathbf{S}) = 5$. Note that $6 = \frac{3}{2}(5-1)$.

- There is a total order on the set of possible length sequences for a $(k+1)$-valent tree with unit edge lengths such that the minimal sequence is guaranteed to come from a traversal.

Open questions

- Is it possible to reconstruct a general simple, edge-weighted tree up to isomorphism from the joint probability distribution of its random length sequence?
- For the purposes of simulations studies in phylogenetics, we would like to have generative models for random trees that produce trees which are "like" biological trees. Are there features of the joint distribution of the random length sequence that are common to many biological trees and can be used to determine which generative models capture features of biological trees?

Open questions

- Is it possible to reconstruct a general simple, edge-weighted tree up to isomorphism from the joint probability distribution of its random length sequence?
- For the purposes of simulations studies in phylogenetics, we would like to have generative models for random trees that produce trees which are “like” biological trees. Are there features of the joint distribution of the random length sequence that are common to many biological trees and can be used to determine which generative models capture features of biological trees?