Discrete Applied Mathematics 285 (2020) 612–620

Contents lists available at ScienceDirect

Discrete Applied Mathematics

journal homepage: www.elsevier.com/locate/dam

On the extremal maximum agreement subtree problem

Alexey Markin

Department of Computer Science, Iowa State University Ames, IA 50011, USA

ARTICLE INFO

Article history:
Received 29 May 2019
Received in revised form 21 May 2020
Accepted 3 July 2020
Available online 13 July 2020

Keywords:
Maximum agreement subtree
Phylogenetic trees
Lower bound
Labeled trees

ABSTRACT

Given two phylogenetic trees with leaf set $\{1, \ldots, n\}$ the maximum agreement subtree problem asks what is the maximum size of the subset $A \subseteq \{1, \ldots, n\}$ such that the two trees are equivalent when restricted to A. The long-standing extremal version of this problem focuses on the smallest number of leaves, $\mathsf{mast}(n)$, on which any two (binary and unrooted) phylogenetic trees with n leaves must agree. In this work, we prove that this number grows asymptotically as $\Theta(\log n)$; thus closing the enduring gap between the lower and upper asymptotic bounds on $\mathsf{mast}(n)$.

1. Introduction

The algorithmic aspects of the maximum agreement subtree problem have been heavily researched for many versions of this problem (see, e.g., [1,5,11]). The extremal problem explored in this work was first addressed more than 25 years ago by Kubicka et al. [7], where they proved the $c_1(\log\log n)^{1/2} \leq \mathsf{mast}(n) \leq c_2 \log n$ bounds for some constants c_1 and c_2. The lower bound was later improved to $\Omega(\log\log n)$ by Steel and Székely [10] and then to $\Omega(\sqrt{\log n})$ by Martin and Thatte [8]. The result by Martin and Thatte originated from their proof that if at least one of the trees is either a caterpillar or a balanced tree (or an almost-balanced tree), then the maximum agreement subtree must be $\Omega(\log n)$. Additionally, Martin and Thatte conjectured that two rooted balanced trees must agree on at least $\sqrt{n}$ leaves. Note that in this conjecture a rooted balanced tree is a tree of height h with exactly $n = 2^h$ leaves. Recently, however, Bordewich et al. [3] claimed that they disproved this conjecture.

In this work we close the gap between the lower and upper asymptotic bounds and demonstrate that $\mathsf{mast}(n) \in \Theta(\log n)$. More precisely, first we prove a 'dual' (weaker) theorem stating that if any two phylogenetic trees with leaf set $\{1, \ldots, n\}$ are arbitrarily rooted, then they either agree as rooted trees on $\Omega(\frac{\log n}{\log\log n})$ leaves or agree as the original unrooted trees on $\Omega(\log n)$ leaves. This theorem was, in part, motivated by the Martin and Thatte constructive lower bound for the constrained case when one of the trees is rooted and balanced. Next, we extend this theorem with a more involved analysis and obtain the main result.

An example of a maximum agreement subtree between two trees is depicted in Fig. 1.

2. Preliminaries

A *(phylogenetic X-)tree* is a binary unrooted tree with all internal nodes of degree three and leaves bijectively labeled by elements of set X; for convenience, we identify leaves with their labels from X. The set of leaves of a tree T is denoted by $\mathsf{Le}(T)$, which is used when set X is not explicitly defined. Two X-trees are identical if there exists a label-preserving graph isomorphism between them. Given a set $Y \subset X$, Y-tree $T|Y$ is defined as the binary unrooted tree such that the

E-mail address: amarkin@iastate.edu.

https://doi.org/10.1016/j.dam.2020.07.007
0166-218X/

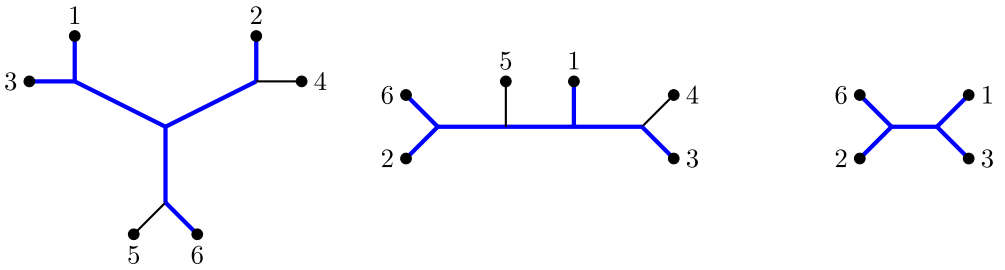

Fig. 1. An example of a maximum agreement subtree (right) between two unrooted trees with six leaves depicted on the left; these trees reduce to the subtree on the right when leaves {4, 5} are removed.

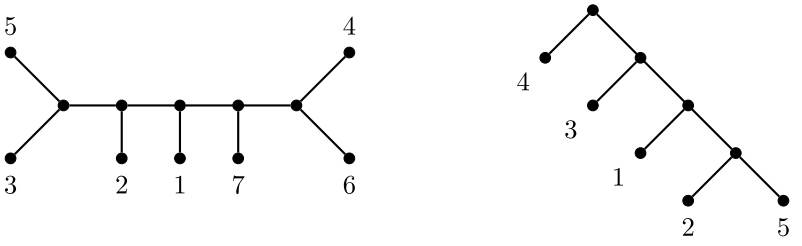

Fig. 2. Unrooted (left) and rooted (right) examples of caterpillar trees.

minimal connected subgraph of T which contains all leaves from Y is a subdivision of $T|Y$. For convenience, we define the *size* of a tree as $|T| := |\mathsf{Le}(T)| = |X|$.

A *rooted (phylogenetic X-)tree* T is a binary rooted tree with a designated root node of degree two, denoted $\rho(T)$, and each internal node having a designated left child and a right child. Given a set $Y \subset X$ a rooted Y-tree $T|Y$ is defined similarly to the unrooted case. Given a node $v \in T$, T_v denotes the subtree of T rooted at v.

A rooted tree T defines a partial order on its nodes: given two nodes x and y we say $x \preceq y$ if x is a descendant of y (and $x \prec y$ if additionally $x \neq y$). Further, we say that x and y are *incomparable* if neither $x \preceq y$ nor $y \preceq x$. For a set $Z \subseteq X$ the *least common ancestor (lca)* of Z, denoted $\mathsf{lca}_T(Z)$, is the lowest node v such that each $l \in Z$ is a descendant of v.

For a rooted X-tree T let $\mathsf{Ord}(T)$ be the left-to-right ordering of leaves induced by the pre-order traversal of nodes of T. For example, $\mathsf{Ord}(T)$ for T being the rooted tree from Fig. 2 (right) is (4, 3, 1, 2, 5). We refer to $\mathsf{Ord}(T)$ as the *leaf ordering* of T. We will often identify the leaves of T with their indices in $\mathsf{Ord}(T)$.

Caterpillar trees. An unrooted or rooted X-tree is a *caterpillar* if every internal node (including, if present, the root) is adjacent to at least one leaf. Fig. 2 demonstrates the structure of caterpillars.

Maximum agreement subtree. For two (unrooted or rooted) trees T and S on $\{1, \ldots, n\}$ leaf-set a *maximum agreement set* is the maximum set $Y \subseteq \{1, \ldots, n\}$, such that $T|Y = S|Y$ up to label-preserving graph isomorphism. The tree $T|Y$ is called a *maximum agreement subtree* and the size of the maximum agreement set/subtree is denoted by $\mathsf{mast}(T, S)$.

Let $\mathcal{P}(n)$ be the set of all unrooted X-trees with $X = \{1, \ldots, n\}$ then

$$\mathsf{mast}(n) := \min_{T,S \in \mathcal{P}(n)} (\mathsf{mast}(T, S)).$$

That is, $\mathsf{mast}(n)$ is the minimum number of leaves on which any two unrooted X-trees must agree. Throughout the work we use $\log x$ to denote $\log_2 x$.

3. Dual lower bound result

In this section we prove a 'dual' (rooted/unrooted) lower bound result for $\mathsf{mast}(n)$ and lay the foundation for our main result.

Given any two unrooted X-trees T and S with $X = \{1, \ldots, n\}$ and $n \geq 4$, let T' and S' be rooted trees obtained from T and S respectively by rooting them at arbitrarily chosen edges $e_T \in E(T), e_S \in E(S)$ (the rooting is performed by subdividing the chosen edge with a new node and designating this node as the root). For each internal node in T' and S' then one of the children is designated to be the left child and the other to be the right child arbitrarily. In this section we prove the following theorem.

Theorem 1. *Either the rooted trees T' and S' have a rooted (caterpillar) agreement subtree of size at least $\frac{1}{4}\frac{\log n}{\log\log n}$ or the original unrooted trees T and S have a (caterpillar) agreement subtree of size at least $\log n$.*

The rest of the section is dedicated to the constructive proof of Theorem 1. To begin with, the following naïve observation is implicitly used throughout the proof.

Observation 1. *If A is an agreement subtree of $T|Q$ and $S|Q$, where $Q \subset \mathsf{Le}(T) = \mathsf{Le}(S)$, then A is an agreement subtree of T and S (in both rooted and unrooted cases).*

Further, Observation 2 helps understanding our construction.

Observation 2. *A rooted X-tree is a caterpillar if and only if there exists an ordering of leaves $1, \ldots, |X| = n$ such that for each $1 \leq i < n$ the least common ancestor of set $R_i := \{i+1, \ldots, n\}$ is strictly below the least common ancestor of set $R_i \cup \{i\}$ (or, equivalently, i is incomparable with $\mathsf{lca}(R)$). Further, if such orderings are identical for two rooted X-trees T' and S' then these trees are equivalent caterpillars.*

Proof. Assume such an ordering exists for an X-tree T. For convenience, we identify the leaves in T with their indices in the ordering. Let x_i be the least common ancestor of the set $R_i = \{i+1, \ldots, n\}$ for each $0 \leq i \leq n-2$. Then by the properties of the ordering we have $x_0 \succ x_1 \succ \ldots \succ x_{n-2}$. Since a binary rooted tree has exactly $n-1$ internal nodes, it follows that $(x_0, x_1, \ldots, x_{n-2})$ is a path in T and x_0 is the root node. Further, note that each x_i has at most one non-leaf child (and therefore it has at least one leaf child). Hence, T is a caterpillar tree by definition. Moreover, x_0 must be adjacent to leaf 1, x_1 to leaf 2 and so on (x_{n-2} is adjacent to $n-1$ and n). That is, a leaf ordering uniquely defines a caterpillar tree. Therefore, if two trees share such ordering, then they are equivalent caterpillar trees.

Finally, given a rooted caterpillar tree, let $(x_0, x_1, \ldots, x_{n-2})$ be the path from the root down to the lowest internal node. Further, let l_i be the leaf child of x_i for each $0 \leq i \leq n-3$ and let l_{n-2} and l_{n-1} be the children of x_{n-2}. Then it is not difficult to see that $l_0, l_1, \ldots, l_{n-1}$ is a proper ordering that satisfies the properties listed in the observation. □

We now turn to the construction.

Set up. Consider the left-to-right leaf orderings $\mathsf{Ord}(T')$ and $\mathsf{Ord}(S')$ of T' and S' respectively and let α then be a common subsequence of $\mathsf{Ord}(T')$ and $\mathsf{Ord}(S')$ (or of $\mathsf{Ord}(T')$ and $\mathsf{Ord}(S')$-reversed) of size at least $\sqrt{n}$. Note that α is guaranteed to exist by the Erdős–Szekeres theorem [6] (see Remark 1). If α is common to $\mathsf{Ord}(T')$ and $\mathsf{Ord}(S')$-reversed, then swap left and right children for all internal nodes in S', which would then make α common to $\mathsf{Ord}(T')$ and $\mathsf{Ord}(S')$. Let $X^{(1)} := \{x \mid x \in \alpha\}$ and let $T^{(1)} := T'|X^{(1)}$ and $S^{(1)} := S'|X^{(1)}$. Note that $\mathsf{Ord}(T^{(1)}) \equiv \mathsf{Ord}(S^{(1)})$.

Remark 1. The original Erdős–Szekeres theorem implies that a sequence of distinct integers of size n has a monotonically increasing or a monotonically decreasing subsequence of size at least $\sqrt{n}$. Then, given two sequences S_1 and S_2 over $\{1, \ldots, n\}$, let us bijectively re-label the elements of the sequences in the way that S_1 becomes $(1, \ldots, n)$. Consider S_2 after the re-labeling; it either has a common subsequence with S_1 of size $\sqrt{n}$ (i.e., S_2 has a monotonically increasing subsequence of size $\sqrt{n}$) or S_2-reversed has a common subsequence with S_1 of size $\sqrt{n}$ (i.e., S_2 has a monotonically decreasing subsequence of size $\sqrt{n}$) by the Erdős–Szekeres theorem.

For convenience of the analysis, we present our construction as an iterative algorithm: on each iteration it either locates a large ($\log n$) agreement caterpillar or adds a new leaf to an agreement set M and proceeds to the next iteration with a restricted leaf-set. Next, we describe it more formally.

Iteration description.

Input: rooted $X^{(i)}$-trees $T^{(i)}$ and $S^{(i)}$ with the same leaf orderings, a set of agreement leaves M with $M \cap X^{(i)} = \emptyset$

Outcome: Either
(i) finds a leaf $x \in X^{(i)}$ and a set $Y \subset X^{(i)}$, such that $\mathsf{lca}_{T^{(i)}}(Y) \prec \mathsf{lca}_{T^{(i)}}(Y \cup \{x\})$, $\mathsf{lca}_{S^{(i)}}(Y) \prec \mathsf{lca}_{S^{(i)}}(Y \cup \{x\})$, and $|Y| \geq \frac{|X^{(i)}|}{2 \log n}$, or
(ii) finds an agreement caterpillar for original trees T and S of size at least $\log n$.
In the former case the construction proceeds to the next iteration by adding x to M and setting $T^{(i+1)} = T^{(i)}|Y$, $S^{(i+1)} = S^{(i)}|Y$. In the latter case the iteration stops, as an agreement subtree satisfying Theorem 1 was located

For convenience, we now define a *good pair* as follows.

Definition 1. We say that a pair $(x \in X^{(i)}, Y \subset X^{(i)})$ is a good pair if it has the properties from Outcome (i) above. That is, $\mathsf{lca}_{T^{(i)}}(Y) \prec \mathsf{lca}_{T^{(i)}}(Y \cup \{x\})$, $\mathsf{lca}_{S^{(i)}}(Y) \prec \mathsf{lca}_{S^{(i)}}(Y \cup \{x\})$, and $|Y| \geq \frac{|X^{(i)}|}{2 \log n}$.

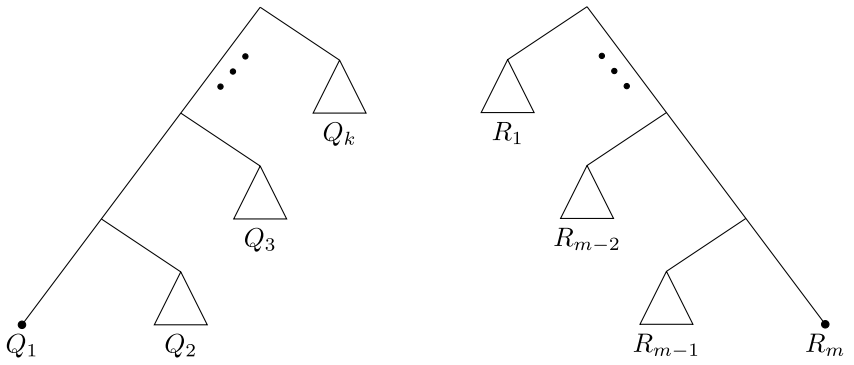

Fig. 3. Schematic definition of $Q_1, \dots, Q_k, R_1, \dots, R_m$ subtrees from trees $T^{(i)}$ (left) and $S^{(i)}$ (right).

Next, we demonstrate that such an iterative algorithm always exists.

Constructive proof. To begin with, without loss of generality assume that the left subtree of $T^{(i)}$ (subtree rooted at the left child of the root) is larger than or equal to the right subtree of $T^{(i)}$ in terms of the number of nodes. If that is not the case, then swap left and right children of all internal nodes in both $T^{(i)}$ and $S^{(i)}$ – that will preserve the equivalence of leaf orderings of $T^{(i)}$ and $S^{(i)}$.

Next, let $P_1 = (u_1, \dots, u_k)$ be the path in $T^{(i)}$ from the left-most leaf to the root and $P_2 = (w_1, \dots, w_m)$ be the path in $S^{(i)}$ from the root to the right-most leaf. Then let $Q_1, \dots, Q_k$ and $R_1, \dots, R_m$ be the subtrees determined by paths P_1 and P_2 respectively (see the illustration on Fig. 3). That is, we define Q_j for some $1 \le j \le k$ (and similarly we define R_j) as follows:

$$Q_j := \begin{cases} T^{(i)}_{u_j} & \text{if } u_j \text{ is the lowest node in } P_1 \text{ (which is } u_1 \text{ in that case);} \\ T^{(i)}_{v} & \text{otherwise, where } v \text{ is the child of } u_j \text{ that is not on the path.} \end{cases}$$

Note that Q_1 and R_m are trivial subtrees that contain only one leaf each. Additionally, note that subtrees are chosen in the way such that leaves in Q_i (or R_i) are to the left of leaves in Q_j (or R_j) in the common leaf ordering if $i < j$.

Lemma 1. *For any fixed constant $C \ge 4$ at least one of the two statements always holds:*

(i) There exists $u \in T^{(i)}$, $v \in S^{(i)}$, and $x \in X^{(i)}$ such that $x \notin \mathsf{Le}(T^{(i)}_u)$, $x \notin \mathsf{Le}(S^{(i)}_v)$, and $\left|\mathsf{Le}(T^{(i)}_u) \cap \mathsf{Le}(S^{(i)}_v)\right| \ge \frac{|X^{(i)}|}{C}$.

(ii) $|Q_j|, |R_l| \le \max\left(\frac{2|X^{(i)}|}{C}, 1\right)$ for all $1 \le j \le k$, $1 \le l \le m$.

Proof. If $\frac{2|X^{(i)}|}{C} < 2$ then the statement is trivially true. In particular, Case (i) holds when $|X^{(i)}| > 1$ (choose u, v to be the same leaf and x to be any other leaf) and Case (ii) must hold otherwise. Assume now that $\frac{2|X^{(i)}|}{C} \ge 2$; it is sufficient to show that if (ii) does not hold, then (i) must hold. Without loss of generality assume that there exists $1 < j \le k$ such that $|Q_j| > \frac{2|X^{(i)}|}{C}$ (note that $j \ne 1$, since $|Q_1| = 1$). Consider now the left and right subtrees of $S^{(i)}$, $S^{(i)}_l = R_1$ and $S^{(i)}_r$ respectively; i.e., subtrees rooted at the children of $\rho(S^{(i)})$. We consider two cases.

- $\mathbf{1 < j < k}$. Assume that the intersection of $\mathsf{Le}(Q_j)$ with $\mathsf{Le}(S^{(i)}_l)$ contains at least $\frac{|X^{(i)}|}{C}$ leaves; then choose $u := \rho(Q_j)$, $v := \rho(S^{(i)}_l)$, and x to be the *right-most* leaf in the leaf ordering. Clearly, x does not belong to Q_j, since $j < k$ and x does not belong to $S^{(i)}_l$ since x is located in the right subtree of $S^{(i)}$. That is, Case (i) of our lemma holds.
  Otherwise, the intersection of $\mathsf{Le}(Q_j)$ with $\mathsf{Le}(S^{(i)}_r)$ must contain at least $\frac{|X^{(i)}|}{C}$ leaves. Then choose $u := \rho(Q_j)$, $v := \rho(S^{(i)}_r)$, and x to be the *left-most* leaf in the leaf ordering. For symmetric arguments Case (i) of our lemma holds again.
- $\mathbf{j = k}$. If the intersection of $\mathsf{Le}(Q_k)$ with $\mathsf{Le}(S^{(i)}_r)$ contains at least $\frac{|X^{(i)}|}{C}$ leaves then clearly Case (i) holds if we choose x to be, for example, the left-most leaf.
  Otherwise, assume that the size of the intersection of $\mathsf{Le}(Q_k)$ with $\mathsf{Le}(S^{(i)}_r)$ is strictly smaller than $\frac{|X^{(i)}|}{C}$. It then follows that $|S^{(i)}_r| < \frac{|X^{(i)}|}{C} \le \frac{|X^{(i)}|}{4}$; hence, $|S^{(i)}_l| \ge \frac{3}{4}|X^{(i)}|$. Given our initial assumption that the left subtree of $T^{(i)}$ is at least as large as its right subtree, it follows that choosing $u := \rho(T^{(i)}_l)$ (root of the left subtree of $T^{(i)}$), $v := \rho(S^{(i)}_l)$, and x to be the right-most leaf satisfies conditions of Case (i) of our lemma. □

Note that when Case (i) holds in the above lemma, the pair $(x, \mathsf{Le}(T^{(i)}_u) \cap \mathsf{Le}(S^{(i)}_v))$ is a good pair for large enough n (i.e., with $2 \log n \ge C \implies n \ge 4$ when $C = 4$). Additionally, note that choosing larger values of C decreases the upper

bound on sizes of $Q_1, \dots, Q_k, R_1, \dots, R_m$ subtrees, when Case (i) of the lemma does not hold. We will exploit this property in the next section, when proving our main result. As for this section, we can consider C to be equal to 4.

Lemma 2. *Let $T^{(i)}$ and $S^{(i)}$ be two $X^{(i)}$-trees of size $n' = |X^{(i)}|$ with the same leaf orderings, and let $Q_1, \dots, Q_k$ and $R_1, \dots, R_m$ be their subtrees as shown in Fig. 3. If each subtree $Q_1, \dots, Q_k$ and $R_1, \dots, R_m$ is of size smaller than $\frac{n'}{\log n}$ then de-rooted $T^{(i)}$ and $S^{(i)}$ agree on a caterpillar tree of size at least $\log n$.*

Proof. For convenience, we identify the leaves in $T^{(i)}$ and $S^{(i)}$ with their indices, $1, \dots, n'$, from the common left-to-right leaf ordering. Then we are going to construct a set $A \subseteq \{1, \dots, n'\}$ such that $|A| \geq \log n$ and A contains at most one leaf from each of the subtrees $Q_1, \dots, Q_k, R_1, \dots, R_m$. It is not then difficult to see that $T^{(i)}|A$ and $S^{(i)}|A$ (and hence $T|A$ and $S|A$) are caterpillars, which must be equivalent after de-rooting, since the leaf orderings of $T^{(i)}$ and $S^{(i)}$ are equivalent.

Observe that the leaves from each subtree $Q_1, \dots, Q_k, R_1, \dots, R_m$ represent an integer interval within $[1, \dots, n']$ of size at most $\frac{n'}{\log n}$; moreover, these intervals are ordered from left to right in the same way as the subtrees are. Now construct set A as follows:

Algorithm 1 $\Theta(\log n)$ agreement set

1: $h := 1, A := \emptyset$;
2: **while** $h \leq n'$ **do**
3: Add h to A;
4: Let Q_j and R_l be the subtrees that contain leaf h;
5: Let r_1 and r_2 be the largest leaves from Q_j and R_l respectively;
6: $h := \max(r_1, r_2) + 1$.
7: **end while**

Note that the above algorithm does not add more than one leaf to A from the same subtree. Further, in Line 6 h increases by at most $\frac{n'}{\log n}$; thus, the size of A in the end of the loop is at least $\log n$. □

Observe that if de-rooted $T^{(i)}$ and $S^{(i)}$ agree on a caterpillar subtree of size at least $\log n$, as implied by Lemma 2, then the original trees T and S also agree on that subtree.

Lemma 3. *Assume that for some fixed $C \geq 4$ we have $|Q_j|, |R_l| \leq \frac{2|X^{(i)}|}{C}$ for all $1 \leq j \leq k$, $1 \leq l \leq m$ (that is, Case (ii) from Lemma 1 holds) and at least one of the subtrees $Q_1, \dots, Q_k, R_1, \dots, R_m$ is of size at least $\frac{|X^{(i)}|}{\log n}$. Then there exists a good pair (x, Y).*

Proof. The proof structure resembles the one of Lemma 1. Without loss of generality assume that Q_j is a tree of size $\geq \frac{|X^{(i)}|}{\log n}$ (note that $j \neq 1$ since $|Q_1| = 1$). We then distinguish two cases.

First, assume that $j < k$; consider the left and right subtrees, $S_l^{(i)}$ and $S_r^{(i)}$, of $S^{(i)}$ and let $F \in \{S_l^{(i)}, S_r^{(i)}\}$ be the subtree with $|\mathsf{Le}(Q_j) \cap \mathsf{Le}(F)| \geq \frac{|\mathsf{Le}(Q_j)|}{2}$. If $F = S_l^{(i)}$ then choose x to be the right-most leaf in the common leaf ordering. Otherwise, when $F = S_r^{(i)}$, choose x to be the left-most leaf. It is then not difficult to see that $(x, \mathsf{Le}(Q_j) \cap \mathsf{Le}(F))$ is a good pair.

Finally, assume that $j = k$; then note that $S_l^{(i)} = R_1$ and by our assumption $|R_1| \leq \frac{2|X^{(i)}|}{C} \leq \frac{|X^{(i)}|}{2}$. Similarly, we have $|Q_k| \leq \frac{|X^{(i)}|}{2}$ and given that $\mathsf{Le}(R_1)$ is 'on the left', while $\mathsf{Le}(Q_k)$ in 'on the right', we have $\mathsf{Le}(R_1) \cap \mathsf{Le}(Q_k) = \emptyset$. Then choose x to be any leaf from $\mathsf{Le}(R_1)$ and $Y = \mathsf{Le}(Q_k)$. Clearly, (x, Y) is a good pair. □

Combining Lemmas 1, 2, and 3 we have the following corollary.

Corollary 1. *At least one of the following statements holds.*

(1) There is a good pair (x, Y) with $|Y| \geq \frac{|X^{(i)}|}{C}$;
(2) There is a good pair (x, Y) (recall that, by definition, this implies that $|Y| \geq \frac{|X^{(i)}|}{2 \log n}$);
(3) de-rooted $T^{(i)}$ and $S^{(i)}$ (and therefore original T and S) agree on a caterpillar of size at least $\log n$.

While it is not necessary for this section, we distinguish Cases (1) and (2) above as we will use them separately later for the proof of our main result. Corollary 1 then implies that we can have an algorithm fitting our original iteration description; Algorithm 2 presents it.

Note that by Observation 2, $T^{(1)}|M$ and $S^{(1)}|M$ (and hence $T'|M$ and $S'|M$) must be equivalent caterpillar trees. We now find the lower bound on the size of the returned set M (given that Lines 7 and 8 are not encountered). Note that we have

Algorithm 2 Locating an agreement caterpillar

Input: rooted $X^{(1)}$-trees $T^{(1)}, S^{(1)}$ with the same leaf orderings ($\mathsf{Ord}(T^{(1)}) = \mathsf{Ord}(S^{(1)})$).
$M := \emptyset, i := 1$;
while $|X^{(i)}| > 1$ **do**
  if there exists a good pair (x, Y) **then**
    Add x to M and set $X^{(i+1)} := Y, T^{(i+1)} := T^{(i)}|Y, S^{(i+1)} := S^{(i)}|Y$;
  else
    There must exist a set A, such that original trees T and S agree on A and $|A| \geq \log n$;
    return A.
  end if
end while
return M.

$|X^{(i+1)}| \geq \frac{|X^{(i)}|}{2\log n}$ and assume that the number of iterations performed is p. Then $|X^{(p+1)}| = 1$ and

$$\begin{aligned} & |X^{(p+1)}| \cdot (2\log n)^p = (2\log n)^p \geq |X^{(1)}| \geq \sqrt{n} \\ \Longrightarrow \quad & p\log(2\log n) \geq \tfrac{1}{2}\log n \\ \Longrightarrow \quad & p \geq \tfrac{1}{2}\frac{\log n}{\log\log n + 1} \geq \tfrac{1}{4}\frac{\log n}{\log\log n}, \end{aligned}$$

with the last inequality holding for $n \geq 4$.

Remark 2. As a corollary of Theorem 1 we have $\mathsf{mast}(n) \in \Omega(\frac{\log n}{\log\log n})$. However, a stronger result can be obtained as we demonstrate in the next section.

4. Asymptotics of mast(n)

We are going to refine the analysis presented in the previous section in order to obtain our main result.

Theorem 2. $\mathsf{mast}(n) \in \Theta(\log n)$.

The upper bound of $\mathsf{mast}(n) \in O(\log(n))$ was shown by Kubicka et al. [7]. To observe this result consider a balanced tree and a caterpillar tree; the maximum agreement subtree must then have the caterpillar shape and the size of such caterpillar is bounded by the length of the longest path in the balanced tree, which is $O(\log(n))$.

We now show that $\mathsf{mast}(n) \in \Omega(\log n)$. To do that we re-use the set up from the previous section. That is, we focus on rooted $X^{(1)}$-trees $T^{(1)}$ and $S^{(1)}$ with the same leaf orderings and of size at least $\sqrt{n}$. Further, we re-use a similar iteration methodology for construction of an agreement tree; and same definitions of Q_j and R_l subtrees for each iteration.

First, we prove the following technical lemma for each iteration i.

Lemma 4. *If a subtree Q_j is of size at least $A > n^c$ for some constant $c > 0$ then either*

(i) There exists a subtree R_p in $S^{(i)}$ such that $|\mathsf{Le}(Q_j) \cap \mathsf{Le}(R_p)| \geq \frac{A}{n^c}$, or
(ii) S and T agree on at least $\frac{1}{3}c\log n$ leaves.

The proof of Lemma 4 uses the following result established by Martin and Thatte [8] and based on the earlier work by Steel and Székely [10].

Proposition 1 (*Martin and Thatte [8]; Steel and Székely [10]*)**.** *Any two unrooted X-trees T and S on n leaves, where T is a caterpillar, have a maximum agreement subtree of size at least $\frac{1}{3}\log n$.*

Proof of Lemma 4. If (i) does not hold, then for all $1 \leq l \leq m$: $|\mathsf{Le}(Q_j) \cap \mathsf{Le}(R_l)| < \frac{A}{n^c}$. In that case m has to be at least n^c; therefore, taking a single leaf from each R_l for $l = 1 \ldots m$ will produce a set M such that $S|M$ is a caterpillar (see Fig. 3) and $|M| \geq n^c$. Hence, by Proposition 1, $S|M$ and $T|M$ must agree on at least $\frac{1}{3}c\log n$ leaves, and Case (ii) of our lemma holds. □

Next, recall that Lemma 1 from the previous section allows us to choose a constant C, which we set to $C := 40$ in this section. We now refine Lemma 2.

Lemma 5. *Assume that $|X^{(i)}| \geq n^{\frac{1}{4}}$ and $|Q_j|, |R_l| \leq \frac{2|X^{(i)}|}{C} = \frac{|X^{(i)}|}{20}$ for all $1 \leq j \leq k$, $1 \leq l \leq m$ (i.e., Case (ii) from Lemma 1 holds); then at least one of the following statements holds (for sufficiently large n).*

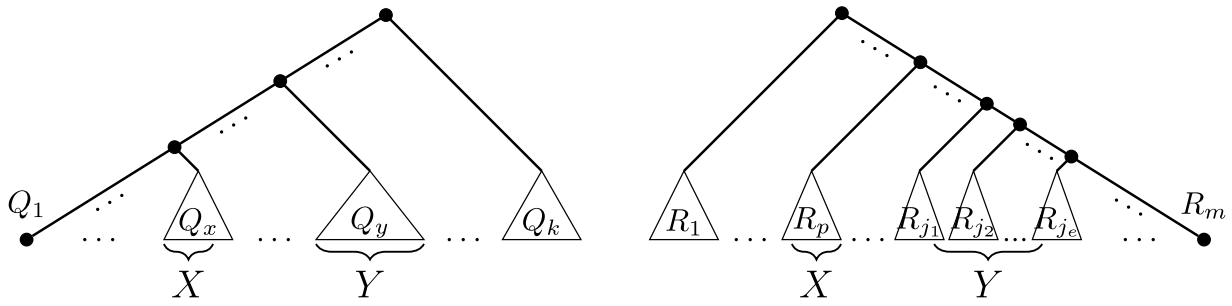

Fig. 4. An illustration of the potential structure of $T^{(i)}$ and $S^{(i)}$ trees for the proof of Lemma 5 – e.g., when Case (ii) from that lemma does not hold.

(i) There exist disjoint sets $X, Y \subset X^{(i)}$ such that $|X| \geq n^{1/16}$, $|Y| \geq \frac{|X^{(i)}|}{10 \log n}$, $\mathsf{lca}_{T^{(i)}}(X)$ is incomparable with $\mathsf{lca}_{T^{(i)}}(Y)$, and $\mathsf{lca}_{S^{(i)}}(X)$ is incomparable with $\mathsf{lca}_{S^{(i)}}(Y)$;

(ii) T and S agree on at least $\frac{1}{48} \log n$ leaves (that induce a caterpillar).

Proof. Similarly to the proof of Lemma 2, we identify leaves in $T^{(i)}$ and $S^{(i)}$ with their indices, $1, \ldots, |X^{(i)}| = n'$, in their common left-to-right leaf ordering. Then each subtree Q_j or R_l induces an integer interval inside $[1, n']$ of size at most $n'/20$. Consider now the leaves in the $I := [\frac{8}{20}n', \frac{12}{20}n']$ interval. Let l be the smallest leaf (integer) from I such that the subtrees Q_{i_1} and R_{h_1} that contain l 'lie' completely within I (that is, $\mathsf{Le}(Q_{i_1})$ and $\mathsf{Le}(R_{h_1})$ are within I). Observe that $l \leq \frac{9}{20}n'$. Similarly, we define r to be the largest leaf (integer) from I such that subtrees Q_{i_s} and R_{h_t} that contain r lie completely within I; then $r \geq \frac{11}{20}n'$. We now focus on subtrees $Q_{i_1}, \ldots, Q_{i_s}$ and $R_{h_1}, \ldots, R_{h_t}$. Observe that $\bigcup_{l=i_1}^{i_s} \mathsf{Le}(Q_l)$ and $\bigcup_{l=h_1}^{h_t} \mathsf{Le}(R_l)$ are supersets of $\{l, \ldots, r\}$ and $|\{l, \ldots, r\}| \geq \frac{2}{20}n'$. By applying Lemma 2 to trees $T^{(i)}|\{l, \ldots, r\}$ and $S^{(i)}|\{l, \ldots, r\}$, we know that either at least one of the $Q_{i_1}, \ldots, Q_{i_s}, R_{h_1}, \ldots, R_{h_t}$ subtrees is of size at least $\frac{(2/20)n'}{\log n}$, or T and S agree on a caterpillar of size at least $\log n$ (i.e., Case (ii) of Lemma 5 holds). Assume now that the former holds and let Q_y be a subtree of size at least $\frac{(2/20)n'}{\log n}$. Note that we can assume that such Q_y exists without loss of generality. That is, in case there is no such Q_y, a subtree of size at least $\frac{(2/20)n'}{\log n}$ must be present among the $R_{h_1}, \ldots, R_{h_t}$ subtrees, and let R_y be such subtree. We then reduce this case to the case when Q_y of size at least $\frac{(2/20)n'}{\log n}$ does exist. In particular, we rotate each tree $T^{(i)}$ and $S^{(i)}$ such that their resulting common leaf ordering becomes the reverse of the original ordering (that is, we swap left and right children of each non-leaf node in the trees). As a result, subtrees $R_1, \ldots, R_m$ virtually take place of the $Q_1, \ldots, Q_k$ subtrees and vice versa. Renaming $R_1, \ldots, R_m$ as $Q_1, \ldots, Q_m$ and swapping labels m and k would then assure that Q_y of size at least $\frac{(2/20)n'}{\log n}$ exists and $\mathsf{Le}(Q_y)$ is within $[\frac{8}{20}n', \frac{12}{20}n']$.

Let us now focus on the $[1, \frac{4}{20}n']$ interval. Applying Lemma 2 to $T^{(i)}|\{1, \ldots, \lceil \frac{4}{20}n' \rceil\}$ and $S^{(i)}|\{1, \ldots, \lceil \frac{4}{20}n' \rceil\}$, we know that there must exist a subtree Q_x (or R_x, which we disregard for symmetry) of size at least $\frac{(4/20)n'}{\log n}$ and $\mathsf{Le}(Q_x)$ lies within $[1, \frac{5}{20}n')$ – otherwise, Case (ii) of our Lemma holds. Recall that $n' \geq n^{\frac{1}{4}}$; then for sufficiently large n we have $\frac{(4/20)n'}{\log n} \geq n^{\frac{1}{8}}$.

Consider Lemma 4 with $Q_j = Q_x$ and $c = \frac{1}{16}$. Then there either exists R_p such that $|\mathsf{Le}(R_p) \cap \mathsf{Le}(Q_x)| \geq |Q_x|/n^{\frac{1}{16}} \geq n^{\frac{1}{16}}$, or S and T agree on at least $\frac{1}{48} \log n$ leaves (i.e., Case (ii) holds). Assume now that such R_p exists. Clearly, $\mathsf{Le}(R_p)$ is within the $[1, \frac{6}{20}n')$ interval and therefore $\mathsf{Le}(R_p)$ does not intersect with $\mathsf{Le}(Q_y)$.

Summing up the above arguments, define $Y := \mathsf{Le}(Q_y)$ and $X := \mathsf{Le}(Q_x) \cap \mathsf{Le}(R_p)$. We claim that these two sets satisfy condition (i) of our lemma. The conditions on size are satisfied by the construction; hence, we only need to confirm the incomparability conditions. Note that $\mathsf{lca}_{T^{(i)}}(X)$ and $\mathsf{lca}_{T^{(i)}}(Y)$ are located within the Q_x and Q_y subtrees respectively ($x \neq y$) and therefore are incomparable. Further, let $R_{j_1}, \ldots, R_{j_e}$ be the subtrees that intersect on leaves with the set Y; given that Y is within $[\frac{8}{20}n', \frac{12}{20}n']$ we have $\mathsf{Le}(R_{j_l})$ lying within $(\frac{7}{20}n', \frac{13}{20}n')$ for all $1 \leq l \leq e$. Note now that $\mathsf{lca}_{S^{(i)}}(X)$ is within the R_p subtree and $\mathsf{Le}(R_p)$ precedes the $(\frac{7}{20}n', \frac{13}{20}n')$ interval. Hence, if v_p is the parent of the root of R_p then $\mathsf{lca}_{S^{(i)}}(Y)$ must be located below it and $\mathsf{lca}_{S^{(i)}}(X)$ is incomparable with $\mathsf{lca}_{S^{(i)}}(Y)$ (see Fig. 4 for an illustration). That is, Case (i) holds. □

If Case (ii) from the above lemma holds, then Theorem 2 clearly holds as well. Otherwise, see Observation 3.

Observation 3. *Assume that Case (i) of Lemma 5 holds. That is, there exist disjoint sets $X, Y \subset X^{(i)}$ such that $|X| \geq n^{1/16}$, $|Y| \geq \frac{|X^{(i)}|}{10 \log n}$, and $\mathsf{lca}(X)$ is incomparable with $\mathsf{lca}(Y)$ in both $T^{(i)}$ and $S^{(i)}$. Then due to Theorem 1 either (1) trees $T^{(i)}|X$ and $S^{(i)}|X$ agree on a rooted caterpillar of size at least*

$$\frac{\log n^{\frac{1}{16}}}{4 \log \log n^{\frac{1}{16}}} = \frac{1}{16 \cdot 4} \frac{\log n}{\log \log n - \log 16} \geq \frac{1}{64} \frac{\log n}{\log \log n}$$

or (2) T and S agree on at least a $\log n^{\frac{1}{16}} = \frac{1}{16} \log n$ caterpillar.

Algorithm 3 then summarizes all the above results.

Algorithm 3 $\Omega(\log n)$ MAST

1: **Input:** rooted $X^{(1)}$-trees $T^{(1)}, S^{(1)}$ with the same leaf orderings.
2: $M := \emptyset, i := 1, C := 40$;
3: **while** $|X^{(i)}| \geq n^{\frac{1}{4}}$ **do**
4: **if** there exists a good pair (x, Y) with $|Y| \geq \frac{|X^{(i)}|}{C}$ **then**
5: Add x to M and set $X^{(i+1)} := Y, T^{(i+1)} := T^{(i)}|Y, S^{(i+1)} := S^{(i)}|Y$;
6: **else if** there exists a pair (X, Y) as described in Lemma 5, Case (i) **then**
7: **if** $T^{(i)}|X$ and $S^{(i)}|X$ agree on $M' \geq \frac{1}{64}\frac{\log n}{\log\log n}$ leaves **then**
8: Add leaves from M' to M and
9: Set $X^{(i+1)} := Y, T^{(i+1)} := T^{(i)}|Y, S^{(i+1)} := S^{(i)}|Y$;
10: **else**
11: By Observation 3 there exists a leaf-set A ($|A| \geq \frac{1}{16}\log n$) such that T and S agree on A;
12: **return** A.
13: **end if**
14: **else**
15: By Lemma 5 there exists a leaf-set A ($|A| \geq \frac{1}{48}\log n$) such that T and S agree on A;
16: **return** A.
17: **end if**
18: **end while**
19: **return** M.

If the construction presented in Algorithm 3 exits on lines 12 or 16 then we directly get at least a $\frac{1}{48}\log n$ agreement subtree. Assume now that these lines are never reached and the algorithm returns the set M. It is not difficult to see that $T^{(1)}|M$ should be equivalent to $S^{(1)}|M$ and therefore T and S agree on M. This can be seen by considering the following observation (a generalization of Observation 2).

Observation 4. *Let $(X_1, \ldots, X_p)$ be an ordered partition of X and let rooted X-trees T' and S' have the following properties:*

- $\mathsf{lca}_{T'}(X_i)$ *is incomparable with* $\mathsf{lca}_{T'}\left(\bigcup_{j=i+1}^{p} X_j\right)$ *for all* $1 \leq i < p$;
- *Similarly,* $\mathsf{lca}_{S'}(X_i)$ *is incomparable with* $\mathsf{lca}_{S'}\left(\bigcup_{j=i+1}^{p} X_j\right)$ *for all* $1 \leq i < p$;
- $T'|X_i = S'|X_i$ *for all* $1 \leq i \leq p$.

Then $T' = S'$.

Let us now determine the lower bound on the size of M. Assume that line 5 is executed p times overall, while line 9 is executed q times. The size of M is then at least $p + q \cdot \frac{1}{64}\frac{\log n}{\log\log n}$. Further, let $X^{(p+q+1)}$ be the set of leaves after the last iteration of the algorithm (i.e., $|X^{(p+q+1)}| < n^{\frac{1}{4}}$). We then have

$$\begin{aligned}
& |X^{(p+q+1)}| \cdot C^p \cdot (10\log n)^q \geq |X^{(1)}| \geq \sqrt{n} \\
\Longrightarrow \quad & n^{\frac{1}{4}} C^p \cdot (10\log n)^q \geq \sqrt{n} \\
\Longrightarrow \quad & p\log C + q(\log\log n + \log 10) \geq \tfrac{1}{4}\log n \\
\Longrightarrow \quad & p \geq \tfrac{1}{\log C}\left(\tfrac{1}{4}\log n - q\log\log n - q\log 10\right).
\end{aligned}$$

Finally,

$$|M| \geq p + q \cdot \frac{1}{64}\frac{\log n}{\log\log n} \geq \frac{\log n}{4\log C} + q\left(\frac{1}{64}\frac{\log n}{\log\log n} - \frac{\log\log n}{\log C} - \frac{\log 10}{\log C}\right).$$

Note that

$$\frac{1}{64}\frac{\log n}{\log\log n} \geq \frac{\log\log n}{\log C} + \frac{\log 10}{\log C}$$

for sufficiently large n; hence $|M| \geq \frac{\log n}{4\log C}$ for large n and Theorem 2 holds.

Observation 5. *Algorithm 3 returns an agreement set of size at least $\frac{1}{48}\log n$. That is, a pair of unrooted trees over the same leaf set of size n must agree on at least $\frac{1}{48}\log n$ leaves for sufficiently large n.*

5. Conclusion

In this work, we solve the long-standing open problem that asks how many leaves two unrooted phylogenetic trees must agree on (asymptotically) and demonstrate that the answer is $\Theta(\log n)$. More precisely, we show that this value is lower bounded by $\frac{1}{48}\log n$ for large n. While we note that the presented arguments can be adjusted to obtain a constant better than 1/48, the question of finding an extremal example and a tight constant remains open. The problem addressed in this work relates to another prominent question that asks what the expected size of the maximum agreement set for a pair of randomly sampled rooted or unrooted trees is. While non-trivial lower- and upper-bounds were established for this problem under the uniform and Yule distributions [2,4], the precise asymptotics are yet to be found. The expected values are conjectured to be $\Theta(\sqrt{n})$ for both distributions. Further, Misra and Sullivant [9] proved that, when a tree shape is fixed, the expected size of a maximum agreement set is indeed $\Theta(\sqrt{n})$.

CRediT authorship contribution statement

Alexey Markin: Conceptualization, Methodology, Writing - Original Draft, Writing - Review & Editing, Visualization.

Acknowledgments

I would like to thank the anonymous reviewers as well as Oliver Eulenstein and Mike Steel for their valuable comments that helped improve the quality of this manuscript. This material is based upon work supported by the National Science Foundation, USA under Grant No. 1617626.

References

[1] A. Amir, D. Keselman, Maximum agreement subtree in a set of evolutionary trees: Metrics and efficient algorithms, SIAM J. Comput. 26 (6) (1997) 1656–1669.
[2] D.I. Bernstein, L.S. Tung Ho, C. Long, M. Steel, K.S. John, S. Sullivant, Bounds on the expected size of the maximum agreement subtree, SIAM J. Discrete Math. 29 (4) (2015) 2065–2074.
[3] M. Bordewich, S. Linz, M. Owen, K.S. John, C. Semple, K. Wicke, On the maximum agreement subtree conjecture for balanced trees, 2020, arXiv preprint arXiv:2005.07357.
[4] D. Bryant, A. McKenzie, M. Steel, The size of a maximum agreement subtree for random binary trees, DIMACS Ser. Discrete Math. Theoret. Comput. Sci. 61 (2003) 55–66.
[5] R. Cole, M. Farach-Colton, R. Hariharan, T. Przytycka, M. Thorup, An $O(n \log n)$ algorithm for the maximum agreement subtree problem for binary trees, SIAM J. Comput. 30 (5) (2000) 1385–1404.
[6] P. Erdős, G. Szekeres, A combinatorial problem in geometry, Compos. Math. 2 (1935) 463–470.
[7] E. Kubicka, G. Kubicki, F. McMorris, On agreement subtrees of two binary trees, Congr. Numer. (1992) 217.
[8] D.M. Martin, B.D. Thatte, The maximum agreement subtree problem, Discrete Appl. Math. 161 (13–14) (2013) 1805–1817.
[9] P. Misra, S. Sullivant, Bounds on the expected size of the maximum agreement subtree for a given tree shape, SIAM J. Discrete Math. 33 (4) (2019) 2316–2325.
[10] M. Steel, L.A. Székely, An improved bound on the maximum agreement subtree problem, Appl. Math. Lett. 22 (11) (2009) 1778–1780.
[11] M. Steel, T. Warnow, Kaikoura tree theorems: Computing the maximum agreement subtree, Inform. Process. Lett. 48 (2) (1993) 77–82.