Tree distances under random walks on tree spaces

Sean Cleary Alejandro Morejon

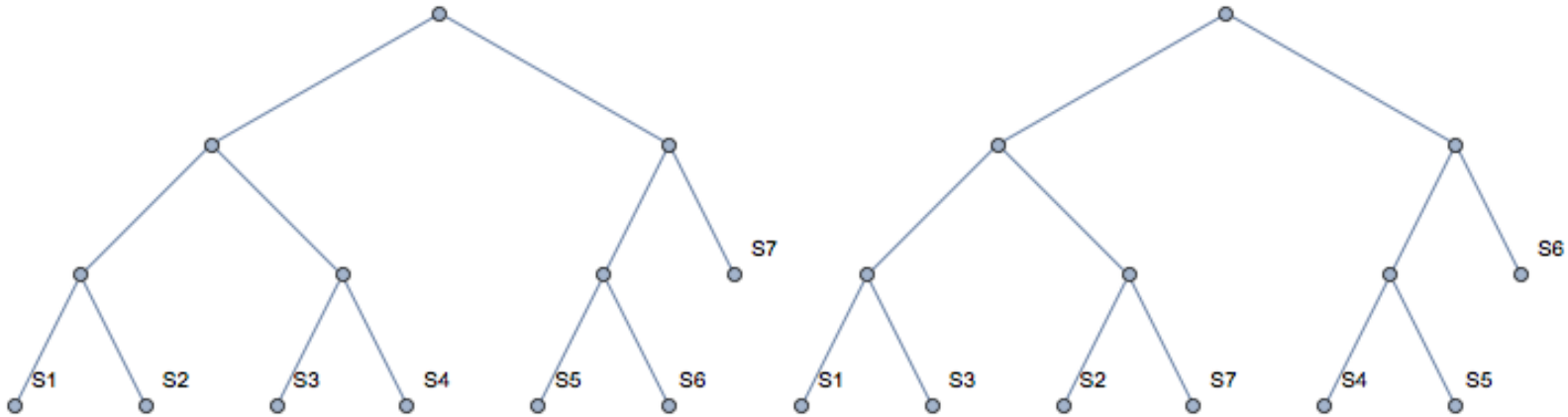
The City College of New York and the CUNY Graduate Center

AMS Special Session, Manoa 2019

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

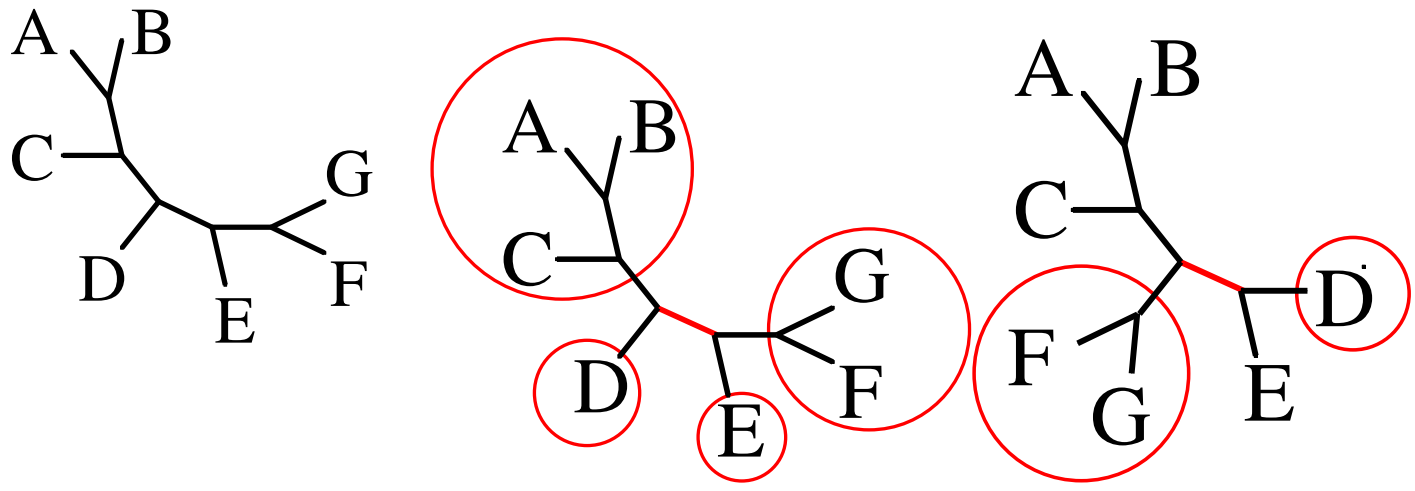
Distances between phylogenetic trees

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

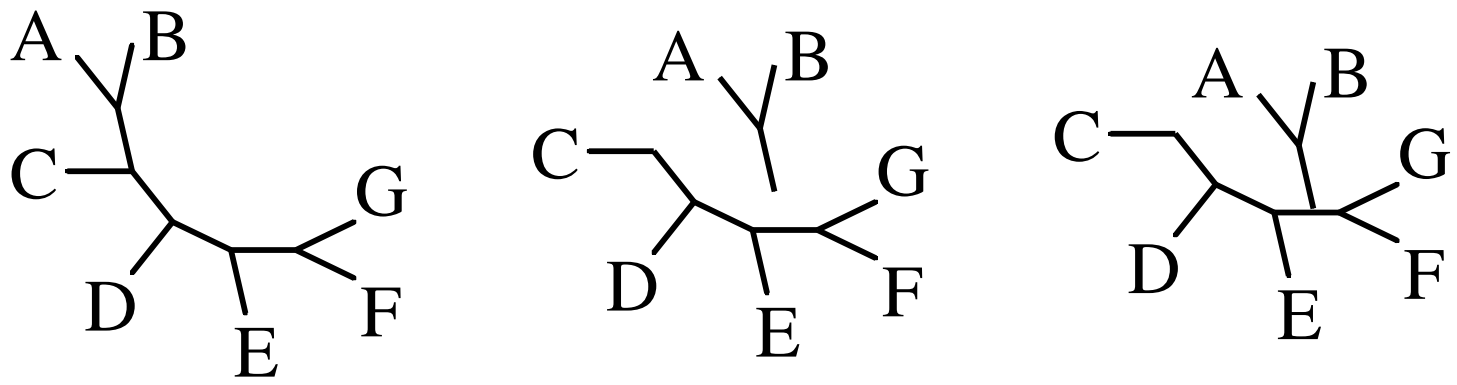


- Metrics on trees: Nearest Neighbor Interchange (NNI), Robinson-Foulds (RF) Δ, Subtree-Prune-Regraft (SPR), Tree Bisection and Reconnection (TBR), Billera-Holmes-Vogtmann (BHV), ...

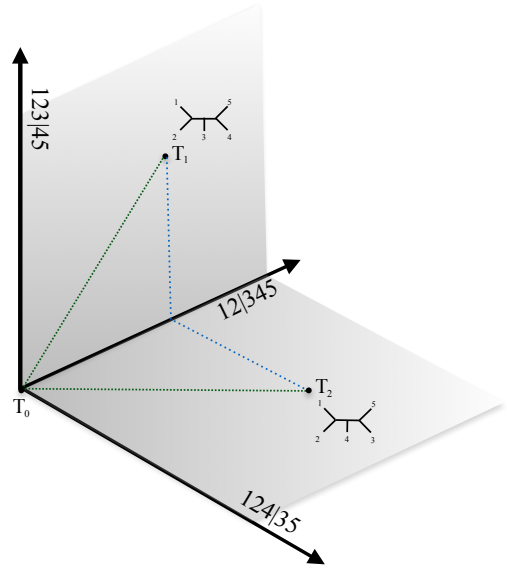
Nearest neighbor interchange move at edge between D and E:



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



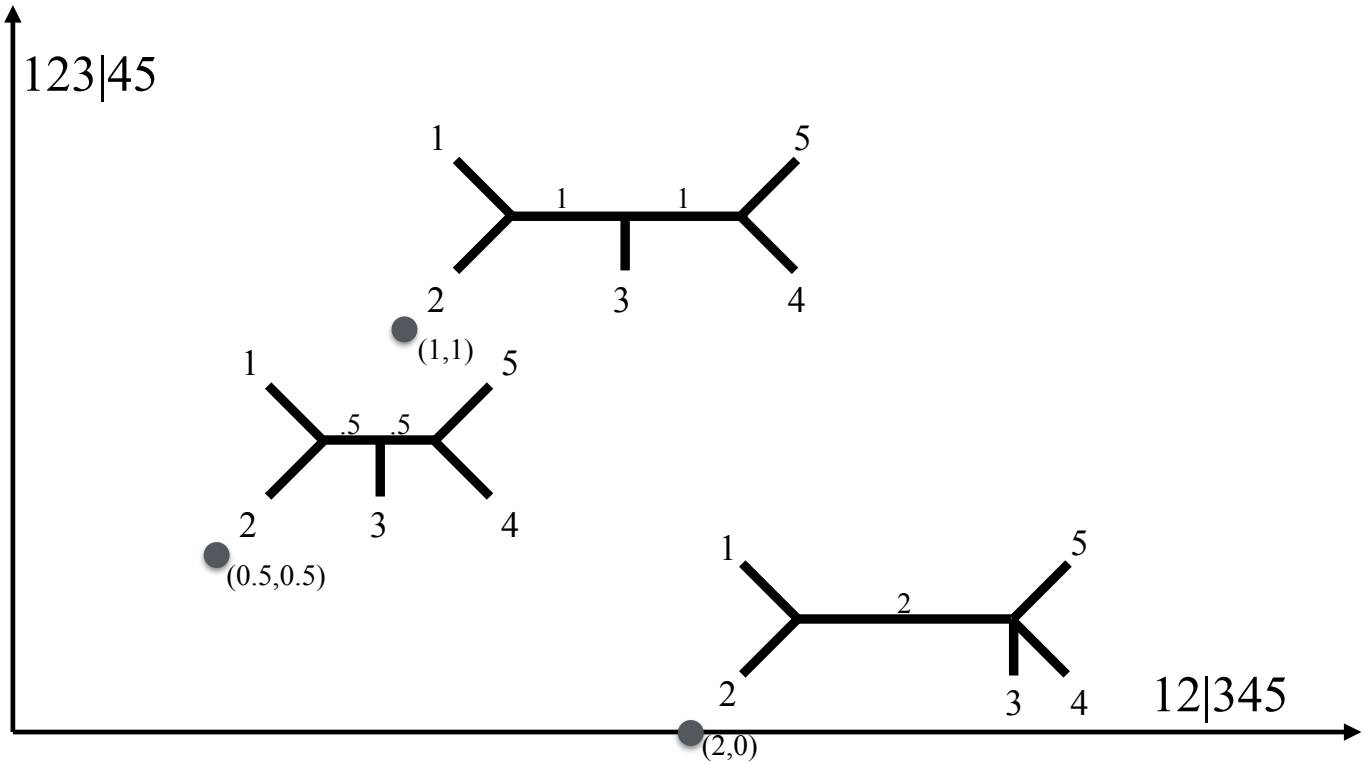
Trees as Vectors



	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

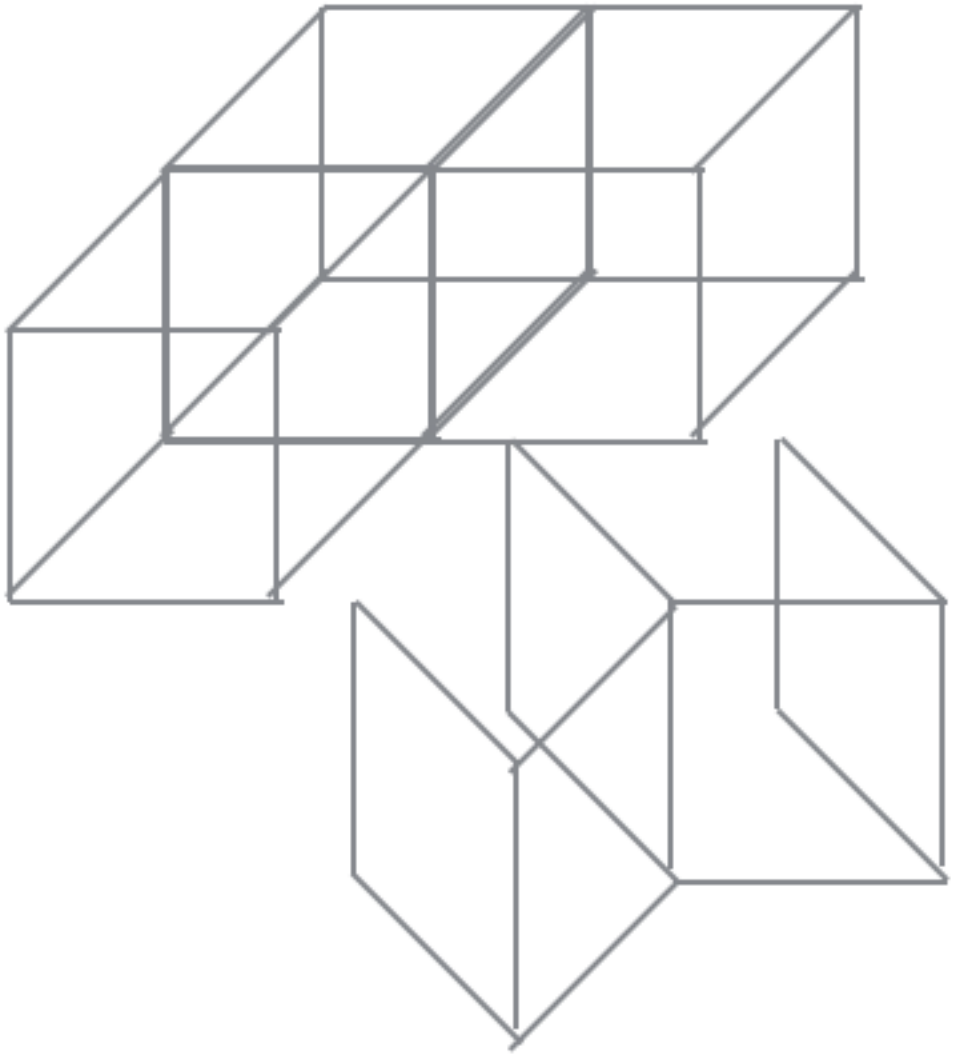
Billeara-Holmes-Vogtmann space

- 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

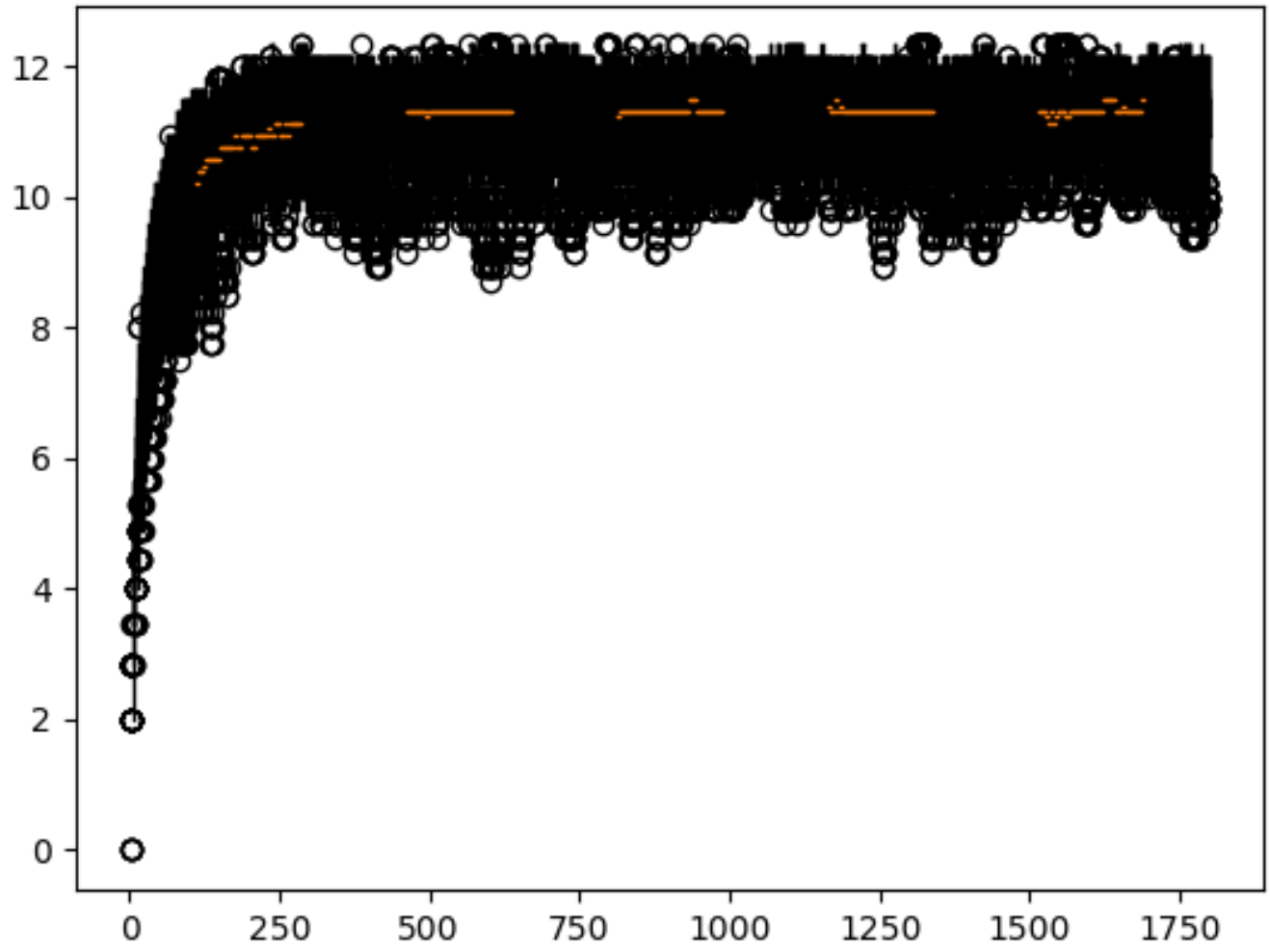
Random walks in treespace

- 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?

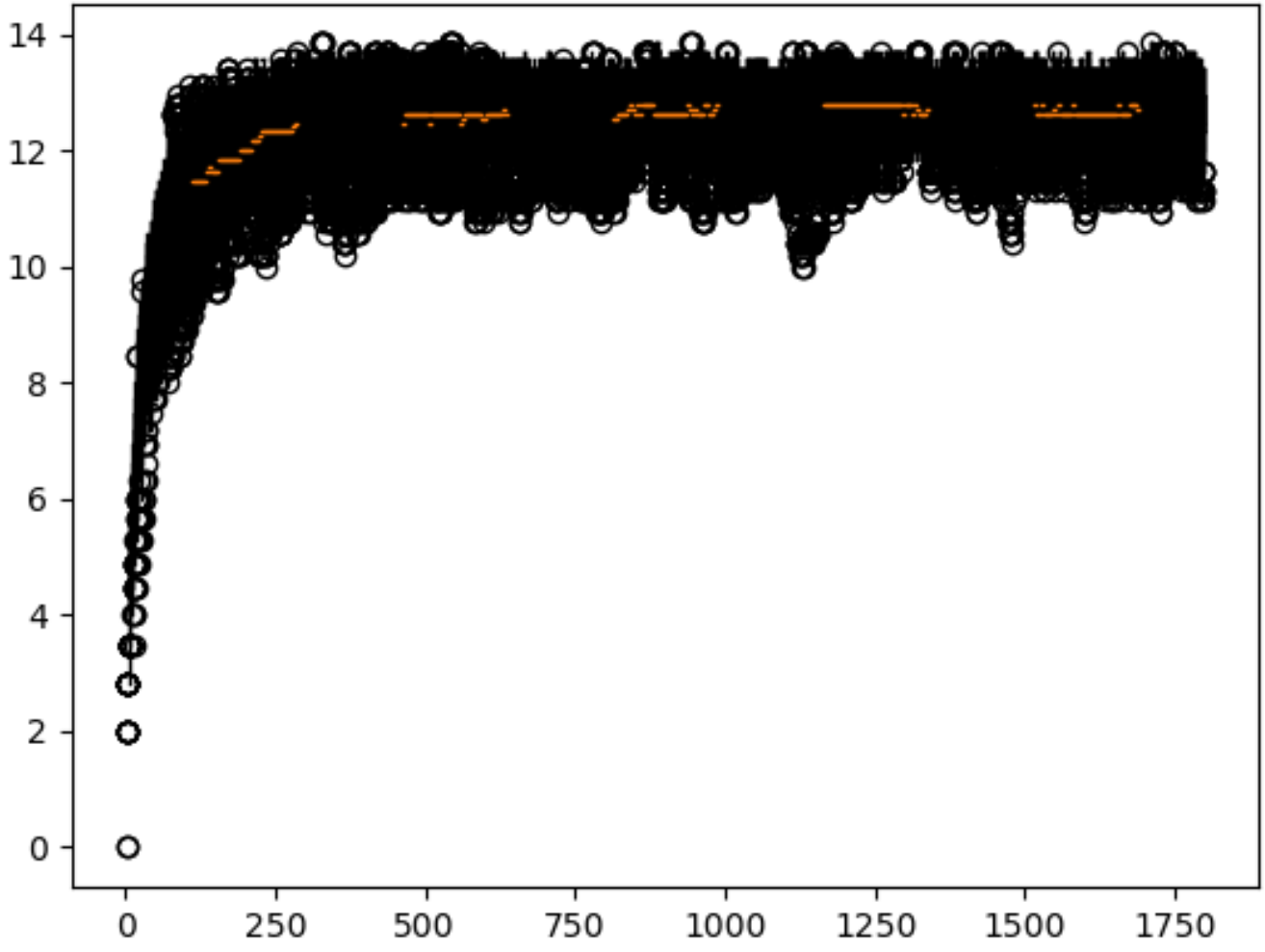
Distances as walk lengthens

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



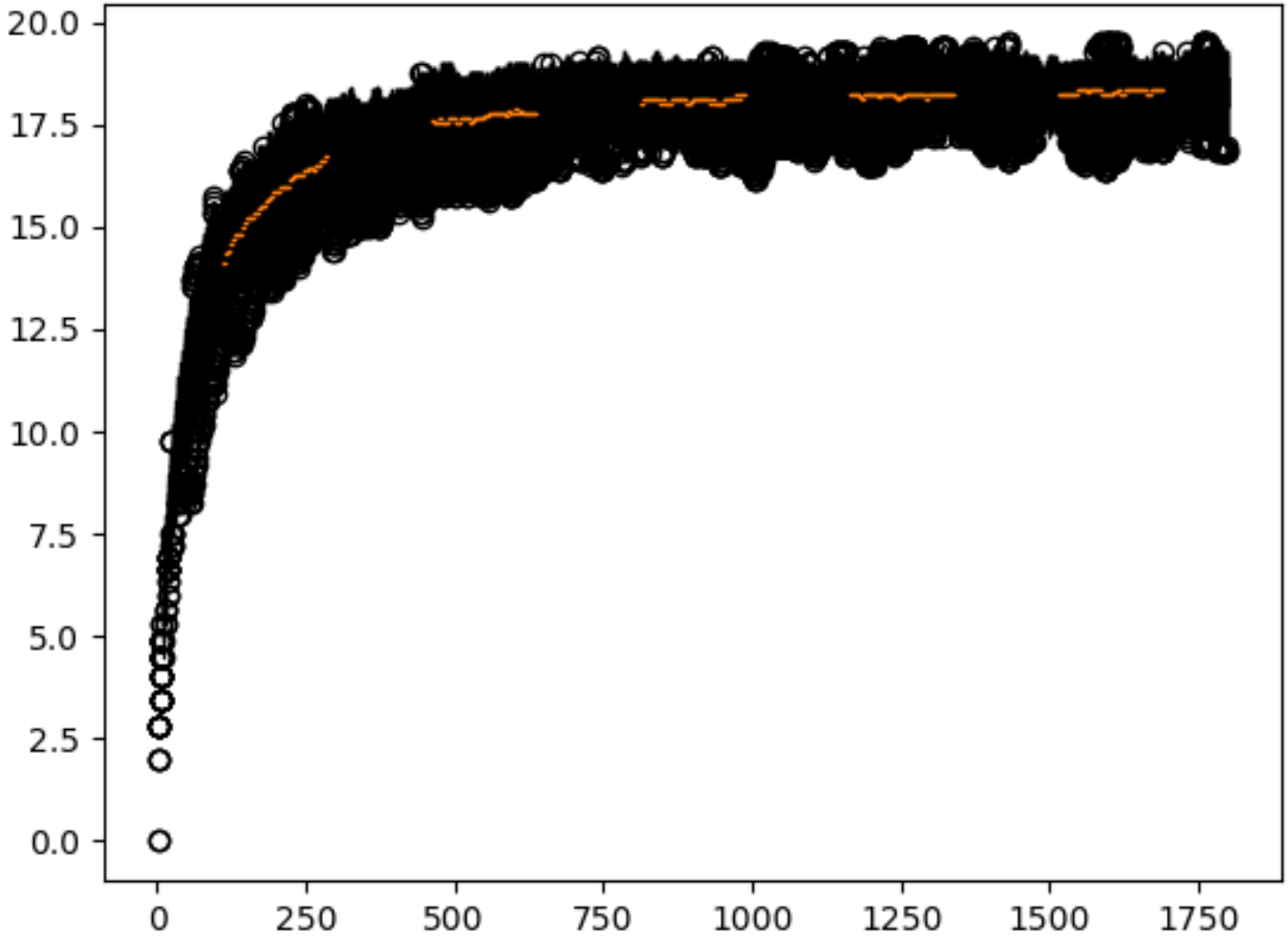
Distances as walk lenghtens

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



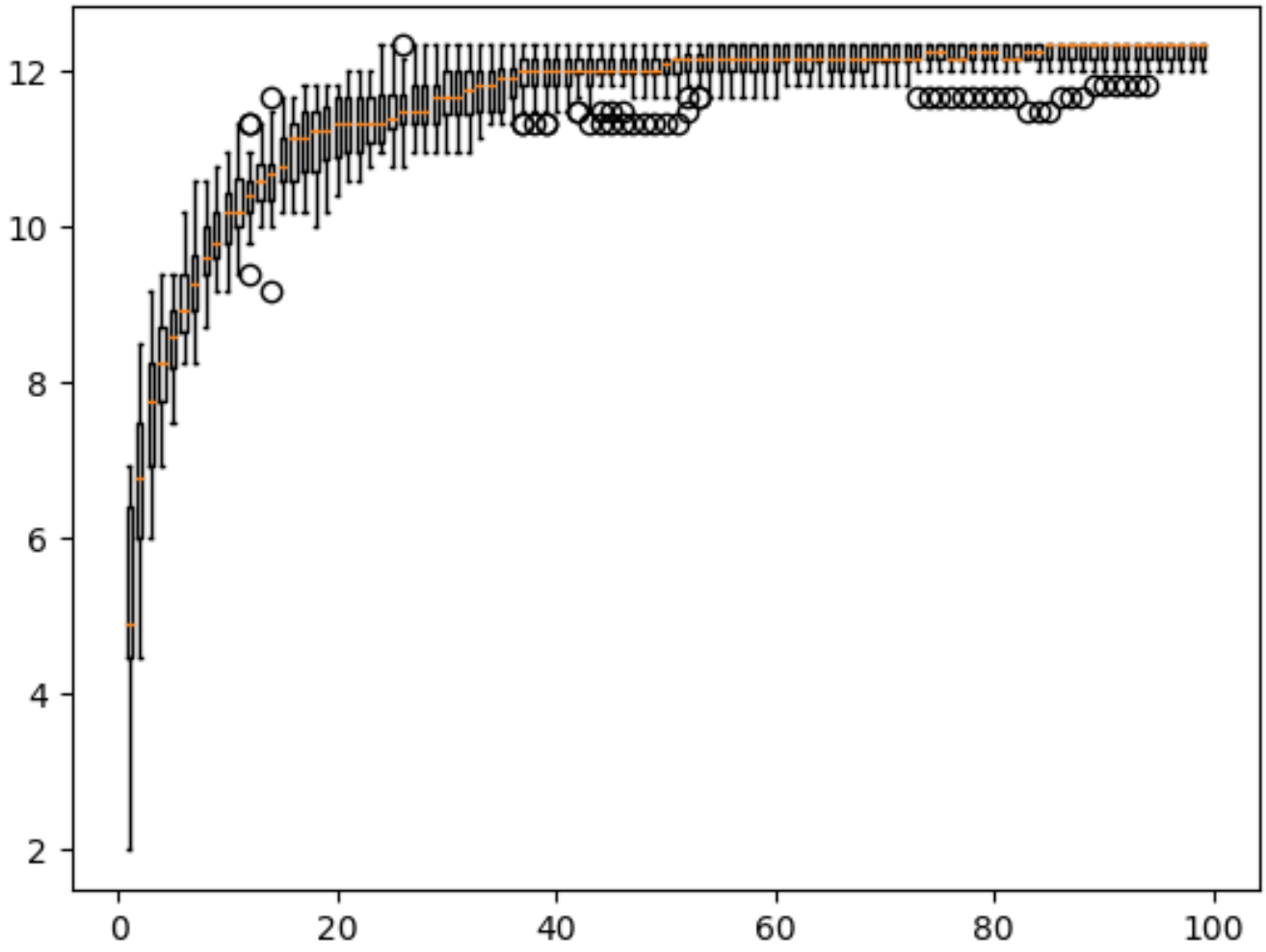
Distances as walk lengthens

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



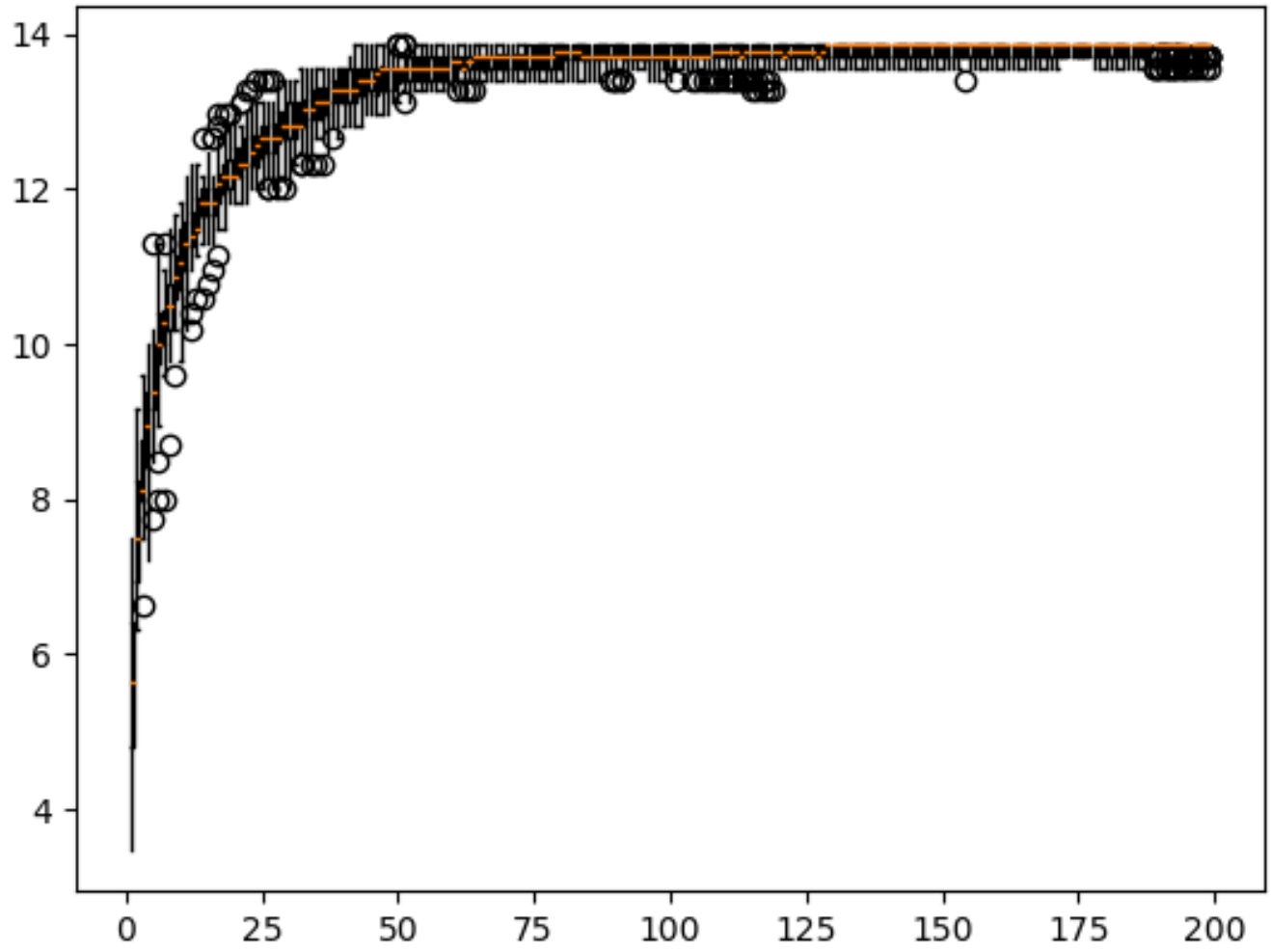
Distances as walk lengthens

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



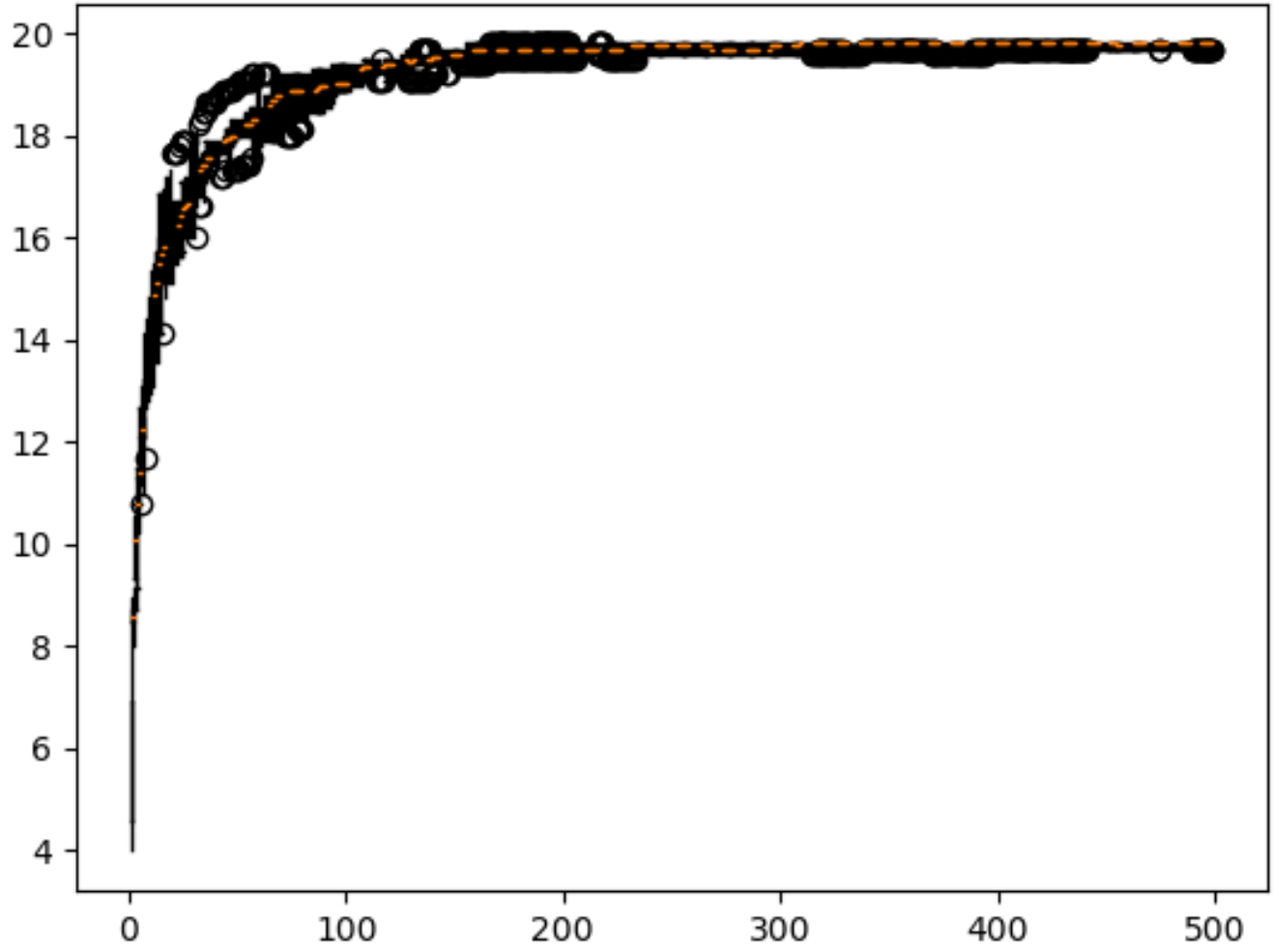
Distances as walk lenghtens

- 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 $n \log n$ 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 will have some common edges already.

Coupon collecting

Number of steps to collect all of a set of N coupons:

$$E(T) = \ldots = \frac{n}{n} + \frac{n}{n-1} + \ldots + \frac{n}{2} + \frac{n}{1}$$

$$\sim n \log n + \gamma n + \frac{1}{2} + O(\frac{1}{n})$$

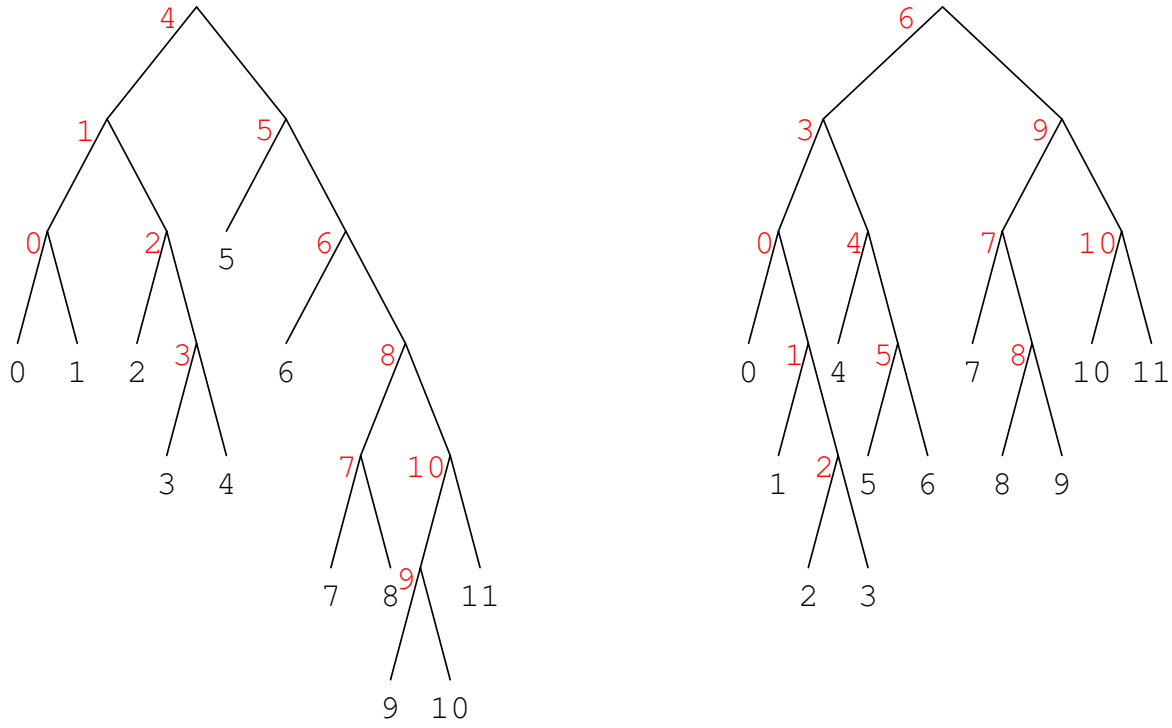
Much of time is spent waiting for the last few coupons.
Randomly generated trees may have some common edges on average.

Other results:

Observations:

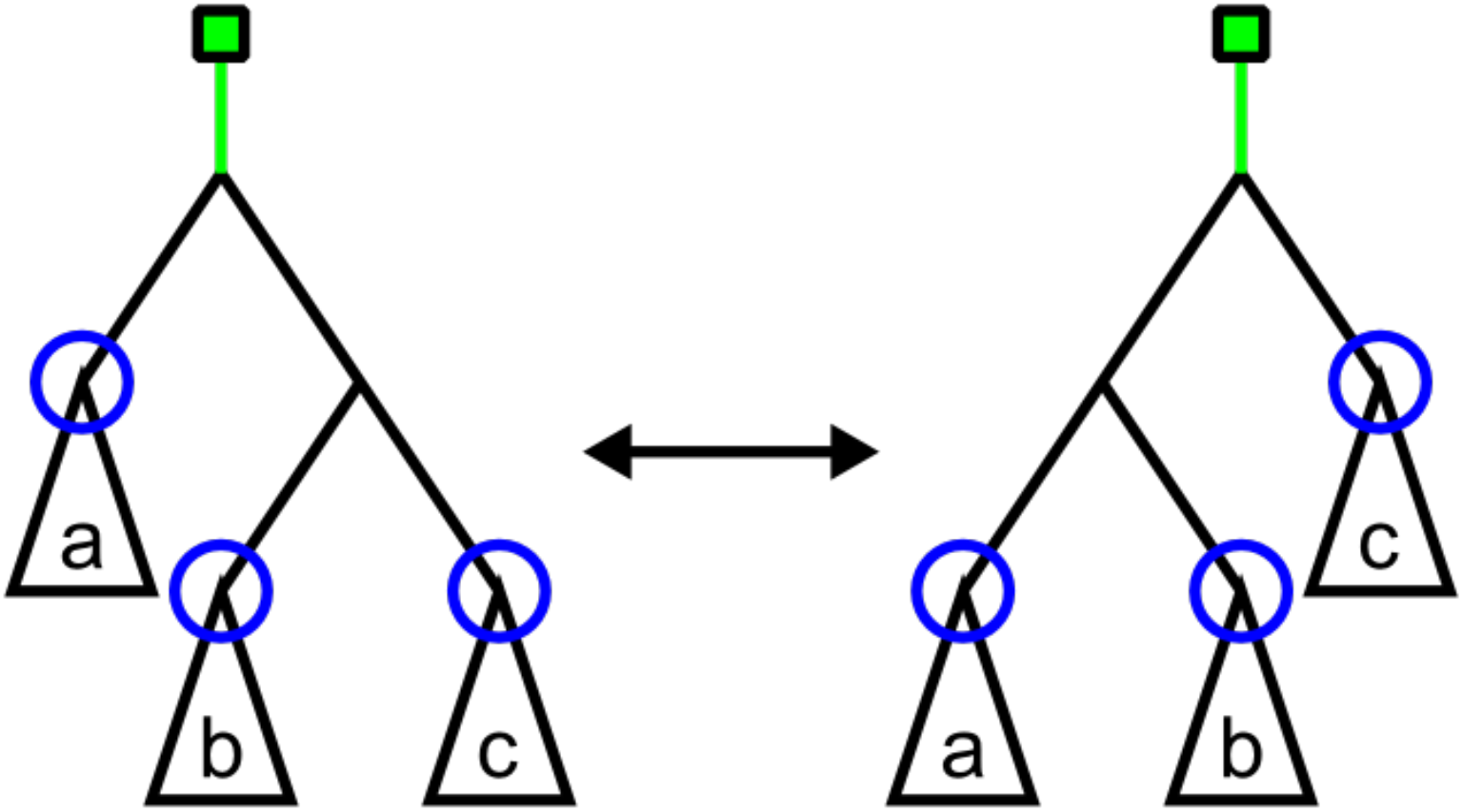
- For weighted trees (edge lengths Poisson distributed, for example) similar behavior
- Random walks with weighted edges- selecting an edge equally or proportional to weight, similar behavior
- NNI moves mixing rate not known even for ordered trees (rotation moves)
- In edge length one case, time to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

Other settings: focus on tree shape, use ordered trees



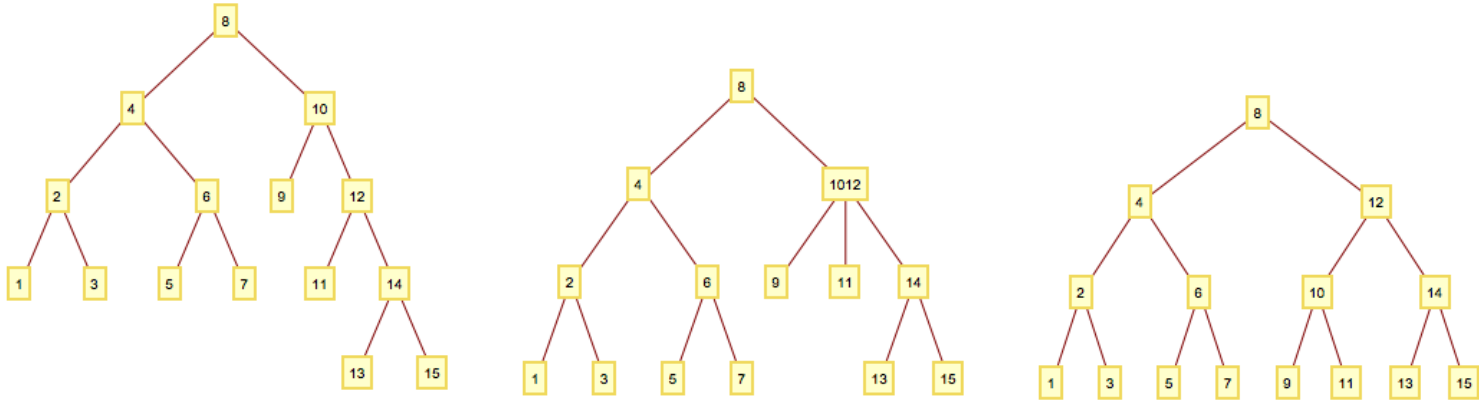
Two trees with left-to-right orders on the leaves (black) and nodes (red).

Rotation operations



Rotation at a node.

Rotation as a slide through a vertex, intermediate temporarily ternary tree



Rooted binary trees and their dual polygonal triangulations

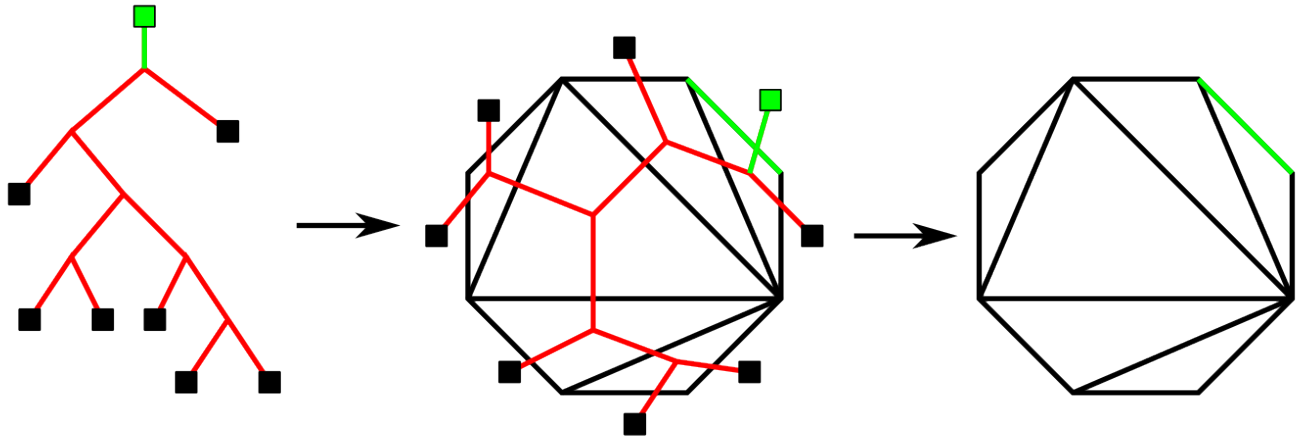
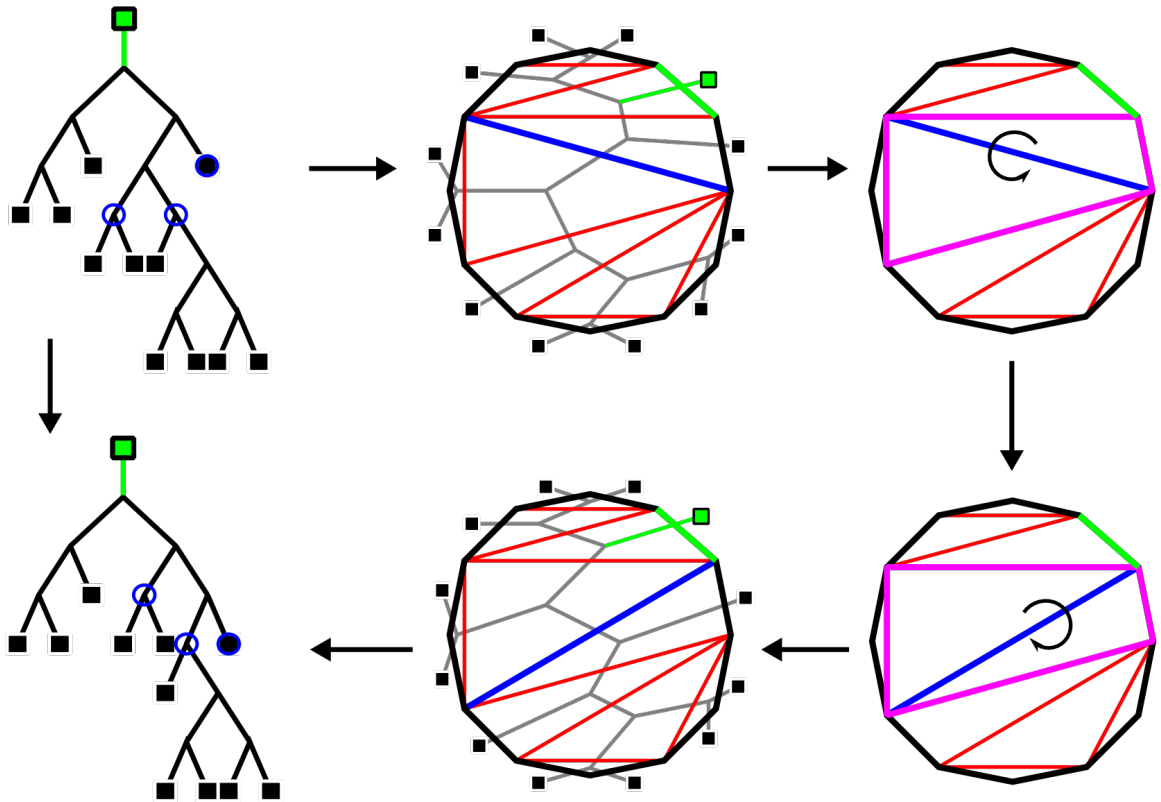


Figure: Converting a rooted binary tree with 7 leaves into its corresponding triangulation of an octagon with a marked edge.

An edge flip in 12-gon and corresponding rooted binary tree and rotation,



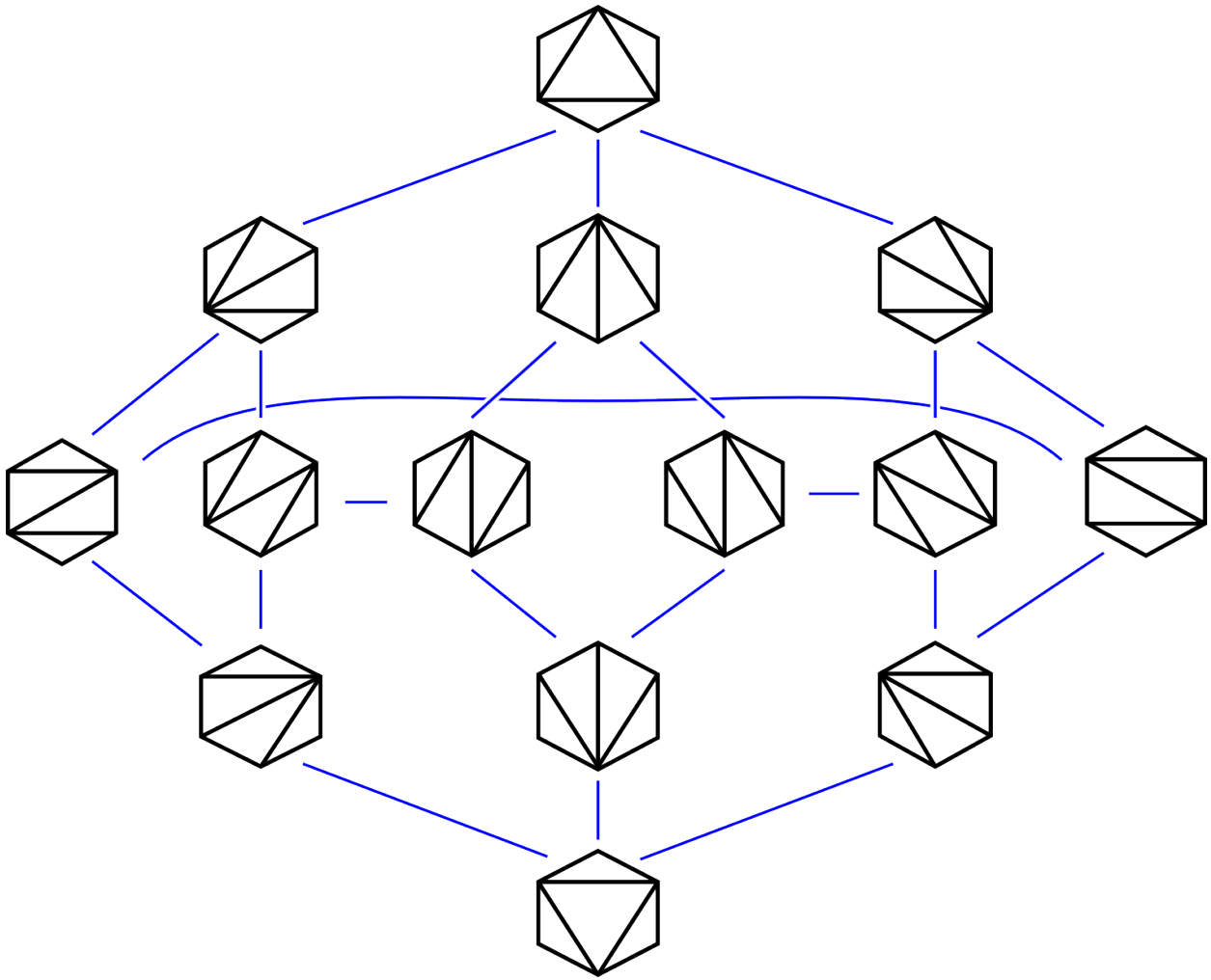
Rotation distance

Def The rotation distance $d_R(S, T)$ between two trees S and T with the same number of nodes is the minimum number rotations needed to transform S to T.

Well-defined? yes, via all-right tree.

Bounds? No more than $2n - 2$ ever, via all-right tree.
Culik-Wood (1982)
No more than $2n - 6$ for $n > 11$, realized!
Sleator-Tarjan-Thurston (1988)

The associahedron for triangulations of the hexagon.

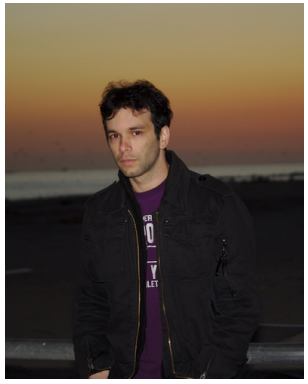


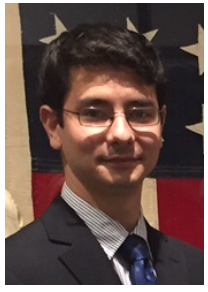
Walks on the associahedron

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

Look at rotation 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?
Cannot compute rotation distance effectively beyond about size 15
Walks in associahedra: appear to mixing with times growing, appears linearly, only for small n possible.

Contributors

Alejandro Morejon: 

Roland Maio: 

Haris Nadeem: 

Treespace Working Group:

sites.google.com/site/treespaceworkinggroup/

Rotation distance bounds

Sleator-Tarjan-Thurston (1983): $d_R(S, T) \leq 2n - 6$ for trees of size 11 or greater.

Sleator-Tarjan-Thurston (1983): There is an N such that for all $n > N$, there are trees of size n such that $d_R(S, T) = 2n - 6$ for trees of size 11 or greater.

Proof uses hyperbolic volume estimates of three manifolds, not very constructive.

Known results from exhaustion

n nodes	1	2	3	4	5	6	7	8	9	10	11	12	13
$n+1$ leaves	2	3	4	5	6	7	8	9	10	11	12	13	14
$n+2$ gons	3	4	5	6	7	8	9	10	11	12	13	14	15
max dist	0	1	2	4	5	7	9	11	12	15	16	18	20
gap from 2n	-2	-3	-4	-4	-5	-5	-5	-5	-6	-5	-6	-6	-6
# trees	1	2	5	14	42	132	429	1430	4862	16796	58786	208k	743k

n is the number of internal nodes in T_i.
Culik-Wood: $d(T_1, T_2) \leq 2n - 2$ for all n
S-T-T: $d(T_1, T_2) \leq 2n - 6$ for all $n \geq 11$
Number of tree pairs: 1, 4, 25, 196, 1764,
17424, 184041, 2044900, 23639044, 282105616
3455793796, 43268992144, 551900410000, 7152629313600,
93990019574025, 1250164827828900

Rotation distance bounds

Sleator-Tarjan-Thurston (1983): $d_R(S, T) \leq 2n - 6$ for trees of size 11 or greater.

Sleator-Tarjan-Thurston (1983): There is an N such that for all $n > N$, there are trees of size n such that $d_R(S, T) = 2n - 6$ for trees of size 11 or greater.

Proof uses hyperbolic volume estimates of three manifolds, not very constructive.

Exhaustive results up to about 20 gave antipodal tree pairs of all sizes tested greater than 11.

Rotation distance bounds

Sleator-Tarjan-Thurston (1983): There is an N such that for all $n > N$, there are trees of size n such that $d_R(S, T) = 2n - 6$ for trees of size 11 or greater.

Dehornoy (2009) There are explicit tree pairs at distances $2n - O(\sqrt{n})$ for large n.

Pournin (2012) Explict examples for all sizes at least 11.

Rotation distance algorithms

No known polynomial time algorithms for exact rotation distance.

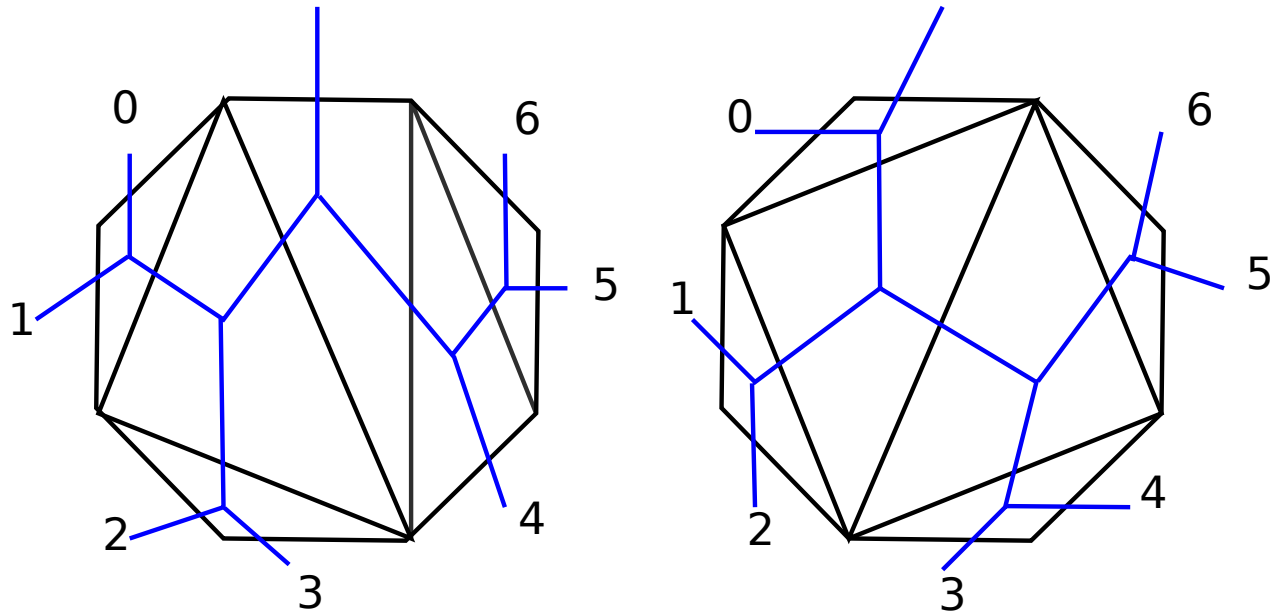
Baril-Pallo (2006) heuristic approximation algorithms

Cleary- St. John (2008) linear time approximation algorithm

Cleary- St. John (2009) rotation distance is fixed-parameter tractable

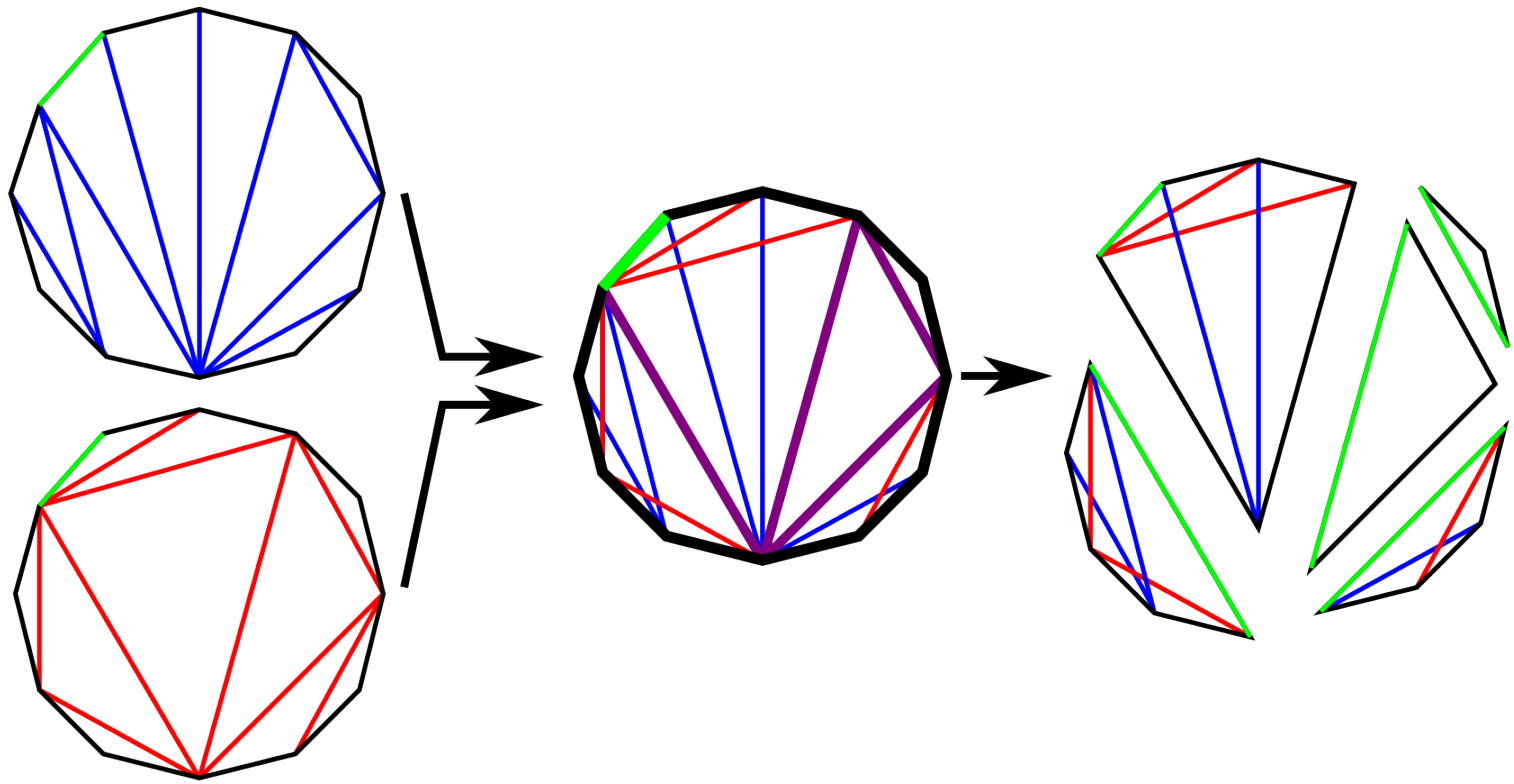
Incidence of common edges

Easier to see from the polygonal perspective, taking trees as duals of triangulations of regular n-gons:
Never undo a matched edge (S-T-T).



Breaking apart along common edges

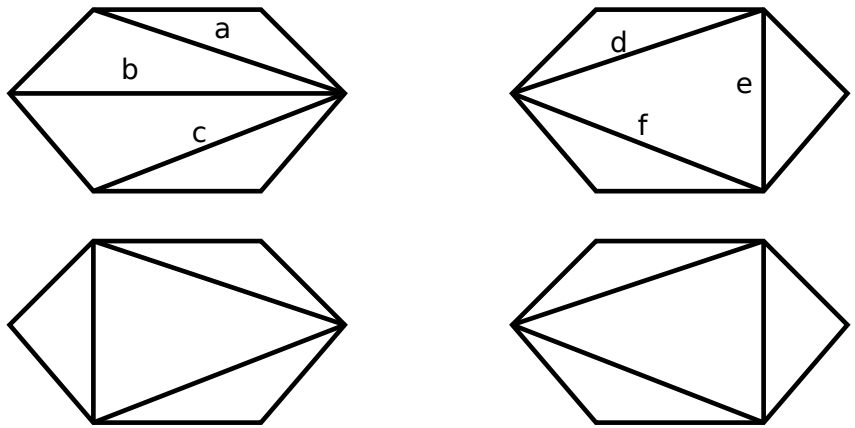
Root edges are marked in green- the purple edges are common to both the blue and the red.



Tree pair is separated into several pieces, and total distance is sum of distances between pairs.

One-off edges are easy but not symmetric

Two one-off edges *a* and *c* in the left triangulation with respect to the right one, but there is just one one-off edge *e* in the right one with respect to the left one:



Two triangulations of the hexagon, with no one-off edges present in either triangulation.

Breaking apart along common edges

Questions:

- How often does a triangulation have a common edge?
- How many common edges on average? Distribution?
- How many sizable pieces result?
- On average, how large is the largest piece?
- How many immediate edge flips are there on average?

In general, interested what happens as the size grows.

(joint with Andrew Rechnitzer and Thomas Wong)

Definition

The generating function for the combinatorial class $\mathscr{P}$ will be denoted $P(z)$, where z is conjugate to the number of non-rooted edges in the polygon between a pair of triangulations. Thus,

$$P(z) = \sum_{n=1}^{\infty} c_n^2 z^{n+1}$$

Define the auxiliary variables m and q on the generating function $P(z)$ with m conjugate to the number of matched edges and q conjugate to the size of the $\mathscr{R}$-piece containing the root edge.

Theorem

The two functions $P(z; m, q)$ and $R(z)$ are related by:

$$P(z; m, q) = R\left(q(z + mP(z; m, 1))\right).$$

Theorem

The two functions $P(z; m, q)$ and $R(z)$ are related by:

$$P(z; m, q) = R\left(q(z + mP(z; m, 1))\right).$$

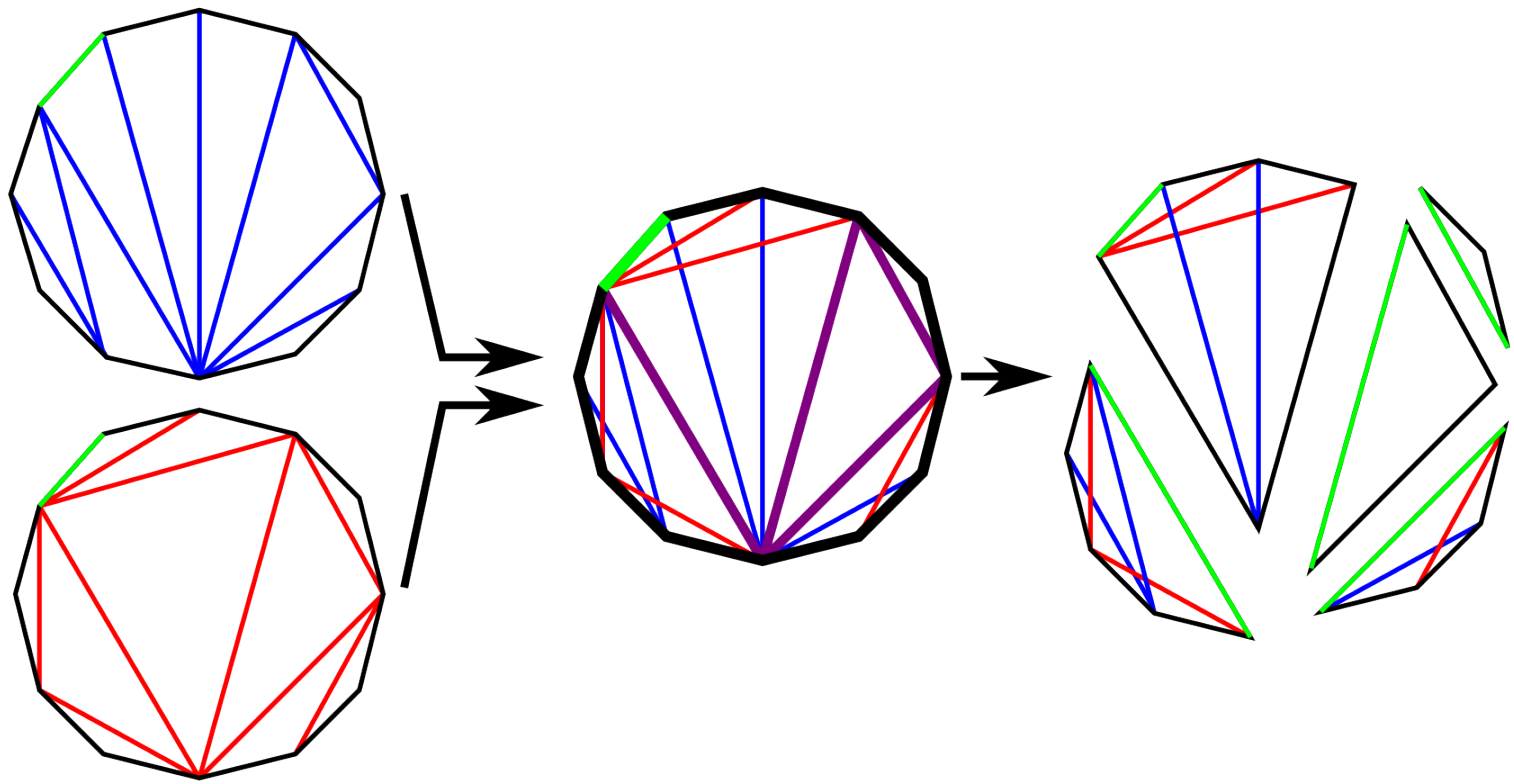
Proof: Recursive construction of $\mathscr{P}$. Each pair of triangulations can be constructed by first picking the corresponding root component in $\mathscr{R}$ (size q) and substituting:

1. If that edge is an edge in the superimposed triangulation, do nothing;
2. If that edge is a common chord, replace it with the corresponding $\mathscr{P}$-piece. This substitution step does not contribute to the root component size q, but it does increase the number of matched edges by 1 since the replaced edge is now a common chord.

Thus $\mathscr{P} = \mathscr{R} \circ (\mathscr{P} + \mathscr{Z})$ where $\mathscr{Z}$ is the atomic class denoting an edge in the root component.

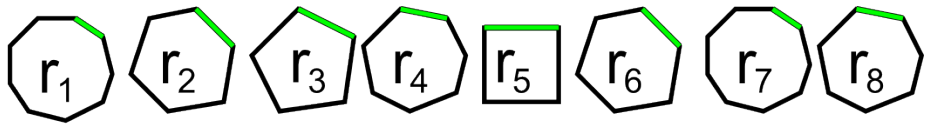
Breaking apart along common edges

Root edges are marked in green- the purple edges are common to both the blue and the red.

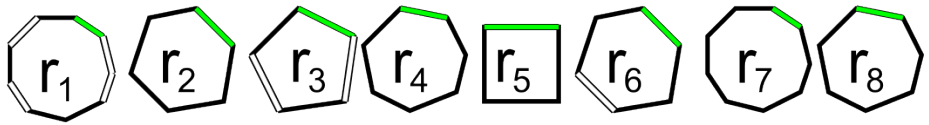


Assembling a sequence of 8 $\mathscr{R}$-pieces

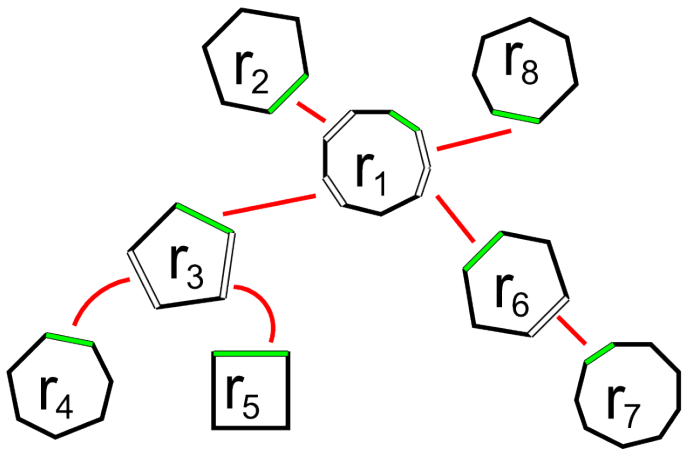
1)



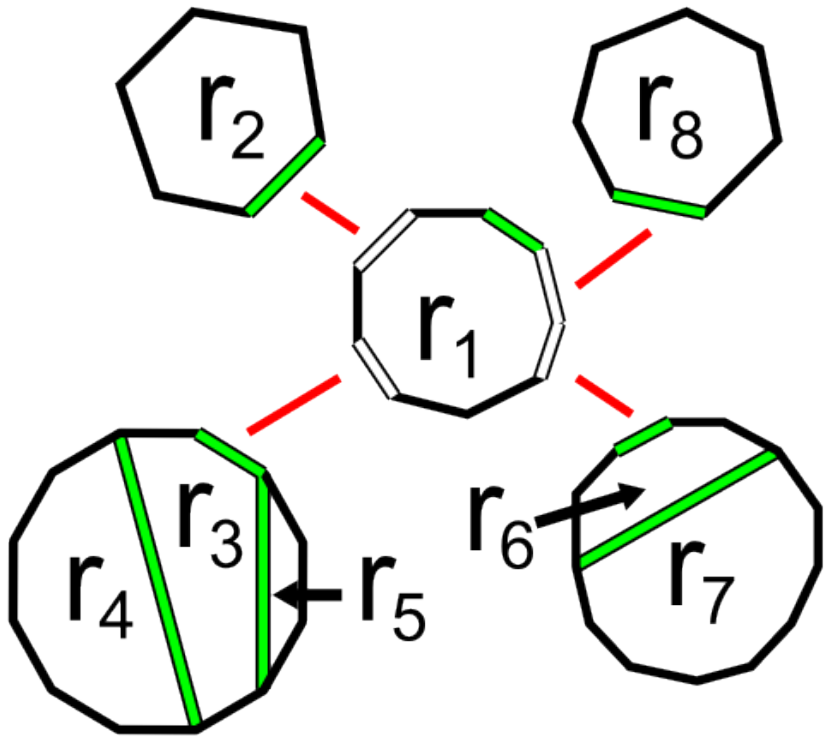
2)



3)



Amalgamating the primitive pieces:



Expected reduction properties

Questions: given a large tree pair diagram:

- How much does it reduce?
- How many common edges are there?
- What are the sizes of the primitive pieces?