JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 31, Number 4, 2024
Mary Ann Liebert, Inc.
Pp. 312–327
DOI: 10.1089/cmb.2023.0338

Open camera or QR reader and scan code to access this article and other resources online.

Asymmetric Cluster-Based Measures for Comparative Phylogenetics

SANKET WAGLE,[1,*] ALEXEY MARKIN,[2] PAWEŁ GÓRECKI,[3] TAVIS K. ANDERSON,[2] and OLIVER EULENSTEIN[1,*]

ABSTRACT

Phylogenetic inference and reconstruction methods generate hypotheses on evolutionary history. Competing inference methods are frequently used, and the evaluation of the generated hypotheses is achieved using tree comparison costs. The Robinson–Foulds (RF) distance is a widely used cost to compare the topology of two trees, but this cost is sensitive to tree error and can overestimate tree differences. To overcome this limitation, a refined version of the RF distance called the Cluster Affinity (CA) distance was introduced. However, CA distances are symmetric and cannot compare different types of trees. These asymmetric comparisons occur when gene trees are compared with species trees, when disparate datasets are integrated into a supertree, or when tree comparison measures are used to infer a phylogenetic network. In this study, we introduce a relaxation of the original Affinity distance to compare heterogeneous trees called the asymmetric CA cost. We also develop a biologically interpretable cost, the Cluster Support cost that normalizes by cluster size across gene trees. The characteristics of these costs are similar to the symmetric CA cost. We describe efficient algorithms, derive the exact diameters, and use these to standardize the cost to be applicable in practice. These costs provide objective, fine-scale, and biologically interpretable values that can assess differences and similarities between phylogenetic trees.

Keywords: Cluster Affinity, phylogenetic tree, Robinson–Foulds, supertrees.

1. INTRODUCTION

PHYLOGENETICS IS A FIELD OF STUDY that investigates the evolutionary relationships between different entities such as genes or species. It involves inferring evolutionary trees or ''phylogenies'' using computational methods. These trees represent hypotheses about the relationships between different groups,

[1]Department of Computer Science, Iowa State University, Ames, Iowa, USA.
[2]National Animal Disease Center, USDA-ARS, Ames, Iowa, USA.
[3]Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, Warsaw, Poland.

© Sanket Wagle, et al., 2024. Published by Mary Ann Liebert, Inc. This Open Access article is distributed under the terms of the Creative Commons License (http://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly credited.

and can be based on data from DNA sequences, protein amino acid sequences, or morphological characters (Yang and Rannala, 2012).

Different datasets and tree inference methods may lead to inconsistent hypotheses of evolutionary relationships. To resolve these inconsistencies, tree comparison costs are used to evaluate the fit between different trees and a given dataset (Allen and Steel, 2001; Estabrook et al., 1985; Robinson and Foulds, 1981; Waterman and Smith, 1978). If the true tree is known, it is possible to compare the different inferred trees and identify those that accurately describe the true tree topology or are close to it (Russo et al., 1996). If the true tree is unknown, which is generally the case in biological studies, cost measures can rank trees and identify areas of incongruence that were generated from different input datasets or tree inference methods (e.g., Shen et al., 2021). Objectively identifying these areas may have utility in phylogenomic studies (Prum et al., 2015; Wickett et al., 2014). Furthermore, one can evaluate the robustness of phylogenetic inference and identify areas of uncertainty by comparing the tree costs associated with different analyses and datasets (Lozano-Fernandez, 2022). Therefore, comparative phylogenetics applies such costs to facilitate phylogenomic studies and provide a metric for assessing tree errors and incongruence.

Tree comparison costs have distinct properties such as their distribution, sensitivity to small changes in the tree structure due to data errors or noise, and computability. Therefore, developing tree comparison costs with beneficial characteristics will enhance research in phylogenetics and other areas that rely upon tree-focused inference. For example, comparison costs are used in epidemiology for comparing transmission trees and virus genealogies (Giardina et al., 2017), to determine horizontal gene transfer events (Bogdanowicz and Giaro, 2017), and in natural language processing to compare and determine an aggregate parse tree (Kulkarni et al., 2022). Consequently, the development and improvement of tree comparison costs is a mature and highly active research area in computational and comparative phylogenetics (e.g., Robinson et al., 2016). The Robinson–Foulds (RF) distance is commonly used to compare the topology of two trees. It measures the cardinality of the symmetric difference of the cluster sets represented by each tree.

Although the RF distance can be computed efficiently, it is highly sensitive to differences in trees, and even small variations can result in a significant RF distance. This property of the RF distance makes it challenging to differentiate between meaningful and random differences, which can lead to incorrect conclusions about the similarity of the compared trees. These issues have been previously discussed by Steel and Penny (1993) and Lin et al. (2011). There have been various approaches proposed in the literature to generalize the RF distance measure. These include matching similar splits between trees (Bogdanowicz and Giaro, 2013; Bogdanowicz and Giaro, 2011), normalizing using the Jaccard Index (Nye et al., 2006), or normalizing through the use of an arbitrary exponent (Böcker et al., 2013). While these methods have improved the application of RF-like distance measures, they do not fully account for the inherent asymmetry in many phylogenetic comparison studies.

Phylogenomic studies often encounter asymmetry issues. For instance, when comparing a gene tree to a species tree, the gene tree may not be fully resolved due to low branch support or insufficient phylogenetic signal in the alignment. As a result, the clusters in the species tree may not be present in the gene tree, which leads to asymmetric comparison costs (Page, 2002; Swenson et al., 2011). Asymmetry in phylogenomics also appears due to gene duplications (i.e., paralogous genes) or horizontal gene transfer, and assessing trees objectively can strengthen phylogenomic studies (Lozano-Fernandez, 2022).

In this work, we introduce the *asymmetric Cluster Affinity* (CA) cost to address the problems of RF. This measure involves determining the minimum cost of a cluster in the source tree for each cluster in the target tree, thereby reducing the issues of RF. We also introduce a biologically interpretable cost related to CA, the Cluster Support (CS) cost that normalizes the individual cluster contributions to the cost by the respective cluster sizes in the source tree. Both costs can be efficiently computed in $O(n \log^3 n)$ time for trees of size n by adopting the algorithm by Truszkowski et al. (2019). We describe an $O(n \log n)$ time algorithm to study the robustness of our costs against small phylogenetic error. Phylogenetic studies often involve comparing different trees using various costs, which must be normalized to provide meaningful results. To facilitate this normalization, we derive the diameters of our costs, which can be beneficial for practitioners. Our comparative studies have shown that our costs have a broader distribution range and are less skewed than the RF distance. These tree comparison costs can provide objective and biologically interpretable values to evaluate the similarities and differences between phylogenetic trees at a fine scale.

1.1. Related work

The RF distance (Robinson and Foulds, 1981) is a standard distance for comparing phylogenies due to its low complexity and interpretability. However, since RF has a narrow distribution and is sensitive to tree

errors, there have been many attempts to generalize RF and overcome its shortcomings. A few notable examples are (1) the information-theoretic generalized RF (Smith, 2020) and (2) Matching Split (MS) and Matching Cluster(MC) distances (Bogdanowicz and Giaro, 2013; Bogdanowicz and Giaro, 2011; Lin et al., 2011) that find a minimum perfect matching between the split representation or the cluster representation of two trees. The CA distance later relaxed the matching requirement in the MC distance, thus allowing a one-to-many mapping from one tree to another. This relaxation allowed for a considerably faster distance computation while preserving the robustness to tree error and the wide distribution range that the MC distance has. The CA distance assigns for a pair of clusters c_1 and c_2 a corresponding cost equal to the total number of changes required to change the cluster c_1 into the cluster c_2. The mapping cost for tree T_1 to tree T_2 is the sum of the minimum mapping of all the clusters in T_1 to the clusters in T_2. The CA distance for T_1 and T_2 is the arithmetic mean between the mapping costs from T_1 to T_2 and vice versa.

Another key property of tree comparison measures is the *diameter*, that is, the maximum distance between two trees over a fixed taxon set of size n. It is crucial to understand the diameter to make tree distances comparable for different pairs of trees through normalization using the diameter. Additionally, normalization enables comparison across different tree comparison measurements. For RF, the diameter is known, but the exact diameters of MS, MC, and CA remain an open problem (although Bogdanowicz and Giaro, 2013; Bogdanowicz and Giaro, 2011; Moon and Eulenstein, 2019 provide bounds).

1.2. Our contribution

In this study, we relax the CA cost, introduce an efficient algorithm for computing updates to the cost in response to Nearest Neighbor Interchange (NNI) tree edit operations, and establish an exact diameter for the cost. We adopt two approaches to relax the cost. First, we allow the cost to be asymmetric by only taking into account the cost of CA mapping from T_1 to T_2. This asymmetry appears naturally when comparing different tree types, such as a gene tree to a species tree in a phylogenomic analysis or an estimated tree to the ground truth in simulation/convergence studies. Second, we permit clusters in T_1 to map to *trivial* clusters in T_2 (i.e., leaves and the root cluster). This was not allowed in the original definition (Moon and Eulenstein, 2019).

Note that the CA cost sums up the raw cluster differences across all clusters. That way, a 5% difference in a bigger cluster with 500 taxa may dominate over a 40% difference in a small cluster with just 10 taxa. To overcome this, we define another measure, where CA is normalized by the respective cluster size, which we denote the *CS* cost. This way all clusters contribute the percentage of the distance rather than the raw numbers as in the original CA cost. A 5% overall CS cost means that clusters in T_1 are, on average, 5% different from clusters in T_2.

We developed an algorithm that efficiently conducts an NNI (Bordewich and Semple, 2005) search with the CA or CS cost. This algorithm requires a single preprocessing step that takes $O(n^2)$ time, and the subsequent NNI iterations only take $O(n\log n)$ time each. Then, Subtree Prune and Regraft (SPR) search can be conducted efficiently using an NNI-graph presented in Chaudhary et al. (2012). As NNI and SPR search strategies are standard for species tree inference heuristics (Bininda-Emonds, 2004), these algorithms make the CA and CS costs directly applicable for phylogenomic inference.

Furthermore, we proved that the CA diameter for trees with n leaves is $\lceil \frac{n^2-2n}{4} \rceil$ and the CS diameter is $n - H_{\lceil \frac{n}{2} \rceil} - H_{\lfloor \frac{n}{2} \rfloor}$, where H_i is the i-th harmonic number. We introduced the concept of a *separation diameter* that measures how asymmetric a cost is; that is, the maximum value of $|CA(T_1, T_2) - CA(T_2, T_1)|$ for all trees T_1 and T_2. We prove that the CA cost is significantly impacted by the topology of T_1, and the separation diameter for CA is in the order of $\Theta(n^2)$. This result implies that, in practice, CA costs need to be normalized by a diameter specific to the topology of T_1. While the exact topology-specific diameter remains an open problem, we provide a practical upper bound on that diameter, supported by our theoretical results. Lastly, we demonstrated that our relaxed CA cost and the novel CS cost are similar to or improve upon the original CA formulation in terms of distribution properties and robustness to error.

2. ASYMMETRIC CA COST

A phylogenetic tree T over a taxon set M is a rooted binary tree where each leaf is bijectively labeled with the elements from M. The vertex set and edge set of T are denoted by $V(T)$ and $E(T)$, respectively. By

$L(T)$ we denote the set of all leaves of T, and by the size of T, usually denoted as n, we define the size of $L(T)$.

An edge (u, v) from $E(T)$ is directed from u to v, where v is a *child* of u, and u is the *parent* of v. For a vertex $w \in V(T)$, $\mathrm{Ch}(w)$ is the set of all children of w. If two distinct vertices u and v have the same parent, they are *siblings*. We also define $T(v)$ as the subtree of T rooted at v.

We define the *height* of a node v as the edge-length of the longest path i.e., the path containing the maximum number of edges from the node v to a leaf l such that $l \in L(T(v))$. We define the *height* of a tree T as the height of the root of T.

For two sets A and B, we define $A\Delta B$ as the symmetric difference between them. That is, $A\Delta B = (A \backslash B) \cup (B \backslash A)$.

A set of leaves $L(T(v))$ is called a cluster of the node v and is denoted by c_v. Note that we identify the leaves in a phylogenetic tree with the respective labels (taxa). A tree G can be represented by a set of clusters $C(G) = \{c_i | i \in V(G)\}$.

For convenience, we assume throughout the text that M is a taxon set, G, S, and T are trees over M, and $c \subseteq M$ is a cluster over M.

Definition 1 (CA cost (Moon and Eulenstein, 2019)). CA cost *from c to S is* $d(c, S) := \min_{x \in C(S)} |c \Delta x|$, *and* CA cost *from G to S is* $d(G, S) := \sum_{c \in C(G)} d(c, S)$, *and the* Symmetric CA cost *between G and S is* $d_{sym}(G, S) := \frac{d(G,S) + d(S,G)}{2}$.

Definition 2 (Diameter of cost function). *The diameter of a cost function between trees is the maximum value that the cost function can have over all trees over the same leaf set.*

2.1. Tree edit operations

We define two classic tree edit operations for rooted trees, namely the SPR operation and the NNI operation (Bordewich and Semple, 2005).

Let T be a phylogenetic tree and let $e = (u, v)$ be an edge in $E(T)$ and $w \in V(T)$. Then we define the SPR operation $SPR(T, v, w) = T'$ as the rooted binary tree obtained by deleting e and then adjoining edge f between v and the component C_u that contains u in one of the two following ways:

1. If w is the root of T, we create a new vertex u' and a new edge from u' to w. Then, we adjoin f between u' and v and suppress the degree two vertex u. Then, u' becomes the new root for the tree T'.
2. Otherwise, we create a new vertex, u' which subdivides the edge whose bottom node is w in C_u and adjoining f between u' and v. Then, we either suppress the degree-two vertex u or if u is the root of T, delete u and the edge incident with u, making the other end-vertex of the edge the root.

The *NNI operations* are SPR operations where the subtree is pruned close to its regrafting position as follows. For a nonroot node v from T, let $NNI(T, v)$ be the SPR operation with the unique edge $e = (u, v)$ from T and w being the sibling of u.

2.2. NNI search using CA heuristic

We present an algorithm for the efficient NNI tree space traversal.

Theorem 1. *Given a cluster c and a tree* T_0 *of size n*, let $T_1, T_2, T_3, \ldots$ *be a sequence of trees, such that* T_i *is obtained from* T_{i-1} *by a single NNI operation. Then, after* $O(n)$ *preprocessing steps on* T_0, *one can compute* $d(c, T_i)$ *for each* $i > 0$ *in* $O(\log n)$ *time.*

Proof. In the previous section, we showed that computing $d(c, T_0)$ can be performed in $O(n)$ time. During this computation, we obtain $|c\Delta c_v|$ values for each $v \in V(T_0)$. For convenience, let $d_v := |c\Delta c_v|$. We then place all d_v values in a binary min-heap. Recall that obtaining the minimum value from a binary min-heap can be performed in $O(1)$ time, and changing an element's value can be performed in $O(\log n)$ time. Building a min-heap requires $O(n)$ time.

Now, we show how to compute $d(c, T_1)$ and update the min-heap in $O(\log n)$ time. Then, computation of $d(c, T_2)$, $d(c, T_3)$, … follows the same algorithm. Let $T_1 = NNI(T_0, u)$, where u is not the root of T. Let v be the parent of u, w be the sibling of u, and x be the sibling of v. Note that the only cluster that changes after

the NNI is c_v. The new cluster in T_1 is $c' = c_u \cup c_x$. Note that we can compute $|c \Delta c'|$ in constant time as $|c \Delta c'| = d_u + d_x - |c|$.

Updating d_v in the binary heap by removing the old value and then replacing it with $|c \Delta c'|$, can be done in $O(\log n)$ time. Then, querying the minimum value from the min-heap will give us $d(c, T_1)$ in constant time. □

2.3. Diameter of the CA cost

In this section, we derive the diameter for the asymmetric CA cost. However, first we require a few additional definitions to obtain the diameter of the CA cost.

A *rooted caterpillar tree* is a rooted tree T, where each internal node has at least one leaf child. We define a caterpillar C_n using the standard nested parenthesis notation as $(n, (n-1, \ldots (2, 1) \ldots))$, where the leaves are numbers. A *perfectly balanced tree* is a rooted tree T where each leaf is at the same distance in the number of edges from the root. *The cherry of a caterpillar tree* is the smallest subtree of the caterpillar tree, which contains two leaf nodes. For a tree T of the size n and a node $v \in V(T)$, we define let $\tau_T(v) := \min(n - |c_v|, |c_v| - 1)$ and $\tau(T) := \sum_{v \in V(T)} \tau(v)$.

Definition 3 (Caterpillar-extend). *For a noncaterpillar tree T let v be an internal node in T such that v has two children u and t, T(u) and T(t) are two caterpillars such that* $|L(T(u))| \geq |L(T(t))| \geq 2$. *By* $T \longrightarrow T'$ *we denote the caterpillar-extend operation* $T' = NNI(T, l)$ *on v, where l is a leaf-child of t. See* Figure 1 *for an example of the caterpillar-extend operation.*

Lemma 1. *For every node v in a tree T,* $d(c_v, S) \leq \tau_T(v)$ *and* $d(T, S) \leq \tau(T)$ *for all trees S with the same leaf set as T.*

Proof. Let n be the size of S. Consider a leaf l on tree S such that $l \in c_v$. Hence $|c_v \Delta \{l\}| = |c_v| - 1$. Similarly, for the cluster M of the root of the tree S, we have $|c_v \Delta M| = |M| - |c_v| = n - |c_v|$. Since $d(c_v, S) = \min_{x \in C(S)} |c_v \Delta x|$ and both clusters $\{l\}$ and M exist in $C(S)$, $d(c_v, S) \leq \min(|c_v| - 1, n - |c_v|)$. Recall that we have $\tau_T(v) = \min(|c_v| - 1, n - |c_v|)$ and $\tau(T) = \sum_{v \in V(T)} (\tau(v))$. Hence $d(c_v, S) \leq \tau_T(v)$ and $\forall T', d(T, T') \leq \tau(T)$ □

Lemma 2. If $T \longrightarrow T'$, then $\tau(T) \leq \tau(T')$.

Proof. Let n be the size of T and T'. Let v, u, and t be the nodes from the caterpillar-extend operation definition. Note, that $|c_u| \geq |c_t|$ and $\tau(T) = \sum_{i \in V(T)} \tau_T(i)$.

Let c_w represent the new cluster that is formed in T'. Note that there are only two different clusters between T and T', namely, c_t and c_w. Hence, if $\tau_{T'}(w) \geq \tau_T(t)$ then $\tau(T) \geq \tau(T')$. Since $|c_u| \geq |c_t|$, $|c_t| \leq \frac{n}{2}$ and $\tau_T(t) = |c_t| - 1$. Then, there are two cases for c_w. If $|c_w| \leq \frac{n}{2}$, we have $\tau_{T'}(w) = |c_w| - 1 = |c_u| > \tau_T(t)$. Otherwise, $|c_w| > \frac{n}{2}$ and it follows from $n \geq |c_t| + |c_u|$ that $\tau_T(w) = n - |c_w| = n - |c_u| - 1 \geq \tau_T(t)$. We conclude $\tau_{T'}(w) \geq \tau_T(t)$, which implies $\tau(T) \geq \tau(T')$. □

Corollary 1. *Any maximal sequence of caterpillar-extend operations that starts in a tree T terminates in a caterpillar tree* T^*. *Moreover,* $\tau(T^*)$ *is maximal in the set of all trees of fixed size, and it does not depend on* T^*, *as long as* T^* *is a caterpillar.*

Lemma 3. *For any n,* $\tau(C_n) = d(C_n, \bar{C}_n)$ *where* $\bar{C}_n = (1, (2, \ldots (n, n-1) \ldots))$.

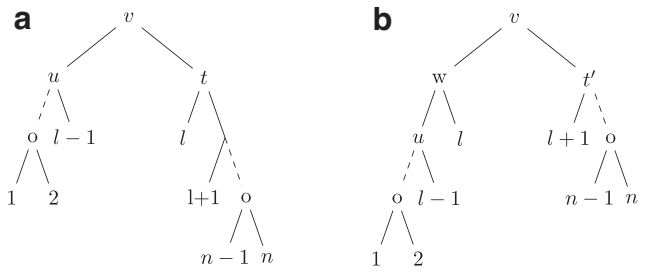

FIG. 1. **(a)** Shows a tree T where v is an internal node with two children u and t in T such that $T(u)$ and $T(t)$ are caterpillar subtrees and $|L(T(u))| \geq |L(T(t))| \geq 2$. **(b)** Shows the tree T' obtained after a caterpillar extend operation on T where $T' = NNI(T, l)$.

Proof. Consider a cluster c from C_n. For any cluster r in $\bar{C}_n$, associated with the vertex $v \in V(\bar{C}_n)$, such that $|c \cap r| \geq 1$ and $|r| > 1$, $|r| \geq n - |c|$ since there are $n - |c|$ taxa that are above c in C_n. More precisely, if $|c \cap r| \geq 1$ and $|r| > 1$, then $|r| = n - |c| + |c \cap r|$. Thus for a cluster r such that $|c \cap r| \geq 1$ and $|r| > 1$,

$$|c \Delta r| = |c| + n - |c| + |c \cap r| - 2|c \cap r| = n - |c \cap r|.$$

Note that $|c \cap r| = |c|$ if r is the cluster of the root. Also, note that there is always at least one leaf node l in $\bar{C}_n$ such that $l \in c$ and thus $|c \Delta \{l\}| = |c| - 1$. For all the remaining cases, when $r \cap c = \oslash$, we have $|c \Delta r| = |c| + |r|$. Hence we have,

$$d(c, \bar{C}_n) = \min(|c| - 1,\ n - |c|) = \tau(c).$$

Hence,

$$d(C_n, \bar{C}_n) = \sum_{v \in V(C_n)} d(c_v, \bar{C}_n) = \sum_{v \in V(C_n)} \tau(c_v) = \tau(C_n).$$

□

Theorem 2 (CA cost diameter). *The maximum CA cost between two trees of size n is* $\lceil \frac{n^2 - 2n}{4} \rceil$.

Proof. We show that the diameter is $\tau(T^*)$ where T^* is a caterpillar. Let T and S be two trees over the same set of leaves M and let $|M| = n$. Then, by Lemma 1 $d(T, S) \leq \tau(T)$. Next, we transform T into a caterpillar T^* by a sequence of caterpillar-extend operations. By Lemma 2, $\tau(T) \leq \tau(T^*)$. By Corollary 1, $\tau(T^*)$ is maximal and does not depend on T^* as long as T^* is a caterpillar. We showed that for any pair of trees, T, S, $d(T, S)$ is bounded by the value $\tau(T^*)$. Since, $\tau(T^*)$ is reached by two caterpillars, by Lemma 3, we conclude that $\tau(T^*)$ is maximal. It remains to derive the exact value:

$$\tau(T^*) = \sum_{c \in C(T)} \min(n - |c|,\ |c| - 1) = \sum_{j=\lceil \frac{n}{2}+1 \rceil}^{n-1} (n - j) + \sum_{i=2}^{\lceil \frac{n}{2} \rceil} (i - 1)$$

$$= \sum_{j=\lceil \frac{n}{2}+1 \rceil}^{n} (n - j) + \sum_{i=1}^{\lceil \frac{n}{2} \rceil} (i - 1) = n * \frac{n}{2} - \sum_{j=1}^{n} (j) + \sum_{j=1}^{\lceil n/2 \rceil} (j) - \left\lceil \frac{n}{2} \right\rceil + \sum_{i=1}^{\lceil n/2 \rceil} (i)$$

$$= \frac{n^2}{2} - \frac{n(n+1)}{2} - \left\lceil \frac{n}{2} \right\rceil + 2 * \sum_{i=1}^{\lceil n/2 \rceil} (i) = \frac{n^2}{2} - \frac{n(n+1)}{2} - \left\lceil \frac{n}{2} \right\rceil + \frac{n}{2}\left(\frac{n}{2} + 1\right)$$

$$= \frac{n^2}{4} - \left\lceil \frac{n}{2} \right\rceil = \left\lceil \frac{n^2 - 2n}{4} \right\rceil.$$

Hence, the diameter for the CA cost is $\lceil \frac{n^2 - 2n}{4} \rceil$. □

2.4. Separation diameter

We define the separation cost for CA cost as the absolute difference between the CA costs between the two trees S and G. That is $\sigma(G, S) = |d(G, S) - d(S, G)|$. Similarly, let the *separation diameter* of the CA cost be the diameter of the separation cost for the CA cost. Below we show how to derive a bound for the separation diameter for the CA cost.

Lemma 4. *The separation diameter of the CA cost is bound above by* $\lceil \frac{n^2 - 2n}{4} \rceil$.

Proof. It follows from Theorem 2. □

Lemma 5. *For any* $n = 2^m$, *there exists a perfectly balanced tree* P_n *such that* $d(C_n, P_n) = \tau(C_n) = \frac{n^2 - 2n}{4}$.

Proof. We construct a perfectly balanced tree P_n such that each nontrivial cluster in P_n is of the form $\{i, \ldots, j, n-j+1, \ldots, n-i+1\}$ for some $1 \le i \le j \le \frac{n}{2}$. We define the labeling in P_n as follows: let the i^{th} cherry in P_n in the prefix order have leaf labels i and $n-i+1$ for $1 \le i \le n/2$.

Consider a cluster $c = \{1, 2, \ldots, k\}$ in C_n and a nonleaf cluster $p \in P_n$. Then $p = \{i, \ldots j, n-j+1, \ldots, n-i+1\}$ for some $1 \le i \le j \le n$. We have two cases as follows:

1. **Case 1:** if $i \le k \le j$, then $|c \cap p| = k-i+1$ and thus $|c \cap p| \le \frac{|p|}{2}$. Hence, $|c \Delta p| = |c| + |p| - 2(|c \cap p|) \ge |c| \ge \tau_G(c)$.
2. **Case 2:** if $n-j+1 \le k \le n-i+1$ then $|c \cap p| = k-(n-j+1)+1+(j-i+1) = k-n+2j-i+1$. Thus,

$$|c \Delta p| = k + 2(j-i+1) - 2(k-n+2j-i+1) \ge n-k+n-2(\frac{n}{2}) = n-k = \tau_G(c).$$

For all the other remaining cases, while $|p|$ increases, $|c \cap p|$ remains constant, and hence they can be ignored. Hence, for every cluster $c \in C(C_n)$, $d(c, P_n) = \tau_{C_n}(c)$. Thus, $d(C_n, P_n) = \tau(C_n) = \frac{n^2-2n}{4}$. □

Lemma 6. *For a perfectly balanced tree T of size $n = 2^m$, $\tau(T) = n \log n - 3n + 2$.*

Proof. Let v be an internal node in $V(T)$ such that v is not the root of T and v has height h. Hence, $|c_v| = 2^h$. Since T is a perfectly balanced tree $|c_v| \le \frac{n}{2}$ and thus $\tau(c_v) = |c_v| - 1$. Thus, the total contribution of nodes with height h to $\tau(T)$ is $(2^h - 1).2^{m-h}$. Hence,

$$\tau(T) = \sum_{c \in C(T)} \tau_T(c) = \sum_{h=0}^{m-1} (2^h - 1) \cdot 2^{m-h}$$
$$= n(m - 1 - 2 \cdot \frac{n-1}{n}) = n \log_2 n - 3n + 2.$$

Thus $\tau(T) = n \log n - 3n + 2$. □

Lemma 7. *For every n, there exists a caterpillar tree G and a tree S such that the separation diameter $\sigma(G, S) \ge \lceil \frac{n^2-4n}{16} \rceil - n \log_2 n - 3n + 2$.*

Proof. It follows from Lemma 6 and Lemma 1 that $d(S, G)) \le \tau(S) = n \log_2 n - 3n + 2$ and hence, $\sigma(G, S) \ge \lceil \frac{n^2-2n}{4} \rceil - (n \log_2 n - 3n + 2)$ when $n = 2^m$ for some m.

For the remaining n's, we prove it by construction. Let $S_n = (n, (n-1, (n-2, \ldots (n-n'+1, P_{n'}) \ldots)))$ where $n' = 2^{\lfloor \log_2 n \rfloor}$ and $P_{n'}$ is a perfectly balanced tree of the size n' from the proof of Lemma 6. Since all clusters of size larger than n' are present in C_n and $n > n' \ge \frac{n}{2}$, we have $\sigma(C_n, S_n) \ge \sigma(C_{n'}, P_{n'}) \ge \lceil \frac{(n')^2-2n'}{4} \rceil - n' \log_2 n' - 3n' + 2 \ge \lceil \frac{n^2-4n}{16} \rceil - n \log_2 n - 3n + 2$. □

Theorem 3. *The separation diameter is $\Theta(n^2)$.*

Proof. From Lemma 7, $\sigma(G, S) \ge \lceil \frac{n^2-4n}{16} \rceil - n \log_2 n - 3n + 2$ and hence $\sigma(G, S) = \Omega(n^2)$. Since the CA distance is bound by $O(n^2)$, the separations affinity distance is also $O(n^2)$. Hence, the separation diameter is $\Theta(n^2)$. □

2.5. CS cost

We define the CS cost and then derive the exact diameter for that cost.

Definition 4 (CS cost). *The* CS cost *from cluster c to tree S is $d'(c, S) := \min_{x \in C(S)} \frac{|c \Delta x|}{|c|}$ and the CS cost from tree G to tree S is $d'(G, S) := \sum_{c \in C(G)} d'(c, S)$.*

For interpretability, in practice, we further divide $d'(G, S)$ by $|C(G)|$. Note that the CS cost can be computed in the same way as the CA in $O(n \log^3 n)$ time, where n is the size of a tree. We further introduce some tree operations to derive the diameter for the CS cost.

Definition 5 (Caterpillar swap). *For a tree T, let v be an internal node in T such that $T(v)$ is a caterpillar tree and u is the sibling of v with two caterpillar subtrees $T(x)$ and $T(y)$ such that $|c_v| < |c_y| < |c_u|$ and $|c_y| \leq |c_x|$. Let $T' = NNI(T, x)$. By the caterpillar-swap on x, we call the transformation of T into T', which is denoted by $T \rightarrow^{\text{swap}} T'$.*

Definition 6 (Caterpillar split). *For a tree T, let v be an internal node in T such that $T(v)$ is a caterpillar tree and u is the sibling of v with two caterpillar subtrees $T(x)$ and $T(y)$ such that $|c_y| \leq |c_v|$ and $|c_y| \leq |c_x|$. Let $T' = SPR(T, l, v)$ where l is the leaf-child of y. By the caterpillar-split on y, we call the transformation of T into T', which is denoted by $T \rightarrow^{\text{split}} T'$.*

Definition 7 (Caterpillar-balance). *For a tree T, let $T(u)$ and $T(v)$ be two caterpillars subtrees of the root such that $|c_v| < |c_u| - 1$. By $T \rightharpoonup T'$ we denote the caterpillar-balance operation $T' = NNI(T, l)$, where l is a leaf-child of u.*

For a tree T with a leaf set M and a node $v \in V(T)$ let $\phi_T(v) := \frac{\tau_T(v)}{|c_v|}$ and $\phi(T) := \sum_{v \in V(T)} \phi_T(v)$.

Lemma 8. *For every node v in a tree T, $d'(c_v, S) \leq \phi_T(v)$ and for every tree S, $d'(T, S) \leq \phi(T)$.*

Proof. We know from Lemma 1 that $d(c_v, S) \leq \tau_T(v)$. Hence, $d'(c_v, S) \leq \phi_T(v)$. Consequently, we have $d'(T, S) \leq \phi(T)$. □

We define a 2-caterpillar tree D_n such that the two subtrees from the root are $C_{\lceil \frac{n}{2} \rceil}$ and $C_{\lfloor \frac{n}{2} \rfloor}$. To derive the diameter of the CS cost, we show that there exists a sequence of operations, see Rules 1–5 below that can transform any tree T into a 2-caterpillar tree and that $\phi(D_n)$ is maximal over all trees of size n.

Let T be a tree of the size n. Let u, v be siblings in $V(T)$ and let x, y be children of u. Assume without loss of generality that $|c_y| \leq |c_x|$ and $|c_v| \leq |c_u|$. We define the following rules to transform T into a 2-caterpillar tree.

Rule 1. If $|c_v| \leq \lceil \frac{n}{2} \rceil$ and $T(v)$ is a noncaterpillar tree, transform T by applying caterpillar extend operations on v until the subtree rooted at v is not a caterpillar.

Rule 2. If $T(x)$ and $T(y)$ are caterpillar subtrees such that $|c_x| < \lceil \frac{n}{2} \rceil$ and $|c_u| > \lceil \frac{n}{2} \rceil$, perform the caterpillar extend operation on u (provided the operation is allowed) repeatedly until one of the child subtrees of u has exactly $\lceil \frac{n}{2} \rceil$ leaves.

Rule 3. If $T(v)$, $T(x)$, $T(y)$ are caterpillar subtrees and $T(u)$ is a noncaterpillar subtree such that $|c_v| < |c_y|$ and $|c_x| \geq \lceil \frac{n}{2} \rceil$, then perform the caterpillar swap operation on x.

Rule 4. If $T(v)$, $T(x)$, $T(y)$ are caterpillar subtrees and $T(u)$ is a noncaterpillar subtree such that $|c_y| \leq |c_v|$ and $|c_x| \geq \lceil \frac{n}{2} \rceil$, then perform the caterpillar split operations on a child of u repeatedly until such a transformation is not allowed. Note that after the application of Rule 4, one child of u becomes a leaf.

Rule 5. If u and v are children of the root such that $T(u)$ and $T(v)$ are caterpillar trees and $|c_v| < |c_u| - 1$, then perform the caterpillar balance operation.

We show in the following lemma that applying these rules to a tree T to obtain a tree T' always results in an increase in the value of ϕ.

Lemma 9. *For a tree T' obtained by applying a rule to a tree T, $\phi(T') > \phi(T)$.*

Proof. Let n be the size of T and T'. Since each rule consists of caterpillar-extend, caterpillar-balance, caterpillar-swap, and caterpillar-split operations, it suffices to prove that for S obtained by applying an operation to T, $\phi(S) > \phi(T)$.

$T \longrightarrow S$**:** Note that T and S only differ by two clusters. Let c_t, c_w be the differing clusters as defined in the caterpillar-extend operation. Note that by the definitions of Rule 1 and Rule 2, $|c_t| < |c_w| \leq \lceil \frac{n}{2} \rceil$, which implies $\tau_S(w) = |c_w| - 1$ and $\tau_T(t) = |c_t| - 1$. Hence, $\phi_S(w) = \frac{|c_w| - 1}{|c_w|} = 1 - \frac{1}{|c_w|} > 1 - \frac{1}{|c_t|} = \phi_T(t)$.

$T \rightarrow^{\text{swap}} S$**:** Using the notation from Figure 2, note that only u and u' change between T and S. Moreover, $|c_u| > |c_{u'}| > \lceil \frac{n}{2} \rceil$ since $|c_x| \geq \lceil \frac{n}{2} \rceil$. Hence, $\phi_S(u') = \frac{n - |c_{u'}|}{|c_{u'}|} > \frac{n - |c_u|}{|c_u|} = \phi_T(u)$.

$T \rightarrow^{\text{split}} S$**:** Using the notation from Figure 3, T and S differ by four clusters $c_u, c_y, c_{v'}, c_{u'}$. Note that $|c_u| > |c_{u'}| > \lceil \frac{n}{2} \rceil$ since $|c_x| \geq \lceil \frac{n}{2} \rceil$. Moreover, $|c_y| < |c_{v'}| < \lceil \frac{n}{2} \rceil$. Hence, $\phi_{T_i}(u') = \frac{n - |c_{u'}|}{|c_{u'}|} > \frac{n - |c_u|}{|c_u|} = \phi_T(u)$ and $\phi_S(v') = \frac{|c_{v'}| - 1}{|c_{v'}|} = 1 - \frac{1}{|c_{v'}|} > 1 - \frac{1}{|c_y|} = \phi_T(y)$.

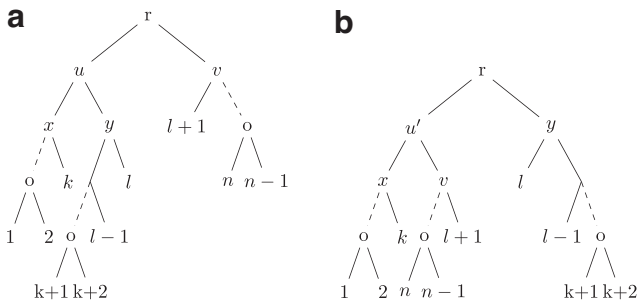

FIG. 2. **(a)** Shows a tree T where u, v are siblings such that $T(v)$ is a caterpillar tree u has two caterpillar subtrees $T(x)$ and $T(y)$ such that $|L(T(y))| \leq |L(T(x))|$ and $|L(T(v))| < |L(T(y))|$. **(b)** Shows the tree T' obtained after a caterpillar swap operation on T where $T' = NNI(T, x)$.

$T \rightharpoonup S$**:** Note that T and S differ by only two clusters. Let c_u and c_w be the two clusters as defined in the caterpillar-balance operation from Figure 4. Hence $|c_w| = |c_v| + 1$, which implies $n = |c_u| + |c_w| - 1$. Moreover, as $|c_v| < \lceil \frac{n}{2} \rceil$, $|c_w| \leq \lceil \frac{n}{2} \rceil < |c_u|$. Hence, we have $\phi_S(w) = \frac{|c_w|-1}{|c_w|} = \frac{n-|c_u|}{|c_w|} > \frac{n-|c_u|}{|c_u|} = \phi_T(u)$.

Hence, if S is obtained by applying an operation to T, $\phi(S) > \phi(T)$, which implies $\phi(T') > \phi(T)$ where T' is obtained by applying a rule to T. □

Hence, for any tree T' obtained from a tree T through an application of one of rules, $\phi(T') > \phi(T)$. We further prove that for any tree T, we can apply a rule to T if and only if T is not a 2-caterpillar tree.

Lemma 10. *T is not a 2-caterpillar tree if and only if there is a rule that can be applied to T.*

Proof. Let n be the size of T.

($\Rightarrow$) If T is a caterpillar tree such that w and t are the children of the root and $|c_w| < |c_t|$ and T has at least three leaves then we can apply Rule 5 to T as $|c_w| = 1$ and $|c_t| = n - 1$. Hence, T cannot be a caterpillar tree. Let u be a node in T such that $T(u)$ is not a caterpillar tree and u has two children, x and y such that $T(x)$ and $T(y)$ are caterpillar trees. Assume without loss of generality that $|c_y| \leq |c_x|$. If u is the root of T, then we have $|c_y| + |c_x| = n$, which implies $|c_y| < |c_x| - 1$ since T is not a 2-caterpillar tree and Rule 5 can be applied.

Hence, u must be a nonroot internal node. If $|c_u| \leq \lceil \frac{n}{2} \rceil$ then Rule 1 can be applied to T. Hence $|c_u| > \lceil \frac{n}{2} \rceil$. Furthermore, $|c_x| \geq \lceil \frac{n}{2} \rceil$ since otherwise Rule 2 can be applied to T. Let v be the sibling of u. Note that $T(v)$ must be a caterpillar tree, since $|c_v| \leq \lceil \frac{n}{2} \rceil$ and Rule 1 could be applied to T otherwise. Hence, if $|c_v| < |c_y|$, Rule 3 can be applied to T and conversely if $|c_v| \geq |c_y|$, Rule 4 can be applied to T.

($\Leftarrow$) Assume conversely that T is a 2-caterpillar tree. Let u be an internal node in T and let x and y be the children of u. Assume without loss of generality that $|c_y| \leq |c_x|$. If u is the root node, $|c_y| \geq |c_x| - 1$. Conversely, if u is not the root node, then $T(u)$ is a caterpillar tree. Hence, there is no rule that can be applied to T if T is a 2-caterpillar tree. □

Hence, for any non 2-caterpillar tree T, there exists a rule that can be applied to T. Conversely, if T is a 2-caterpillar tree, there is no rule that can be applied to T. We showed by Lemma 9 that if we derive a tree T' by applying a rule to a tree T, then $\phi(T') > \phi(T)$. This allows us to repeatedly apply the rules to derive the maximum value of ϕ in the set of all trees of fixed size.

We say that a tree T is transformed into T' by a sequence of rule applications if there is a sequence of trees $T = T_0, T_1, \ldots, T_l = T'$ with $l \geq 0$ such that T_{i+1} is obtained from T_i by an application of a rule. We show that for any tree T, there is a maximal sequence of rule applications that can transform T into a 2-caterpillar tree T'.

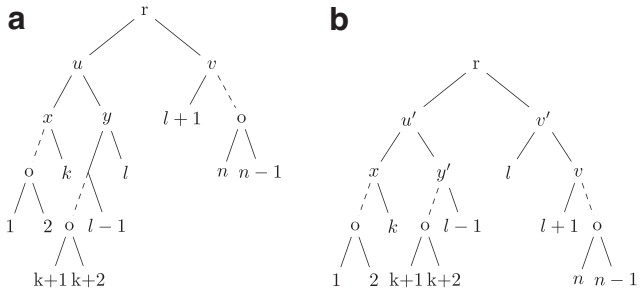

FIG. 3. **(a)** Shows a tree T where u, v are siblings such that $T(v)$ is a caterpillar tree and u has two caterpillar subtrees $T(x)$ and $T(y)$ such that $|L(T(y))| \leq |L(T(v))|$ and $|L(T(y))| \leq |L(T(x))|$. **(b)** Shows the tree T' obtained after a caterpillar-split operation on T where $T' = SPR(T', l-1, v)$.

Lemma 11. *Every maximal sequence of rule applications is finite and terminates with a 2-caterpillar tree.*

Proof. By Lemma 10, for any tree T of size n, if T is not a 2-caterpillar tree then there exists a rule that can be applied to T. Moreover, by Lemma 9, for a tree T' derived by an application of a rule to T, $\phi(T) < \phi(T')$. Note that for any tree T, $\phi(T) \leq 2n-1$ since for any node $v \in V(T)$, $\phi_T(v) \leq 1$. We first show that for S obtained by applying an operation to T, $\phi(S) - \phi(T) > \frac{1}{n^2} > 0$. We use the same notation as described in Lemma 9.

$T \longrightarrow S$: For c_w and c_t, we have $\phi_S(w) - \phi_T(t) = 1 - \frac{1}{|c_w|} - 1 + \frac{1}{|c_t|} = \frac{1}{|c_t|} - \frac{1}{|c_w|} = \frac{|c_w| - |c_t|}{|c_w|.|c_t|} > \frac{1}{n^2}$.

$T \rightarrow^{\text{swap}} S$: For c_u and $c_{u'}$, we have $\phi_S(u') - \phi_T(u) = \frac{n - |c_{u'}|}{|c_{u'}|} - \frac{n - |c_u|}{|c_u|} = n \cdot \left(\frac{1}{|c_{u'}|} - \frac{1}{|c_u|}\right) = n \cdot \frac{|c_u| - |c_{u'}|}{|c_u|.|c_{u'}|} > \frac{1}{n^2}$.

$T \rightarrow^{\text{split}} S$: For c_u and $c_{u'}$, we have the same result as above. Similarly for $c_{v'}$ and c_y, we have $\phi_S(v') - \phi_T(y) = \frac{1}{|c_y|} - \frac{1}{|c_{v'}|} = \frac{|c_{v'}| - |c_y|}{|c_y|.|c_{v'}|} > \frac{1}{n^2}$.

$T \rightharpoonup S$: For $|c_u|$ and $|c_w|$, we have $\phi_S(w) - \phi_T(u) = \frac{|c_w| - 1}{|c_w|} - \frac{n - |c_u|}{|c_u|} = \frac{n - |c_u|}{|c_w|} - \frac{n - |c_u|}{|c_u|} = (n - |c_u|) \cdot (\frac{|c_u| - |c_w|}{|c_u|.|c_w|}) > \frac{1}{n^2}$.

We proved that every rule application increases $\phi(T)$ by at least $\frac{1}{n^2}$. Thus, every sequence of rule applications has at most $(2n-1) \cdot n^2$ elements. We conclude that the sequence is finite and must terminate with a 2-caterpillar by Lemma 10. □

Corollary 2. *$\phi(T^*)$ is maximal in the set of all trees of fixed size and it does not depend on T^*, as long as T^* is a 2-caterpillar.*

Hence, we have for every tree T, a maximal finite sequence of transformations S_T that can transform a tree T into a 2-caterpillar tree T^*. Note that since the transformations are defined by the rules above, we can show that $\phi(T^*) > \phi(T)$. It remains to show that there exists a scenario that would lead to $\phi(T^*)$. We show that for any 2-caterpillar tree T^*, we can construct such a tree $\bar{T}^*$ where $d'(T^*, \bar{T}^*) = \phi(T^*)$.

Lemma 12. *For any n, $\phi(D_n) = d'(D_n, \bar{D}_n)$ where $\bar{D}_n$ is a 2-caterpillar tree such that the subtrees of the root are $\bar{C}_{\lceil \frac{n}{2} \rceil}$ and $\bar{C}_{\lfloor \frac{n}{2} \rfloor}$.*

Proof. Let u, w be the two children of the root of D_n. For D_n, $\phi(D_n) = \sum_{v \in V(D_n)} \phi_{D_n}(v) = \sum_{v \in V(D_n(u))} \phi_{D_n}(v) + \sum_{v \in V(D_n(w))} \phi_{D_n}(v) + \phi_{D_n}(r)$ where r is the root node of D_n. Note $\phi_{D_n}(r) = 0$. By Lemma 3, for every node $v \in V(D_n(u))$, $d(c_v, \bar{D}_n) = \tau_{D_n}(v)$, which implies $d'(c_v, \bar{D}_n) = \phi_{D_n}(v)$. Similarly, for every node $v \in V(D_n(w))$, $d(c_v, \bar{D}_n) = \tau_{D_n}(v)$, which implies $d'(c_v, \bar{D}_n) = \phi_{D_n}(v)$. Hence,

$$d'(D_n, \bar{C}'_n) = \sum_{v \in V(D_n(u))} \phi_{D_n}(v) + \sum_{v \in V(D_n(w))} \phi_{D_n}(v) = \phi(D_n).$$ □

Hence, for any tree T we are able to convert it into a 2-caterpillar tree T^* by the sequence of operations defined in Lemma 11. Moreover, by Lemma 9 we know that $\phi(T^*) > \phi(T)$ and by Lemma 12 we can construct a tree T' such that $d'(T^*, T') = \phi(T^*)$. Thus, as $\phi(T^*)$ is the maximal value of ϕ, we derive the diameter for the CS cost as the value of $\phi(T^*)$.

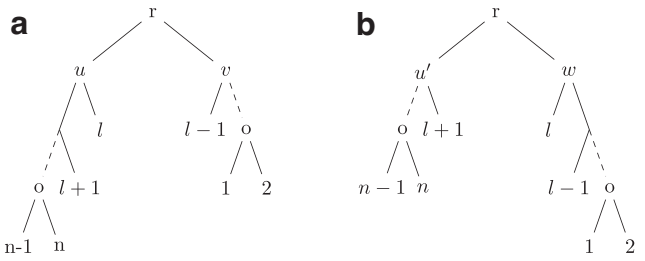

FIG. 4. **(a)** Shows a tree T where u, v are children of the root of T such that $T(u)$ and $T(v)$ are caterpillar subtrees and $|L(T(v))| < |L(T(u))| - 1$. **(b)** Shows the tree T' obtained after a caterpillar-balance operation on T where $T' = NNI(T, l)$.

Theorem 4 (CA support cost diameter). *The maximum CS cost between two trees of size n is* $n - H_{\lceil \frac{n}{2} \rceil} - H_{\lfloor \frac{n}{2} \rfloor}$, *where* H_i *is the i-th harmonic number.*

Proof. We show that the diameter is $\phi(T^*)$ where T^* is a 2-caterpillar. Let T and S be two trees over the same set of leaves M and let $|M| = n$. Then, by Lemma 8, $d'(T, S) \leq \phi(T)$. Next, we transform T into a 2-caterpillar tree T^* by the sequence described in Lemma 11. By Corollary 2, we showed that for any pair of trees, T, S, $d(T, S)$ is bounded by the value $\phi(T^*)$. Since, $\phi(T^*)$ is reached by two 2-caterpillar trees by Lemma 12, we conclude that $\phi(T^*)$ is maximal. It remains to derive the exact value. Let u, w be the children of the root of T^*.

$$\phi(T^*) = \sum_{v \in V(T^*)} \frac{\tau(v)}{|c_v|} = \sum_{v \in V(T^*(u))} \frac{\tau(v)}{|c_v|} + \sum_{v \in V(T^*(w))} \frac{\tau(v)}{|c_v|} = \sum_{i=1}^{\lceil \frac{n}{2} \rceil} \frac{i-1}{i} + \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} \frac{i-1}{i} = n - H_{\lceil \frac{n}{2} \rceil} - H_{\lfloor \frac{n}{2} \rfloor}$$

Hence the diameter for the CS cost is $n - H_{\lceil \frac{n}{2} \rceil} - H_{\lfloor \frac{n}{2} \rfloor}$. □

3. EMPIRICAL STUDY

We compare the distribution and the robustness of the CA and CS costs (defined in this work) with the classic RF distance. Since the CA and CS are asymmetric costs, we use the asymmetric (one-sided) version of RF for comparison, also known as the false negative rate. We define the one-sided RF between trees T_1 and T_2 as the number of clusters in T_1 that are not present in T_2.

3.1. CA and CS display a broad distribution range

Table 1 shows the descriptive statistics of the (asymmetric) CA, CS, and RF costs on pairs of random trees. We generated 10,000 pairs of trees each containing 100 taxa. We used the birth–death model to generate the trees with a birth rate of 1.0 and a death rate of 0.5. The birth–death process was terminated when the tree had the required number of extant taxa.

We observe that CA and CS both have a broader distribution range (standard deviation and the min–max range) than RF. Out of all three costs, CA has the broadest range and has the least skewed distribution, while RF is most skewed toward the maximum. These results are similar to the comparison between the original (symmetric) CA distance and RF (Moon and Eulenstein, 2019); thus, demonstrating that our cost relaxation maintained the key properties of the original CA distance. Figure 5 visually captures the distributions for CA and CS. Note that we truncate the data for RF to preserve the details of the histograms for CA and CS. Moreover, due to the large diameter of the CA cost, the discretization of the bins on the histogram results in multiple peaks being observed.

Table 1. Asymmetric Robinson–Foulds, Cluster Affinity, and Cluster Support Distribution Statistics for Random Tree Pairs with $n \in \{100, 1000\}$ Leaves

Leaves		*RF*	*CA*	*CS*
100	Mean	0.997	0.816	0.958
	SD	0.004	0.056	0.017
	Median	1.000	0.810	0.961
	Min	0.969	0.687	0.865
1000	Mean	0.999	0.883	0.993
	SD	4.857×10^{-4}	0.036	0.002
	Median	1.000	0.878	0.994
	Min	0.996	0.807	0.982

All values are normalized by the maximum observed cost, respectively. Note that both CA and CS display a broader distribution range than RF. Moreover, out of all the costs, CA is the least skewed toward the maximum while RF is the most skewed toward the maximum.

CA, Cluster Affinity; CS, Cluster Support; RF, Robinson–Foulds.

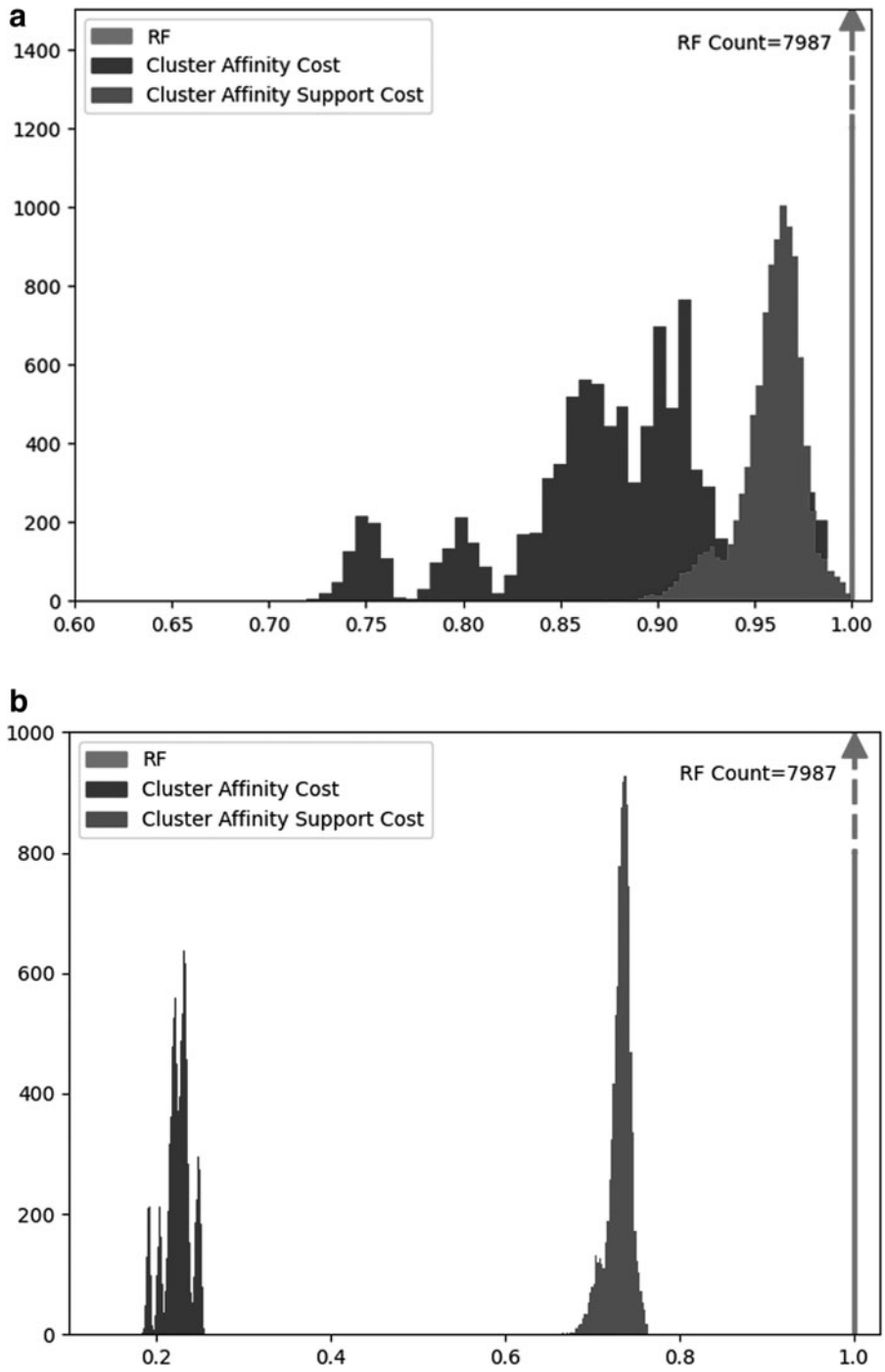

FIG. 5. **(a, b)** Show the distribution of the three costs when normalized by the observed maximum or the theoretical diameter, respectively. Note that both the CA cost and the CS cost have a broader distribution range than the RF cost. Furthermore, under RF, 80% of tree pairs had a cost of 1.0 (the theoretical maximum). **(a)** RF, CA, and CS costs for repeated NNI operations normalized by the observed maxima. **(b)** RF, CA, and CS costs for repeated SPR operations normalized by the observed maxima. CA, Cluster Affinity; CS, Cluster Support; NNI, Nearest Neighbor Interchange; RF, Robinson–Foulds; SPR, Subtree Prune and Regraft.

Additionally, we quantified the CA asymmetry between the tree pairs. For a pair of trees (T_1, T_2) we computed the separation cost $|d(T_1, T_2) - d(T_2, T_1)|$ and normalized by the maximum observed CA cost in our dataset. The average separation cost was 0.057 with a maximum value of 0.270 and a standard deviation of 0.041. That is, the asymmetry between birth–death trees was 6% on average and 27% in the worst case.

We observed that for a pair of trees with $n \in \{100, 1000\}$, it required an average of 1.5 ms and 17 ms to compute the asymmetric RF distance, while it required an average of 123 ms and 13,595 ms to compute the CA cost. This is due to the higher computational complexity and resolution of the CA cost. In this study, we used a quadratic algorithm for computing the CA and CS costs. This algorithm computes the symmetric difference between each pair of clusters in the two trees and is required for the NNI algorithm from Section 2.2. The experiments were run on Ubuntu 22.04 with an Intel i7 CPU and 64GB of RAM with Python 3.10 and Dendropy 4.5.2.

3.2. CA and CS are robust to tree edit operations

We demonstrate that CA and CS are significantly more robust to tree edit operations (and, hence, tree-error) than RF. Our experimental setup follows (Moon and Eulenstein, 2019) for comparison between the CA, CS, and RF distances.

3.2.1. Dataset. We generated a set of random trees $T_1, T_2, \ldots, T_{100}$ where each tree had 100 leaves using the birth–death model with birth rate 1.0 and death rate 0.5. For each tree T_i, we generated a sequence of trees $P(i) := (p(i)_0, p(i)_1, \ldots, p(i)_{2000})$, where each $p(i)_{j+1}$ was obtained by an NNI operation on $p(i)_j$ and the edge for the NNI operation was chosen uniformly and independently at random. Similarly, for each tree T_i, we also generated a sequence of trees $Q(i) := (q(i)_0, q(i)_1, \ldots, q(i)_{500})$ where each $q(i)_{j+1}$ was obtained by an SPR operation on $q(i)_j$ where the edges for the SPR operation were chosen uniformly and independently at random. Note that $p(i)_0 = q(i)_0 = T_i$.

3.2.2. Experimental setting. Distances were computed between tree pairs $(p(i)_0, p(i)_j)$ and $(q(i)_0, q(i)_j)$ for each j and averaged over all $i \in \{1, \ldots, 100\}$ under the RF cost, CA cost, and CS cost.

3.2.3. Results. The respective RF, CA, and CS costs over consecutive NNI and SPR operations are shown in Figure 6. We observe that for both NNI and SPR edit operations, RF approaches the maximum value very rapidly. In contrast, CA and CS costs are significantly more robust to the tree edits and demonstrated better resolution than RF. The CS cost was most robust in terms of both NNI and SPR edit operations.

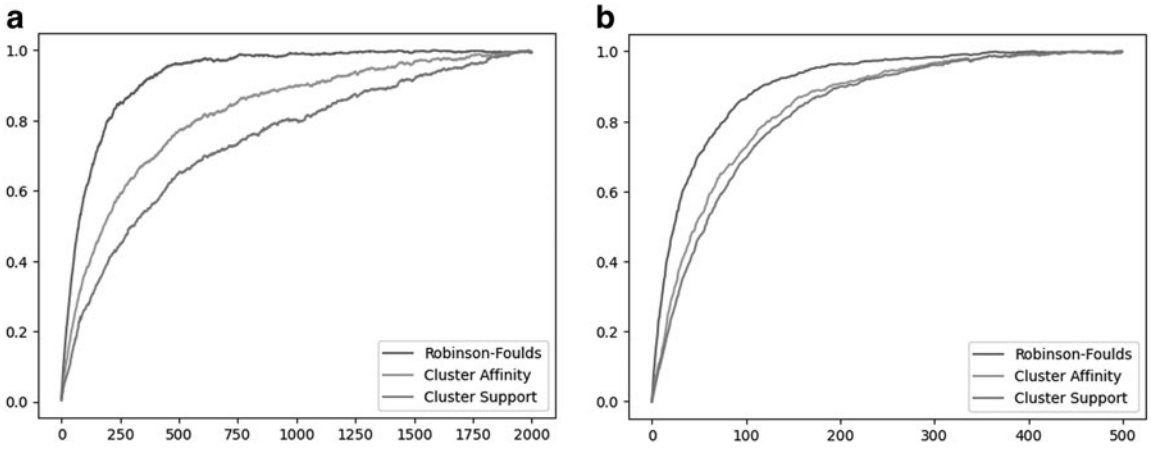

FIG. 6. **(a, b)** Show the changes in RF, CA, and CS costs for subsequent NNI and SPR operations, respectively. Note that the RF distance approaches the maximum value more rapidly than the CA and CS costs.

4. CONCLUSION

A key objective in phylogenetics is to construct a plausible hypothesis of the evolutionary history of an organism. Merging different datasets and gene trees into larger species trees and supertrees can improve the resolution of evolutionary inference. To achieve this, it is necessary to develop methods that can estimate uncertainty and identify conflicts in different gene trees and datasets. Our study introduces an asymmetric CA cost, which is based on determining the minimum cost for a cluster in the source tree for every cluster in the target tree. We have also developed efficient algorithms for calculating the CA cost and determining its diameter. To measure the asymmetry of CA cost, we define a separation diameter as the maximum difference between the two directions that CA cost can be computed. Furthermore, we present a more interpretable cost definition in which cluster differences are normalized by their respective cluster sizes. We have also derived a theoretical diameter for the CS cost and experimentally demonstrated its robustness to tree error.

A promising implication of developing this measure is its future use in the inference of phylogenetic networks (Huson and Bryant, 2006). Some approaches to phylogenetic network inference use tree distance measures to analyze a collection of input trees to seek a minimum reticulation network with the smallest number of reticulation vertices into which the input trees can be embedded (e.g., Markin et al., 2019). The use of RF distance in clustering algorithms often leads to many shallow reticulation events, as the algorithm tries to match clusters fully. To overcome this, we suggest applying CA/CS costs in phylogenetic network inference. This approach allows minor cluster mismatches in smaller clusters to be overlooked, enabling the algorithm to focus on identifying significant reticulation events higher up on the phylogeny (Kong et al., 2022).

The software for computing and visualizing the CA cost is being prepared for release and available upon request from the authors.

ACKNOWLEDGMENTS

This contribution is an extension of an article that appeared in RECOMB-CG (https://doi.org/10.1007/978-3-031-36911-7_9) (Wagle et al., 2023). The authors thank Dr. Swenson for referring them to Truszkowski et al., (2019) and discussions benefiting this work.

AUTHOR DISCLOSURE STATEMENT

The authors declare they have no conflicting financial interests.

FUNDING INFORMATION

This work was supported in part by the U.S. Department of Agriculture (USDA) Agricultural Research Service (ARS) (project number 5030-32000-231-000-D, 5030-32000-231-095-S, and 5030-32000-231-103-A); the USDA-ARS Research Participation Program of the Oak Ridge Institute for Science and Education (ORISE) through an interagency agreement between the U.S. Department of Energy and USDA-ARS (contract number DE-AC05-06OR23100); and a grant from the National Science Center of Poland 2017/27/B/ST6/02720. The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication. USDA is an equal opportunity provider and employer.

REFERENCES

Allen BL, Steel M. Subtree transfer operations and their induced metrics on evolutionary trees. Ann Comb 2001;5:1–15.

Bininda-Emonds OR. Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life, vol. 4. Springer Science & Business Media; 2004.

Böcker S, Canzar S, Klau GW. The generalized robinson-foulds metric. In: Algorithms in Bioinformatics: 13th International Workshop, WABI 2013, Sophia Antipolis, France, September 2–4, 2013. Proceedings 13. Springer; 2013; pp. 156–169.

Bogdanowicz D, Giaro, K. Matching split distance for unrooted binary phylogenetic trees. IEEE/ACM Trans Comput Biol Bioinform 2011;9(1):150–160.

Bogdanowicz D, Giaro K. On a matching distance between rooted phylogenetic trees. Int J Appl Math Comput Sci 2013;23(3):669–684.

Bogdanowicz D, Giaro K. Comparing phylogenetic trees by matching nodes using the transfer distance between partitions. J Comput Biol 2017;24(5):422–435.

Bordewich M, Semple C. On the computational complexity of the rooted subtree prune and regraft distance. Ann Comb 2005;8:409–423.

Chaudhary R, Burleigh JG, Eulenstein O. Efficient error correction algorithms for gene tree reconciliation based on duplication, duplication and loss, and deep coalescence. In: BMC Bioinformatics, vol. 13. BioMed Central; 2012; pp. 1–10.

Estabrook GF, McMorris F, Meacham CA. Comparison of undirected phylogenetic trees based on subtrees of four evolutionary units. Syst Zool 1985;34(2):193–200.

Giardina F, Romero-Severson EO, Albert J, et al. Inference of transmission network structure from HIV phylogenetic trees. PLoS Comput Biol 2017;13(1):e1005316.

Huson DH, Bryant D. Application of phylogenetic networks in evolutionary studies. Mol Biol Evol 2006;23(2):254–267.

Kong S, Pons JC, Kubatko L, et al. Classes of explicit phylogenetic networks and their biological and mathematical significance. J Math Biol 2022;84(6):47.

Kulkarni A, Sabetpour N, Markin A, et al. CPTAM: Constituency parse tree aggregation method. In: SDM. 2022.

Lin Y, Rajan V, Moret BM. A metric for phylogenetic trees based on matching. IEEE/ACM Trans Comput Biol Bioinformatics 2011;9(4):1014–1022.

Lozano-Fernandez J. A practical guide to design and assess a phylogenomic study. Genome Biol Evol 2022;14(9):evac129.

Markin A, Anderson TK, Vadali VSKT, et al. Robinson-foulds reticulation networks. In: Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics. Association for Computing Machinery: Niagara Falls, NY, USA; 2019; pp. 77–86.

Moon J, Eulenstein O. The cluster affinity distance for phylogenies. In: Bioinformatics Research and Applications. (Cai Z, Skums P, Li M. eds.) Springer International Publishing: Cham; 2019; pp. 52–64. ISBN 978-3-030-20242-2.

Nye TM, Lio P, Gilks WR. A novel algorithm and web-based tool for comparing two alternative phylogenetic trees. Bioinformatics 2006;22(1):117–119.

Page RD. Modified mincut supertrees. In: Algorithms in Bioinformatics: Second International Workshop, WABI 2002 Rome, Italy, September 17–21, 2002 Proceedings 2. Springer; 2002; pp. 537–551.

Prum RO, Berv JS, Dornburg A, et al. A comprehensive phylogeny of birds (aves) using targeted next-generation DNA sequencing. Nature 2015;526(7574):569–573.

Robinson DF, Foulds LR. Comparison of phylogenetic trees. Math Biosci 1981;53(1–2):131–147.

Robinson O, Dylus D, Dessimoz C. Phylo. io: Interactive viewing and comparison of large phylogenetic trees on the web. Mol Biol Evol 2016;33(8):2163–2166.

Russo C, Takezaki N, Nei M. Efficiencies of different genes and different tree-building methods in recovering a known vertebrate phylogeny. Mol Biol Evol 1996;13(3):525–536.

Shen X-X, Steenwyk JL, Rokas A. Dissecting incongruence between concatenation-and quartet-based approaches in phylogenomic data. Syst Biol 2021;70(5):997–1014.

Smith MR. Information theoretic generalized robinson–foulds metrics for comparing phylogenetic trees. Bioinformatics 2020;36(20):5007–5013.

Steel MA, Penny D. Distributions of tree comparison metrics—Some new results. Syst Biol 1993;42(2):126–141.

Swenson M, Suri R, Linder CR, et al. An experimental study of quartets maxcut and other supertree methods. Algorithms Mol Biol 2011;6(1):1–11.

Truszkowski J, Gascuel O, Swenson KM. Rapidly computing the phylogenetic transfer index. In: 19th International Workshop on Algorithms in Bioinformatics (WABI 2019), vol. 143 of Leibniz International Proceedings in Informatics (LIPIcs). (Huber KT, Gusfield D. eds.) Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik: Dagstuhl, Germany; 2019; pp. 20:1–20:12; doi: 10.4230/LIPIcs.WABI.2019.20. ISBN 978-3-95977-123-8.

Wagle S, Markin A, Górecki P, et al. The asymmetric cluster affinity cost. In: RECOMB International Workshop on Comparative Genomics. Springer; 2023; pp. 131–145.

Waterman MS, Smith TF. On the similarity of dendrograms. J Theor Biol 1978;73(4):789–800.

Wickett NJ, Mirarab S, Nguyen N, et al. Phylotranscriptomic analysis of the origin and early diversification of land plants. In: Proc Natl Acad Sci U S A 2014;111(45):E4859–E4868.

Yang Z, Rannala B. Molecular phylogenetics: principles and practice. Nat Rev Genet 2012;13(5):303–314.

Address correspondence to:
Prof. Oliver Eulenstein
Department of Computer Science
Iowa State University
Ames, IA 50014
USA

E-mail: oeulenst@iastate.edu

Sanket Wagle
Department of Computer Science
Iowa State University
Ames, IA 50014
USA

E-mail: swagle@iastate.edu