A Tight Kernel for Computing the Tree Bisection and Reconnection Distance between Two Phylogenetic Trees

SIAM J. DISCRETE MATH.
Vol. 33, No. 3, pp. 1556–1574

A TIGHT KERNEL FOR COMPUTING THE TREE BISECTION AND RECONNECTION DISTANCE BETWEEN TWO PHYLOGENETIC TREES[*]

STEVEN KELK[†] AND SIMONE LINZ[‡]

Abstract. In 2001 Allen and Steel showed that, if subtree and chain reduction rules have been applied to two unrooted phylogenetic trees, the reduced trees will have at most $28k$ taxa where k is the tree bisection and reconnection distance between the two trees. Here we reanalyze Allen and Steel's kernelization algorithm and prove that the reduced instances will in fact have at most $15k-9$ taxa. Moreover we show, by describing a family of instances which have exactly $15k-9$ taxa after reduction, that this new bound is tight. These instances also have no common clusters, showing that a third commonly encountered reduction rule, the cluster reduction, cannot further reduce the size of the kernel in the worst case. To achieve these results we introduce and use "unrooted generators" which are analogues of rooted structures that have appeared earlier in the phylogenetic networks literature. Using similar arguments we show that, for the minimum hybridization problem on two rooted trees, $9k-2$ is a tight bound (when subtree and chain reduction rules have been applied) and $9k-4$ is a tight bound (when, additionally, the cluster reduction has been applied) on the number of taxa, where k is the hybridization number of the two trees.

Key words. fixed-parameter tractability, tree bisection and reconnection, generator, kernelization, phylogenetic network, phylogenetic tree, hybridization number

AMS subject classifications. 68W01, 68W40, 68R05, 92B10

DOI. 10.1137/18M122724X

1. Introduction.

In the study of evolution *phylogenetic trees* are often used to depict the evolution of a set of species (or more abstractly, *taxa*) X. These are trees in the usual graph-theoretical sense where the leaves are bijectively labeled by X and the internal vertices represent hypothetical (common) ancestors of X [25]. Phylogenetic trees are typically constructed from genetic markers, such as DNA alignments, and under many objective functions the goal of constructing the "best" phylogenetic tree is NP-hard [22]. This has stimulated the development of heuristics which explore the space of phylogenetic trees using topological rearrangement moves [11, 17]. One popular such move is the *tree bisection and reconnection* (TBR) move which, informally, deletes some edge of the tree and then reattaches the two induced components by a newly introduced edge. In attempting to understand the connectivity of phylogenetic tree space under the action of TBR moves, it is natural to ask the following: what is the smallest number of TBR moves required to transform one tree T into another tree T'? This is known as the TBR *distance* of the two trees, denoted $d_{\mathrm{TBR}}(T,T')$. It is NP-hard to compute [1, 15]. In 2001 Allen and Steel showed the following kernelization result: if *common pendant subtrees* and *common chains* in the two trees are repeatedly collapsed, then this preserves d_{TBR}, and upon termination the reduced trees will contain at most $28 \cdot d_{\mathrm{TBR}}(T,T')$ taxa [1]. This result was used to prove that the computation of d_{TBR} is fixed-parameter tractable (FPT); see [10] for an introduction to FPT.

[*] Received by the editors November 16, 2018; accepted for publication July 9, 2019; published electronically September 5, 2019.
https://doi.org/10.1137/18M122724X
Funding: The second author was supported by the New Zealand Marsden Fund.

[†] Department of Data Science and Knowledge Engineering (DKE), Maastricht University, The Netherlands (steven.kelk@maastrichtuniversity.nl).
[‡] Department of Computer Science, University of Auckland, New Zealand (s.linz@auckland.ac.nz).

Here we strengthen the bound given by Allen and Steel. We show that the reduced trees obtained from their kernelization algorithm in fact have at most $15 \cdot d_{\mathrm{TBR}} - 9$ taxa, and that this result is "best possible": for every $k \geq 2$, we demonstrate two trees with TBR distance k such that, after the kernelization procedure has terminated, they have *exactly* $15 \cdot d_{\mathrm{TBR}} - 9$ taxa. This proves that, if smaller kernels are to be obtained, additional reduction rules will be required. One natural candidate for a third reduction rule is the *common cluster* reduction rule [5]. However, using a slightly modified construction, we show that this rule (when applied in addition to the subtree and chain reduction rules) cannot improve upon the $15 \cdot d_{\mathrm{TBR}} - 9$ bound.

A novel feature of our proofs is that they leverage recent insights from the phylogenetic *networks* literature, where networks are essentially the generalization of phylogenetic trees to graphs [16]. Specifically, it was recently shown that if one embeds two trees T and T' into an unrooted phylogenetic network $N = (V, E)$, then the minimum value of $|E| - (|V| - 1)$ ranging over all such N will be equal to $d_{\mathrm{TBR}}(T, T')$ [29]. This is significant because it attaches a static, graph-based interpretation to the TBR distance: it allows us to view its computation as a graph/network-construction problem. In turn, this allows us to define and use unrooted analogues of *generators* (i.e., backbone topologies) [19, 26] which have been used extensively in the FPT literature on rooted phylogenetic networks (see, e.g., [27] and references therein). Once viewed this way, the strengthened $15 \cdot d_{\mathrm{TBR}} - 9$ bound can be derived via a fairly straightforward counting argument. The generators also turn out to be invaluable in proving the tightness of the bound.

As a spin-off to these results we show that the earlier-identified upper bound of $9k - 2$ [21] on the size of the standard hybridization number weighted kernel [7] for rooted trees is also tight, and that in this case the cluster reduction can only improve the bound slightly, to $9k - 4$, which as we demonstrate is also tight.

In the final part of the article we devote a discussion section to summarizing the (new) state of the algorithmic landsdcape for TBR distance and reflect upon the broader consequences of our strengthened bound on the size of the TBR kernel. We also list a number of related phylogenetics problems where there is still quite some potential for improving bounds on kernel sizes.

2. Preliminaries.

Unrooted phylogenetic trees and networks. Throughout this paper X denotes a finite set (of *taxa*) with at least two elements. An *unrooted binary phylogenetic network* on X is a simple, connected, and undirected graph N with $|X|$ vertices, called *leaves*, of degree one and bijectively labeled with X, and all other vertices of degree 3. We define the *reticulation number* of N as $r(N) = |E| - (|V| - 1)$, where E and V are the edge and vertex sets of N, respectively. If $r(N) = 0$, then N is called an *unrooted binary phylogenetic tree* on X.

Subtrees, chains, and clusters. Let N be an unrooted binary phylogenetic network on X. A *pendant subtree* of N is an unrooted binary phylogenetic tree on a proper subset of X that can be obtained from N by deleting a single edge. For $n \geq 1$, let $C = (\ell_1, \ell_2, \ldots, \ell_n)$ be a sequence of distinct leaves in X and, for each $i \in \{1, 2, \ldots, n\}$, let p_i denote the unique neighbor of ℓ_i in N. We call C an n-chain of N if there exists a path $p_1, p_2, \ldots, p_n$ in N such that $p_2, \ldots, p_{n-1}$ is a simple path. That is, we optionally allow that $p_1 = p_2$ (i.e., ℓ_1 and ℓ_2 have a common parent) and/or $p_{n-1} = p_n$ (i.e., ℓ_{n-1} and ℓ_n have a common parent). Furthermore, n is referred to as the *length* of C. By definition, note that each element in X is a chain of length 1 in N. Last, for $Y \subset X$, we say that Y is a *cluster* of N if there exists a single edge in N whose deletion disconnects N into two parts such that the leaves of one part are bijectively labeled by elements in Y while the leaves of the other part are bijectively labeled by elements in $X - Y$. If $|Y|$=1, then the cluster is called *trivial* and, otherwise, it is called *nontrivial.* Note that, if Y is a cluster of N, then $X - Y$ is also a cluster of N. We say that $Y, X - Y$ is a *bipartition* of N.

Generators. Rooted generators have played an important role in establishing kernelization results for problems on rooted trees [21, 28, 30], but have only received very little attention [14] as a tool to tackle problems on unrooted trees. Here we give a definition of an unrooted generator that we will subsequently use to establish an improved kernel for the problem of computing the TBR distance (formally defined below) between two trees. Let k be a positive integer. For $k \geq 2$, a *k-generator* (or short *generator* when k is clear from the context) is a connected cubic multigraph with edge set E and vertex set V such that $k = |E| - (|V| - 1)$. Furthermore, for $k = 1$, we define the graph that consists of a single vertex u and a loop edge $\{u, u\}$ to be the unique 1-*generator.* The edges of a generator are also called its *sides.* Intuitively, the sides are the places where leaves can be attached in order to obtain an unrooted binary phylogenetic network from a generator. We now formalize this concept. Let G be a k-generator, let $\{u, v\}$ be a side of G, and let Y be a set of leaves. The operation of subdividing $\{u, v\}$ with $|Y|$ new vertices and, for each such new vertex w, adding a new edge $\{w, \ell\}$, where $\ell \in Y$ such that Y bijectively labels the new leaves, is referred to as *attaching* Y to $\{u, v\}$. Additionally, if G is the 1-generator, then the degree-2 vertex u is suppressed after attaching Y to $\{u, u\}$. Moreover, if at least two new leaves are attached to G in a way that at least one new leaf is attached to each loop and to each pair of parallel edges, then the resulting graph is an unrooted binary phylogenetic network N with $r(N) = k$. Note that N has no pendant subtree with more than a single leaf. Conversely, we obtain G from N by deleting all leaves and, repeatedly, suppressing any resulting degree-2 vertices. We say that G *underlies* N. In summary, we make the following observation.

Observation 1. Let N be an unrooted binary phylogenetic network with $r(N) = k \geq 2$ and with no pendant subtree of size at least two. Then the graph G that is obtained from N by deleting all leaves and, repeatedly, suppressing any resulting degree-2 vertices is a k-generator. Conversely, we obtain N from G by attaching to each side $s = \{u, v\}$ of G a (possibly empty) set of leaves.

Tree bisection and reconnection. Let T be an unrooted binary phylogenetic tree on X. Apply the following three-step operation to T:

1. Delete an edge in T and suppress any resulting degree-2 vertex so that two new unrooted binary phylogenetic trees, say T_1 and T_2, are obtained.
2. If T_1 (resp., T_2) has at least one edge, subdivide an edge in T_1 (resp., T_2) with a new vertex v_1 (resp., v_2) and otherwise set v_1 (resp., v_2) to be the single isolated vertex of T_1 (resp., T_2).
3. Add a new edge $\{v_1, v_2\}$ to obtain a new unrooted binary phylogenetic tree T' on X.

We say that T' has been obtained from T by a single *TBR* operation. Furthermore, we define the TBR distance between two unrooted binary phylogenetic trees T and T' on X, denoted by $d_{\text{TBR}}(T, T')$, to be the minimum number of TBR operations that is required to transform T into T'. It is well known that, for any such pair of trees, one can always obtain one from the other by a sequence of TBR operations. In particular, d_{TBR} is a metric [1]. By building on an earlier result by Hein et al. [15, Theorem 8], Allen and Steel [1] established NP-hardness of computing the TBR distance.

Unrooted minimum hybridization. In [29], it was shown that computing the TBR distance for a pair of unrooted binary phylogenetic trees T and T' is equivalent to a problem that is concerned with computing the minimum number of extra edges required to simultaneously explain T and T'. To describe this problem precisely, let N be an unrooted binary phylogenetic network on X, and let T be an unrooted binary phylogenetic tree on X. We say that N *displays* T if T can be obtained from a subtree of N by suppressing degree-2 vertices. Furthermore, for two unrooted binary phylogenetic trees T and T' on X, we set

$$h^u(T,T') = \min_N\{r(N)\},$$

where the minimum is taken over all unrooted binary phylogenetic networks on X that display T and T'. The value $h^u(T,T')$ is known as the *hybridization number* of T and T' [29].

Unrooted-Hybridization-Number (UHN)
Input. Two unrooted binary phylogenetic trees T and T' on X.
Output. An unrooted binary phylogenetic network N on X that displays T and T' and such that $r(N) = h^u(T,T')$.

We are now in a position to formally state the aforementioned equivalence that was established in [29, Theorem 3].

Theorem 1. *Let T and T' be two unrooted binary phylogenetic trees on X. Then*

$$d_{\mathrm{TBR}}(T,T') = h^u(T,T').$$

Reductions and kernelization. While computing the TBR distance is NP-hard, it was also shown in [1] that the problem is FPT when parameterized by d_{TBR}. For two unrooted binary phylogenetic trees T and T' on X, the authors used the following two reductions to kernelize the problem.

Subtree reduction. Replace a maximal pendant subtree with at least two leaves that are common to T and T' by a single leaf with a new label.

Chain reduction. Replace a maximal n-chain with $n \geq 4$ that is common to T and T' by a 3-chain with three new leaf labels correctly oriented to preserve the direction of the chain. For an illustration of this reduction, see Figure 1.

If T and T' cannot be reduced under the subtree (resp., chain) reduction, we say that T and T' are *subtree* (resp., *chain*) *reduced*.

The next theorem, which is due to [1, Theorem 3.4], shows that both reductions are safe, i.e., they do not change the TBR distance.

Theorem 2. *Let T and T' be two unrooted binary phylogenetic trees on X, and let S and S' be two trees obtained from T' and T', respectively, by applying a single subtree or chain reduction. Then $d_{\mathrm{TBR}}(T,T') = d_{\mathrm{TBR}}(S,S')$.*

Repeated applications of the previous lemma allow us to obtain two trees, say S and S' on X', from T and T', respectively, that are subtree and chain reduced such that $d_{\mathrm{TBR}}(T,T') = d_{\mathrm{TBR}}(S,S')$. The importance of S and S' lies in the following kernelization result that we have alluded to in the introduction and that is a direct consequence of [1, Lemmas 3.6 and 3.7].

Theorem 3. *Let T and T' be two unrooted binary phylogenetic trees on X, and let S and S' be a subtree and chain reduced tree pair on X' that has been obtained*

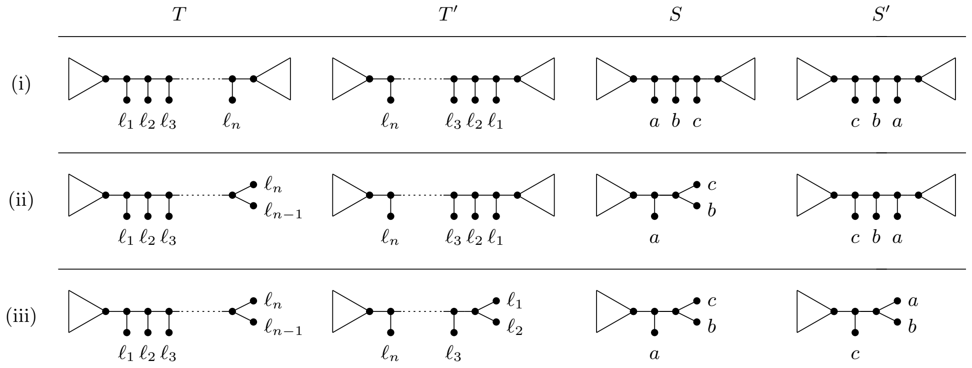

FIG. 1. *A chain reduction applied to the common n-chain $C = (\ell_1, \ell_2, \ldots, \ell_n)$ with $n \geq 4$ of two unrooted binary phylogenetic trees T and T'. In the reduced trees S and S', C is replaced with the 3-chain (a, b, c). Triangles indicate subtrees of T, T', S, and S'. As shown in* (ii) *and* (iii)*, C may be pendant in one or both of T and T' and, hence, ℓ_1 and ℓ_2 have the same parent and/or ℓ_{n-1} and ℓ_n have the same parent. Note that, if T and T' both have the property that ℓ_1 and ℓ_2 have the same parent, then C is in fact a common pendant subtree and reduced under the subtree reduction.*

from T and T' by repeated applications of the subtree and chain reduction. Then $|X'| \leq 28 d_{\mathrm{TBR}}(T, T')$.

Noting that the subtree and chain reduction can be applied in time that is polynomial in the size of X, it immediately follows that computing the TBR distance is FPT when parameterized by this minimum distance.

Cluster reduction. In [5] it is shown that if T and T' have a common nontrivial cluster, computation of $d_{\mathrm{TBR}}(T, T')$ can be reduced to the computation of d_{TBR} on two new smaller pairs of trees obtained by decomposing T and T' around the common cluster. (Compared to the subtree and chain reductions the cluster reduction involves a number of subtleties; for brevity, we do not describe them here.) This decomposition procedure can be continued until trees are obtained without nontrivial common clusters. If T and T' do not have a nontrivial common cluster, we say that T and T' are *cluster reduced.*

Maximum parsimony distance. A *character* f on X is a function $f : X \to C$, where $C = \{c_1, c_2, \ldots, c_r\}$ is a set of *character states* for some positive integer r. Let T be an unrooted phylogenetic tree on X with vertex set V, and let f be a character on X whose set of character states is C. An *extension* g of f to V is a function $g : V \to C$ such that $g(\ell) = f(\ell)$ for each $\ell \in X$. Given an extension g of f, let $l_g(T)$ denote the number of edges $\{u, v\}$ in T such that $g(u) \neq g(v)$. Then the *parsimony score* of f on T, denoted by $l_f(T)$, is obtained by minimizing $l_g(T)$ over all possible extensions g of f. Last, for two unrooted phylogenetic trees T and T' on X, we define the *maximum parsimony distance* d_{MP} as

$$d_{\mathrm{MP}}(T, T') = \max_f |l_f(T) - l_f(T')|.$$

Importantly, for two unrooted binary phylogenetic trees, the maximum parsimony distance is a lower bound on the TBR distance as noted in the discussion of [12]. We will freely use this fact throughout the rest of this paper. For more details on the maximum parsimony distance, we refer the interested reader to [12, 18, 23].

3. Tight kernels for computing the TBR distance.

In this section, we use generators to reanalyze the TBR distance kernelization result by Allen and Steel [1] and establish a kernel for the problem whose size is significantly smaller. Let T and T' be two unrooted binary phylogenetic trees on X, and let N be an unrooted phylogenetic network on X that displays T and T'. Furthermore, let $C = (\ell_1, \ell_2, \ldots, \ell_n)$ be an n-chain of N, and for each $i \in \{1, 2, \ldots, n\}$, let p_i be the unique neighbor of ℓ_i in N. Since each embedding of T (resp., T') into N uses either all or all but one edge in $\{\{p_1, p_2\}, \{p_2, p_3\}, \ldots, \{p_{n-1}, p_n\}\}$, it is straightforward to check that exactly one of the following three cases holds.

1. C is a chain of T and T'
2. There exists a *breakpoint* $i \in \{1, 2, \ldots, n-1\}$ such that
$$C_1 = (\ell_1, \ell_2, \ldots, \ell_i) \text{ and } C_2 = (\ell_{i+1}, \ell_{i+2}, \ldots, \ell_n)$$
are chains of T and C is a chain of T', or C is a chain of T and C_1 and C_2 are chains of T'.
3. There exist two not necessarily distinct breakpoints $i, j \in \{1, 2, \ldots, n-1\}$ such that
$$(\ell_1, \ell_2, \ldots, \ell_i) \text{ and } (\ell_{i+1}, \ell_{i+2}, \ldots, \ell_n)$$
are chains of T and
$$(\ell_1, \ell_2, \ldots, \ell_j) \text{ and } (\ell_{j+1}, \ell_{j+2}, \ldots, \ell_n)$$
are chains of T'.

We say that C has 0, 1, or 2 breakpoints relative to T and T' depending on whether C is not cut (Case (1)), cut once (Case (2)), or cut twice (Case (3)). Intuitively, the number of breakpoints indicates how many trees of T and T' do not have C as a chain.

LEMMA 1. *Let N be an unrooted binary phylogenetic network with $r(N) = k \geq 2$ and with no pendant subtree of size at least two. Furthermore, let G be the generator that underlies N. Then G has $3(k-1)$ sides.*

Proof. Let E be the edge set of G, and let V be the vertex set of G. By construction of G from N, recall that each vertex of G has degree 3 and that $|E|-|V|+1 = k$. Hence,
$$2|E| = 3|V| = 3(|E| - k + 1),$$
where the first equality is due to the handshaking lemma. Now, solving for $|E|$, we have $|E| = 3(k-1)$. Since E is equal to the set of sides of G, the lemma follows. □

LEMMA 2. *Let S and S' be two unrooted binary phylogenetic trees X, and let N be an unrooted phylogenetic network on X that displays S and S'. Furthermore, let C be an n-chain of N. Depending on the number of breakpoints of C relative to S and S', n is bounded from above in the following way.*

1. *Suppose that S and S' are subtree and chain reduced. Then $n \leq 3$ if C has 0 breakpoints, $n \leq 6$ if C has 1 breakpoint, and $n \leq 9$ if C has 2 breakpoints.*
2. *Suppose that S and S' are subtree, chain, and cluster reduced. Then $n \leq 3$ if C has 0 breakpoints, $n \leq 6$ if C has 1 breakpoint, and $n \leq 7$ if C has 2 breakpoints.*

Proof. We establish the lemma for when C has two breakpoints. The other cases are similar and omitted. Let $C = (\ell_1, \ell_2, \ldots, \ell_n)$, and let i and j be the two breakpoints of C relative to S and S'. Without loss of generality, we may assume that $i \leq j$. Since C has two breakpoints, one of the following two cases applies.

(a) If $i = j$, then the chains $(\ell_1, \ell_2, \dots, \ell_i)$ and $(\ell_{i+1}, \ell_{i+2}, \dots, \ell_n)$ are common to S and S'.

(b) If $i < j$, then the chains $(\ell_1, \ell_2, \dots, \ell_i), (\ell_{i+1}, \ell_{i+2}, \dots, \ell_j), (\ell_{j+1}, \ell_{j+2}, \dots, \ell_n)$ are common to S and S'.

First, suppose that S and S' are subtree and chain reduced and, towards a contradiction, assume that $n > 9$. Regardless of whether (a) or (b) applies, it follows by the pigeonhole principle that one of the smaller chains has length at least 4 and is common to S and S', thereby contradicting that S and S' are chain reduced. Second, suppose that S and S' are subtree, chain, and cluster reduced and, towards a contradiction, assume that $n > 7$. If $i < j$, observe that $\{\ell_{i+1}, \ell_{i+2}, \dots, \ell_j\}$ is the leaf set of a pendant subtree in S and S'. Hence $\{\ell_{i+1}, \ell_{i+2}, \dots, \ell_j\}$ is a cluster that is common to S and S'. Since S and S' are cluster reduced, this implies that $i + 1 = j$. Now, regardless of whether (a) or (b) applies, it follows again by the pigeonhole principle that one of the other chains that is common to S and S' has length at least 4; a contradiction. □

Let T and T' be two unrooted binary phylogenetic trees. Using generators, the next lemma exploits the equivalence between computing the TBR distance and UHN to establish a new and improved upper bound on the number of leaves of a pair of subtree and chain reduced trees.

Lemma 3. *Let S and S' be two unrooted binary phylogenetic trees on X that are subtree and chain reduced, and $d_{\mathrm{TBR}}(S, S') \geq 2$. Then, $|X| \leq 15d_{\mathrm{TBR}}(S, S') - 9$.*

Proof. Let N be an unrooted binary phylogenetic network on X with edge set E and vertex set V that displays S and S' such that

$$r(N) = h^u(S, S') = d_{\mathrm{TBR}}(S, S') = k \geq 2.$$

By Theorem 1, such a network exists. Let G be the generator that underlies N. Furthermore, let D and D' be two subdivisions of S and S', respectively, in N. Since N displays S and S', such subdivisions exist. If D is a spanning tree of N, then set $B = D$ and, otherwise, set B to be a spanning tree of N obtained from D by greedily adding edges. Similarly, if D' is a spanning tree of N, then set $B' = D'$ and, otherwise, set B' to be a spanning tree of N obtained from D' by greedily adding edges. (Observe that B and B' may have unlabeled leaves.) Moreover, as $|E| = |V| - 1 + k$, observe that each of B and B' has $|V| - 1 = |E| - k$ edges, where the left-hand side of the equation follows from the definition of a spanning tree.

Now, as S and S' are subtree reduced, it is sufficient to attach leaves to the sides of G in the process of obtaining N from G (see Observation 1). Let $s = \{u, v\}$ be a side of G. We next assign a *cut count* c_s to s. First, if s is decorated with at least one leaf in obtaining N from G, let $Y = \{\ell_1, \ell_2, \dots, \ell_n\}$ be the set of leaves that are attached to s. By construction, N has a maximal n-chain C whose leaves are bijectively labeled with the elements in Y. Let P be the path from u to v in N that has length $n+1$ and whose vertices (except for u and v) are neighbors of the elements in Y. We define c_s to be the number of trees in $\{B, B'\}$ that do not use all edges of P. Since B and B' both span N, note that there is at most one edge that is not used by B and at most one (not necessarily distinct) edge that is not used by B'. Hence, relative to C, we have $b_C \leq c_s$, where b_C is the number of breakpoints of C relative to S and S'. It follows from Lemma 2.1 that, if $c_s = 0$, then at most 3 taxa can be attached to s in obtaining N from G. Similarly, if $c_s = 1$, then at most 6 taxa can be attached to s and, if $c_s = 2$, then at most 9 taxa can be attached to s in this process.

Second, if s is not decorated with any leaf when obtaining N from G, then $\{u, v\}$ is an edge in N, and we define c_s to be the number of trees in $\{B, B'\}$ that do not use $\{u, v\}$. Since we are establishing an upper bound on $|X|$, we may assume for the remainder of this proof that the bounds established in Lemma 2 also apply for when s is not decorated. By assigning a cut count to each side of G, and recalling that both B and B' contain $|E| - k$ edges, it is easily checked that

$$\sum_s c_s = 2k,$$

where the sum is taken over all sides of G.

Now, let $0 \leq q \leq k$ be the number of sides of G whose cut count is equal to two. Consequently, by leveraging the above equality, there are $2(k - q)$ sides whose cut count is equal to 1 and, by Lemma 1, $3(k - 1) - (k + k - q)$ sides whose cut count is zero. Hence, as S and S' are subtree and chain reduced, this implies that

$$|X| \leq 9q + 6 \cdot 2(k - q) + 3(3(k - 1) - (k + k - q)) = 15k - 9 = 15d_{\mathrm{TBR}}(S, S') - 9,$$

thereby establishing the lemma. □

Remark. Let S and S' be two unrooted binary phylogenetic trees on X that are subtree, chain, and cluster reduced. Clearly, Lemma 3 holds for S and S'. Alternatively, by using an argument that is similar to that in the proof of Lemma 3 as well as the bounds derived in Lemma 2.2, we obtain

$$|X| \leq 7q + 6 \cdot 2(k - q) + 3(3(k - 1) - (k + k - q)) = -2q + 15k - 9.$$

Now noting that the right-hand side of the last inequality is maximized for $q = 0$, we again have

$$|X| \leq -2q + 15k - 9 \leq 15k - 9 = 15d_{\mathrm{TBR}}(S, S') - 9.$$

For $k \geq 4$, we will see in the next section that $15k - 9$ is a tight upper bound on the size of the leaf set X of two unrooted binary phylogenetic trees S and S' with $d_{\mathrm{TBR}}(S, S') = k$ and that are reduced under all three reductions. To establish this result, the fact that no side has a cut count of two (i.e., $q = 0$) gives us some important clues about the properties of S and S'. In particular, it turns out that S and S' are displayed by an unrooted binary phylogenetic network N on X with $r(N) = k$ such that N has no chain of length 7. For full details, see section 4.

In comparison to Theorem 3 that was first established in [1], the next theorem, which is an immediate consequence of Theorem 2 and Lemma 3, establishes a significantly improved linear kernel for computing the TBR distance.

THEOREM 4. *Let T and T' be two unrooted binary phylogenetic trees on X, where $d_{\mathrm{TBR}}(T, T') \geq 2$, and let S and S' be a subtree and chain reduced tree pair on X' that has been obtained from T and T' by repeated applications of the subtree and chain reduction. Then $|X'| \leq 15d_{\mathrm{TBR}}(T, T') - 9$.*

4. Tight examples.

In this section, we show that the upper bounds on the size of the leaf set of two unrooted binary phylogenetic trees that do not contain any common subtree and chain (and cluster) as established in Lemma 3 are tight. To this end, we provide two families of constructions. Throughout this section, we attach leaves to a side s of a generator that is depicted in Figures 2 or 3. If s connects two vertices of a generator that lie on a horizontal line, we attach leaves to s from left to right. Otherwise, we attach leaves to s from top to bottom. Furthermore, for a set

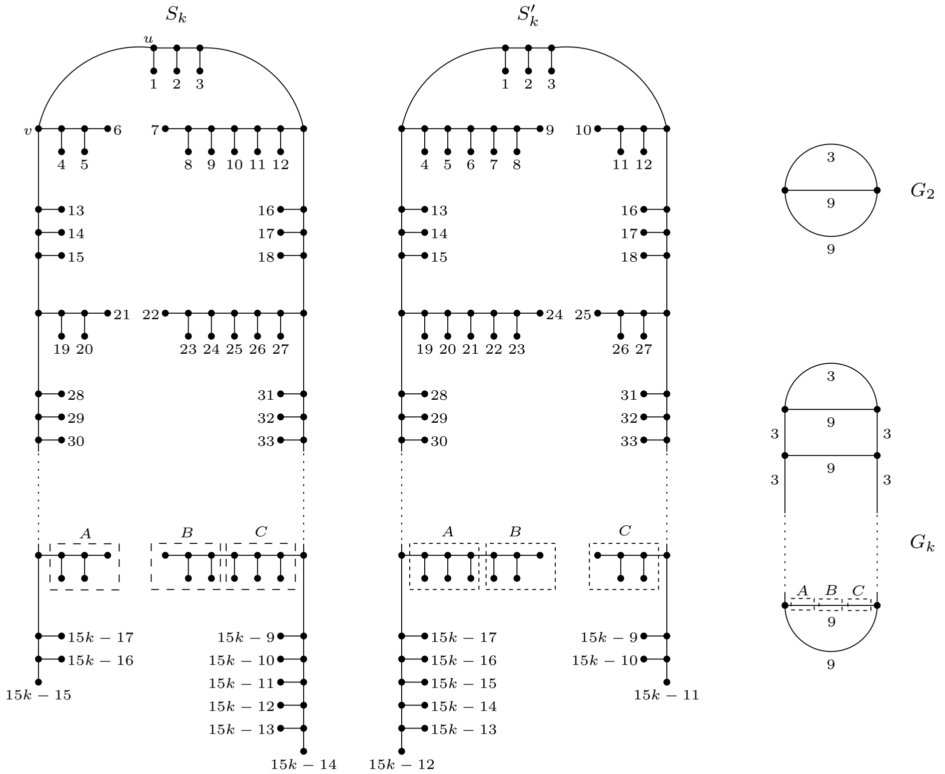

FIG. 2. *The two unrooted binary phylogenetic trees S_k and S'_k as well as the generator G_k (and G_2) that are used to show that the upper bound given in Lemma* 3 *is tight for a pair of subtree and chain reduced trees. Blocks A, B, and C indicate three chains each of length* 3. *For example, for $k = 4$, block A contains taxa* 34, 35, *and* 36, *block B contains taxa* 37, 38, *and* 39, *and block C contains taxa* 40, 41, *and* 42. *The edge $\{u, v\}$ of S_k is used to argue that $k = d_{\text{TBR}}(S_k, S'_k)$ in the proof of Lemma* 5.

of leaves that is attached to s, we order the elements from small to large and then attach from left to right or top to bottom in such a way that preserves the ordering of the elements.

We now begin with the description of a family of pairs of unrooted binary phylogenetic trees that are subtree and chain reduced (but not cluster reduced). For $k \geq 2$, consider the two unrooted binary phylogenetic trees S_k and S'_k with leaf set X_k and $|X_k| = 15k - 9$ that are shown in Figure 2. It is easy to check that S_k and S'_k do not contain any common subtree of size at least 2 or any common n-chain with $n \geq 4$. Note however that S_k and S'_k do contain k common clusters of size three. In particular, S_k and S'_k contain the common cluster $\{15i - 8, 15i - 7, 15i - 6\}$ for each $i \in \{1, 2, \ldots, k-1\}$ and, additionally, the common cluster $\{15k-14, 15k-13, 15k-12\}$.

LEMMA 4. *For $k \geq 2$, let S_k and S'_k be the two unrooted binary phylogenetic trees that are shown in Figure* 2, *and let G_k be the generator that is shown in the same figure. There exists an unrooted binary phylogenetic network N_k with $r(N_k) = k$ that displays S_k and S'_k and whose underlying generator is G_k.*

Proof. The proof is constructive, i.e., starting with G_k, we attach leaves to the sides of G_k to obtain a phylogenetic network that displays S_k and S'_k. The lemma then follows from Observation 1. Obtain an unrooted binary phylogenetic network N_k in the following way.

1. Attach $\{1, 2, 3\}$ to the topmost horizontal side of G_k.
2. For each $i \in \{1, 2, \dots, k-2\}$ in increasing order, perform the following three steps.
   (a) Attach $\{15i-11, 15i-10, \dots, 15i-3\}$ to the topmost horizontal side of G_k which is still undecorated.
   (b) Attach $\{15i-2, 15i-1, 15i\}$ to the topmost left-vertical side of G_k which is still undecorated.
   (c) Attach $\{15i+1, 15i+2, 15i+3\}$ to the topmost right-vertical side of G_k which is still undecorated.
3. Attach $\{15k-26, 15k-25, \dots, 15k-18\}$ to the second-to-lowest horizontal side of G_k.
4. Attach $\{15k-17, 15k-16, \dots, 15k-9\}$ to the lowest horizontal side of G_k.

Observe that the label of each side s of G_k as depicted in Figure 2 refers to the number of leaves that is attached to s. Furthermore, note that N_k has k chains of length 9. Regarding each such chain as a sequence of three blocks where each block contains three leaves of the chain, which is indicated by A, B, and C in Figure 2, S_k can be obtained from N_k by breaking each 9-chain between A and B and suppressing all resulting degree-2 vertices, and S'_k can be obtained from N_k by breaking each 9-chain between B and C and suppressing all resulting degree-2 vertices. Hence, N_k displays S_k and S'_k. Moreover, by construction, G_k underlies N_k. Let E_k and V_k be the edge and vertex set of G_k, respectively. We complete the proof by noting that, as $|E_k| - |V_k| + 1 = k$, it again follows by construction that $r(N_k) = k$. □

LEMMA 5. *For $k \geq 2$, let S_k and S'_k be the two unrooted binary phylogenetic trees that are shown in Figure 2. Then $d_{\mathrm{TBR}}(S_k, S'_k) = k$.*

Proof. Let N_k be the unrooted binary phylogenetic network whose construction is described in the proof of Lemma 4. Then it follows from the same lemma and Theorem 1 that

$$d_{\mathrm{TBR}}(S_k, S'_k) = h^u(S_k, S'_k) \leq r(N_k) = k. \tag{1}$$

We complete the proof by showing that $d_{\mathrm{TBR}}(S_k, S'_k) \geq k$. Let X_k be the leaf set of S_k and S'_k, and let $C = \{0, 1\}$. Furthermore, let $f : X_k \to C$ be the character on X_k defined as follows. Consider the bipartition of X_k induced by removing the edge $\{u, v\}$ as labeled in Figure 2. Give one side of the partition state 0 and the other state 1. Clearly, $l_f(S_k) = 1$. Moreover, by applying Fitch's algorithm [13], we see that $l_f(S'_k) = k + 1$. Hence, by definition of the maximum parsimony distance, we have

$$k = |1 - (k+1)| \leq d_{\mathrm{MP}}(S_k, S'_k) \leq d_{\mathrm{TBR}}(S_k, S'_k). \tag{2}$$

Combining inequalities (1) and (2) establishes the lemma. □

We now turn to a construction for two unrooted binary phylogenetic trees that are subtree, chain, and cluster reduced. For $k \geq 4$, consider the two trees S_k and S'_k on leaf set X_k that are shown in Figure 3. As with the trees depicted in Figure 2, note that $|X_k| = 15k - 9$.

LEMMA 6. *For $k \geq 4$, let S_k and S'_k be the two unrooted binary phylogenetic trees that are shown in Figure 3. Then S_k and S'_k are subtree, chain, and cluster reduced.*

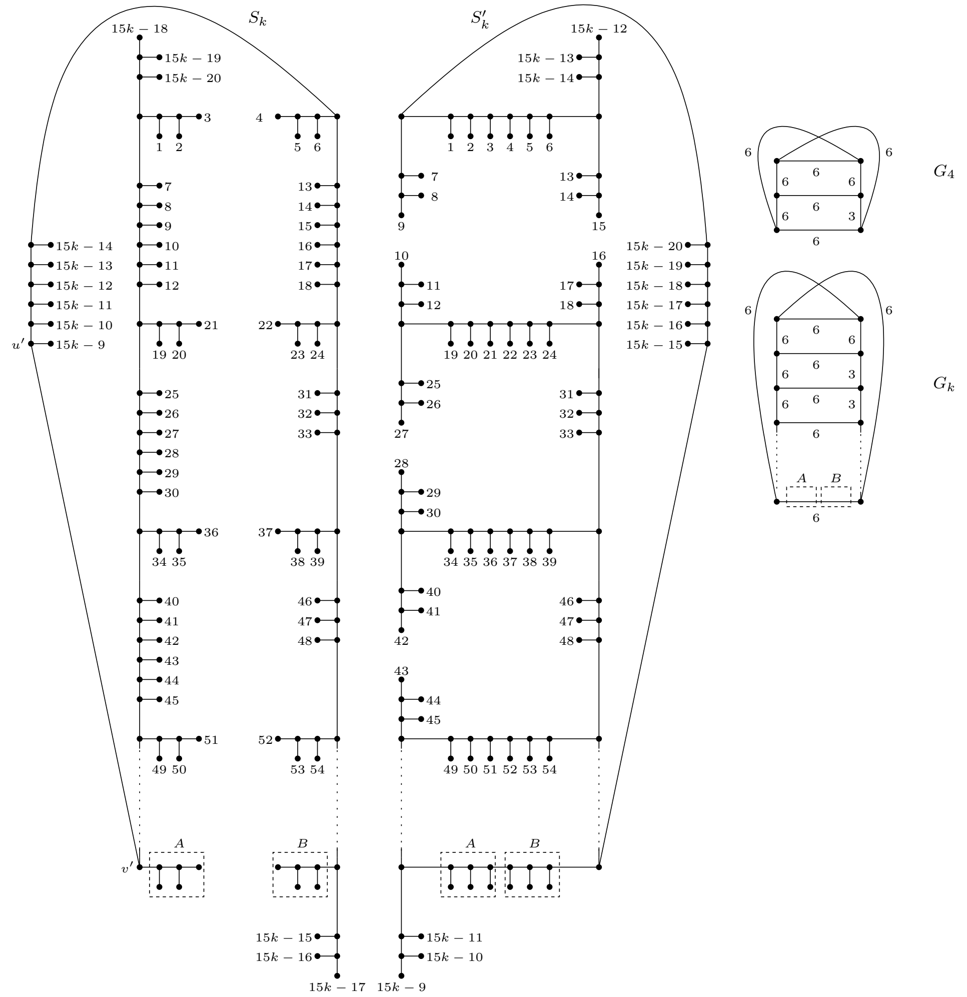

FIG. 3. *The unrooted binary phylogenetic trees S_k and S'_k as well as the generator G_k (and G_4) that are used to show that the upper bound given in Lemma* 3 *is tight for a pair of subtree, chain, and cluster reduced trees. Blocks A and B indicate two chains each of length* 3*. For example, for $k = 6$, block A contains taxa* 64, 65, *and* 66, *and block B contains taxa* 67, 68, *and* 69*. The edge $\{u', v'\}$ of S_k is used to argue that $k = d_{\mathrm{TBR}}(S_k, S'_k)$ in the proof of Lemma* 8*.*

Proof. It is straightforward to check that S_k and S'_k are subtree reduced. To see that they are also chain reduced, notice that S'_k has k maximal 6-chains. However, none of these chains (or a subchain of size at least 4) is also a chain of S_k. We complete the proof by showing that S_k and S'_k are also cluster reduced. Let P be the path from $15k - 18$ to $15k - 17$ in S_k. Furthermore, let e be an arbitrary edge of S_k, and let Y_1, Y_2 be the bipartition of X_k such that one subtree obtained from S_k by deleting e has leaf set Y_1 while the other subtree has leaf set Y_2. It suffices to show that, if Y_1 and Y_2 both have size at least two, then Y_1, Y_2 is not a bipartition of S'_k, i.e., there exists no edge in S'_k whose removal results in two subtrees with leaf sets Y_1 and Y_2, respectively. First, suppose that e is an edge that does not lie on P. Then for some $i \in \{1, 2\}$, we have $|Y_i| \leq 3$. Moreover, if $|Y_i| \in \{2, 3\}$, then it is easily verified that Y_1, Y_2 is not a bipartition of S'_k. Second, suppose, that e lies on P. Then $|\{15k-18, 15k-17\} \cap Y_1| = 1$ and $|\{15k-18, 15k-17\} \cap Y_2| = 1$. Now, let f be an edge of S'_k whose removal results in two subtrees that both have at least two leaves, the leaf set of one subtree contains $15k-18$ and the leaf set of the other subtree contains $15k-17$. Clearly, the unique edge f in S'_k that satisfies all three properties is the second edge on the path from $15k-18$ to $15k-17$. Let Z_1, Z_2 be the bipartition of X_k such that Z_1 and Z_2 are the leaf sets of the two trees resulting from the deletion of f in S'_k. Without loss of generality, we may assume that $\{9, 15\} \subset Z_1$ and $\{10, 16\} \subset Z_2$. But Z_1, Z_2 is not a bipartition of S_k because we can either separate 9 and 10, or 15 and 16 by removal of a single edge in S_k but not both. It now follows that S_k and S'_k are cluster reduced. This completes the proof of the lemma. □

LEMMA 7. *For $k \geq 4$, let S_k and S'_k be the two unrooted binary phylogenetic trees that are shown in Figure* 3, *and let G_k be the generator that is shown in the same figure. There exists an unrooted binary phylogenetic network N_k with $r(N_k) = k$ that displays S_k and S'_k and whose underlying generator is G_k.*

Proof. The proof is similar to that of Lemma 4. Starting with G_k, obtain an unrooted binary phylogenetic network N_k in the following way.

1. Attach $\{1, 2, \ldots, 6\}$ to the topmost horizontal side of G_k.
2. Attach $\{7, 8, \ldots, 12\}$ to the topmost left-vertical side of G_k.
3. Attach $\{13, 14, \ldots, 18\}$ to the topmost right-vertical side of G_k.
4. For each $i \in \{2, \ldots, k-2\}$ in increasing order, perform the following three steps.
   (a) Attach $\{15i-11, 15i-10, \ldots, 15i-6\}$ to the topmost horizontal side of G_k which is still undecorated.
   (b) Attach $\{15i-5, 15i-4, \ldots, 15i\}$ to the topmost left-vertical side of G_k which is still undecorated.
   (c) Attach $\{15i+1, 15i+2, 15i+3\}$ to the topmost right-vertical side of G_k which is still undecorated.
5. Attach $\{15(k-2)+4, 15(k-2)+5, \ldots, 15(k-2)+9\}$ to the lowest horizontal side of G_k.
6. Attach $\{15(k-2)+10, 15(k-2)+11, \ldots, 15(k-2)+15\}$ to the undecorated curved side of G_k that connects the top-left vertex with the bottom-right vertex.
7. Attach $\{15(k-2)+16, 15(k-2)+17, \ldots, 15(k-2)+21\}$ to the undecorated curved side of G_k that connects the top-right vertex with the bottom-left vertex.

With $15(k-2)+21 = 15k-9$, it follows that N_k has $15k-9$ leaves. Observe that N_k has $2k$ chains of length 6. Regard each such chain as a sequence of two blocks where each block contains three leaves of the chain, which is indicated by A and B in Figure 3. In the following, we refer to the process of deleting the unique edge of a 6-chain that has one vertex in block A and one in block B as *breaking a chain.* Now, S_k can be obtained from N_k by breaking each of the $k-1$ 6-chains whose leaves decorate the $k-1$ horizontal sides of G_k, breaking the 6-chain whose leaves decorate the curved side of G_k that connects the top-left vertex with the bottom-right vertex, and suppressing all resulting degree-2 vertices. Similarly, S'_k can be obtained from N_k by breaking each of the $k-2$ 6-chains whose leaves decorate the $k-2$ left-vertical sides of G_k, breaking the 6-chain whose leaves decorate the topmost right-vertical side of G_k, breaking the 6-chain whose leaves decorate the curved side of G_k that connects the top-right vertex with the bottom-left vertex, and suppressing all resulting degree-2 vertices. Hence, N_k displays S_k and S'_k. Moreover, by construction and Observation 1, G_k underlies N_k. We complete the proof by noting that, as $|E_k| - |V_k| + 1 = k$, it again follows by construction that $r(N_k) = k$, where V_k and E_k is the vertex and edge set of G_k, respectively. □

Lemma 8. *For $k \geq 4$, let S_k and S'_k be the two unrooted binary phylogenetic trees that are shown in Figure* 3. *Then $d_{\mathrm{TBR}}(S_k, S'_k) = k$.*

Proof. By considering the unrooted binary phylogenetic network N_k that is described in the proof of Lemma 7 (instead of the one described in the proof of Lemma 4), the proof of this lemma can be established in exactly the same way as the proof of Lemma 5. Note that the edge $\{u', v'\}$ as depicted in S_k of Figure 3 plays the role of the edge $\{u, v\}$ that is used in the proof of Lemma 5. □

We are now in a position to establish the main result of this section.

Theorem 5. *Let S and S' be two unrooted binary phylogenetic trees on X. If S and S' are subtree and chain reduced and $d_{\mathrm{TBR}}(S, S') \geq 2$, then $|X| \leq 15d_{\mathrm{TBR}}(S, S') - 9$ is a tight bound. Moreover, if S and S' are subtree, chain, and cluster reduced and $d_{\mathrm{TBR}}(S, S') \geq 4$, then $|X| \leq 15d_{\mathrm{TBR}}(S, S') - 9$ is again a tight bound.*

Proof. Suppose that S and S' are subtree and chain reduced. It immediately follows from Lemma 3 that $|X| \leq 15d_{\mathrm{TBR}}(S, S') - 9$. To establish that the bound is tight, it is sufficient to choose S and S' such that they are subtree and chain reduced, and $|X| = 15d_{\mathrm{TBR}}(S, S') - 9$. For $k \geq 2$, set $S = S_k$ and $S' = S'_k$, where S_k and S'_k are the two trees shown in Figure 2. By construction, we have $|X| = 15k - 9$. Moreover, by Lemma 5, we have $k = d_{\mathrm{TBR}}(S, S')$. The proof for when S and S' are subtree, chain, and cluster reduced can be established analogously by setting S and S' to be the two trees that are shown in Figure 3 for $k \geq 4$ and considering Lemma 8 instead of Lemma 5. □

The next corollary is an immediate consequence of the last theorem.

Corollary 1. *The linear kernel as presented in Theorem* 4 *is tight.*

5. Tight kernels for computing the rooted variant of the minimum hybridization problem.

In this section, we turn to rooted phylogenetic trees and networks, and show that a previously published kernelization result that is concerned with the rooted analogue of UHN is also tight. Before formally stating the problem, we introduce some new definitions some of which are the rooted versions of their counterparts introduced in section 2.

A *rooted binary phylogenetic network* N on X is a rooted acyclic digraph with no edges in parallel and satisfying the following properties:

(i) the (unique) root ρ has out-degree two;

(ii) the set X labels the set of vertices of out-degree zero, each of which has in-degree one (i.e., the leaves); and

(iii) all other vertices either have in-degree one and out-degree two or in-degree two and out-degree one.

For two vertices u and v in N, we say that u is a *parent* of v and v is a *child* of u if (u, v) is an edge in N. For a leaf ℓ, we denote its unique parent by p_ℓ. Furthermore, a vertex of in-degree two and out-degree one is called a *reticulation*, and we use $r(N)$ to denote the number of reticulations in N. Lastly, N is called a rooted binary phylogenetic tree on X if $r(N) = 0$.

Let T be a rooted binary phylogenetic tree on X, and let $Y \subset X$. A subtree of T is *pendant* in T if it can be obtained from T by deleting a single edge. Now, for $n \geq 2$, let $C = (\ell_1, \ell_2, \dots, \ell_n)$ be a sequence of distinct leaves in X. We call C an *n-chain* of T if $p_{\ell_1} = p_{\ell_2}$ or p_{ℓ_1} is a child of p_{ℓ_2} and, for all $i \in \{2, 3, \dots, n-1\}$, we have that p_{ℓ_i} is a child of $p_{\ell_{i+1}}$. Furthermore, we say that $Y \subset X$ is a *nontrivial cluster* of T if $|Y| \geq 2$ and there exists a vertex in T whose set of descendants is precisely Y.

Now, let N be a rooted binary phylogenetic network on X, and let T be a rooted binary phylogenetic tree on X. We say that T is *displayed* by N if T can be obtained from a subtree of N by suppressing vertices with in-degree one and out-degree one. Furthermore, for two rooted binary phylogenetic trees T and T' on X, we set

$$h(T, T') = \min_N \{r(N)\},$$

where the minimum is taken over all rooted binary phylogenetic networks on X that display T and T'. Historically, this number is referred to as the hybridization number for T and T' (see, e.g., [8, 28, 30, 31] and references therein).

We now formally state the rooted version of UHN.

Rooted-Hybridization-Number (RHN)
Input. Two rooted binary phylogenetic trees T and T' on X.
Output. A rooted binary phylogenetic network N on X that displays T and T' and such that $r(N) = h(T, T')$.

Solving RHN for a pair of rooted binary phylogenetic trees is NP-hard [8] and FPT [7], when parameterized by the hybridization number. To establish a fixed-parameter tractability result, the authors of the latter paper used rooted variants of the subtree and chain reduction to kernelize the problem. Without detailing the reductions for two rooted binary phylogenetic trees, we next give definitions of what it means for two rooted binary phylogenetic trees to be reduced under any of the two reductions. Specifically, for two rooted binary phylogenetic trees S and S' on X, we say that S and S' are subtree reduced if they do not have a common pendant subtree with at least two leaves and chain reduced if they do not have a common n-chain with $n \geq 3$. By building on the results of [7], the authors of [21] improved the kernel size by using (rooted) generators. In the language of our paper, they established the following lemma.

Lemma 9 (see [21, Theorem 3.2]). *Let S and S' be two rooted binary phylogenetic trees on X that are subtree and chain reduced and $h(S, S') \geq 1$. Then $|X| \leq 9h(S, S') - 2$.*

We next show, by describing a specific family of pairs of rooted binary phylogenetic trees, that the bound given in Lemma 9 is tight. For $k \geq 1$, consider the two rooted binary phylogenetic trees S_k and S'_k on X_k that are shown in Figure 4. By construction, we have $|X_k| = 9k - 2$; particularly, $|X_1| = 7$ and $|X_k| = 7 + 9(k-1)$ for each $k \geq 2$. Moreover, it is easy to see that S_k and S'_k do not have any common pendant subtree with at least two leaves or any common n-chain with $n \geq 3$. Thus, S_k and S'_k are subtree and chain reduced.

Lemma 10. *For $k \geq 1$, let S_k and S'_k, be the two rooted binary phylogenetic trees on X_k that are shown in Figure* 4. *Then $h(S_k, S'_k) = k$.*

Proof. Let N_k be the rooted binary phylogenetic network that is shown in Figure 4. Note that the leaf set of N_k is X_k. Moreover, observe that S_k can be obtained from N_k by deleting the set $\{e_1, e_2, \dots, e_k\}$ of edges and, subsequently, suppressing all resulting vertices with in-degree one and out-degree one. Similarly, S'_k can be obtained

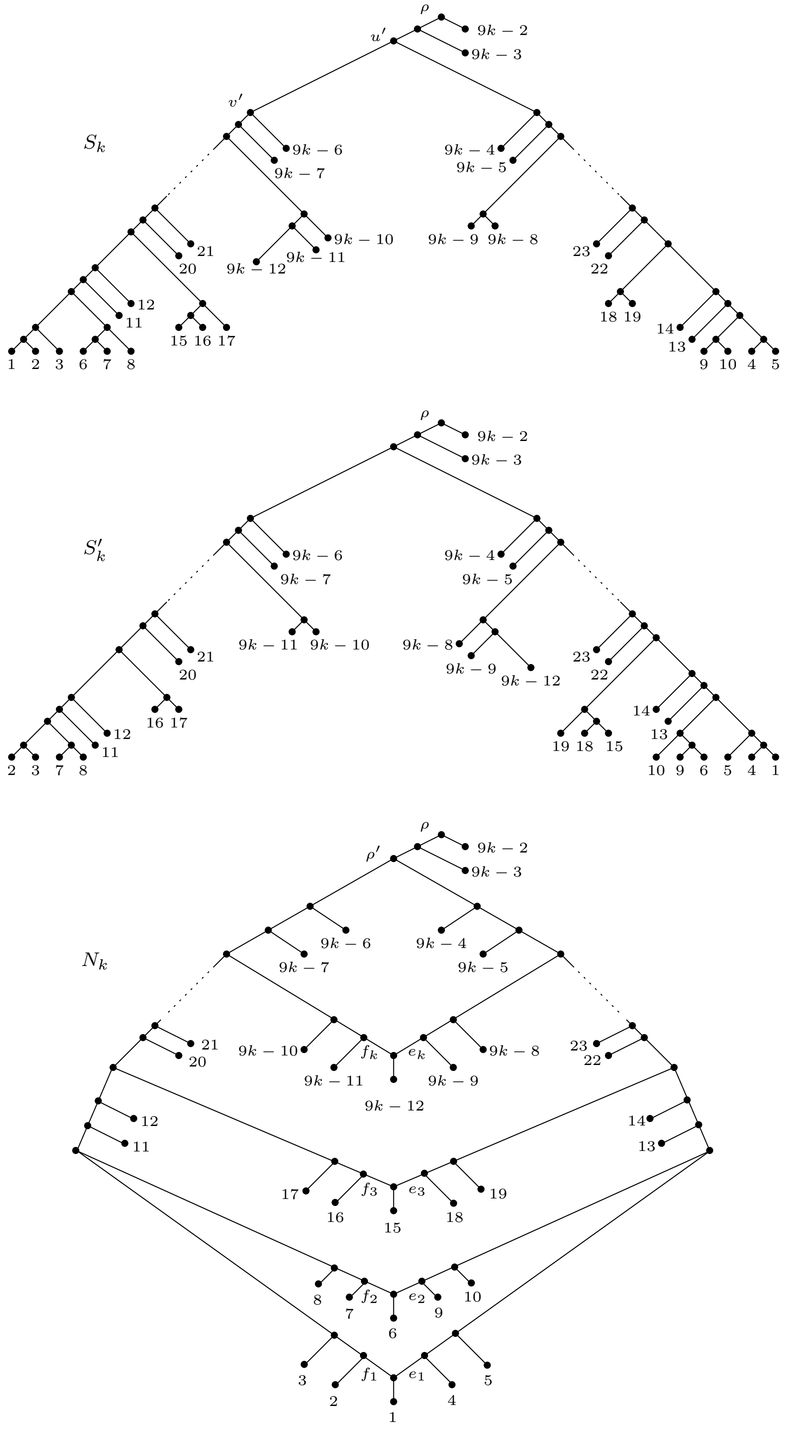

FIG. 4. *The rooted binary phylogenetic trees S_k and S'_k as well as the rooted binary phylogenetic network that are used to show that the upper bound given in Lemma* 9 *is tight for a pair of subtree and chain reduced trees. The edge (u', v') of S_k is used to argue that $k = h(S_k, S'_k)$ in the proof of Lemma* 10.

from N_k by deleting the set $\{f_1, f_2, \ldots, f_k\}$ of edges and, subsequently, suppressing all resulting vertices with in-degree one and out-degree one. Hence, N_k displays S_k and S'_k. Since $r(N_k) = k$, we have

$$h(S_k, S'_k) \leq r(N_k) = k. \tag{3}$$

We complete the proof by showing that $k \leq h(S_k, S'_k)$. Let N^*_k be a rooted binary phylogenetic network on X_k such that $r(N^*_k) = h(S_k, S'_k)$. Furthermore, let $\bar{S}_k$ and $\bar{S}'_k$ be the two unrooted binary phylogenetic trees on X_k that are obtained from S_k and S'_k, respectively, by suppressing the root ρ and ignoring the directions on the edges. Then $\bar{S}_k$ and $\bar{S}'_k$ are displayed by the unrooted binary phylogenetic network $\bar{N}^*_k$ on X_k that is obtained from N_k by suppressing its root and ignoring the directions on the edges. As $r(N^*_k) = r(\bar{N}^*_k)$, it follows that $h^u(\bar{S}_k, \bar{S}'_k) \leq h(S_k, S'_k)$. To show that $k \leq h^u(\bar{S}_k, \bar{S}'_k)$, we use the same approach as in the second half of the proof of Lemma 5, where the edge (u', v') as depicted in S_k of Figure 4 plays the role of the edge $\{u, v\}$. Hence, we have

$$k = |1 - (k+1)| \leq d_{\mathrm{MP}}(\bar{S}_k, \bar{S}'_k) \leq d_{\mathrm{TBR}}(\bar{S}_k, \bar{S}'_k) = h^u(\bar{S}_k, \bar{S}'_k) \leq h(S_k, S'_k)$$

which, in combination with (3), establishes the lemma. □

The main result of this section is the following theorem whose proof can be established in the same way as was the proof of Theorem 5.

Theorem 6. *Let S and S' be two rooted binary phylogenetic trees on X and $h(S, S') \geq 1$. If S and S' are subtree and chain reduced, then $|X| \leq 9h(S, S') - 2$ is a tight bound.*

Similarly to UHN and Corollary 1, the last theorem implies that the upper bound on the size of a kernel for RHN, as established in [21, Theorem 3.2], is also tight.

We now turn to the rooted version of the cluster reduction [2]. Informally, this reduction breaks an instance, say T and T', of RHN into a number of smaller tree pairs such that the sum of the hybridization number over all tree pairs equates to this number for T and T'. In what follows, we say that two rooted binary phylogenetic trees are cluster reduced if they do not have any nontrivial cluster in common. Observe that the two phylogenetic trees S_k and S'_k as shown in Figure 4 have cluster $X_k - \{9k-3, 9k-2\}$ in common. By deleting $9k-3$ and $9k-2$ from the trees and network of Figure 4 and their respective parents, we obtain two rooted binary phylogenetic trees, say R_k and R'_k, and a rooted binary phylogenetic network, say M_k. Clearly, R_k and R'_k are subtree, chain, and cluster reduced. Furthermore, M_k displays R_k and R'_k and $r(M_k) = r(N_k) = k$ for any $k \geq 1$. Reworking the proof of [21, Theorem 3.2], we note that the edge side incident with ρ (this is a particular side of the generator used in that proof, corresponding to the path from ρ to ρ' as shown in Figure 3) cannot be decorated with any leaf. This is because any two distinct trees displayed by the network would then have a nontrivial common cluster. Hence, the counting argument from [21, Theorem 3.2] goes through, with the exception that leaves $9k-3$ and $9k-2$ are no longer present in the network. This yields the following lemma.

Lemma 11. *Let S and S' be two rooted binary phylogenetic trees on X that are subtree, chain, and cluster reduced and $h(S, S') \geq 1$. Then $|X| \leq 9h(S, S') - 4$.*

Moreover, by repeating the argument to establish Theorem 6 but using R_k, R'_k, and M_k instead of S_k, S'_k, and N_k as shown in Figure 4, it follows that the bound given in Lemma 11 is tight.

6. Discussion.

Following the results in this article, the algorithmic state of knowledge about computation of TBR distance can be summarized as follows. The problem is NP-hard, but permits a polynomial-time 3-approximation [9], a branching FPT algorithm with running time $O(3^k \cdot \text{poly}(n))$ (where k is the TBR distance and $n = |X|$) [31], a kernel of size $15k-9$, and an exponential-time algorithm with running time $O(2.619^n \cdot \text{poly}(n))$ [20]. An interesting consequence of the strengthened $15k-9$ bound is that results which indirectly use the size of the TBR kernel to compute bounds, automatically improve. One concrete example of this is the result in [18] which proves that after application of subtree and chain reduction rules, reduced instances of d_{MP} contain at most $\frac{4}{3} \cdot 28 \cdot d_{\text{TBR}}$ taxa. The 28 now improves to 15. Our result implies a similar improvement on the kernel for computing the so-called subtree prune and regraft (SPR) distance [32].

Returning to TBR distance, the natural question is whether through the introduction of new polynomial-time reduction rules (in addition to subtree, chain, and cluster reduction rules) the size of the kernel can be further reduced and, if so, what the limit is of such an approach. This ties in with the FPT literature on (complexity-theoretic) lower bounds on kernel size (see, e.g., [3]), which have not yet been explored in the phylogenetics literature. Towards an easy lower bound we note that, if computation of TBR distance is APX-hard, then there exists a constant $c > 1$ such that a polynomial-time c-approximation for TBR distance is not possible (assuming P $\neq$ NP). Such a result would exclude the existence of a kernel of size $c \cdot k$ for TBR distance (assuming P $\neq$ NP). This is because the TBR distance of two trees on n taxa is at most $O(n)$ [1]; so simply returning a trivial solution for the kernelized instance would yield a c-approximation.[1] However, although it is likely that computing the TBR distance is APX-hard, to the best of our knowledge the result has never been proven. This is an interesting hole to close in the literature.

Beyond TBR distance we can ask whether existing bounds on kernels for other phylogenetic distances and incongruency measures are tight and, if not, whether they can be improved. Many such kernelizations use subtree reductions and variations of chain reductions. The reductions typically have a common core but details differ from case to case depending on the specific combinatorial nature of the problem at hand: there are many subtle differences between TBR distance, SPR distance [4, 6, 32], hybridization number [7, 29, 30], and agreement forests [24, 31], for example. Other relevant factors include whether the input trees are unrooted or rooted, whether the input trees are binary or nonbinary, and the number of trees allowed in the input (see earlier references and [28]). In obtaining the tight $15k - 9$ bound we were greatly helped by our ability to reformulate the problem as a phylogenetic network construction problem, which in turn allowed us to make use of generators. Interestingly, the generators not only helped us improve the upper bound, they also gave strong hints concerning the topology of tight instances. Once discovered, we could use the maximum parsimony distance to argue lower bounds on d_{TBR} distance. In how far do these three ingredients exist simultaneously for other problems and, where they do not exist, in which direction do we have to advance our knowledge to obtain tight bounds on kernel sizes?

Acknowledgment. Both authors would like to thank the Lorentz Center in the Netherlands for hosting the workshop *Distinguishability in Genealogical Phylogenetic Networks*, where this work was initiated.

[1] Note that the kernel for RHN is weighted [21] and, so, returning a trivial solution for a kernelized instance of RHN does not yield a c-approximation for this problem.

REFERENCES

[1] B. Allen and M. Steel, *Subtree transfer operations and their induced metrics on evolutionary trees*, Ann. Comb., 5 (2001), pp. 1–15.

[2] M. Baroni, C. Semple, and M. Steel, *Hybrids in real time*, Syst. Biol., 55 (2006), pp. 46–56.

[3] H. L. Bodlaender, B. M. P. Jansen, and S. Kratsch, *Kernelization lower bounds by cross-composition*, SIAM J. Discrete Math., 28 (2014), pp. 277–305.

[4] M. Bonet and K. S. John, *On the complexity of uSPR distance*, IEEE/ACM Trans. Comput. Biol. Bioinform., 7 (2010), pp. 572–576.

[5] M. Bordewich, C. Scornavacca, N. Tokac, and M. Weller, *On the fixed parameter tractability of agreement-based phylogenetic distances*, J. Math. Biol., 74 (2017), pp. 239–257.

[6] M. Bordewich and C. Semple, *On the computational complexity of the rooted subtree prune and regraft distance*, Ann. Comb., 8 (2005), pp. 409–423.

[7] M. Bordewich and C. Semple, *Computing the hybridization number of two phylogenetic trees is fixed-parameter tractable*, IEEE/ACM Trans. Comput. Biol. Bioinform., 4 (2007), pp. 458–466.

[8] M. Bordewich and C. Semple, *Computing the minimum number of hybridization events for a consistent evolutionary history*, Discrete Appl. Math., 155 (2007), pp. 914–928.

[9] J. Chen, J.-H. Fan, and S.-H. Sze, *Parameterized and approximation algorithms for maximum agreement forest in multifurcating trees*, Theoret. Comput. Sci., 562 (2015), pp. 496–512.

[10] M. Cygan, F. Fomin, L. Kowalik, D. Lokshtanov, D. Marx, M. Pilipczuk, M. Pilipczuk, and S. Saurabh, *Parameterized Algorithms*, 1st ed., Springer, Cham, Switzerland, 2015.

[11] J. Felsenstein, *Inferring Phylogenies*, Sinauer Associates, Sunderland, MA, 2004.

[12] M. Fischer and S. Kelk, *On the maximum parsimony distance between phylogenetic trees*, Ann. Comb., 20 (2016), pp. 87–113.

[13] W. M. Fitch, *Toward defining the course of evolution: Minimum change for a specific tree topology*, Syst. Biol., 20 (1971), pp. 406–416.

[14] P. Gambette, V. Berry, and C. Paul, *Quartets and unrooted phylogenetic networks*, J. Bioinf. Comput. Biol., 10 (2012), 1250004.

[15] J. Hein, T. Jiang, L. Wang, and K. Zhang, *On the complexity of comparing evolutionary trees*, Discrete Appl. Math., 71 (1996), pp. 153–169.

[16] D. Huson, R. Rupp, and C. Scornavacca, *Phylogenetic Networks: Concepts, Algorithms and Applications*, Cambridge University Press, Cambridge, 2011.

[17] K. S. John, *The shape of phylogenetic treespace*, Syst. Biol., 66 (2017), e83–e94.

[18] S. Kelk, M. Fischer, V. Moulton, and T. Wu, *Reduction rules for the maximum parsimony distance on phylogenetic trees*, Theoret. Comput. Sci., 646 (2016), pp. 1–15.

[19] S. Kelk and C. Scornavacca, *Constructing minimal phylogenetic networks from softwired clusters is fixed parameter tractable*, Algorithmica, 68 (2014), pp. 886–915.

[20] S. Kelk and G. Stamoulis, *A note on convex characters, Fibonacci numbers and exponential-time algorithms*, Adv. Appl. Math., 84 (2017), pp. 34–46.

[21] S. Kelk, L. van Iersel, N. Lekić, S. Linz, C. Scornavacca, and L. Stougie, *Cycle killer... Qu'est-ce que c'est? On the comparative approximability of hybridization number and directed feedback vertex set*, SIAM J. Discrete Math., 26 (2012), pp. 1635–1656.

[22] D. Money and S. Whelan, *Characterizing the phylogenetic tree-search problem*, Syst. Biol., 61 (2012), pp. 228–239.

[23] V. Moulton and T. Wu, *A parsimony-based metric for phylogenetic trees*, Adv. Appl. Math., 66 (2015), pp. 22–45.

[24] F. Shi, J. Chen, Q. Feng, and J. Wang, *A parameterized algorithm for the maximum agreement forest problem on multiple rooted multifurcating trees*, J. Comput. System Sci., 97 (2018), pp. 28–44.

[25] M. Steel, *Phylogeny: Discrete and Random Processes in Evolution*, CBMS-NSF Regional Conf. Ser. in Appl. Math. 89, SIAM, Philadelphia, 2016.

[26] L. van Iersel, J. Keijsper, S. Kelk, L. Stougie, F. Hagen, and T. Boekhout, *Constructing level-2 phylogenetic networks from triplets*, IEEE/ACM Trans. Comput. Biol. Bioinform., 6 (2009), pp. 667–681.

[27] L. van Iersel, S. Kelk, N. Lekić, C. Whidden, and N. Zeh, *Hybridization number on three rooted binary trees is EPT*, SIAM J. Discrete Math., 30 (2016), pp. 1607–1631.

[28] L. van Iersel, S. Kelk, and C. Scornavacca, *Kernelizations for the hybridization number problem on multiple nonbinary trees*, J. Comput. System Sci., 82 (2016), pp. 1075–1089.

[29] L. van Iersel, S. Kelk, G. Stamoulis, L. Stougie, and O. Boes, *On unrooted and root-uncertain variants of several well-known phylogenetic network problems*, Algorithmica, 80 (2018), pp. 2993–3022.

[30] L. van Iersel and S. Linz, *A quadratic kernel for computing the hybridization number of multiple trees*, Inform. Process. Lett., 113 (2013), pp. 318–323.

[31] C. Whidden, R. G. Beiko, and N. Zeh, *Fixed-parameter algorithms for maximum agreement forests*, SIAM J. Comput., 42 (2013), pp. 1431–1466.

[32] C. Whidden and F. Matsen, *Calculating the unrooted subtree prune-and-regraft distance*, IEEE/ACM Trans. Comput. Biol. Bioinform., 16 (2019), pp. 898–911.