# Tree distances under random walks on tree spaces

Sean Cleary Alejandro Morejon

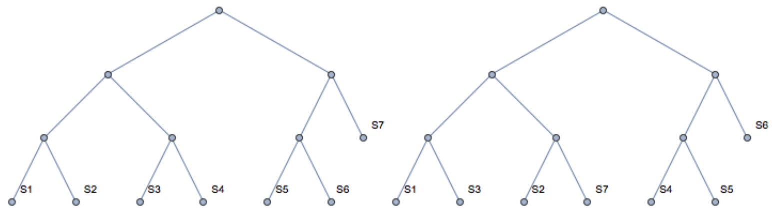
The City College of New York and the CUNY Graduate Center

AMS Special Session, Manoa 2019

This material is based upon work supported by the National Science Foundation under grant no. DMS-1417820

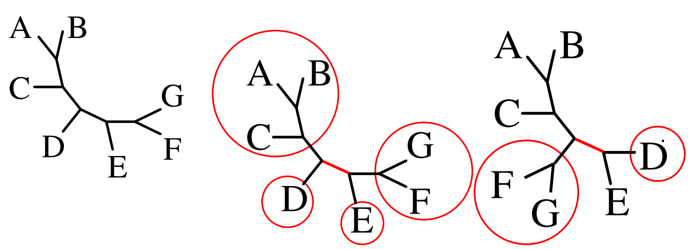
# Distances between phylogenetic trees

- Given two trees (rooted or unrooted) on the same set of leaves (taxa), to what extent to they differ?

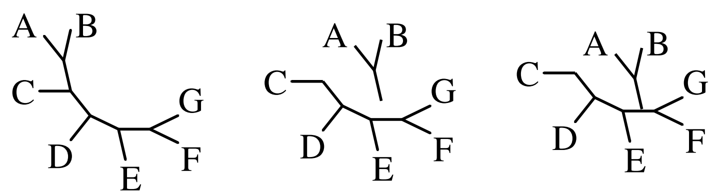



- Metrics on trees: Nearest Neighbor Interchange (NNI), Robinson-Foulds (RF) $\Delta$, Subtree-Prune-Regraft (SPR), Tree Bisection and Reconnection (TBR), Billera-Holmes-Vogtmann (BHV), ...

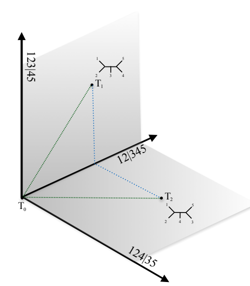
Nearest neighbor interchange move at edge between D and E:



Subtree-prune-regraft move at edge between subtree containing AB and the rest of the tree:



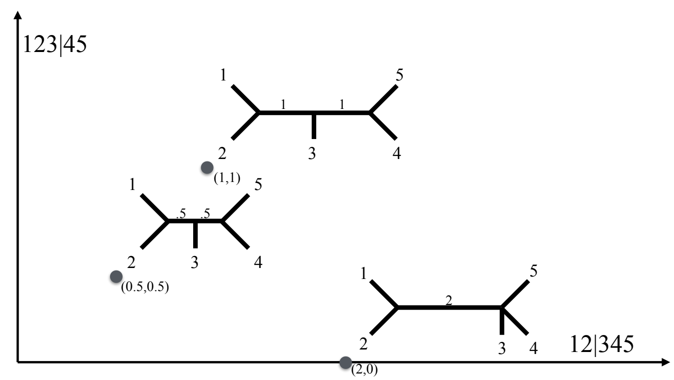
# Trees as Vectors




| | 1 \| 2345 | 2 \| 1345 | 3 \| 1245 | 4 \| 1235 | 5 \| 1234 | 12 \| 345 | 13 \| 245 | 14 \| 235 | 15 \| 234 | 23 \| 145 | 24 \| 135 | 25 \| 134 | 34 \| 125 | 35 \| 124 | 45 \| 123 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $T_0 = (1, 2, 3, 4, 5)$ | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| $T_1 = ((1, 2), (3, (4, 5))$ | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| $T_2 = ((1, 2), (4, (3, 5))$ | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |

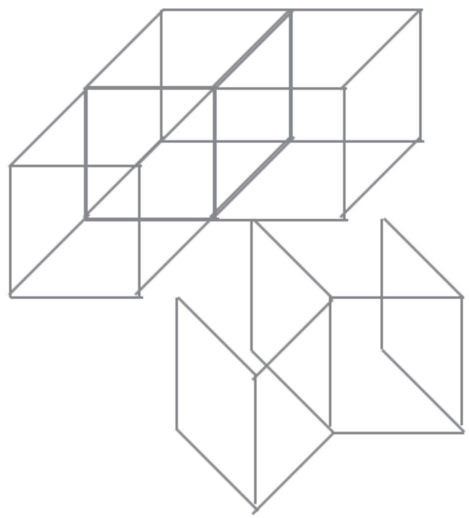
## Billeara-Holmes-Vogtmann space

- Trees on fixed *n* taxa, edge lengths positive reals
- Edges to leaves have fixed length 1
- Form a space with a metric which is locally Euclidean



## Billera-Holmes-Vogtmann treespace

- Adjacency of orthants: two orthants of dimension $k$ share a face of dimension $l$ if they have $l$ edges in common.
- Gives gluing of orthants to obtain space



## Geodesics in treespace

- Piecewise Euclidean segments in a sequence of orthants
- Many cells to consider
- First algorithms exponential
- Polynomial-time algorithm of order $O(n^4)$ of Owen-Provan (2011) uses incompatibility graph of edges to compute which edges to drop and introduce successively.
- Owen-Provan algorithm foundational for computing means, interpolations, variances, principal components, ...

## Random walks in tree spaces

Typical search algorithm:

- Generate a set of random trees
- Score them with respect to some optimality criterion
- Take the best or some of the best
- Perturb these trees with some local adjustments
- Take the best scoring and repeat.

## Random walks in tree spaces

Effectiveness of this process depends upon

- How well random walks visit regions of tree space
- How well hill-climbing works
- Length of process

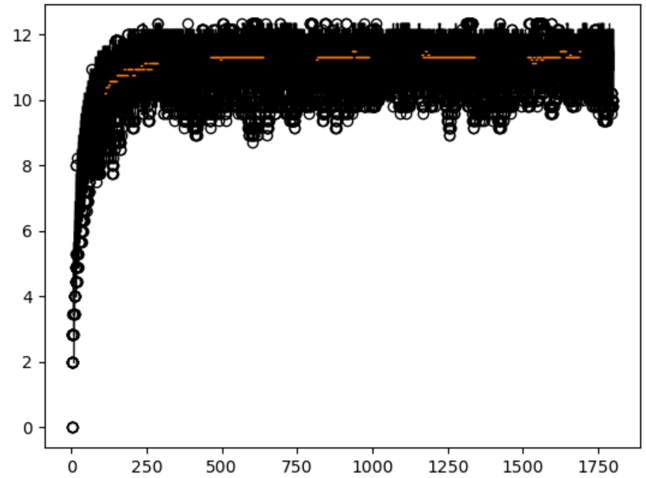
## Random walks in treespace

- Randomly generate an initial tree $T_0$.
- Select a location to perform a move (NNI, SPR) at random
- Apply move to get $T_{i+1}$
- Iterate to get sequence $\{T_0, T_1, T_2, \dots\}$

Look at BHV distances from $T_0$, gives sequence $d_i = d(T_0, T_i)$
Sequence of trees of generally increasing distance- how long until essentially $T_i$ is unrelated to $T_0$?

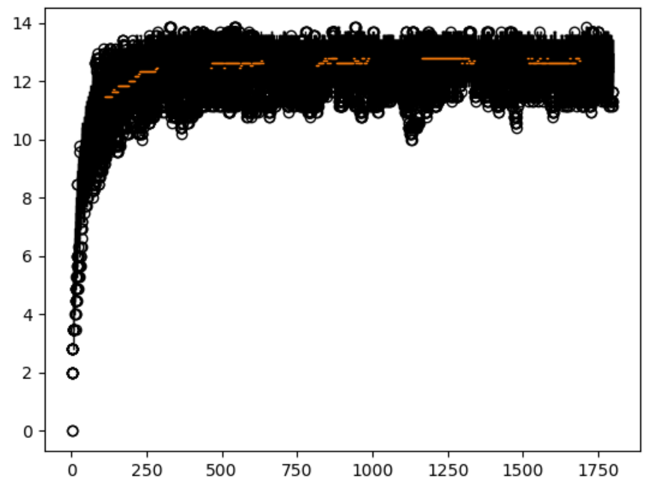
## Distances as walk lengthens

- Example: compute distance from an initial tree of size 40 under random NNI walks:



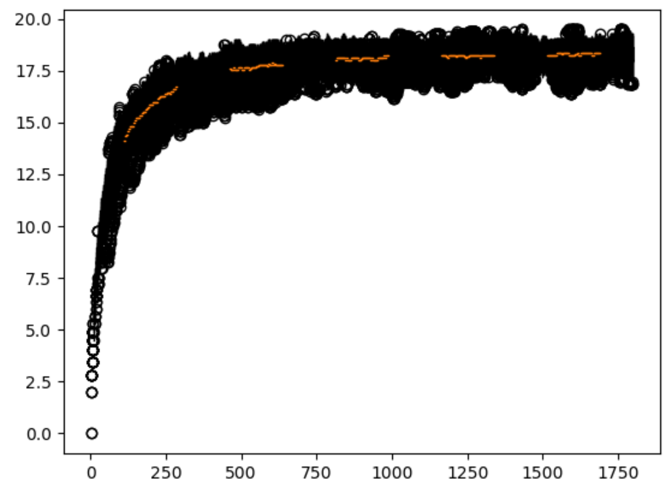
## Distances as walk lenghtens

- Example: compute distance from an initial tree of size 50 under random NNI walks:



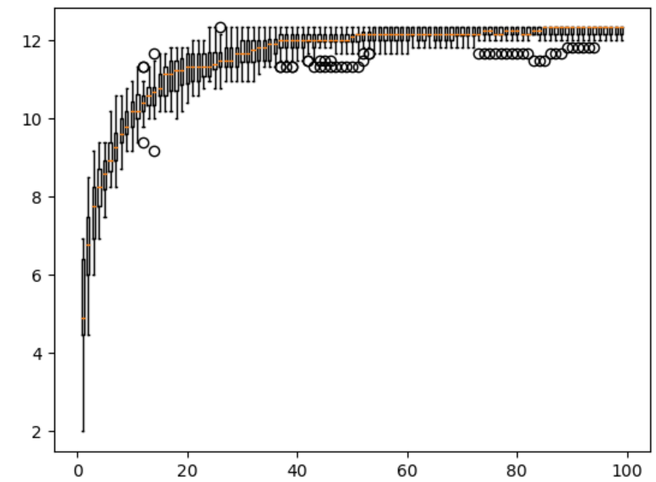
## Distances as walk lengthens

- Example: compute distance from an initial tree of size 100 under random NNI walks:



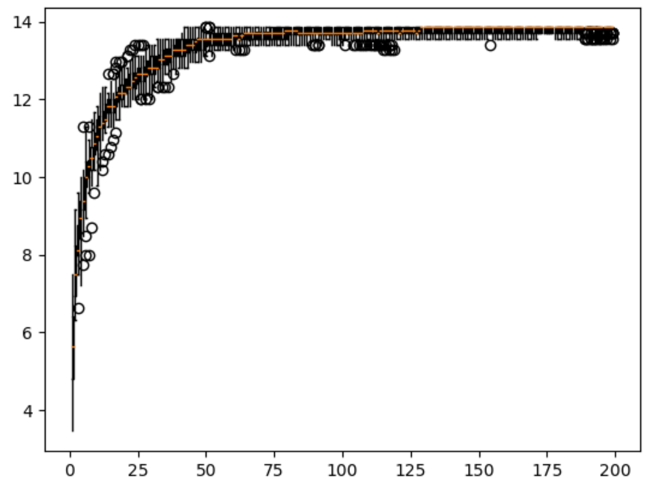
## Distances as walk lengthens

- Example: compute distance from an initial tree of size 40 under random SPR walks:



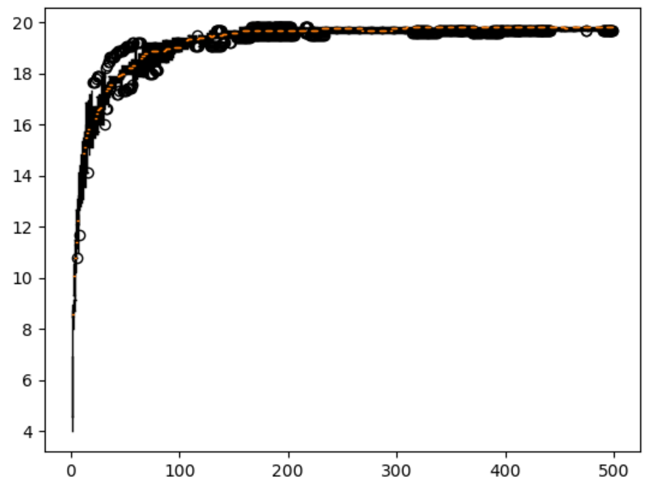
## Distances as walk lenghtens

- Example: compute distance from an initial tree of size 50 under random SPR walks:



## Distances as walk lengthens

- Example: compute distance from an initial tree of size 100 under random SPR walks:



## Walk for mixing

Observations:

- Diameter of treespace is $n$ for both NNI and SPR.
- Neighborhoods larger for SPR than NNI ($n^2$ vs. $n$.)
- These are for undirected random walks
- Walks used for searching are generally directed by optimality criterion

## Walk duration to lose initial information

Observations:

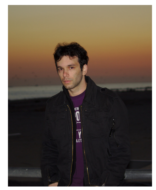
- Seems to grow about linearly with size of tree.
- Faster for SPR than NNI
- NNI can reverse earlier progress away from $T_0$ more readily than SPR
- Coupon-collecting arguments give $n \log n$ bound for all edges to be affected
- Don't need complete coupon collecting as random trees may have some common edges already.

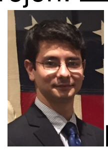
## Other results:

Observations:

- For weighted trees (edge lengths Poisson distributed, for example) similar behavior
- Random walks with weighted edges- selecting an edge equally or proportional to weight, similar behavior
- NNI moves mixing rate not known even for ordered trees (rotation moves)
- In edge length one case, time to a maximally distant tree (distance $2\sqrt{n-2}$) seems to grow as $n \log n$.

## Contributors

Alejandro Morejon: 

Roland Maio:  Haris Nadeem: 

Treespace Working Group:

sites.google.com/site/treespaceworkinggroup/