Combinatorics of spaces of trees: an application of topology to phylogenetics

Curran N. McConnell

Dalhousie University

Categorical Approaches to Topology and Geometry, CMS Summer Meeting 2019

How phylogenetics works

- Discover when species branched apart by comparing their genomes.
- Determine pairwise "evolutionary time" distance between gene sequences.
- Build the evolutionary tree that best reflects these pairwise distances.
- This uses the theory of maximum-likelihood estimation.

How phylogenetics breaks down

Different subsequences can suggest different evolutionary histories. Anomalies occur because of:

- Statistical artefacts
- Model inadequacy
- Cross-species transfer of genetic material

How phylogenetics breaks down

Detecting non-tree phenomena is hard!

Biologists analyze gene sequences in terms of trees. How to detect non-tree phenomena, like when distantly-related plankton pass each other DNA directly?

How phylogenetics breaks down

Idea: use topological data analysis (TDA)

Topology can complement statistics to better distinguish between kinds of anomalies.

Where my research begins

- Use persistent homology to analyze evolutionary-tree datasets.

Where my research begins

- Use persistent homology to analyze evolutionary-tree datasets.
- Understand combinatorial and topological properties of the spaces these datasets live in.

n-trees

Definition

A rootless binary tree is an acyclic connected graph in which every vertex is either order 1 or order 3.

n-trees

Definition

A rootless binary tree is an acyclic connected graph in which every vertex is either order 1 or order 3.

Definition

A leaf in a rootless binary tree is a vertex that has exactly one neighbour.

n-trees

Definition

A rootless binary tree is an acyclic connected graph in which every vertex is either order 1 or order 3.

Definition

A leaf in a rootless binary tree is a vertex that has exactly one neighbour.

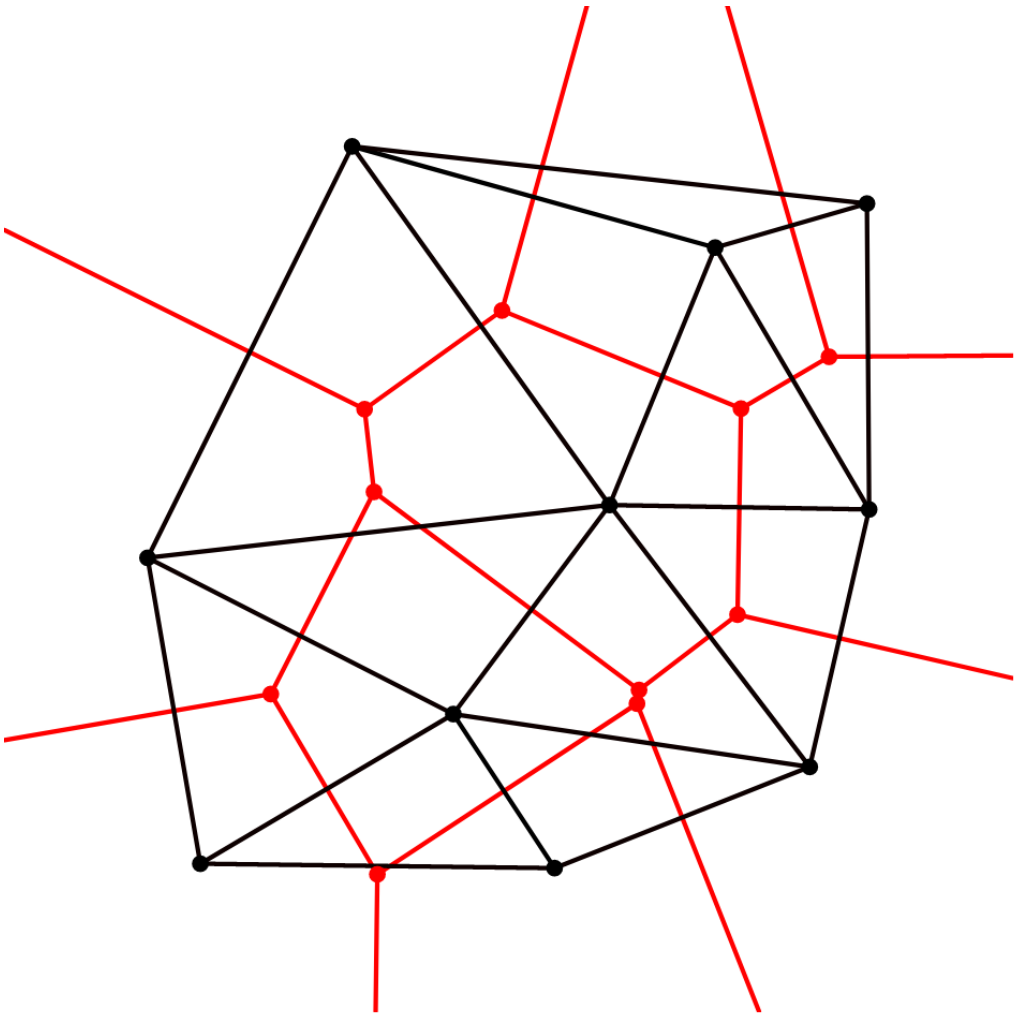
Definition

An n-tree is a rootless binary tree with n labelled leaves. I will later mention rooted n-trees as well.

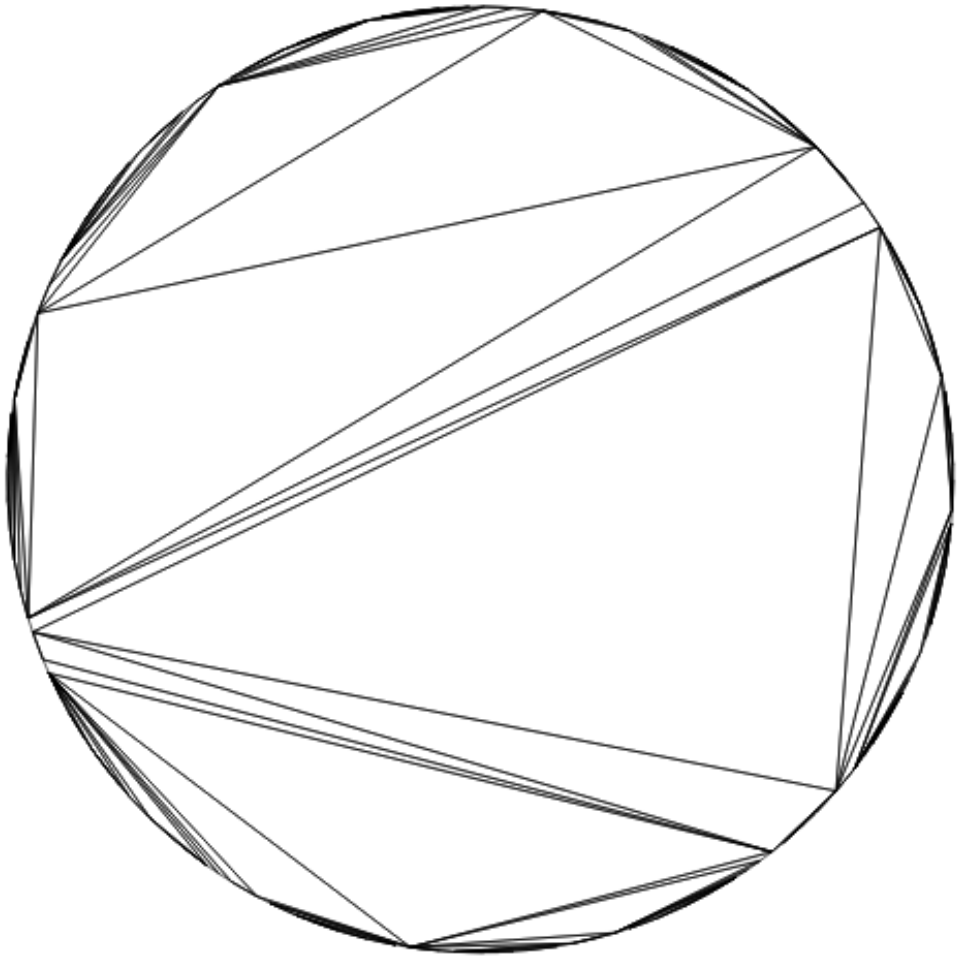
Properties of n-trees

- There are $(2n-5)!! = (2n-5)(2n-7)\cdot \ldots \cdot 5 \cdot 3 \cdot 1$ n-trees for each $n \geq 3$.
- n-trees have a dual interpretation as triangulations of convex polygons with labelled sides.

Dual interpretation of n-trees



The collection of ∞-trees



Tree metrics

- A plethora of metrics are used.
- Reliable and fast-ish: quartet distance.

Quartet distance

Definition

A pair of pairs of vertices $\{\{a, b\}, \{c, d\}\}$ is a quartet in a tree T if there exists an edge e in T such that deleting e from T causes $\{a, b\}$ and $\{c, d\}$ to lie in separate components.

Quartet distance

Definition

A pair of pairs of vertices $\{\{a, b\}, \{c, d\}\}$ is a quartet in a tree T if there exists an edge e in T such that deleting e from T causes $\{a, b\}$ and $\{c, d\}$ to lie in separate components.

Definition

Symmetric difference of sets $\triangle$ is given by

$$A \triangle B = (A \cup B) \setminus (A \cap B).$$

Quartet distance

Definition

A pair of pairs of vertices $\{\{a, b\}, \{c, d\}\}$ is a quartet in a tree T if there exists an edge e in T such that deleting e from T causes $\{a, b\}$ and $\{c, d\}$ to lie in separate components.

Definition

Symmetric difference of sets $\triangle$ is given by

$$A \triangle B = (A \cup B) \setminus (A \cap B).$$

Definition

Quartet distance between two trees S and T is defined by

$$d(S, T) = |Q(S) \triangle Q(T)|$$

where Q gives the set of quartets in a tree.

Tree spaces

- Let T_n be the set of n-trees, for every $n \in \mathbb{N}$.
- Let T_∞ be the set of binary trees with infinitely many leaves.
- Let Q_n be T_n with quartet distance.

Dual interpretation of tree metrics

- Quartet distance $\mapsto$ counting certain label-preserving homotopies.
- Contract exterior edges down to a point, one at a time.
- If you can finish at a pair of triangles glued to one another, one with sides a and b and the other with sides c and d, then $\{\{a, b\}\{c, d\}\}$ is a quartet in your tree.

Homology of a simplicial complex

- Construct C_n as free module with n-simplices of the complex as its basis.
- Software frequently uses $\mathbb{Z}/2\mathbb{Z}$ as the module ring for computational reasons.

Homology of a simplicial complex

- Construct C_n as free module with n-simplices of the complex as its basis.
- Software frequently uses $\mathbb{Z}/2\mathbb{Z}$ as the module ring for computational reasons.
- Construct $Z_n = \ker \partial_n$, the module of n-cycles.

Homology of a simplicial complex

- Construct C_n as free module with n-simplices of the complex as its basis.
- Software frequently uses $\mathbb{Z}/2\mathbb{Z}$ as the module ring for computational reasons.
- Construct $Z_n = \ker \partial_n$, the module of n-cycles.
- Construct $B_n = \operatorname{im} \partial_{n+1}$, the module of n-boundaries.
- Construct $H_n = Z_n/B_n$, the homology module.

Homology of a simplicial complex

- Construct C_n as free module with n-simplices of the complex as its basis.
- Software frequently uses $\mathbb{Z}/2\mathbb{Z}$ as the module ring for computational reasons.
- Construct $Z_n = \ker \partial_n$, the module of n-cycles.
- Construct $B_n = \operatorname{im} \partial_{n+1}$, the module of n-boundaries.
- Construct $H_n = Z_n/B_n$, the homology module.

Homology of a simplicial complex

- H_n is occupied by equivalence classes of n-cycles that surround each $n+1$-dimensional hole in the complex.
- For H_0, a better intuition is that elements represent connected components of the complex.

Vietoris-Rips complex

Definition

Given a subset S of a metric space X, the Vietoris-Rips complex $\mathcal{R}_\varepsilon$ contains every simplex σ constructed from points in S that satisfies the following condition:
For every $a, b \in \sigma$, $B_\varepsilon(a) \bigcap B_\varepsilon(b) \neq \varnothing$.

Vietoris-Rips complex

Definition

Given a subset S of a metric space X, the Vietoris-Rips complex $\mathcal{R}_\varepsilon$ contains every simplex σ constructed from points in S that satisfies the following condition:
For every $a, b \in \sigma$, $B_\varepsilon(a) \bigcap B_\varepsilon(b) \neq \varnothing$.

- The homology of a filtered Vietoris-Rips complex approximates the homology of a filtered Čech complex.

Vietoris-Rips complex

Definition

Given a subset S of a metric space X, the Vietoris-Rips complex $\mathcal{R}_\varepsilon$ contains every simplex σ constructed from points in S that satisfies the following condition:
For every $a, b \in \sigma$, $B_\varepsilon(a) \bigcap B_\varepsilon(b) \neq \varnothing$.

- The homology of a filtered Vietoris-Rips complex approximates the homology of a filtered Čech complex.
- Under certain conditions, a Čech complex will have homology isomorphic to the singular homology of X.

Persistent homology

- Begin with point cloud data.

Persistent homology

- Begin with point cloud data.
- Inflate a ε-ball at each point.

Persistent homology

- Begin with point cloud data.
- Inflate a ε-ball at each point.
- Draw an edge between points when their ε-balls intersect.

Persistent homology

- Begin with point cloud data.
- Inflate a ε-ball at each point.
- Draw an edge between points when their ε-balls intersect.
- Draw an n-simplex wherever possible.

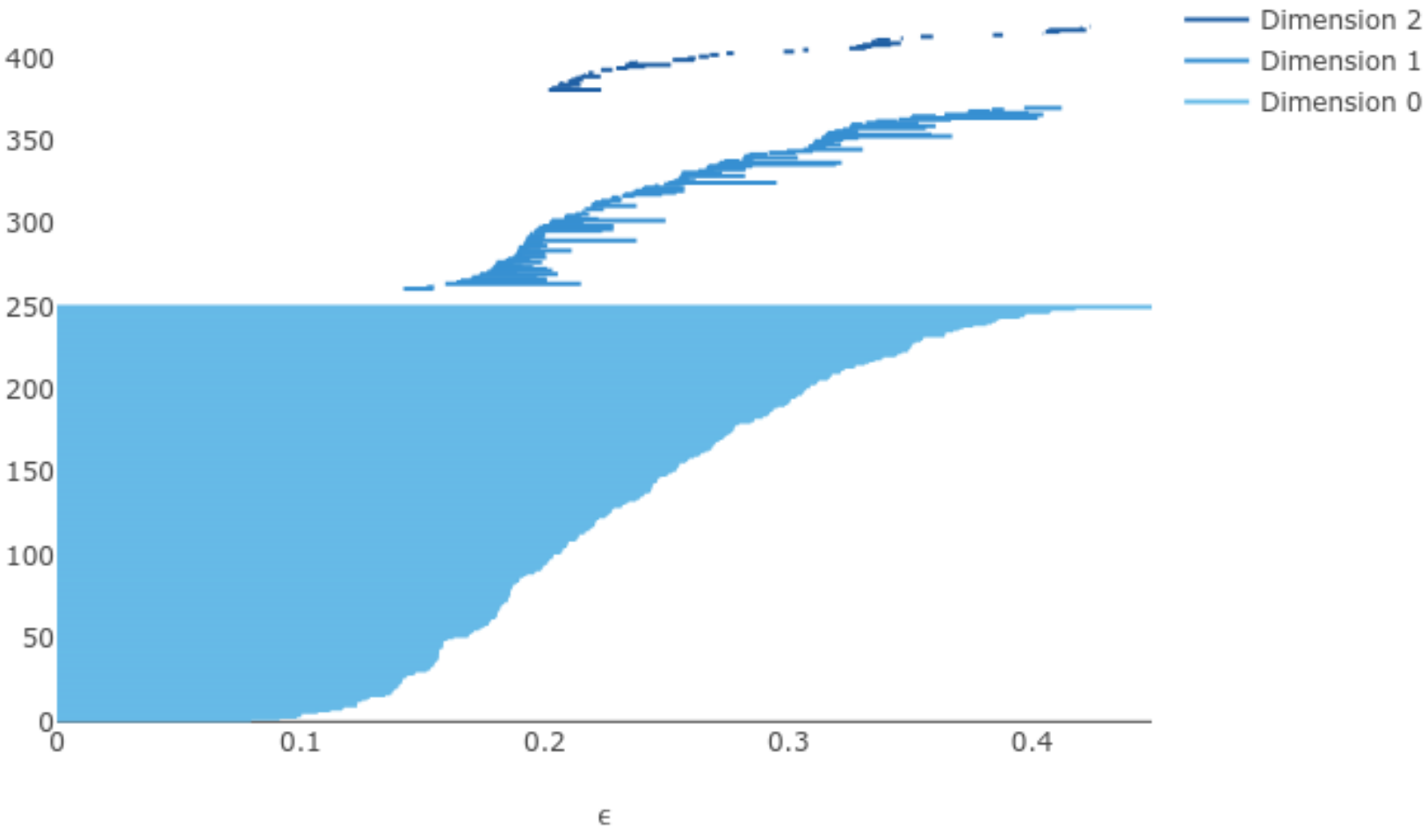
Persistent homology

- Begin with point cloud data.
- Inflate a ε-ball at each point.
- Draw an edge between points when their ε-balls intersect.
- Draw an n-simplex wherever possible.
- Compute the homology of this complex as ε changes.

Persistent homology

- Begin with point cloud data.
- Inflate a ε-ball at each point.
- Draw an edge between points when their ε-balls intersect.
- Draw an n-simplex wherever possible.
- Compute the homology of this complex as ε changes.
- Track when generators appear/disappear.

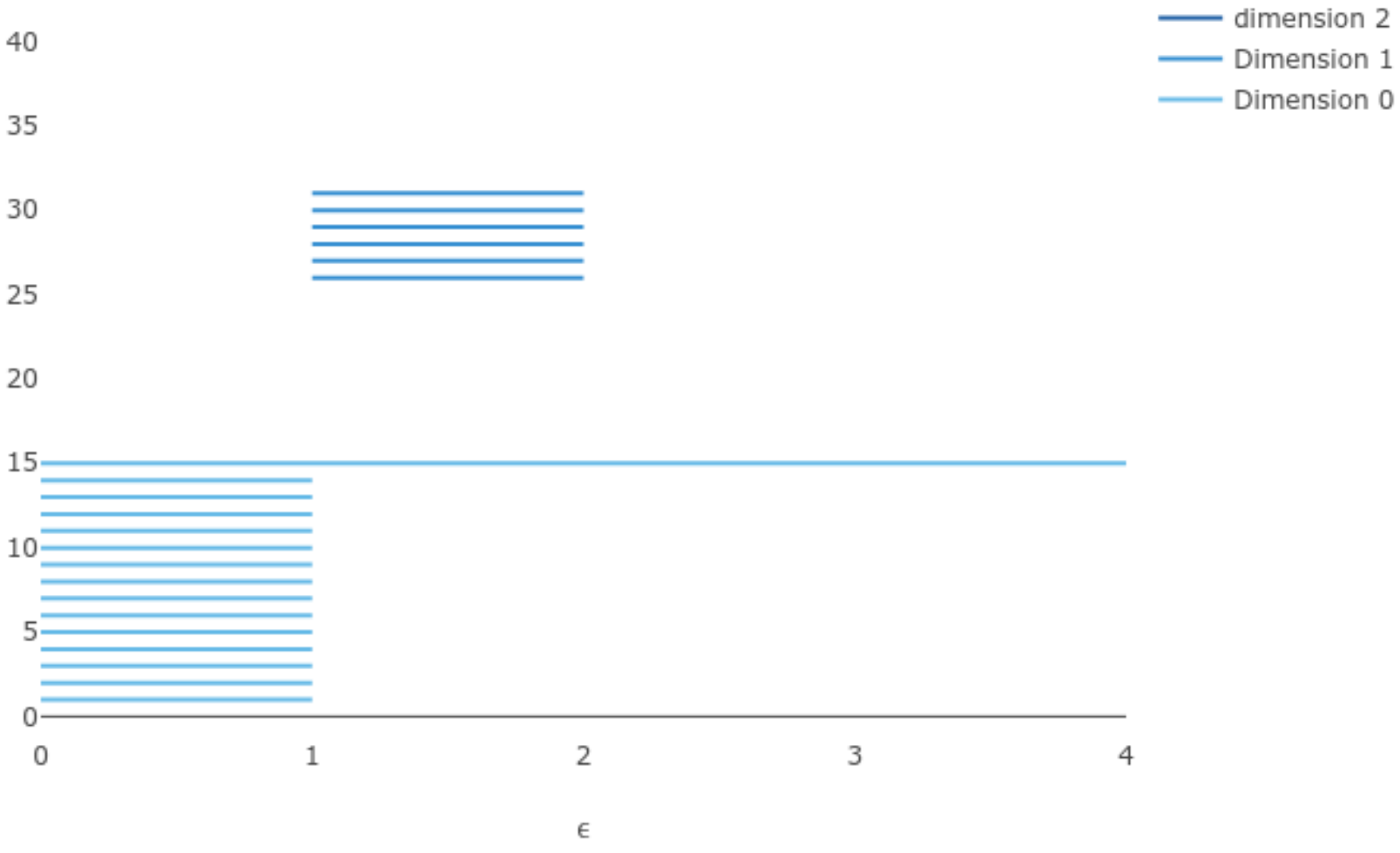
Persistent homology in quartet space



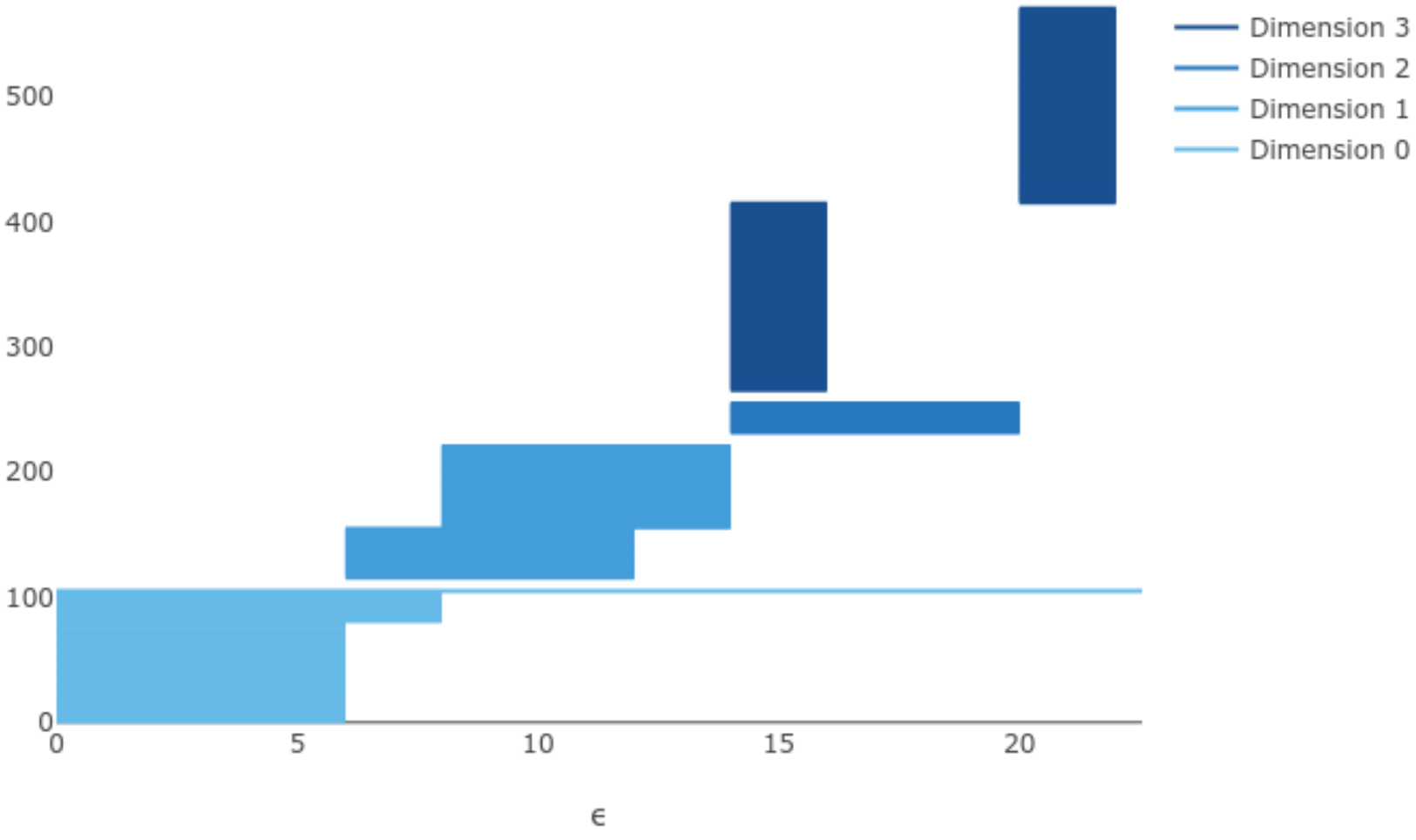
Persistent homology in quartet space

- Are topological features due to the dataset, or the ambient space?
- Never a problem for data embedded in $\mathbb{R}^n$.

Filtration of Q_5 complex



Filtration of Q_6 complex



The category of tree spaces

Consider the category $\mathbf{Q}$.
Objects: Q_n for $n = 1, 2, ...$
(Quartet metric is technically undefined until Q_4.)
Arrows: generated from insertion maps and deletion maps.

Deletion and insertion maps

- Deletion maps are easy: there are only n of them $Q_n \to Q_{n-1}$.
- Insertion maps are not easy because there is no neutral way to choose an insertion site.

Uniform graftings

- Write $S \star T$ to graft T onto S uniformly.

Uniform graftings

- Write $S \star T$ to graft T onto S uniformly.
- Non-commutative and non-associative.

Uniform graftings

- Write $S \star T$ to graft T onto S uniformly.
- Non-commutative and non-associative.
- We are interested in grafting subtrees in non-uniformly as well.

Uniform graftings

Distance under uniform grafting

For n-trees S and T, and for a rooted k-tree R, we have

$$d(g_R(S), g_R(T)) = k^4 d(S, T).$$

Distance under uniform grafting

Proof.

(Sketch.) Every quartet in $g_R(S)$ will either lie entirely within one subtree equivalent to R, or will be split across two to four such subtrees. Quartets which are split across fewer than four subtrees are shared by both $g_R(S)$ and $g_R(T)$, so do not contribute to quartet distance. A quartet that is split across four subtrees exists in $g_R(S)$ whenever the leaves in S to which those subtrees were grafted formed a quartet. So there are $d(S, T)$ possible subtree-quartet choices in which it is possible to form a quartet unique to $g_R(S)$ or $g_R(T)$. There are k^4 leaf choices for each such subtree-quartet choice.

Thus $d(g_R(S), g_R(T)) = k^4 d(S, T)$. □

“Factoring” quartet space?

- This means that there will be scaled, disjoint copies of Q_k in Q_n whenever $k|n$.
- Upper bound for the number of copies:

$$(2\frac{n}{k} - 3)!! \frac{n!}{\frac{n}{k}! \cdot k!^{n/k}}$$

“Factoring” quartet space?

- I am trying to work out how the presence of these copies of Q_k lying Q_n affects the persistent homology of Q_n.

“Factoring” quartet space?

- I am trying to work out how the presence of these copies of Q_k lying Q_n affects the persistent homology of Q_n.
- I conjecture that some important features of the persistent homology of Q_n depend on the factors of n.

"Factoring" quartet space?

- I am trying to work out how the presence of these copies of Q_k lying Q_n affects the persistent homology of Q_n.
- I conjecture that some important features of the persistent homology of Q_n depend on the factors of n.
- Knowing the persistent homology of Q_n will help to interpret the barcode diagrams for natural datasets in Q_n.

“Factoring” quartet space?

- I am trying to work out how the presence of these copies of Q_k lying Q_n affects the persistent homology of Q_n.
- I conjecture that some important features of the persistent homology of Q_n depend on the factors of n.
- Knowing the persistent homology of Q_n will help to interpret the barcode diagrams for natural datasets in Q_n.
- Approximate Q_n for highly-coprime n using Q_m using highly divisible m close to n.

Changing metrics

- Quartet metric has some drawbacks, and now that I have a better understanding of the kinds of problems that are arising, I might choose something different.

Changing metrics

- Quartet metric has some drawbacks, and now that I have a better understanding of the kinds of problems that are arising, I might choose something different.
- Possibility: use a metric that is especially nice with respect to general graftings.

Changing metrics

- Quartet metric has some drawbacks, and now that I have a better understanding of the kinds of problems that are arising, I might choose something different.
- Possibility: use a metric that is especially nice with respect to general graftings.
- Possibility: use a metric that is at least partially-defined on T_∞ and consider whether there are interesting features there that can be described in terms of its role in a category like $\mathbf{Q}$.

Future research directions

- Look for better bounds on the number of copies of Q_k in Q_n when $k|n$.

Future research directions

- Look for better bounds on the number of copies of Q_k in Q_n when $k|n$.
- Determine how these copies of Q_k interact with each other and surrounding space under persistent homology.

Future research directions

- Look for better bounds on the number of copies of Q_k in Q_n when $k|n$.
- Determine how these copies of Q_k interact with each other and surrounding space under persistent homology.

Acknowledgements

- I thank my research supervisors Dr Dorette Pronk and Dr Andrew Irwin.
- Dr Ed Susko explained the statistical aspects of phylogenetics to us. He also generously preprocessed data and provided it to us. Researchers at the Roger Lab at Dalhousie had conducted earlier stages of preprocessing.
- Thanks to the CMS for their travel funding for this conference.
- I thank the funding agencies that make my work possible, NSERC and the Simons Foundation.