Ricci-Ollivier Curvature of the Rooted Phylogenetic Subtree-Prune-Regraft Graph

Conference Paper · January 2016

DOI: 10.1137/1.9781611974324.6 · Source: arXiv

2 authors:

Chris Whidden
Fred Hutchinson Cancer Research Center
26 PUBLICATIONS **354** CITATIONS
SEE PROFILE

Frederick Matsen IV
Fred Hutchinson Cancer Research Center
47 PUBLICATIONS **335** CITATIONS
SEE PROFILE

Ricci-Ollivier Curvature of the Rooted Phylogenetic Subtree-Prune-Regraft Graph

Chris Whidden
Program in Computational Biology
Fred Hutchinson Cancer Research Center
Seattle, WA 98109
Email: cwhidden@fredhutch.org

Frederick A Matsen IV
Program in Computational Biology
Fred Hutchinson Cancer Research Center
Seattle, WA 98109
Email: matsen@fredhutch.org

Abstract

Statistical phylogenetic inference methods use tree rearrangement operations to perform either hill-climbing local search or Markov chain Monte Carlo across tree topologies. The canonical class of such moves are the subtree-prune-regraft (SPR) moves that remove a subtree and reattach it somewhere else via the cut edge of the subtree. Phylogenetic trees and such moves naturally form the vertices and edges of a graph, such that tree search algorithms perform a (potentially stochastic) traversal of this *SPR graph*. Despite the centrality of such graphs in phylogenetic inference, rather little is known about their large-scale properties. In this paper we learn about the rooted-tree version of the graph, known as the rSPR graph, by calculating the Ricci-Ollivier curvature for pairs of vertices in the rSPR graph with respect to two simple random walks on the rSPR graph. By proving theorems and direct calculation with novel algorithms, we find a remarkable diversity of different curvatures on the rSPR graph for pairs of vertices separated by the same distance. We confirm using simulation that degree and curvature have the expected impact on mean access time distributions, demonstrating relevance of these curvature results to stochastic tree search. This indicates significant structure of the rSPR graph beyond that which was previously understood in terms of pairwise distances and vertex degrees; a greater understanding of curvature could ultimately lead to improved strategies for tree search.

I. Introduction

Molecular phylogenetic methods, which reconstruct evolutionary trees from DNA or RNA data, are of fundamental importance to modern biology. Statistical phylogenetics forms the currently most popular means of reconstructing phylogenetic trees, in which the tree is viewed as an unknown parameter in a likelihood-based statistical inference problem. The likelihood function in this setting is the likelihood of generating the observed sequences via a continuous time Markov chain (CTMC) evolving down the tree starting from a sequence assumed to be sampled from the stationary distribution [1]. The lengths of the branches of the phylogenetic tree give the "time" parameter in the CTMC, where the generated sequence accrues mutations, typically in an IID manner across sites. Thus likelihood-based phylogenetics gives an optimality criterion that comes from both branch length and topology. Within this framework researchers typically choose either a Bayesian approach, in which an algorithm (typically Markov chain Monte Carlo or MCMC) approximates the posterior distribution of trees and their associated parameters, or a maximum likelihood (ML) approach, in which an algorithm searches across tree topologies and continuous parameters in order to find the combination with the highest likelihood.

In order to get accurate such estimates, MCMC samplers must sufficiently explore the collection of trees, or heuristic local search algorithms must find a way to the ML tree, avoiding and escaping local peaks. In both settings, one needs a means of describing the trees that are "near" to the current tree. Phylogenetic search algorithms typically explore the collection of trees by rearranging subtrees with subtree-prune-regraft (SPR) moves (Figure 1(d)) or the subset of SPR moves called nearest neighbor interchanges (NNI) [2]. Thus, phylogenetic searches can be viewed as traversing the graph of phylogenetic trees induced by SPR adjacencies—the *SPR graph*.

It has become increasingly clear that the structure of the SPR graph plays an important role in determining the accuracy of tree searches. We will call these "graph effects"—one can't easily move from one tree to another due to the structure of the graph. Researchers have noticed SPR graph effects in MCMC [3]–[6]. We recently showed that graph structure has a significant effect on MCMC mixing applying a popular inference program to real data [7]. Probabilists have also approached the problem using related frameworks that are more amenable to proving theorems, both for other sets of moves on trees with a finite number of leaves [8], [9] or for SPR and related moves on a continuous tree-like object which formalizes the notion of a tree with infinite leaves [10], [11].

Phylogenetic maximum-likelihood (ML) inference is also impacted by graph structure, which can manifest itself as local maxima. There have been investigations of local maxima, though mainly from the perspective of branch lengths rather than topology [12]–[15]. The work on SPR moves in the ML context is primarily found in the literature comparing various SPR variants to one another, usually in the context of comparing one phylogenetic software package to another [16]–[19]. We are not aware of any general conclusions concerning hill-climbing on the SPR or related graphs that has come from this work.

Although the SPR graph is thus very important in determining the degree of success of phylogenetic inference procedures, still little is known about the rooted or unrooted versions of the SPR graph itself, or random walks thereupon. [20] developed a recursive procedure on a tree to find the degree of the corresponding vertex in the rooted SPR (rSPR) graph, and corresponding bounds on degree. [21] showed that the diameter Δ_{rSPR} of the rSPR graph is $n - \Theta(\sqrt{n})$, and for the unrooted case they show

$$n - 2\lceil\sqrt{n}\rceil + 1 \leq \Delta_{\text{uSPR}}(n) \leq n - 3 - \left\lfloor \frac{\sqrt{n-2}-1}{2} \right\rfloor.$$

We are not aware of any other work investigating large-scale properties of the SPR graph.

A new approach to quantifying properties of random walks on discrete and continuous spaces has recently been pioneered by Ollivier and colleagues [22], [23]. They define a notion of curvature which formalizes the notion of to what extent random walking brings points together, and is a generalization of the classical notion of Ricci curvature on manifolds. Here the term *random walk* on a space X simply denotes a family of probability measures parameterized by points of X satisfying a couple of reasonable assumptions, which includes biased walks such as MCMC.

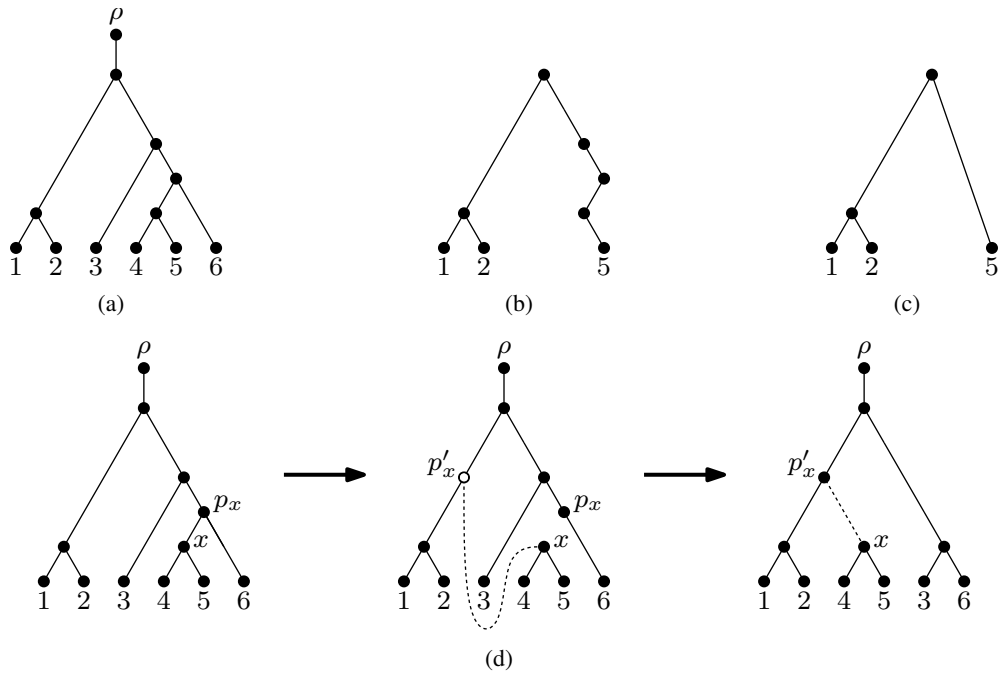


Fig. 1. (a) An X-tree T. (b) $T(V)$, where $V = \{1, 2, 5\}$. (c) $T|V$. (d) An rSPR operation transforms T into a new tree T' by *pruning* a *subtree* and *regrafting* it in another location.

In this paper, we investigate curvature of the rSPR graph with respect to two random walks and connect those results to access times (also known as hitting times) for those random walks. For this first effort we investigate random walks defined only in terms of the graph itself: the uniform random walk and MCMC sampling from the uniform prior on trees. We present a fast new algorithm for computing rSPR graphs from a set of trees, reducing the time to do so from $\mathrm{O}(m^2n)$ to $\mathrm{O}(mn^3)$ for a set of m trees with n leaves. As the full rSPR graph contains $(2n-3)!!$ trees, this is a significant improvement in practice for exploring large subsets of the graph. By exploiting symmetries in the rSPR graph, we were able to calculate all of the curvatures for pairs of trees with up to seven leaves. By carefully examining the overlap in rSPR moves, we present a new method for computing the degree of a tree in the rSPR graph that allows one to select an rSPR neighbor uniformly at random in linear-time without explicitly generating the graph. Using this method to simulate these random walks, we find that the distribution of access times between pairs of trees can be described by distance between the trees, the degrees of the trees, and the curvature. By getting a more fine-tuned understanding of the rSPR neighborhood of pairs of vertices, we are able to give bounds on curvatures under these random walks. In particular, we present a full characterization of the change in rSPR degree that occurs from a given rSPR move and find that even though they each count as one move, rSPR moves which move large subtrees may have greater graph effects than those that move small subtrees. Finally, we prove that the two related notions of Ricci-Ollivier curvature (coarse and asymptotic) are close, and in fact identical in the limit of many taxa, when applied to the rSPR graph. In summary, we extend knowledge about an important graph for phylogenetics, specifically in a way that models phylogenetic MCMC search.

The automated computational analysis code can be found at https://github.com/matsengrp/curvature.

II. Preliminaries

We follow the definitions and notation from [7], [25], [26]. A (rooted binary phylogenetic) X-tree is a rooted tree T whose nodes have zero or two children such that the leaves of T are bijectively labelled with the members of a label set X. As in [7], [25], [26], the tree is augmented with a labelled root node ρ and ρ is considered a member of X (Fig. 1(a)). We generally use n to refer

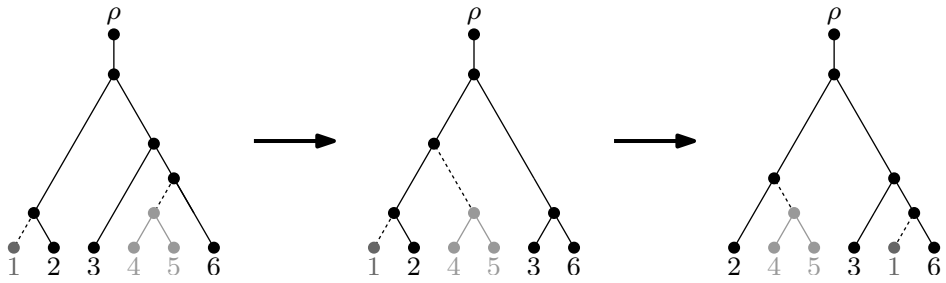


Fig. 2. Two rSPR operations, each of which moves one grey subtree. The leftmost and rightmost trees are rSPR distance two apart.

to the number of leaves in an X-tree. For a subset V of X, $T(V)$ is the smallest subtree of T that connects all nodes in V (Fig. 1(b)). The V-tree induced by T is the smallest tree $T|V$ that can be obtained from $T(V)$ by suppressing unlabelled nodes with fewer than two children (Fig. 1(c)). For the rest of the paper, **we will assume that all phylogenetic trees are binary and rooted**, and thus that tree inclusion is rooted tree inclusion.

A *parent (sub)tree* of a subtree U is the smallest subtree strictly containing U. A *parent edge* of a subtree U is the edge connecting U to the rest of the tree. The *internal edges* of a tree are the edges that do not contact a leaf or ρ. A *ladder tree* (also known as a *caterpillar tree*) is a tree such that every internal node has a leaf as a direct descendant. A *balanced tree* is a tree such that the sum of the depths of internal nodes is minimum over all trees with the same number of leaves. The *least common ancestor* (LCA) of a set R of two or more nodes is the unique node that is an ancestor of each node $r \in R$ and at maximum depth. Similarly, the LCA of two or more subtrees is the LCA of their parent nodes.

A (rooted) *subtree-prune-regraft* (rSPR) operation on an X-tree T cuts an edge $e = (x, p_x)$ where p_x denotes the parent of node x. T is divided into two subtrees T_x and T_{p_x} containing x and p_x, respectively. Then the operation adds a new node p'_x to T_{p_x} by subdividing an edge of T_{p_x} and adding a new edge (x, p'_x), making x a child of p'_x. Finally, p_x is suppressed, joining the two edges on either side of that node. See Figure 1(d) for an example. Note that the inclusion of ρ allows for rSPR moves that move subtrees to the root of the tree.

rSPR operations give rise to a distance measure between X-trees: $d_{\mathrm{SPR}}(T_1, T_2)$ is the minimum number of rSPR operations required to transform an X-tree T_1 into T_2. For example, the trees in Figure 2 are separated by two rSPR operations. Moreover, rSPR operations naturally give rise to a graph on the set of X-trees for which this distance is simply the shortest-path graph distance. Let $\mathcal{T}_n$ be the set of trees with n leaves and label set $X = \{1, 2, \ldots n, \rho\}$. Then the rSPR graph G of $\mathcal{T}_n$ is the graph with vertex set $V(G) = \mathcal{T}_n$ and edge set $E(G) = \{(T, S) \mid d_{\mathrm{SPR}}(T, S) = 1, T \in V, S \in V\}$.

To avoid confusion between the two types of graph structures considered here, we refer to vertices of the rSPR graph as *vertices* and vertices of individual trees (i.e. leaves and internal nodes) as *nodes*. Let $N(T)$ be the set of rSPR neighbors of a tree T (this does not include T). For example, the tree T with 4 leaves in Figure 3 has 10 neighbors. We say that the degree of T is $|N(T)|$, that is, the number of trees which can be obtained from T by a single rSPR operation. Because we assume that all trees are bifurcating, we use degree to refer only to the degree of rSPR graph vertices.

Ricci-Ollivier curvature provides a rigorous yet intuitive formalization of the shape of a metric space with respect to a random walk. For the purposes of this paper, we will specialize to that space being a graph equipped with the shortest-path distance, and we will describe curvature intuitively. For a more rigorous presentation in the more general setting of a Polish metric space, see [22] or the survey [27].

Let m_x and m_y be probability densities of the position of a specified random walk after a fixed period of time starting at points x and y of a graph $G = (V, E)$, respectively. The transportation distance [28] (equivalently Wasserstein distance, or "earth movers distance" [29]) between m_x and

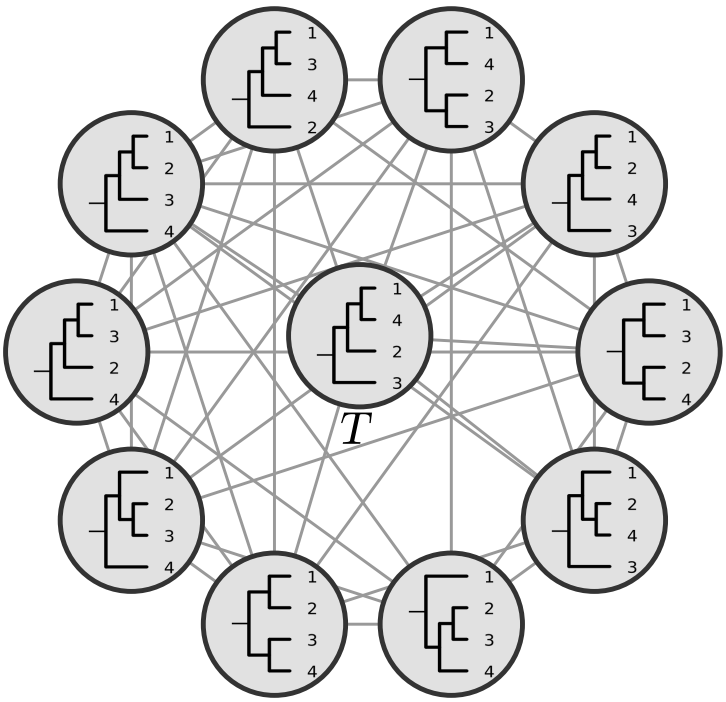


Fig. 3. The neighborhood of an X-tree T with 4 leaves, showing connections between neighbors.

m_y is the minimum amount of "work" required to move m_x to m_y along edges of the graph, that is

$$W_1(m_x, m_y) := \min_{\xi \in \Pi(m_x, m_y)} \sum_{\{z,w\} \subset V} d(z,w)\xi(z,w),$$

where $d(z,w)$ is the graph shortest-path distance ($d_{\mathrm{SPR}}(z,w)$ in our case) and $\Pi(m_x, m_y)$ is the set of mass transport programs moving the mass in m_x to the configuration described by m_y (more formally, a density on $V \times V$ which is m_x after projecting on the first component and m_y after projecting on the second).

The *coarse Ricci-Ollivier curvature* of x and y is then defined as:

$$\kappa(m; x, y) := 1 - \frac{W_1(m_x, m_y)}{d(x,y)} \tag{1}$$

For the purposes of this paper, we will use $\kappa(x,y)$ to denote the simple (uniform choice of neighbor) random walk, and use $\kappa(\mathrm{MH}; x, y)$ to indicate curvature with respect to the Metropolis-Hastings random walk sampling the uniform distribution (described in detail in Section III-C). Positive curvature implies that m_x and m_y are "closer" than x and y, zero curvature implies that they are neither closer nor farther, and negative curvature implies that m_x and m_y are on average "more distant" than x and y. Curvature thus provides an intuitive measure of the difficulty of moving between regions of the graph with the given random walk.

Lin et al. [30] defined a variant definition of curvature in terms of lazy random walks which Loisel and Romon [31] dubbed the *asymptotic Ricci-Olivier curvature*. The lazy random walk only travels according to m_x with probability p and otherwise stays put. Thus the lazy mass assignment m_x^p is the sum of $p\, m_x$ and a point mass of $1-p$ on x. We denote the coarse curvature of the p-lazy random walk between two vertices x and y with respect to a random walk m by $\kappa_p(m; x, y)$. For example, $\kappa_{1/4}(m; x, y)$ describes the curvature of the lazy random walk that follows the given random walk m with probability $1/4$ and remains stationary with probability $3/4$. The asymptotic Ricci-Ollivier curvature of x and y is then:

$$\mathrm{ric}(m; x, y) := \lim_{p \to 0} \frac{\kappa_p(m; x, y)}{p} \tag{2}$$

As above for κ, we use $\mathrm{ric}(x, y)$ as shorthand for $\mathrm{ric}(m; x, y)$ when m is the uniform lazy random walk, and $\mathrm{ric}(\mathrm{MH}; x, y)$ when m is the Metropolis-Hastings random walk sampling the uniform distribution (Section III-C). This definition of curvature is invariant of p for small enough p [31] and can be used to avoid parity problems on graphs where the uniform random walk is periodic without choosing a specific laziness parameter (e.g. Ollivier often considered $\kappa_{\frac{1}{2}}(x, y)$ for this purpose). As we prove in Lemma V.7 the two notions of rough and asymptotic curvature differ only by a small factor bounded by $\frac{2}{\max(|N(x)|,|N(y)|)}$ between adjacent vertices and are equal for nonadjacent vertices.

III. Access times of random walks on the rSPR graph can be understood using distance, degree, and curvature

A. Computing rSPR graphs

We computed the rSPR graph explicitly for trees on four to seven leaves. This required a fast new algorithm for rSPR graph construction, as previous algorithms required $\mathrm{O}(m^2 n)$ time, where m is the number of trees in the graph ($(2n-3)!!$ in this case) and n the number of leaves. Here we reduce that time to $\mathrm{O}(mn^3)$. In previous work [7], we constructed (unrooted) SPR graphs from subsets of m high probability trees sampled from phylogenetic posteriors to compare mixing and identify graph effects. Although the SPR distance (rooted and unrooted) is NP-hard to compute [25], [32], it is fixed-parameter tractable with respect to the distance in the rooted case [25]. In particular, one can determine in $\mathrm{O}(n)$-time whether two rooted phylogenetic trees are adjacent in the rSPR graph ($\mathrm{O}(n^2)$-time for unrooted trees) using the algorithms of Whidden et al. [7], [26], [33], [34]. We applied this method comparing each of the m trees pairwise to identify adjacencies, requiring a total of $\mathrm{O}(m^2 n)$-time ($\mathrm{O}(m^2 n^2)$-time in the unrooted case). However, applying this method to complete rSPR graphs, with the $\mathrm{O}(m^2)$ factor where $m = (2n-3)!!$, requires an enormous amount of computation.

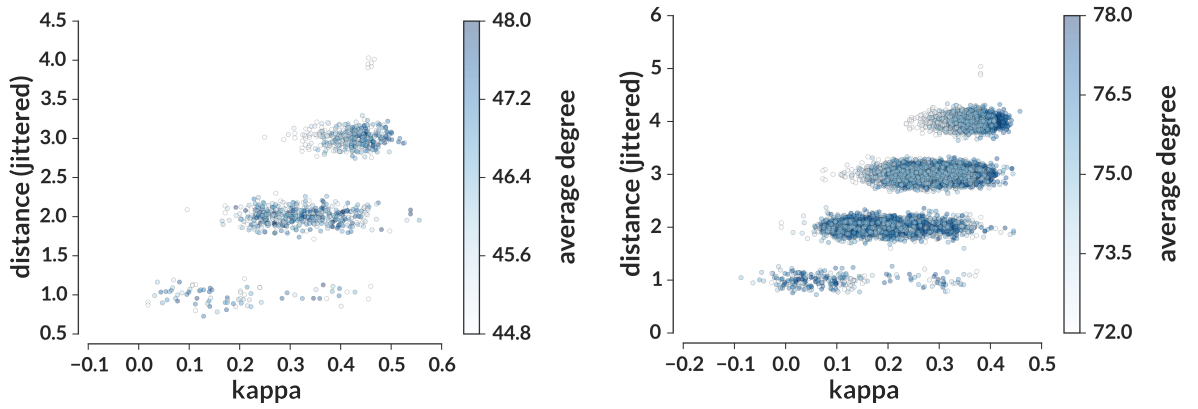
To quickly compute dense rSPR graphs (those containing a significant portion of the full rSPR graph) we avoid pairwise computations by successively adding trees to the graph and efficiently identifying their neighbors:

1. Let G be an empty graph.
2. For each of the m trees in a specified but arbitrary order:
   a) Add a vertex i to the graph representing the current tree T where i is T's order index.
   b) For each of the $\mathrm{O}(n^2)$ neighbors of T:
      i) If the current neighbor S's index is in G then add an edge (j, i) to the graph, where j is S's order index.

Theorem III.1. *The subgraph of the rSPR graph induced by a set $\mathcal{T}$ of m trees with n leaves can be constructed in $\mathrm{O}(mn^3)$-time.*

Proof. The correctness of the procedure follows by induction on the number of trees already processed, i, by observing that the procedure has constructed the subgraph of vertices $1, 2, \ldots i$ and will construct the subgraph of vertices $1, 2, \ldots i+1$.

We implement the graph with an adjacency list representation with integer-labelled vertices that supports $\mathrm{O}(\log n)$ edge insertions and lookups (with e.g. red-black trees [35], as the vertex degrees are $\mathrm{O}(n^2)$). As described above, the integer labels are simply the order of the input trees. Adding the vertices to the graph requires $\mathrm{O}(m)$-time, as they are added in ascending order to the end of the vertex list, which can be stored as a fixed-size array. Adding the $\mathrm{O}(mn^2)$ edges to the graph requires $\mathrm{O}(mn^2 \log n)$-time. Enumerating the neighbors of T requires $\mathrm{O}(n^3)$-time for each T, for a total of $\mathrm{O}(mn^3)$-time. We discuss below, in Section III-C how to do so efficiently without considering duplicate neighbors. We store the tree to index mappings for current vertices of G in a trie [36] using the $\mathrm{O}(n)$-length Newick [37] representation of the tree as a string. These representations are made unique by ordering the tree so that leftmost subtrees contain the smallest alphanumeric label of descendants. This requires only $\mathrm{O}(n)$-time for each tree (i.e. a total of $\mathrm{O}(mn^3)$-time) using a standard nodes-and-pointers representation of the tree and assuming integer leaf labels (a simple $\mathrm{O}(mn \log n)$ leaf preprocessing step could be applied to extend this procedure to phylogenetic trees with string



(a) 6 taxa (b) 7 taxa

Fig. 4. Scatter plot of $\kappa(\mathrm{MH}; T_1, T_2)$ values versus $d_{\mathrm{SPR}}(T_1, T_2)$ for the rSPR graph. Color displays the average degree of T_1 and T_2. Distance values randomly perturbed ("jittered") a small amount to avoid superimposed points.

labels). Similarly, it takes $\mathrm{O}(n)$-time to determine the index of each of the $\mathrm{O}(mn^2)$ S's. Therefore the graph can be constructed in $\mathrm{O}(mn^3)$-time. □

This procedure has been implemented in the C++ program `dense_spr_graph` of the software package `spr_neighbors` [38], which outputs an edge list format graph suitable for input to other software. The construction procedure reduced the time required to compute the 10,395-vertex 7-taxon rSPR graph from 2,104.68 seconds to 12.71 seconds on an Intel Core 2 Duo E7500 desktop running Ubuntu 14.04. Moreover, although we do not study the 135,135-vertex 8-taxon rSPR graph in this paper, our algorithm required only 303.45 seconds to construct it on the same hardware. Constructing the 8-taxon rSPR graph using the previous method required 377,395 seconds (more than 4 days), and thus that method is infeasible for constructing larger rSPR tree graphs. Thus, we believe our fast graph construction procedure will itself be useful for further studies of rSPR graph subsets similar to [7].

B. Computing curvature values

To compute curvature values, we first used `dense_spr_graph` to compute the rSPR graph for four to seven leaves, as discussed in Section III-A. We then computed curvatures for given pairs of trees directly, by using linear programming [31] to compute the minimal mass transport W_1 using the SAGE [39] front-end to the GLPK [40] solver; code can be found in [41] which grew from the code described in [31].

This would have required an enormous amount of computation to directly compute curvatures for the $((2n-3)!!)^2$ pairs of trees with n leaves, even for the small values of n we consider here. We instead exploited the fact that pairs of trees which are equivalent modulo label renumbering are symmetric in the rSPR graph and therefore guaranteed to have the same curvature. For example, the pairs $\{(((1,2),3),4),((1,2),(3,4))\}$ and $\{(((1,4),2),3),((1,4),(2,3))\}$ are the same after relabeling, so their curvatures are the same. We thus directly computed curvature values for one representative pair from each such equivalence class, or *tanglegram* [42]; the group-theoretic enumeration methods are described in a manuscript in preparation, and the SAGE [39] and GAP4 [43] code can be found online [44].

We find a wide variation in curvature values among tanglegrams (Figure 4). Curvature values tended to increase with increasing rSPR distance, and their variance decreased with increasing distance. Neighboring trees had minimum curvature values for a given number of leaves, and we found maximum curvature values between trees at maximum distance or one rSPR move closer than the maximum. This suggests that the increased difficulty of moving between trees with a random walk due to distance may be tempered somewhat by curvature in the highly connected rSPR graph.

Larger rSPR graphs tended to have smaller curvature values. Indeed, the 7-leaf rSPR graph contained adjacent pairs of trees with negative curvature. Such pairs are possible bottlenecks in phylogenetic searches which may be exacerbated by likelihood or branch length constraints.

C. Simulating random walks on the rSPR graph

We next explored the correspondence between curvature and graph effects. To do so, we first simulated Metropolis-Hastings random walks on the rSPR graphs with up to seven leaves. The uniform random walk moves from the current vertex to one of its neighbors uniformly at random, which makes this walk more likely to sample higher degree vertices. In contrast, the Metropolis Hastings (MH) random walk with constant likelihood function proposes a move from a tree T to a neighbor tree S uniformly at random and then accepts the move according to the Hastings ratio, $\min\left(1, \frac{|N(T)|}{|N(S)|}\right)$. The MH random walk is guaranteed to sample each tree uniformly at random and is therefore representative of a phylogenetic MCMC program sampling trees under a uniform prior.

To efficiently simulate the MH random walk, we developed a linear-time algorithm for proposing rSPR moves that does not require the rSPR graph to be explicitly built and stored in memory. A naïve approach would require $\mathrm{O}(n^3)$ time: $\mathrm{O}(n)$ time to generate each of the $\mathrm{O}(n^2)$ neighbors of a given tree so that one could be picked uniformly at random. To eliminate an $\mathrm{O}(n^2)$ factor, we developed a deterministic ordering of rSPR moves with a one-to-one correspondence to rSPR neighbors, as described in the next paragraph. Given such an order, a uniform neighbor can be selected by its index in $\mathrm{O}(n)$ time. We note that the recursive formula of Song [20] for the degree of a tree does not group rSPR moves that move a particular subtree, and thus would still require $\mathrm{O}(n^2)$ time to select a specific rSPR neighbor by index.

We consider the distribution of rSPR moves in terms of the number of nodes contained within a subtree. Recall that a tree with n leaves has $2n-1$ total nodes (ignoring the artificial ρ node). Given a subtree R with x nodes, observe that there are $2n-1-x$ possible locations to regraft R. However, some of these moves will result in the same neighboring tree as other rSPR moves. In particular, where we call the edge connecting the subtree rooted at that node to the rest of the tree the "node's edge", we have:

i) Moving R to its sibling edge results in the same tree, not a neighboring tree,
ii) Moving R to its parent edge results in the same tree,
iii) Moving R to its grandparent edge is the same as moving its aunt to its sibling edge, and
iv) Moving R to its aunt edge is the same as moving its aunt to R's edge.

We prove in Lemma III.2 that this list is exhaustive. We can thus assign $2n-1-x-4$ moves to any node at a depth greater than 1 (from the original non-ρ root) and $2n-1-x-2$ moves to depth 1 nodes (lacking both an aunt and grandparent). Call the portion of the neighborhood of a tree T that can be obtained by moving a subtree R rooted at a node u, and that is assigned to u, $N(T,u)$. In this manner, we achieve an alternative solution for computing the degree of a tree in the rSPR graph:

Lemma III.2. *For a tree T with n leaves,*

$$|N(T)| = \sum_{u \in T} |N(T,u)|,$$

for nodes u of T, where $N(T,u)$ is as defined above, and:

$$|N(T,u)| = \begin{cases} 2n-1-x-4 & \textit{if depth(u)} > 1, \\ 2n-1-x-2 & \textit{if depth(u)} = 1 \\ 0 & \textit{if depth(u)} \leq 0 \end{cases}.$$

Moreover, each tree of $N(T,u)$ can be obtained by an rSPR operation moving the subtree rooted at u.

Proof. The statement follows if each of the neighbor assignments are disjoint, that is $N(T,u) \cap N(T,v) = \emptyset$, for all nodes u, v of T. So, suppose, for the purpose of obtaining a contradiction, that there exist two nodes u and v of T such that there exists a tree $S \in (N(T,u) \cap N(T,v))$. Then S

can be obtained from T by moving the subtrees rooted at u or v. Call these U and V, respectively. This implies that both $T \setminus U = S \setminus U$ and $T \setminus V = S \setminus V$ by the definition of an rSPR operation. Then the rSPR moves that move U or V to obtain S must be nearest neighbor interchanges (NNIs), that is, rSPR moves which move their subtree to one of four locations: their grandparent edge, aunt edge, sibling's left child edge or sibling's right child edge. This implies that, without loss of generality, U is moved to its grandparent edge and V to $U's$ sibling (i.e. (iii) above) or U is moved to its aunt edge and V to U's edge (i.e. iv above), a contradiction. Therefore the claim holds. □

In particular, this formulation groups moves that move the same subtree. We can thus apply the following algorithm to select a neighbor uniformly at random for a tree T:

1. Compute the degree of T, $|N(T)|$.
2. Pick a random integer r in the range $[1, |N(T)|]$.
3. Label each node u of T by its preorder number and compute the number of nodes in the subtree rooted at each u.
4. For each tree node u and while $r > 0$:
   a) Decrease r by $|N(T, u)|$.
   b) If $r < 0$, let S be the $|r|$ member of $N(T, u)$ and terminate the for loop.
5. Return the neighbor S.

Lemma III.3. *An rSPR neighbor of a tree T can be chosen uniformly at random in $\mathrm{O}(n)$-time using $\mathrm{O}(n)$ space.*

Proof. We apply the above procedure. We use a standard nodes-and-pointers representation of the trees, which can be constructed in $\mathrm{O}(n)$-time from a Newick string representation and uses linear space in n. We can compute the degree of T in linear time and space using Lemma III.2. To efficiently compute $|N(T, u)|$ for each node u of T, we require the number of nodes x in the subtree rooted at u. We pre-compute these by (1) labeling each node with its preorder number in a preorder traversal and (2) summing the number of descendant nodes in a postorder traversal and storing the results in an array indexed by preorder number. Both of these traversals require $\mathrm{O}(n)$-time. There are $2n - 1 = \mathrm{O}(n)$ nodes of T, and $|N(T, u)|$ can be computed in constant time using the subtree sizes. Moreover, the tree S can be found in $\mathrm{O}(n)$-time by iterating over the edges of T that are not contained within u's subtree to select the corresponding rSPR destination. Finally, we require linear time to apply the chosen rSPR operation which entails removing a node, adding a node, and updating a constant number of pointers. Thus, the for loop requires linear time. By Lemma III.2 the chosen tree is an rSPR neighbor of T and is chosen uniformly at random. Therefore, the procedure uses linear time and space and selects an rSPR neighbor of T uniformly at random. □

Observe that this procedure can be easily adapted to explore the full neighborhood of a tree in $\mathrm{O}(n^3)$ time, which we use for Theorem III.1. We thus have the following corollary:

Corollary III.4. *The rSPR neighbors of a tree T can be enumerated in $\mathrm{O}(n^3)$-time.*

We implemented this procedure in the C++ package `random_spr_walk` [45]. We sampled a 200,000-iteration random walk on the 4-leaf rSPR graph and a 50,000-iteration random walk on the 5-leaf rSPR graph.

D. Access time simulation

The access time for a pair of vertices in a graph is the (random) number of iterations required to go from one of the vertices to the other in a random walk [46]; we were interested in the connection between curvature and access time. In previous work, we computed mean access times (MAT) between pairs of trees in the MCMC random walks: the mean number of iterations required to move from one tree to the other. We applied this work to demonstrate graph effects in uSPR subgraphs of real MCMC samples using MrBayes [7] using the `sprspace` [47] software.

Here, to gain more insight, we used simulation to approximate the entire access time distribution. As before, calculating these distributions for all pairs of trees would have required a tremendous

TABLE I
ORDINARY LEAST SQUARES LINEAR MULTIPLE REGRESSION OF RSPR MEAN ACCESS TIME AGAINST DEGREE AND DISTANCE.

(a) 5 taxa

	coefficient	p-value
T_1 degree	2.799	2.425e-07
T_2 degree	0.9424	0.04367
d_{rSPR}	8.915	5.026e-09

(b) 6 taxa

	coefficient	p-value
T_1 degree	12.15	2.726e-55
T_2 degree	6.629	4.302e-21
d_{rSPR}	41.19	1.104e-44

TABLE II
ORDINARY LEAST SQUARES LINEAR MULTIPLE REGRESSION OF RSPR δ_1 AGAINST DEGREE, DISTANCE, AND κ.

(a) 5 taxa

	coefficient	p-value
T_1 degree	634	9.376e-05
T_2 degree	-167.8	0.2366
d_{rSPR}	-2859	5.151e-06
κ (uniform prior MH)	-1.331e+04	4.462e-06

(b) 6 taxa

	coefficient	p-value
T_1 degree	9.427	2.944e-07
T_2 degree	-2.523	0.1432
d_{rSPR}	-34.19	0.0007557
κ (uniform prior MH)	-690.4	1.436e-22

amount of memory and computational power in order to get good estimates. Again we use the insight that the access time for a pair of trees with a simple random walk does not depend on the actual labeling of those trees, but rather only on their relative labeling. Thus rather than enumerate access times between trees, we enumerate times between pairs of trees in a tanglegram. To calculate the empirical distributions of access times we aggregate all access times for the same tanglegram using our group-theoretic methods [44].

We find that the mean access time between trees T_1 and T_2 is determined by $|N(T_1)|$ and $|N(T_2)|$ (Table I). Furthermore, plotting the distribution of access times between pairs of trees with respect to their distance and curvature hints that smaller κ slightly shifts the distribution of access times towards larger access times (Fig. 5(a)). We quantify this effect by defining δ_1 to be the difference between the first pair of access time counts such that the second entry in the pair is nonzero. For example, δ_1 for distance 1 pairs (green lines in Fig. 5) is the count for time 1 minus the count for time 2, while δ_1 for distance 3 pairs (blue lines in Fig. 5) is the count for time 2 minus the count for time 3. Regression finds a clear influence of κ on δ_1 (Table II). This confirms the intuitive interpretation of $\kappa(T_1, T_2)$ as quantifying the propensity of a random walk to go from T_1 to T_2 relatively directly, certainly before the random walk achieves stationarity. On the other hand, if the random walk starting from T_1 does not quickly arrive at T_2 and instead achieves stationarity, the original position of the random walk is forgotten, and the access time is then a standard exponentially distributed waiting time for an event in a Poisson process (Fig. 5(b)).

The analysis can be reproduced by invoking the SCons (http://scons.org/) build tool and running the cells in an IPython notebook; instructions are in the repository README file.

IV. ROOTED SPR NEIGHBORHOODS

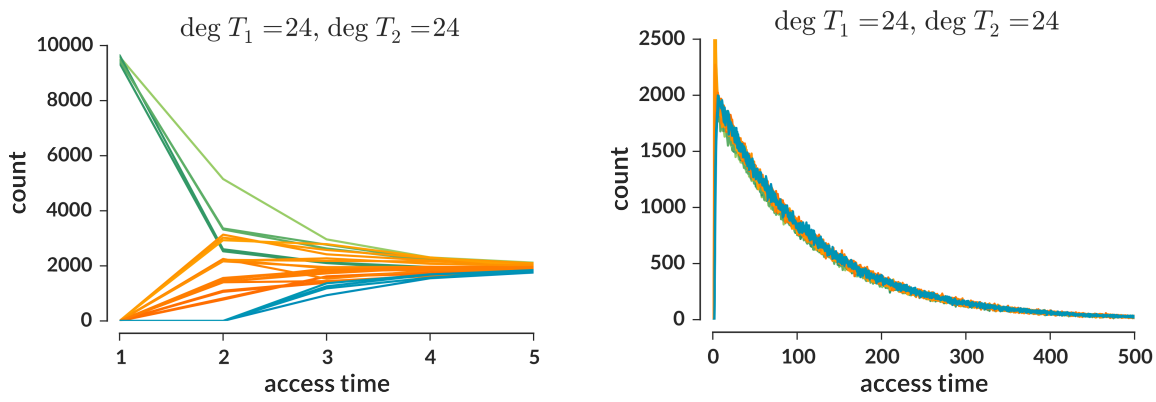
Having made the connection between curvature values and access times on the rSPR graphs, we now consider curvature theoretically. We begin by bounding differences between degrees, and then continue by enumerating squares and triangles in the rSPR graph.

Lemma IV.1 (Song [20])**.** *Let T be a rooted phylogenetic tree with n leaves. Then*

i) $|N(T)| = 3n^2 - 13n + 14$*, if T is a ladder tree,*
ii) $|N(T)| = 4(n-2)^2 - 2\sum_{m=1}^{n-2} \lfloor \log_2(m+1) \rfloor$*, if T is a balanced tree, and*
iii) $3n^2 - 13n + 14 \leq |N(T)| \leq 4(n-2)^2 - 2\sum_{m=1}^{n-2} \lfloor \log_2(m+1) \rfloor$*, otherwise.*

□

We now bound the ratio and difference of rSPR degree between two trees with n leaves.



(a) (b)

Fig. 5. Distribution of rSPR MH access times for those pairs of 5-taxon trees with degree 24 that are not simple inclusions of 4-taxon pairs of trees. Color signifies rSPR distance between the trees, with green, orange, and blue signifying distances of 1, 2, and 3, respectively; the saturation of the color shows coarse curvature $\kappa(\mathrm{MH}; \cdot, \cdot)$, such that increased saturation (i.e. darker color) indicates a smaller κ.

Lemma IV.2. *Let T and S be rooted phylogenetic trees with $n \geq 3$ leaves, and assume without loss of generality that $|N(T)| \leq |N(S)|$. Then:*

i) $\frac{|N(T)|}{|N(S)|} \geq 3/4$, *and*

ii) $|N(S)| - |N(T)| \leq n^2 - 5n + 6$.

Proof. To prove (i), we simply note from Lemma IV.1 that the ladder tree achieves the minimum degree, and the balanced tree achieves the maximum degree:

$$\begin{aligned}
\frac{|N(T)|}{|N(S)|} &\geq \frac{3n^2 - 13n + 14}{4(n-2)^2 - 2\sum_{m=1}^{n-2} \lfloor \log_2(m+1) \rfloor} \\
&\geq \frac{3n^2 - 13n + 12}{4(n-2)^2 - 2(n-2)} \\
&= \frac{3n^2 - 13n + 12}{4n^2 - 16n + 16 - 2(n-2)} \\
&= \frac{3n^2 - 13n + 12}{4n^2 - 18n + 20} \\
&\geq \frac{3n^2 - 13n + 12}{4n^2 - 17\frac{1}{3}n + 18} \qquad \forall n \geq 3
\end{aligned}$$

which is greater than 3/4 when $n \geq 3$. Similarly for (ii):

$$\begin{aligned}
|N(S)| - |N(T)| &\leq (4(n-2)^2 - 2\sum_{m=1}^{n-2} \lfloor \log_2(m+1) \rfloor) - (3n^2 - 13n + 14) \\
&\leq (4(n-2)^2 - 2(n-2)) - (3n^2 - 13n + 14) \\
&= 4n^2 - 16n + 16 - 2n + 4 - 3n^2 + 13n - 14 \\
&= n^2 - 5n + 6.
\end{aligned}$$

□

We can improve these bounds in the case of adjacent trees. To do so, we require the following lemma that characterizes exactly how the degree of a tree changes after an rSPR operation. See Figure 6 for an illustration of the Lemma.

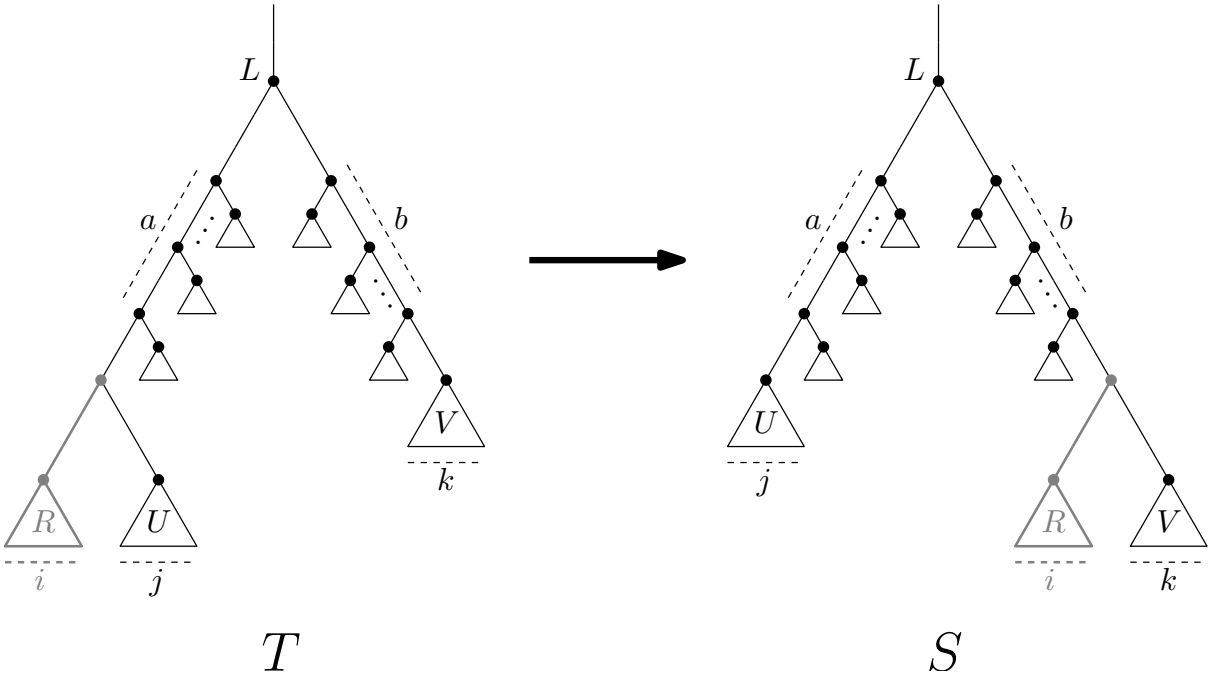


Fig. 6. An rSPR move labelled as in Lemma IV.3. Moving the grey subtree R from its position adjacent to U in tree T to its position adjacent to V in tree S changes the rSPR degree by $2\left(k(a-b)+i-j\right)$.

Lemma IV.3. *Let T and S be phylogenetic trees such that S can be obtained from T by moving a subtree R with k leaves from its position adjacent to subtree U to a location adjacent to subtree V. Let L be the LCA(U,V) in T. Let a be the number of intermediate nodes on the path from the parent of R to L in T, excluding endpoints. Similarly, let b be the number of intermediate nodes on the path from V to L in T, excluding endpoints. Let i be the number of leaves in U and j be the number of leaves in V, excluding any leaves of R. Then the degrees of T and S differ by:*

$$2\left(k(a-b)+i-j\right).$$

Proof. The set of permissible rSPR moves changes in four different ways due to the movement of R: (i) subtrees that include nodes on the path from U to L may now be moved into R and its newly introduced parent node, (ii) subtrees that include nodes on the path from V to L may no longer be moved into R and its parent node, (iii) R's parent subtree may now be moved into U, and (iv) R's parent subtree may no longer be moved into V. No additional moves are introduced or blocked by the original rSPR operation on R.

Recall that a rooted tree with k leaves has $2(k-1)$ internal edges(recall that we are excluding any "root edge" in these calculations). In the first case there are a subtrees that can now be moved onto the $2k$ edges in R (including its newly introduced parent edge and one of the newly subdivided root edges of V) for a total gain of $2ka$ distinct moves. Similarly, we lose $2kb$ moves in the second case. In the third case, R's parent subtree may now make $2(i-1)$ moves into U. Similarly, we lose $2(j-1)$ moves in the fourth case.

Thus the difference in rSPR degree is $2ka-2kb+2(i-1)-2(j-1)$ as claimed. □

Moreover, we can use these ideas to determine the number of rSPR moves that are, in some respects, independent of a given rSPR move. That is, for two trees S and T differing by a single rSPR move, we wish to know the number of rSPR moves that are applicable to both trees rather than unique to one of the trees. To formalize this concept, we consider pairs of trees $T' \in N(T)$ and $S' \in S(T)$ such that $d_{\mathrm{SPR}}(T',S')=1$. The number of such "squares" involving two adjacent trees will play a key role in our bounds on curvature, as they push the curvature of those trees towards 0.

Corollary IV.4. *Continuing with the setting and notation in Lemma IV.3, at least*

$$o := \deg(T) - 2kb - 2(j-1) = \deg(S) - 2ka - 2(i-1)$$

trees in the neighborhood of T can be paired with o trees in the neighborhood of S such that the pairings are disjoint and $d_{\mathrm{SPR}}(T', S') = 1$ for each (T', S') pair.

Proof. By the same arguments as in the proof of Lemma IV.3, o rSPR moves can be applied to T and S with the same source and target nodes. For each such (T', S') pair, we can move R in either tree to obtain the other member of the pair. □

We can now use Lemma IV.3 to improve the bounds in Lemma IV.2 for the case of two adjacent trees.

Lemma IV.5. *Let T and S be rooted phylogenetic trees with $n \geq 3$ leaves, such that $|N(T)| \leq |N(S)|$ and $d_{\mathrm{SPR}}(T, S) = 1$. Then:*

i) $|N(S)| - |N(T)| \leq 2\lfloor \frac{n-2}{2} \rfloor \lceil \frac{n-2}{2} \rceil \leq \frac{1}{2}(n-2)^2$,

ii) $\frac{|N(T)|}{|N(S)|} \geq \frac{5}{6}$, $\forall n \geq 4$, *and*

iii) $\lim_{n\to\infty} \frac{|N(T)|}{|N(S)|} = \frac{6}{7}$.

Proof. We first prove (i). By Lemma IV.3, $|N(S)| - |N(T)| = 2(k(a-b) + i - j)$. This value is maximized by making L the root and minimizing b, namely by setting $b = 0$. The resulting equation $2(ka + i - j)$ is similarly maximized by setting $i = 1$ (which allows us to increase a) then maximally balancing the terms in the product ka as follows.

There are two cases, depending on whether the subtree of k leaves is moved to the root or not. If not, then we set $j = 1$ and split the remaining $n - b - i - j = n - 2$ leaves between k and a in as balanced a way as possible, giving (i). Note that this corresponds to moving the bottom subtree of $\lfloor \frac{n-2}{2} \rfloor$ or $\lceil \frac{n-2}{2} \rceil$ leaves in a ladder tree to the root-most leaf of the tree.

If the subtree of k leaves is moved to the root, then we do not need to exclude the target branch from k and a, gaining an additional leaf to balance the product ka at the cost of increasing j. This corresponds to moving the bottom subtree of $\lfloor \frac{n}{2} \rfloor$ or $\lceil \frac{n}{2} \rceil$ leaves in a ladder tree to the root. Namely, we have $2(ka + 1 - j)$, where $j = n - k = a + 1$. If we move the additional leaf, we have:

$$|N(S)| - |N(T)| \leq 2\left(\left\lceil \frac{n}{2} \right\rceil \left\lfloor \frac{n-2}{2} \right\rfloor + 1 - \left(\left\lfloor \frac{n-2}{2} \right\rfloor + 1\right)\right) = 2\left\lfloor \frac{n-2}{2} \right\rfloor \left\lceil \frac{n-2}{2} \right\rceil$$

like before. Similarly, if we do not move the additional leaf, we also have:

$$|N(S)| - |N(T)| \leq 2\left(\left\lceil \frac{n-2}{2} \right\rceil \left\lfloor \frac{n}{2} \right\rfloor + 1 - \left(\left\lceil \frac{n-2}{2} \right\rceil + 1\right)\right) = 2\left\lfloor \frac{n-2}{2} \right\rfloor \left\lceil \frac{n-2}{2} \right\rceil,$$

proving (i).

The relative change in degree, $\frac{|N(T)|}{|N(S)|}$, can also be written as $\frac{|N(T)|}{|N(T)|+(|N(S)|-|N(T)|)}$. By (i), we have that $|N(S)| - |N(T)| \leq \frac{1}{2}(n-2)^2$, so $\frac{|N(T)|}{|N(S)|} \geq \frac{|N(T)|}{|N(T)|+\frac{1}{2}(n-2)^2}$. This bound is minimized when $|N(T)|$ is minimized, and recall by Lemma IV.1 that $|N(T)|$ is bounded below by $3n^2 - 13n + 14$. Thus

$$\begin{aligned}\frac{|N(T)|}{|N(S)|} &\geq \frac{3n^2 - 13n + 14}{3n^2 - 13n + 14 + \frac{1}{2}(n-2)^2} \\ &\geq \frac{3n^2 - 13n + 14}{3.5n^2 - 15n + 16}.\end{aligned}$$

Statements (ii) and (iii) follow from this bound.

□

Next, we bound the number of neighbors shared by two adjacent trees. The number of such "triangles" involving two adjacent trees has a key role in determining whether the curvature of two adjacent trees is positive or negative.

Lemma IV.6. *Let T and S be rooted X-trees such that $d_{\mathrm{SPR}}(T,S)=1$. Then $|N(T)\cap N(S)|\leq 6n-17$.*

Proof. T and S differ by one rSPR move that moves a subtree R. Pick a neighbor $U\in N(T)\cap N(S)$ of both T and S (this intersection is not empty: T and S are different, so R contains at most $n-2$ of the leaves, thus there must be at least one other tree U obtained by moving R in T and S). Then either (i) T and U differ in the location of R, or (ii) T and U differ in the location of another subtree Q. In the latter case, $T|(X\setminus L(Q))=S|(X\setminus L(Q))$ because T and S differ only in the location of R and $d_{\mathrm{SPR}}(T,U)=d_{\mathrm{SPR}}(S,U)=1$. Then leaves $r'\in R$, $q'\in Q$, and $u'\in U$, for some subtree U, form a triple of T and a different triple in S. This incompatible triple can be resolved in at most $6n-17$ ways, the maximum of which is reached when Q, U, and R are themselves a "triple" of subtrees. By Lemma III.2, each of the subtrees is assigned to at most $2n-6$ unique moves. Moreover, one additional overlapping move also moves one of the subtrees (that of the aunt of the LCA of the three subtrees). The number of shared neighbors is thus at most $3(2n-6)+1=6n-17$. Note that this bound is tight when, for example, T and S are ladders with a different configuration of 3 leaves at maximum depth. □

V. Curvature

For the purposes of this paper, κ is the coarse Ricci-Ollivier curvature (1) on the n-taxon rSPR graph with respect to a specified random walk and $W_{1,n}$ is the corresponding mass transport term from (1).

We first consider properties of the uniform (a.k.a. isotropic) random walk on the rSPR graph. Recall that the uniform random walk begins at a tree T and moves to a tree uniformly at random from $N(T)$. We denote the coarse curvature of the uniform random walk between two trees T and S with $\kappa(T,S)$. Recall that $\kappa(T,S):=1-\frac{W_1(m_T,m_S)}{d(T,S)}$. For the uniform random walk, m_T is simply the probability measure assigning a mass of $\frac{1}{|N(T)|}$ to each of T's neighbors. Our results follow from the lemmas of Section IV.

Proposition V.1. *Fix a positive integer k and let R be a tree with k leaves. Let $\{T_n \mid n>k\}$ be a sequence of phylogenetic trees all containing R, and let $\{S_n \mid n>k\}$ be the same sequence T_n but with R cut off and attached at a different location. Then $\lim_{n\to\infty}\kappa(T_n,S_n)=0$ for the uniform random walk on the rSPR graph.*

Proof. Because $d(T_n,S_n)=1$, we will prove the proposition by showing that the mass transport term $W_{1,n}$ sits between two bounds, each of which has limit 1 as n goes to infinity.

To start we demonstrate the proposition in the case that T_n and S_n have the same number of neighbors. First we claim that $W_{1,n}$ is bounded above by $(|N(T_n)|+\mathrm{O}(kn))/|N(T_n)|$ by exhibiting a mass transport program satisfying that bound. Let (T_n',S_n') be any of the o pairs of neighbors of (T_n,S_n) which are one rSPR move apart as per Corollary IV.4. We pair these trees in the mass transport. There are $\mathrm{O}(kn)$ trees unmatched by this pairing, and we can pair each of them arbitrarily with another tree of distance at most 3. Thus, $W_{1,n}$ is bounded above by $(|N(T_n)|+\mathrm{O}(kn))/|N(T_n)|$.

A lower bound is also available because we can't do better than distance 1 for all trees except for shared neighbors, of which there are $\mathrm{O}(n)$ by Lemma IV.6. By ignoring these trees we get a lower bound of $(|N(T_n)|-(\mathrm{O}(n)))/|N(T_n)|$ for $W_{1,n}$.

The desired control of $W_{1,n}$ is thus obtained because $|N(T_n)|$ is quadratic in n.

Now we prove the proposition when the number of neighbors differ. Assume without loss of generality that $|N(T_n)|<|N(S_n)|$. By Lemma IV.3, $|N(S_n)|-|N(T_n)|=2(k(a-b)+i-j)$, where each of $\{a,b,i,j\}$ is less than n. Thus, $|N(S_n)|-|N(T_n)|=\mathrm{O}(kn)$. We again pair neighbor T_n' of T with neighbor S_n' of S such that $d_{\mathrm{SPR}}(T_n',S_n')=1$ but, as $|N(T_n)|<|N(S_n)|$ we can only account for at most $|N(T_n)|/|N(S_n)|$ of the mass directly and may have to move the $(|N(S_n)|-|N(T_n)|)/|N(S_n)|$ remainder to trees a distance at most 3. Thus, $W_{1,n}$ is bounded above by $(|N(T_n)|+\mathrm{O}(kn))/|N(S_n)|=(|N(S_n)|+\mathrm{O}(kn))/|N(S_n)|$. We again bound $W_{1,n}$ from below with $(|N(T_n)|-\mathrm{O}(n))/|N(T_n)|$ by ignoring the mass in common neighbors of T_n and S_n. The proposition again follows because $|N(T_n)|$ is quadratic in n. □

Next we note a simple and rough bound on the curvature of two trees with respect to their distance, then obtain a tighter bound on the maximum curvature of two adjacent trees.

Lemma V.2. *Let T and S be two trees. Then:*

$$\frac{-2}{d_{\mathrm{SPR}}(T,S)} \leq \kappa(T,S) \leq \frac{2}{d_{\mathrm{SPR}}(T,S)}.$$

Proof. Observe that the distance between neighbors of T and S is bounded between $d_{\mathrm{SPR}}(T,S)-2$ and $d_{\mathrm{SPR}}(T,S)+2$. For the curvature upper bound, we then have $\kappa(T,S) \leq 1 - \frac{d_{\mathrm{SPR}}(T,S)-2}{d_{\mathrm{SPR}}(T,S)} = \frac{2}{d_{\mathrm{SPR}}(T,S)}$. The lower bound follows similarly. □

Lemma V.3. *The maximum curvature of the uniform random walk between two adjacent trees with n leaves is $\frac{6n-17}{3n^2-13n+14}$.*

Proof. The maximum curvature between adjacent trees T and S occurs when their neighborhoods have maximum overlap and all other tree pairs are at distance 1. By Lemma IV.6 the maximum overlap is $6n-17$. The amount of overlapping mass in the shared neighbors of T and S is thus $\frac{6n-17}{\max(|N(T)|,|N(S)|)}$. The minimum mass transfer cost is thus $1 - \frac{6n-17}{\max(|N(T)|,|N(S)|)}$. This is minimized when $|N(T)| = |N(S)|$ are as small as possible, that is T, S are ladders and $|N(T)| = 3n^2-13n+14$.

The maximum curvature is thus $1 - \frac{|N(T)|-(6n-17)}{|N(T)|} = \frac{6n-17}{|N(T)|} = \frac{6n-17}{3n^2-13n+14}$. □

This bound is tight and has been verified computationally for $n \leq 7$.

It is more difficult to obtain a closer bound on the maximum curvature of nonadjacent trees. Lemma V.2 suggests that more distant pairs of trees should have smaller curvatures than close trees as neighborhood effects decrease with respect to the increasing distance. However, our experiments with $n \leq 7$ suggest that maximum curvature tends to increase with distance (with respect to a fixed n), as a far greater fraction of the neighbors approach each other as the distance increases. Indeed, for $5 \leq n \leq 7$ the maximum curvature is obtained by pairs of trees at one less than the maximum distance. Moreover, nearly all of the neighbors of these pairs approach each either. We thus conjecture the following:

Conjecture V.4. *Let k_n be the maximum curvature for uniform random walks on trees of n leaves. Then:*

i) $k_n \leq \frac{2}{\Delta_{rSPR}(n)-1}$, *and*

ii) $\lim_{n\to\infty} k_n = \frac{2}{\Delta_{rSPR}(n)-1}$.

Proving or disproving this conjecture would go a long way toward understanding the effects of relative distance on curvature. However, we suspect that this will require a greater understanding of the distribution of tree neighborhoods with respect to one another than is currently known. Next, we bound the minimum curvature of two adjacent trees.

Lemma V.5. *The curvature of the uniform random walk between adjacent trees with n leaves is at least*

$$\frac{-n^2+2n}{3.5n^2-15n+16}.$$

Proof. In light of Corollary IV.4, the optimal mass transport cost is maximized (and therefore curvature minimized) across adjacent trees T and S by a combination of two effects: trees that cannot be paired at distance 1 and mass that must be moved between unpaired trees due to differing degrees of T and S. As we will show, these effects can be optimized simultaneously. To bound these effects, let m be the maximum (across T and S) proportion of mass that cannot be moved between adjacent neighbors of those trees. We can bound the mass transport cost from above by $1+2m$ because pairs of neighbors of adjacent trees are at most distance 3 apart. This gives a lower bound of $1-(1+2m)/1) = -2m$ on the curvature.

By Lemmas IV.3 and IV.5, the latter effect is maximized when the relative degree change is maximized. By Corollary IV.4, there are at most $o := |N(T)| - 2ka - 2(i-1)$ paired trees, bounding

the former effect. We now construct a pair of trees that maximizes both effects. Let S be the ladder tree with degree $3n^2 - 13n + 14$ and T be the adjacent tree constructed by moving the lower $\lfloor \frac{n}{2} \rfloor$ leaves of S to the root. T has degree at most $3.5n^2 - 15n + 16$. There are thus $2ka + 2(i-1) = 2\left(\lceil \frac{n-2}{2} \rceil \lfloor \frac{n}{2} \rfloor + (1-1)\right) \leq \frac{1}{2}n^2 - n$ unpaired neighbors, the maximum possible. Moreover, as shown by Lemma IV.3 this pair of trees obtains the maximum (absolute and relative) degree change. Thus, the maximum m is:

$$\frac{\frac{1}{2}n^2 - n}{3.5n^2 - 15n + 16}.$$

The claim follows from multiplying this value by -2. □

We further observe that the limit of our curvature lower bound is $-\frac{2}{7}$. Complete enumeration with $n \leq 7$ show that no pair of trees have curvature less than $-\frac{2}{5}$ and our bound meets or exceeds this value for $n > 7$. Moreover, the rSPR distance is a metric, so this bounds the curvature for arbitrary pairs of trees (Proposition 19 of [22]). This directly leads to the following Corollary:

Corollary V.6. *The curvature of the uniform random walk between two phylogenetic trees is at least* $-\frac{2}{5}$.

Note that this bound is not tight (at least for small n) as it is rarely necessary to transport mass the maximum distance between unpaired trees. We also note that the lower bounds in this section do not follow from the more general setting described in [**?**].

We next bound the difference between the coarse and asymptotic curvatures. Recall that $\kappa_p(T, S)$ is the coarse Ricci-Ollivier curvature between trees T and S with respect to the lazy walk that remains at a given tree with probability $1-p$ and transitions with probability p. For the lazy uniform random walk, m_T is now $T \cup N(T)$, with each neighbor assigned mass $\frac{p}{|N(T)|}$ and T assigned the remaining $1-p$ mass. The asymptotic Ricci-Ollivier curvature $\mathrm{ric}(T, S)$ is $\lim_{p\to 0} \kappa_p(T, S)/p$. As we now prove, these two notions of curvature differ only by a small factor inversely proportional to the maximum degree of T and S.

Lemma V.7. *Let T and S be rooted phylogenetic trees with n leaves. Then:*

i) $\mathrm{ric}(T, S) = \kappa(T, S)$, *if* $d_{\mathrm{SPR}}(T, S) > 1$,

ii) $\kappa(T, S) \leq \mathrm{ric}(T, S) \leq \kappa(T, S) + \frac{2}{\max(|N(T)|, |N(S)|)}$, *if* $d_{\mathrm{SPR}}(T, S) = 1$.

Proof. We first prove the lower bound in the uniform case, that is $\kappa(T, S) \leq \mathrm{ric}(T, S)$. Let $W_1(T, S)$ be the mass transport cost in the uniform case, and $W_1'(T, S)$ be the same for the lazy uniform case with parameter p. Recall that $\kappa(T, S) = \kappa_1(T, S) = 1 - \frac{W_1(T,S)}{d_{\mathrm{SPR}}(T,S)}$, and $\kappa_p(T, S)/p = \left(1 - \frac{W_1'(T,S)}{d_{\mathrm{SPR}}(T,S)}\right) \Big/ p$. Observe that

$$W_1'(T, S) \leq pW_1(T, S) + (1-p)\, d_{\mathrm{SPR}}(T, S),$$

by the simple mass transport program obtained by treating the mass at T and S as separate from that of the neighbors. Then:

$$\begin{aligned}
\kappa_p(T, S)/p &= \left(1 - \frac{W_1'(T, S)}{d_{\mathrm{SPR}}(T, S)}\right) \Big/ p \\
&\geq \left(1 - \frac{pW_1(T, S) + (1-p) d_{\mathrm{SPR}}(T, S)}{d_{\mathrm{SPR}}(T, S)}\right) \Big/ p \\
&= \frac{1}{p} - \frac{W_1(T, S)}{d_{\mathrm{SPR}}(T, S)} - \frac{1-p}{p} \\
&= 1 - \frac{W_1(T, S)}{d_{\mathrm{SPR}}(T, S)} \\
&= \kappa(T, S).
\end{aligned}$$

For the upper bound, we observe that

$$W_1'(T, S) \geq pW_1(T, S) + (1-p)\, d_{\mathrm{SPR}}(T, S) - \frac{2}{\max(|N(T)|, |N(S)|)},$$

as at most $1/\max(|N(T)|, |N(S)|)$ of the mass can remain at each of T and S, paired with the lazy remainder. The upper bound then follows analogously to the lower bound. Moreover, no mass can remain at T or S when $d_{\mathrm{SPR}}(T,S) > 1$, in which case the curvatures are equal. □

We now bound the difference between the curvature of the uniform random walk $\kappa(T,S)$ and that of the Metropolis-Hastings (MH) random walk $\kappa(\mathrm{MH};T,S)$. Recall that this random walk proposes a move from a tree T to a neighbor tree S uniformly at random and then accepts the move according to the Hastings ratio, which in this case is $\min\left(1, \frac{|N(T)|}{|N(S)|}\right)$. The mass distribution for the MH random walk thus leaves a portion of mass at the origin tree, proportional to the relative degree difference of its higher degree neighbors. Note that the same statement and proof of Lemma V.7 holds with $\kappa(T,S)$ and $\mathrm{ric}(T,S)$ replaced by the MH curvatures $\kappa(\mathrm{MH};T,S)$ and $\mathrm{ric}(\mathrm{MH};T,S)$, respectively.

Lemma V.8. *Let T and S be phylogenetic trees with n leaves. Then:*

$$\kappa(T,S) - \frac{1}{3d_{\mathrm{SPR}}(T,S)} \leq \kappa(\mathrm{MH};T,S) \leq \kappa(T,S) + \frac{1}{3d_{\mathrm{SPR}}(T,S)}, \text{ and .}$$

$$\kappa(T,S) - 1/6 \leq \kappa(\mathrm{MH};T,S) \leq \kappa(T,S) + 1/6$$

Proof. We first prove the lower bound. By Lemma IV.5, the quotient of degrees for two adjacent trees $\geq \frac{5}{6}$. Thus, the Hastings ratio is always $\geq \frac{5}{6}$. This implies that at most $\frac{1}{6}$ of the mass remains at tree T in the mass distribution. Let $W_1(T,S)$ be the cost of an optimal mass transport for the uniform random walk from T to S, and $W_1'(T,S)$ the cost for the MH random walk. Moreover, let $m_T(z)$ and $m_S(w)$ be the mass assigned for the uniform random walk and $m_T'(z)$ and $m_S'(w)$ be the mass assigned for the MH random walk, for each vertex $z \in N(T)$ and $w \in N(S)$. We construct an upper bound on $W_1'(T,S)$ by moving mass according to W_1 where possible, and moving the remainder either from T to S, from T to a neighbor of S, or from a neighbor of T to S. That is, for each W_1 assignment $\xi(z,w)$, we send $\xi'(z,w) = \xi(z,w)\min\left(\frac{m_T'(z)}{m_T(z)}, \frac{m_S'(w)}{m_S(w)}\right)$ of the mass from z to w. The remaining $\xi(z,w) - \xi'(z,w)$ of the mass is moved from T to S, T to w, and z to S in the respective proportions $\xi(z,w)\max\left(\frac{m_T'(z)}{m_T(z)}, \frac{m_S'(w)}{m_S(w)}\right) - \xi'(z,w)$, $\xi(z,w)\min\left(0, \frac{m_T'(z)}{m_T(z)} - \frac{m_S'(w)}{m_S(w)}\right)$, and, $\xi(z,w)\min\left(0, \frac{m_S'(w)}{m_S(w)} - \frac{m_T'(w)}{m_T(w)}\right)$. The maximum possible mass that is not moved according to W_1 is $\frac{1}{6}$. Moreover, the affected mass must be moved through at most two additional trees. Then, $W_1' \leq W_1 + \frac{2}{6}$. We now have:

$$\begin{aligned}\kappa(\mathrm{MH};T,S) &\geq 1 - \frac{W_1 + \frac{1}{3}}{d_{\mathrm{SPR}}(T,S)} \\ &\geq \kappa(T,S) - \frac{1}{3d_{\mathrm{SPR}}(T,S)}\end{aligned}$$

In the case that $d_{\mathrm{SPR}}(T,S) = 1$, the affected mass must be moved through only at most one additional tree, as T and S are adjacent. We thus obtain the lower bound of $\kappa(T,S) - \frac{1}{6}$ in this case.

We obtain the upper bounds similarly to the lower bounds, by observing that the affected at most $\frac{1}{6}$ of the mass may move through at most two fewer trees (i.e. directly between T and S rather than a pair of neighbors at distance $d_{\mathrm{SPR}}(T,S) + 2$ from each other). Again, this is at most one fewer tree when $d_{\mathrm{SPR}}(T,S) = 1$. □

VI. Conclusion and future work

In summary, we have advanced understanding of the phylogenetic rSPR graph substantially beyond what was previously known, which concerned graph diameter and vertex degree. We did so by developing the first theoretical and computational frameworks to bound and compute Ricci-Ollivier curvature of the rSPR graph. We found that curvature, along with degree and distance, determine the early dynamics of hitting times for random walks. Moreover, we proved that rSPR graph degree changes depend quadratically on the product of the size of the regrafted subtree with its change in depth, as well as that the rSPR graph tends toward flatness with respect to rSPR moves that move

asymptotically small subtrees. Finally, we proved that the coarse and asymptotic definitions of Ricci-Ollivier curvature are closely related with respect to uniform and Metropolis-Hastings walks on the rSPR graph.

In this data-free setting the stationary distribution is, unlike with real data, quite evenly spread over all trees. Correspondingly, we found that the influence of curvature is small in this case (Fig. 5(a)) and that the probability of the target node in the stationary distribution predominantly determines access times for pairs of trees (Fig. 5(b)). However, it is well known that MCMC takes a long time to approximate real phylogenetic posterior distributions even when the Bayesian credible set is small, and in fact our previous work showed significant graph effects in the mixing time for phylogenetic MCMC for credible sets that had tens, hundreds or thousands of trees [7]. Thus, our next step will be to investigate curvature of MCMC with nontrivial likelihood functions, which will significantly reduce the posterior distribution to a more realistic effective size, and in certain cases will lead to "bottlenecks" like those we have observed in real data. In those cases the curvature between two trees at either end of a bottleneck will describe how difficult it is to traverse the bottleneck. Indeed, in both the setting of challenging real data and in simulation with a nontrivial likelihood function, the stationary distribution takes a long time to achieve, and thus the curvature will have a substantially greater impact on the overall hitting times rather than being a short prelude to waiting for a Poisson process event as it is here (Fig. 5).

Now that we have established the foundations of using curvature to understand graphs relevant for phylogenetic inference, we will extend this work in several directions in addition to adding a nontrivial likelihood function as just described. We will next consider the case of unrooted SPR, which is more commonly used in phylogenetic search. We will also explore random walks on ranked trees [48] and graphical models of tree space relevant for phylogenetic algorithms such as BEAST [49] that infer rooted "time-trees." Random walks on other discrete structures such as partitions [50] that can be expressed as certain types of trees may also form interesting subjects for future work.

VII. Acknowledgements

The authors would like to thank Alex Gavruskin, Vladimir Minin, and Bianca Viray for helpful discussions. They are also grateful to the authors of the SAGE and GAP4 software, especially Alexander Hulpke. This work was funded by National Science Foundation award 1223057.

REFERENCES

[1] J. Felsenstein, "Evolutionary trees from DNA sequences: a maximum likelihood approach," *Journal of molecular evolution*, vol. 17, no. 6, pp. 368–376, 1981.

[2] D. F. Robinson, "Comparison of labeled trees with valency three," *Journal of Combinatorial Theory, Series B*, vol. 11, no. 2, pp. 105–119, 1971.

[3] E. Mossel and E. Vigoda, "Phylogenetic MCMC algorithms are misleading on mixtures of trees," *Science*, vol. 309, no. 5744, pp. 2207–2209, 30 Sep. 2005.

[4] ——, "Limitations of Markov chain Monte Carlo algorithms for bayesian inference of phylogeny," *Ann. Appl. Probab.*, vol. 16, no. 4, pp. 2215–2234, 1 Nov. 2006.

[5] F. Ronquist, B. Larget, J. P. Huelsenbeck, J. B. Kadane, D. Simon, and P. van der Mark, "Comment on "phylogenetic MCMC algorithms are misleading on mixtures of trees"," *Science*, vol. 312, no. 5772, p. 367; author reply 367, 21 Apr. 2006.

[6] D. Štefankovič and E. Vigoda, "Fast convergence of Markov chain Monte Carlo algorithms for phylogenetic reconstruction with homogeneous data on closely related species," *SIAM J. Discrete Math.*, vol. 25, no. 3, pp. 1194–1211, 2011.

[7] C. Whidden and F. A. Matsen IV, "Quantifying MCMC exploration of phylogenetic tree space," *Syst. Biol.*, 27 Jan. 2015. [Online]. Available: http://dx.doi.org/10.1093/sysbio/syv006

[8] D. J. Aldous, "Mixing time for a Markov chain on cladograms," *Comb. Probab. Comput.*, vol. 9, no. 03, pp. 191–204, 1 May 2000.

[9] P. Diaconis and S. Holmes, "Random walks on trees and matchings," *Electronic Journal of Probability*, vol. 7, no. 6, pp. 1–17, 2002.

[10] S. N. Evans and A. Winter, "Subtree prune and regraft: A reversible real tree-valued Markov process," *Ann. Probab.*, vol. 34, no. 3, pp. 918–961, May 2006.

[11] S. Athreya, W. Löhr, and A. Winter, "Invariance principle for variable speed random walks on trees," *arXiv preprint*, 24 Apr. 2014. [Online]. Available: http://arxiv.org/abs/1404.6290

[12] K. Fukami and Y. Tateno, "On the maximum likelihood method for estimating molecular trees: uniqueness of the likelihood point," *J. Mol. Evol.*, vol. 28, no. 5, pp. 460–464, May 1989.

[13] M. Steel, "The maximum likelihood point for a phylogenetic tree is not unique," *Syst. Biol.*, vol. 43, no. 4, pp. 560–564, 1 Dec. 1994.

[14] B. Chor, A. Khetan, and S. Snir, "Maximum likelihood on four taxa phylogenetic trees: Analytic solutions," in *Proceedings of the Seventh Annual International Conference on Research in Computational Molecular Biology*, ser. RECOMB '03. New York, NY, USA: ACM, 2003, pp. 76–83.

[15] B. Chor, M. D. Hendy, B. R. Holland, and D. Penny, "Multiple maxima of likelihood in phylogenetic trees: an analytic approach," *Mol. Biol. Evol.*, vol. 17, no. 10, pp. 1529–1541, Oct. 2000.

[16] W. Hordijk and O. Gascuel, "Improving the efficiency of SPR moves in phylogenetic tree search methods based on maximum likelihood," *Bioinformatics*, vol. 21, no. 24, pp. 4338–4347, 15 Dec. 2005.

[17] A. Stamatakis, "RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models," *Bioinformatics*, vol. 22, no. 21, pp. 2688–2690, 23 Aug. 2006.

[18] M. N. Price, P. S. Dehal, and A. P. Arkin, "FastTree 2–approximately maximum-likelihood trees for large alignments," *PLoS One*, vol. 5, no. 3, p. e9490, 2010.

[19] S. Guindon, J.-F. Dufayard, V. Lefort, M. Anisimova, W. Hordijk, and O. Gascuel, "New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0," *Syst. Biol.*, vol. 59, no. 3, pp. 307–321, May 2010.

[20] Y. S. Song, "On the combinatorics of rooted binary phylogenetic trees," *Ann. Comb.*, vol. 7, no. 3, pp. 365–379, 1 Dec. 2003.

[21] Y. Ding, S. Grünewald, and P. J. Humphries, "On agreement forests," *J. Combin. Theory Ser. A*, vol. 118, no. 7, pp. 2059–2065, Oct. 2011.

[22] Y. Ollivier, "Ricci curvature of Markov chains on metric spaces," *J. Funct. Anal.*, vol. 256, no. 3, pp. 810–864, 1 Feb. 2009.

[23] A. Joulin and Y. Ollivier, "Curvature, concentration and error estimates for Markov chain Monte Carlo," *Ann. Probab.*, vol. 38, no. 6, pp. 2418–2442, Nov. 2010.

[24] C. Whidden and F. A. Matsen IV, "Ricci-Ollivier curvature of the rooted phylogenetic subtree-prune-regraft graph," *arXiv preprint*, 4 2015. [Online]. Available: http://arxiv.org/abs/1504.00304

[25] M. Bordewich and C. Semple, "On the computational complexity of the rooted subtree prune and regraft distance," *Ann. Comb.*, vol. 8, no. 4, pp. 409–423, 2005.

[26] C. Whidden, R. G. Beiko, and N. Zeh, "Fixed-parameter algorithms for maximum agreement forests," *SIAM J. Comput.*, vol. 42, no. 4, pp. 1431–1466, 2013.

[27] Y. Ollivier, "A survey of Ricci curvature for metric spaces and Markov chains," *Probabilistic approach to geometry*, vol. 57, pp. 343–381, 2010.

[28] C. Villani, *Topics in Optimal Transportation*, ser. Graduate studies in mathematics. Providence: American Mathematical Society, 2003.

[29] Y. Rubner, C. Tomasi, and L. J. Guibas, "The earth mover's distance as a metric for image retrieval," *International journal of computer vision*, vol. 40, no. 2, pp. 99–121, 2000.

[30] Y. Lin, L. Lu, and S.-T. Yau, "Ricci curvature of graphs," *Tohoku Mathematical Journal*, vol. 63, no. 4, pp. 605–627, 2011.

[31] B. Loisel and P. Romon, "Ricci curvature on polyhedral surfaces via optimal transportation," *arXiv preprint*, 4 Feb. 2014. [Online]. Available: http://arxiv.org/abs/1402.0644
[32] G. Hickey, F. Dehne, A. Rau-Chaplin, and C. Blouin, "SPR distance computation for unrooted trees," *Evolutionary Bioinformatics*, vol. 4, pp. 17–27, 2008.
[33] C. Whidden and N. Zeh, "A unifying view on approximation and FPT of agreement forests," in *Proceedings of the 9th International Workshop, WABI 2009*, ser. Lecture Notes in Bioinformatics, vol. 5724. Springer-Verlag, 2009, pp. 390–401.
[34] C. Whidden, R. G. Beiko, and N. Zeh, "Fast FPT algorithms for computing rooted agreement forests: Theory and experiments," in *Experimental Algorithms*, ser. Lecture Notes in Computer Science, P. Festa, Ed. Springer Berlin Heidelberg, 2010, vol. 6049, pp. 141–153.
[35] L. J. Guibas and R. Sedgewick, "A dichromatic framework for balanced trees," in *Proceedings of the 19th Annual Symposium on Foundations of Computer Science*. IEEE Computer Society, 1978, pp. 8–21.
[36] E. Fredkin, "Trie memory," *Communications of the ACM*, vol. 3, no. 9, pp. 490–499, 1960.
[37] Wikipedia, "Newick format," 2015, [Online; accessed 30-March-2015]. [Online]. Available: http://en.wikipedia.org/wiki/Newick_format
[38] C. Whidden, "spr_neighbors," https://github.com/cwhidden/spr_neighbors, 2015, http://dx.doi.org/10.5281/zenodo.16543.
[39] W. Stein and D. Joyner, "SAGE: System for algebra and geometry experimentation," *ACM SIGSAM Bulletin*, vol. 39, no. 2, pp. 61–64, 2005, http://sagemath.org/.
[40] "GNU linear programming kit," http://www.gnu.org/software/glpk/glpk.html.
[41] F. A. Matsen IV, "gricci," https://github.com/matsengrp/gricci, 2015, http://dx.doi.org/10.5281/zenodo.16428.
[42] B. Venkatachalam, J. Apple, K. St John, and D. Gusfield, "Untangling tanglegrams: comparing trees by their drawings," *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 7, no. 4, pp. 588–597, Oct. 2010.
[43] *GAP – Groups, Algorithms, and Programming, Version 4.7.7*, The GAP Group, 2015, http://www.gap-system.org.
[44] F. A. Matsen IV, "tangle," https://github.com/matsengrp/tangle, 2015, http://dx.doi.org/10.5281/zenodo.16427.
[45] C. Whidden, "random_spr_walk," https://github.com/cwhidden/random_spr_walk, 2015, http://dx.doi.org/10.5281/zenodo.16541.
[46] L. Lovász, "Random walks on graphs: a survey," *Combinatorics, Paul Erdös is Eighty*, vol. 2, no. 1, pp. 1–46, 1993.
[47] C. Whidden, "sprspace," https://github.com/cwhidden/sprspace, 2015, http://dx.doi.org/10.5281/zenodo.16542.
[48] Y. S. Song, "Properties of subtree-prune-and-regraft operations on totally-ordered phylogenetic trees," *Ann. Comb.*, vol. 10, no. 1, pp. 147–163, 1 Jun. 2006.
[49] A. J. Drummond, M. A. Suchard, D. Xie, and A. Rambaut, "Bayesian phylogenetics with BEAUti and the BEAST 1.7," *Mol. Biol. Evol.*, vol. 29, no. 8, pp. 1969–1973, Aug. 2012.
[50] D. Gusfield, "Partition-distance: A problem and class of perfect graphs arising in clustering," *Inf. Process. Lett.*, vol. 82, no. 3, pp. 159–164, 16 May 2002.