# Phylogenetic Network Dissimilarity Measures That Take Branch Lengths Into Account

Berk A. Yakici[1 [0000−0002−2225−3645]], Huw A. Ogilvie[1 [0000−0003−1589−6885]], and Luay Nakhleh[1 [0000−0003−3288−6769]]

Rice University, Houston, TX 77005, USA
{bay,huw.a.ogilvie,nakhleh}@rice.edu

**Abstract.** Dissimilarity measures for phylogenetic trees have long been used for analyzing inferred trees and understanding the performance of phylogenetic methods. Given their importance, a wide array of such measures have been developed, some of which are based on the tree topologies alone, and others that also take branch lengths into account. Similarly, a number of dissimilarity measures of phylogenetic networks have been developed in the last two decades. However, to the best of our knowledge, all these measures are based solely on the topologies of phylogenetic networks and ignore branch lengths. In this paper, we propose two phylogenetic network dissimilarity measures that take both topology and branch lengths into account. We demonstrate the behavior of these two measures on pairs of related networks. Furthermore, we show how these measures can be used to cluster a set of phylogenetic networks obtained by an inference method, illustrating this application on the posterior sample of phylogenetic networks. Both measures are implemented in the publicly available software package PhyloNet.

**Keywords:** Phylogenetic Networks · Dissimilarity · Topology · Branch Lengths

## 1 Introduction

Phylogenetic trees and networks are widely used to model the evolution of genes, species and languages. In the case of genomes and species, a deluge of data is available in the form of whole genomes being assembled and made available for phylogenetic inference [13, 18, 22]. However, the space of phylogenetic trees is notoriously complex due to a mix of discrete and continuous parameters. Therefore, this complexity must be confronted, as probability and likelihood distributions over phylogenetic trees often lack closed form solutions, requiring exploration of this space using algorithms such as hill climbing or Markov-chain Monte Carlo. Worse still, the discrete and continuous parameters are highly dependant, making this exploration fiendishly difficult. Phylogenetic network distributions are more complex again, among other reasons because dimensionality of the problem is no longer fixed.

Nonetheless, phylogenetic trees and networks remain successful models due to their strong and direct relationship to actual evolution. Evolving units, such

as species split and diverge over time, and patterns of splits are encoded by the nodes of phylogenetic trees. Each node represents a species (or other kind of evolving unit), and has one parent and a certain number of children (typically zero or two) to connect with its immediate ancestor and descendants respectively. Units which we have data for are typically represented as external nodes, often called taxa for trees of species or populations. The degree of divergence can be encoded as either the branch lengths or, for ultrametric phylogenies, as node ages. The pattern of splits and reticulations without continuous parameters is known as the tree topology and is an unordered discrete variable that grows hyperexponentially with the number of taxa [9].

Because phylogenetic trees do not account for reticulate evolution, such as introgression, hybridization, or horizontal gene transfer between species, or loanwords borrowed between languages [12], phylogenetic networks were developed in order to model splitting and reticulation. These networks may contain reticulation nodes, with two parents and one child, in addition to the nodes used to encode splits. The probability of inheriting evolutionary subunits, such as genes or words, may be encoded as additional continuous parameters associated with the reticulation nodes. By permitting two immediate ancestors with inheritance probabilities, phylogenetic networks can effectively encode the aforementioned examples of reticulate evolution [19, 8]. The addition of reticulation nodes makes the number of possible phylogenetic network topologies far greater than the number of trees. Deriving these numbers are non-trivial, and so far have been restricted to specific classes of networks. For level-2 networks, there are 1,143 network topologies for three external nodes, compared with only three tree topologies [5].

The complexity of phylogenetic tree and network space has some immediate implications. For example, to quantify the dissimilarity between sets of continuous parameter values we can choose among the $L^p$-norms (e.g. the Euclidean norm), but this is not possible for phylogenies. However, we often wish to measure this dissimilarity to compare different inference methods with each other, to compare them with ground truths, or to study how different sets of evolutionary units co-evolve [2]. Furthermore, novel efficient proposal kernels such as zig-zag cannot be applied to phylogenetic trees without thorough preliminary theoretical work [15].

These implications have motivated the development of myriad measures and metrics of dissimilarity and distances to compare phylogenetic trees, which have enabled the comparison of different phylogenies and improved our understanding of the algorithms used in phylogenetic inference [29]. We can classify many of these measures and metrics into three broad categories: clade-based, move-based, and geometric.

Clade-based measures are based on the presence or absence of clades (or, for unrooted phylogenies, splits). These may be limited to the difference in topology as in the Robinson-Foulds (RF) distance [23], or incorporate branch lengths as in weighted RF and branch score (BS) distances [16]. Move-based measures are based on the number of random-walk moves, such as Nearest-Neighbor In-

terchange (NNI), needed to modify one topology to be identical with another topology [1]. Geometric measures rely on the embedding of phylogenies in geometric spaces, such as the Billera-Holmes-Voghtmann (BHV), $\tau$ and $t$ spaces [3, 10]. The distances between phylogenies is then the shortest path between the corresponding points.

The additional complexity of phylogenetic networks means that available measures and metrics are far less developed. However, the need to accurately model evolution with reticulation demands much greater development and motivates our present study. Presently, beyond simple identity, there are several measures of topological dissimilarity with proofs of constituting metrics on subspaces of phylogenetic networks, e.g., [20, 6, 7].

None of these existing measures considers branch lengths or node ages, despite the importance of these distances in evolutionary biology. To address this absence, in this paper we propose rooted network branch score (rNBS) and average path distance (APD), two novel measures to compute dissimilarity in between two rooted phylogenetic networks $\Psi_1$ and $\Psi_2$ that are sensitive to branch lengths in addition to the topology. When comparing pairs of simulated networks that undergo reticulation elimination and branch-length scaling, we observe an increase in dissimilarity value from both measures with respect to the amount of distortion applied to one of the pairs. Additionally, when we use rNBS, APD, and the topological distance of [20] to cluster networks obtained from a Bayesian Markov chain Monte Carlo (MCMC) sample based on their dissimilarity, we find that both rNBS and APD can highlight structure within searches of tree space that may not be obvious from other parameters and statistics. Thus, we believe both rNBS and APD are suitable in evaluating, prototyping and refining network inference. Both measures are implemented in PhyloNet [26, 28] and are publicly available to download and use. The source code is available at https://github.com/NakhlehLab/PhyloNet/.

## 2 Methods

Our focus in this paper is binary evolutionary (or, explicit) phylogenetic networks.

**Definition 1.** *The topology of a phylogenetic network $\Psi$ is a rooted directed acyclic graph $(V, E)$ such that $V$ contains a unique node with in-degree 0 and out-degree 2 (the root) and every other node has either in-degree 1 and out-degree 2 (an internal tree node), in-degree 1 and out-degree 0 (an external tree node, or leaf), or in-degree 2 and out-degree 1 (a reticulation node). Edges incident into reticulation nodes are referred to as reticulation edges. The leaves are bijectively labeled by a set $\mathcal{X}$ of taxa, with $|\mathcal{X}| = n$. The phylogenetic network $\Psi$ has branch lengths $\lambda : E \to \mathbb{R}^+$, where $\lambda_b$ denotes the branch length of branch $b$ in $\Psi$.*

In this section, we propose two different methods of measuring the dissimilarity of a pair of phylogenetic networks $\Psi_1$ and $\Psi_2$ while taking their branch lengths into account.

### 2.1 Rooted Network Branch Score (rNBS)

In this subsection, we describe introduce a dissimilarity measure based on viewing a network in terms of the trees it displays, similar to the tree-based measure for topological comparison of phylogenetic networks implemented in PhyloNet [26].

**Definition 2.** *Let $\Psi$ be a phylogenetic network. A tree $T$ is displayed by $\Psi$ if it can be obtained from $\Psi$ by removing for each reticulation node exactly one of the edges incident into it followed by repeatedly applying forced contractions until no nodes of in- and out-degree 1 remain. A forced contraction of a node $u$ of in-degree* 1 *and out-degree* 1 *consists of (i) adding an edge from $u$'s parent to $u$'s child, and (ii) deleting node $u$ and the two edges that connect it to its parent and child. Finally, any edges above the root of the resulting tree are removed, so that the root has in-degree* 0 *and out-degree* 2*. We denote by $\mathcal{U}(\Psi)$ the set of all trees displayed by $\Psi$.*

It is important to note here that since branch lengths are taken into account, the set $\mathcal{U}(\Psi)$ can have trees with identical topologies but different branch lengths. This is illustrated in Fig. 1.

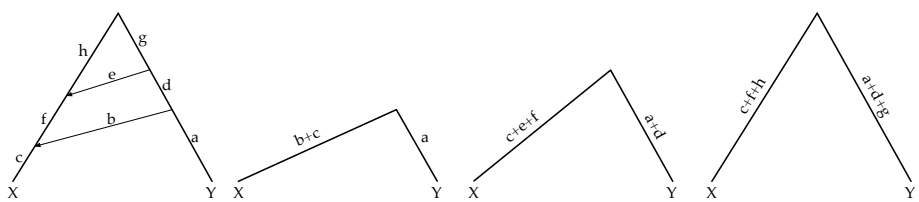



Fig. 1: A phylogenetic network (left) and its displayed trees. Since branch lengths are taken into account, the network displays three different trees. In the case of topology alone, the network displays a single tree.

We build a weighted complete bipartite graph $g = (\mathcal{U}(\Psi_1) \cup \mathcal{U}(\Psi_2), E)$ where the weight of edge $(t_1, t_2) \in E$ equals the rooted branch score of $t_1$ and $t_2$ [11], which is the rooted equivalent of the branch score of [16]. The rNBS of $\Psi_1$ and $\Psi_2$ is then computed as the minimum-weight edge cover of $g$ normalized by the number of edges in the edge cover (this is computable in polynomial time). Fig. 2 illustrates the rNBS of two phylogenetic networks.

Another important note is that the rNBS is not a metric, in particular failing to satisfy the condition that $\text{rNBS}(\Psi_1, \Psi_2) = 0$ if and only if $\Psi_1$ and $\Psi_2$ are isomorphic (while respecting the leaf labeling). Fig. 3 shows two networks that display the same set of trees (even when branch lengths are included). One setting of the network branch lengths that would lead to this scenario is given by the following Rich Newick strings of the two networks:

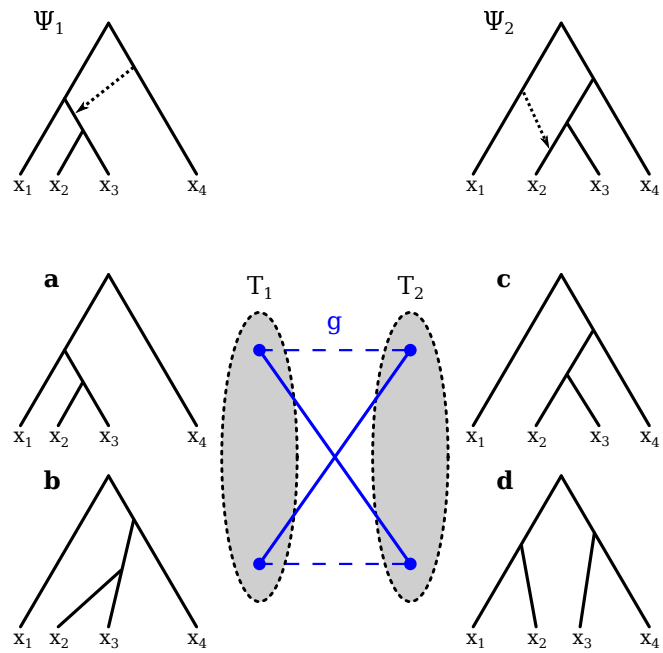



Fig. 2: Illustrating the rooted network branch score (rNBS) of two phylogenetic networks $\Psi_1$ and $\Psi_2$. $T_1 = \mathcal{U}(\Psi_1)$ and consists of the two trees (**a**-**b**) with distinct branch lengths and $T_2 = \mathcal{U}(\Psi_2)$ and consists of the two trees (**c**-**d**) with distinct branch lengths. The complete bipartite graph $g = (T_1 \cup T_2, E)$ is shown. Assuming the minimum-weight edge cover consists of the two edges depicted by the blue solid lines, then $\text{rNBS}(\Psi_1, \Psi_2) = w(\mathbf{a}, \mathbf{d}) \cdot 0.5 + w(\mathbf{b}, \mathbf{c}) \cdot 0.5$.

```
Psi1 = (((((X2:7.0)#H2:4.0)#H1:6.0,X3:8.0):4.0,(#H1:4.0,
X4:5.0):2.0):3.0,(#H2:2.0,X1:2.0):1.0)Root;
Psi2 = ((((X2:4.0)#H1:13.0,X3:8.0):4.0,((#H1:3.0)#H2:8.0,
X4:5.0):2.0):3.0,(#H2:2.0,X1:2.0):1.0)Root;
```

In this case, $\text{rNBS}(\Psi_1, \Psi_2) = 0$ even though the two networks are different.

### 2.2 Average Path Distance (APD)

Here we present a dissimilarity measure that is based directly on the networks, rather than the trees they display. We view phylogenetic networks $\Psi_1$ and $\Psi_2$ as two $n \times n$ matrices $M_1$ and $M_2$, respectively, where entries $[i, j]$ in $M_1$ and $M_2$ correspond to the average path distance between the two leaves labeled $i$ and $j$ in $\Psi_1$ and $\Psi_2$, respectively.

In a phylogenetic tree, the path distance between two leaves is the sum of weights of edges on the unique (simple) path between those two leaves. In a phylogenetic network, there could be more than one path between two leaves.

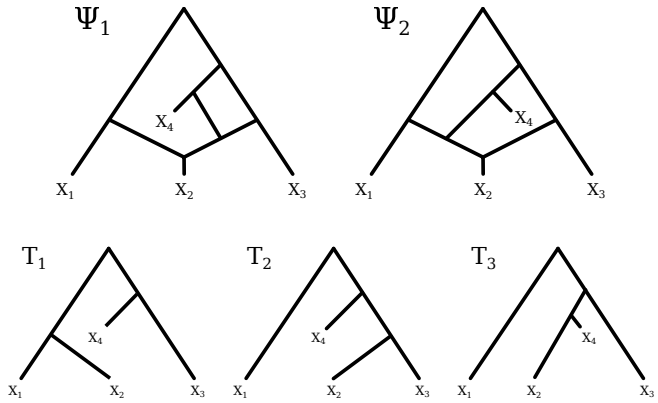



Fig. 3: Two networks $\Psi_1$ and $\Psi_2$ that display the same set of trees $\{T_1, T_2, T_2\}$ (adapted from [21]).

let $\mathcal{M}(\Psi, i, j)$ be the set of all most recent common ancestors (MRCAs) of $i$ and $j$ in network $\Psi$. Here, an MRCA is a node from which there is a path to $i$ and a path to $j$ and these two paths do not share any edge. For example, in the network shown in Fig. 4, there are three MRCAs of $X_2$ and $X_3$. A path length between two leaves $i$ and $j$ is the sum of branch lengths on the paths from $i$ and $j$ to an MRCA. The average path distance (APD) between two leaves is the average of all such paths between the two leaves.

After building matrices $M_1$ and $M_2$ for networks $\Psi_1$ and $\Psi_2$, respectively, APD can be computed by taking the Frobenius norm of the matrix difference:

$$\mathrm{APD}(\Psi_1, \Psi_2) = \|M_1 - M_2\|_F. \tag{1}$$

We can also obtain a normalized APD measure (NormAPD) as follows, which assumes $\Psi_1$ is a reference network and is useful in settings where an inferred network is compared to a reference one:

$$\mathrm{NormAPD}(\Psi_1, \Psi_2) = \frac{\|M_1 - M_2\|_F}{\|M_1\|_F} = \frac{\mathrm{APD}(\Psi_1, \Psi_2)}{\|M_1\|_F}. \tag{2}$$

It is important to note that the APD of the two networks of Fig. 3 is also 0 (under the branch length settings above); therefore, APD is not a distance metric either.

## 3 Results and Discussion

### 3.1 Dissimilarity Under Various Network Perturbations

To study the behavior of our dissimilarity measures on pairs of related networks, we simulated rooted phylogenetic networks that are leaf-labeled by the same

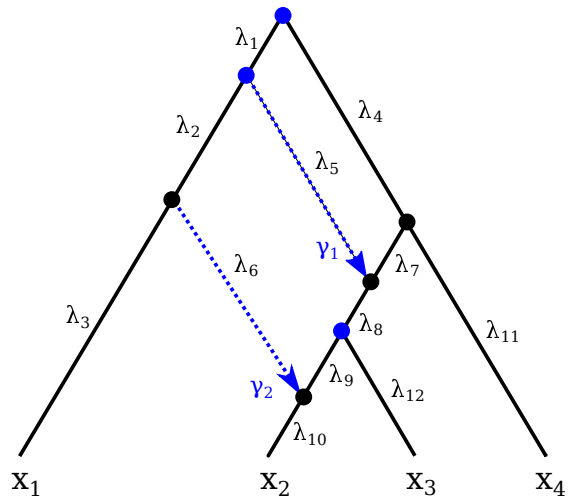
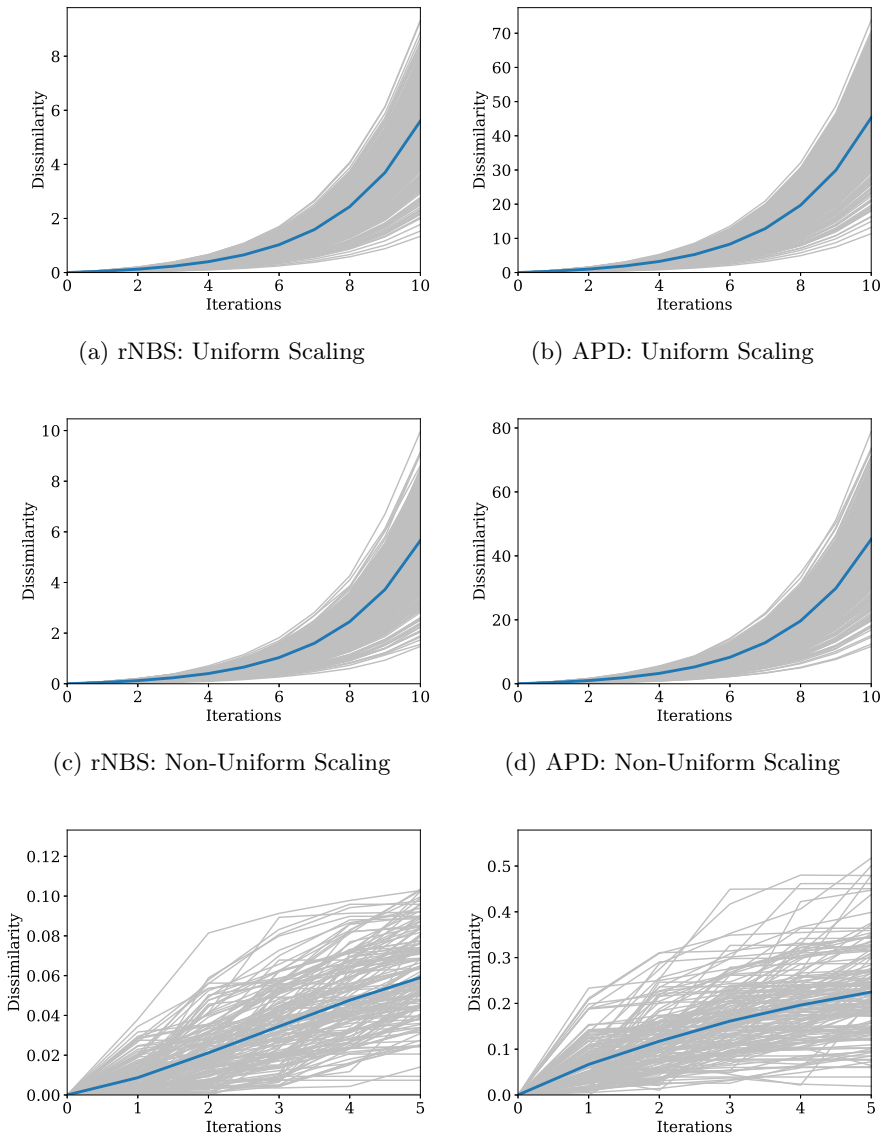



Fig. 4: Illustrating the average path distance (APD) on a phylogenetic network. The blue solid circles correspond to the three MRCAs of $X_2$ and $X_3$. The distances between $X_2$ and $X_3$ that go through these three MRCAs are: $(\lambda_{10}+\lambda_6+\lambda_2+\lambda_1)+(\lambda_{12}+\lambda_8+\lambda_7+\lambda_4)$ (through the root node as the MRCA), $(\lambda_{10}+\lambda_6+\lambda_2)+(\lambda_{12}+\lambda_8+\lambda_5)$ (through the MRCA that is a child of the root), and $(\lambda_{10}+\lambda_9)+\lambda_{12}$ (through the MRCA that is the parent node of $X_3$). The average path distance (APD) of $X_2$ and $X_3$ is the average of these three distances.

set of 8 taxa using the SpeciesNetwork [30] add-on in BEAST 2.5 [4]. For the simulation, we set the origin to 0.1, birth rate to 30, and hybridization rate to 5. We then filtered out networks that contain more than 7 reticulations to limit the number of trees displayed by the networks. This is particularly important for the running time of rNBS, as the Hungarian matching algorithm for computing the minimum-weight edge cover in our implementation takes $\mathcal{O}(|V|^3)$ time, where $V$ is the set of vertices in the bipartite graph. Our final data set contained 500 rooted phylogenetic networks with at most 7 reticulation nodes.

For each network $\Psi_0$ in the set of simulated networks, we generated perturbed versions of $\Psi_0$ and calculated the dissimilarity between the perturbed networks and $\Psi_0$.

**Uniform scaling.** Here we obtain $\Psi_i$ by scaling each branch of $\Psi_0$ by a factor of 1.5 for 10 iterations. At the end of iteration $i$, we have network $\Psi_i$ whose branch lengths are $(1.5)^i\lambda_0$, where $\lambda_0$ is the branch length of the particular branch on the original network $\Psi_0$. We computed $\mathrm{rNBS}(\Psi_0,\Psi_i)$ and $\mathrm{APD}(\Psi_0,\Psi_i)$, for $i = 1, 2, \ldots, 10$, and plotted the results as a function of the iteration number. The results for rNBS and APD are shown in Figs. 5a and 5b, respectively.




(a) rNBS: Uniform Scaling

(b) APD: Uniform Scaling

(c) rNBS: Non-Uniform Scaling

(d) APD: Non-Uniform Scaling

(e) rNBS: Removing Reticulations

(f) APD: Removing Reticulations

Fig. 5: Dissimilarity measures on pairs of networks using three different perturbations. (a-b) Perturbed networks are obtained by uniform scaling of all branches and comparing perturbed networks to a reference network. (c-d) Perturbed networks are obtained by non-uniform scaling of all branches and comparing perturbed networks to a reference network. (e-f) Perturbed networks are obtained by removing reticulation edges and comparing perturbed networks to a reference network. Each gray line shows the relationship between number of iterations and the amount of dissimilarity between the original and perturbed networks. Blue line shows the average dissimilarity value at each iteration.

Both figures show an exponential relationship between the number of iterations and the dissimilarity between the original and perturbed networks measured by rNBS and APD. Furthermore, we observe that neither of the measures have a consistent rate of increase in the dissimilarity values across pairs of networks. In fact, without normalization, both measures report higher values of dissimilarity when the network contains more edges.

**Non-uniform scaling.** Here, we obtain $\Psi_i$, for $i = 1, 2, \ldots, 10$ from $\Psi_0$ in a similar fashion to what we did in the case of uniform scaling, except that in each of the 10 iterations, we scaled each branch by a value drawn from $\mathcal{N}(1.5,\ 0.1^2)$. The results for rNBS and APD are shown in Figs. 5c and 5d, respectively.

Even though scale factors are sampled from a distribution, the rates of increase of dissimilarity values computed by both rNBs and APD are consistent with the results from uniform scaling.

**Reticulation elimination** Here, we further filter out networks that have fewer than 5 reticulation nodes, thus limiting our dataset to networks with at least 5 and at most 7 reticulations, which accounts for 157 networks. We produce $\Psi_i$ at each iteration as follows: $\Psi_1$ is obtained by removing a random reticulation edge from $\Psi_0$, and then $\Psi_i$ is obtained by removing a random reticulation edge from $\Psi_{i-1}$, for $i = 2, \ldots, 5$. The results for rNBS and APD are shown in Figs. 5e and 5f, respectively.

As the figure shows, the dissimilarity values increase as more reticulations are removed, but that increase slows down for the APD values with the number of removed reticulations, which is apparent in Fig. 5f. For the rNBS values (Fig. 5e), observe that the change in values is slow as the average value goes from 0 to around 0.06 when 5 reticulations are removed. The reason for this is that while reticulations are removed, the rest of the topologies and branch length are unperturbed. Therefore, it is natural that the removal of reticulations would have much less of an effect than, say, having all branches differ in lengths between the reference and perturbed networks.

### 3.2 Analyzing Posterior Samples

Phylogenetic analyses oftentimes produce thousands of candidates. To summarize the tree candidates, a consensus tree is often computed, e.g., [25, 14]. As an alternative to a single consensus tree, clustering was offered as early as 1991 by W. Maddison [17]. Subsequently, it was shown that clustering indeed provides a more powerful and informative summary than single consensus trees [24]. Here, we explore the use of our dissimilarity measures for clustering networks in the posterior sample of a Bayesian inference method. We also compare them to clustering based on the topological dissimilarity measure of [20], referred to hereafter as TD.

MCMC_SEQ [27] is a method in PhyloNet [26, 28] for Bayesian inference of phylogenetic network topology, divergence times, and inheritance probabilities,

along with various other parameters from sequence alignment data. The method samples from the posterior distribution of these parameters. This posterior sample can be summarized and analyzed using clustering based on our dissimilarity measures.

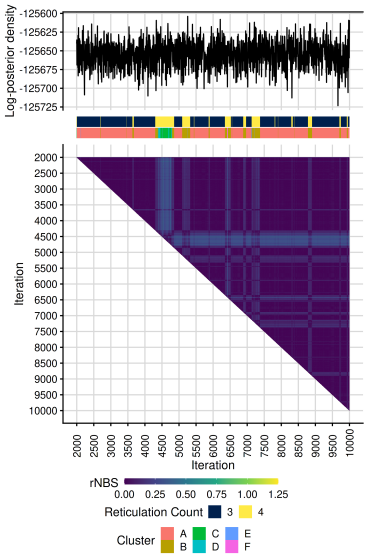
We analyze the posterior sample obtained on a simulated data set of 5 species and a single individual per species. The Markov chain Monte Carlo (MCMC) chain length was set to $50,000,000$, the burn-in period to $10,000,000$, and the sample frequency to 5000. The networks in the produced posterior sample vary in their number of reticulation nodes between 0 (i.e., trees) and 4. However, if we only consider the samples after burn-in period is completed (that is, ignoring the first 2,000 samples), the number of reticulations alternate between 3 and 4. Since clustering would require computing all pairwise distances between the networks in the sample, we further reduced the sample by keeping only every 5th sample, i.e., the 2000th, 2005th, ..., and 9995th samples (the first $2,000$ samples were discarded as part of the burn-in period). We then computed pairwise dissimilarity matrices using rNBS, APD, and topological dissimilarity (TD) [20]. Afterwards, we clustered the posterior networks in two different ways:

1. Clustering I: We partitioned the samples based on the number of reticulation nodes in the networks, thus resulting in two clusters, one consisting of all networks with 3 reticulations and another consisting of all networks with 4 reticulations.
2. Clustering II: We applied agglomerative clustering to the pairwise distance matrices, and set the linkage criterion to "complete," which makes the clustering algorithm use the maximum distances between all observations of the two sets when merging pairs of clusters. This resulted in 6 clusters, labeled $A$-$F$ below.

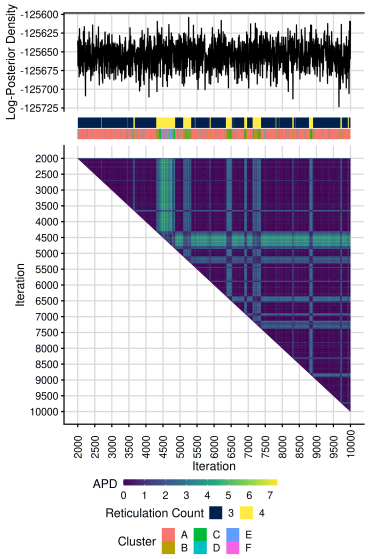
As the number of reticulations in networks is a major distinguishing factor when contrasting networks inferred on biological data, the rationale of the two ways of clustering is to understand (i) how the different iterations of MCMC_SEQ correlated in terms of the clustering, and (ii) whether clustering II is a refinement of clustering I.

For visualization, we plot the log-posterior density per iteration, number of reticulations per iteration (which corresponds to clustering I), the assigned cluster per iteration (which corresponds to clustering II), and the pairwise dissimilarity matrix computed by rNBS, APD, and topological dissimilarity (TD) [20]. Here, we did not visualize the clustered heatmaps, but rather kept them ordered according to the sampling order since we are focused on understanding the correlation, in terms of dissimilarity, between adjacent samples. The results are shown in Fig. 6.

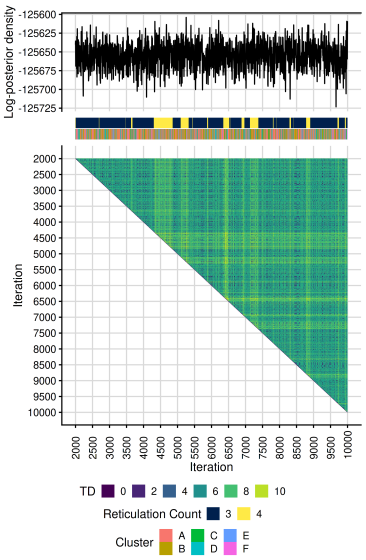
From Fig. 6a we can observe that rNBS, APD, and TD all cluster the networks with respect to the number of reticulations, shown by the purple vs blue shades in Fig. 6a for rNBS, purple vs green shades in Fig. 6b for APD, and green vs yellow shades inn Fig. 6c for TD. While all heat-maps support the clustering by the number of reticulations, the clusters are best defined based on APD




(a) rNBS




(b) APD




(c) Topological Dissimilarity (TD)

Fig. 6: Clustering of samples (networks and their branch lengths) from the posterior distribution obtained by MCMC_SEQ. Each panel shows the log-posterior density, the number of reticulations in each sampled network, the assigned clustering label obtained by agglomerative clustering, and the heatmap of the dissimilarity values arranged by the number of MCMC iterations. Panels (a)—(c) correspond to the three dissimilarity measures rNBS, APD, and TD, respectively.

pairwise distances, followed by rNBS pairwise distances, and then least defined based on TD pairwise distances.

Further, the agglomerative clustering obtained from rNBS labels 79.6% of the networks as $A$, 14.5% of the networks as $B$, 3.6% of the networks as $C$, 1.7% of the networks as $D$, 0.3% of the networks as $E$, and 0.3% of the networks as $F$.

The agglomerative clustering obtained from APD labels 76.5% of the networks as $A$, 8.5% of the networks as $B$, 6.1% of the networks as $C$, 5.6% of the networks as $D$, 2.5% of the networks as $E$, and 0.8% networks as $F$.

The agglomerative clustering obtained from TD labels 42.2% of the networks as $A$, 25.6% of the networks as $B$, 10.6% of the networks as $C$, 10.4% of the networks as $D$, 9.3% of the networks as $E$, and 1.9% of the networks as $F$.

More interestingly, agglomerative clustering obtained based on APD or rNBS pairwise distances shows distinct sub-clusters within the two clusters based on the number of reticulations. For example, the sampled networks from iterations $4400 - 4500$ are labeled by APD-based clustering as $C$, the sampled networks from iterations $4700 - 4800$ as $D$, and the sampled networks from iterations $6950 - 7000$ as $B$ even though all contain 4 reticulations. Similarly, the iterations $4400 - 4500$ in rNBS contain samples labeled $B$, $C$, and $D$ even though they only contain 4 reticulations.

This demonstrates that our measures can help discern patterns of similarity in terms of the posterior values and obtain a refined clustering of the posterior sample. The fact that these structures were only sampled in one segment of the posterior chain also suggests it is less converged than the log-posterior probability trace suggests, and that APD- and rNBS-based clustering are more powerful diagnostic tools for convergence.

### 3.3 Runtime Comparison

In this section, we report on the runtimes of rNBS, APD, and topological dissimilarity (TD) with respect to the number of reticulations in the pairs of networks. All experiments were run on a desktop running Linux Mint 20.3 on a single AMD Ryzen 9 5900X 12-Core Processor and 31.3 GiB available memory. Fig. 7 summarizes the results.

We find that APD is one fold slower than topological dissimilarity. Additionally, while the runtime increases as the number of reticulations increases, the rate of increase is very small. When there are no reticulations, APD can compute approximately $2,082$ dissimilarities per second, and when there are 10 reticulations, APD can compute approximately 649 dissimilarities per second. In contrast, rNBS slows down significantly as the number of reticulations increases, which supports the asymptotic time complexity of the edge covering algorithms used for bipartite graphs formulated for rNBS ($11,278$ per second when there are no reticulations vs. 0.03 per second when there are 10 reticulations). It should be noted that our implementation of APD is written in Java and does not currently use any libraries or hardware acceleration to transform matrices.

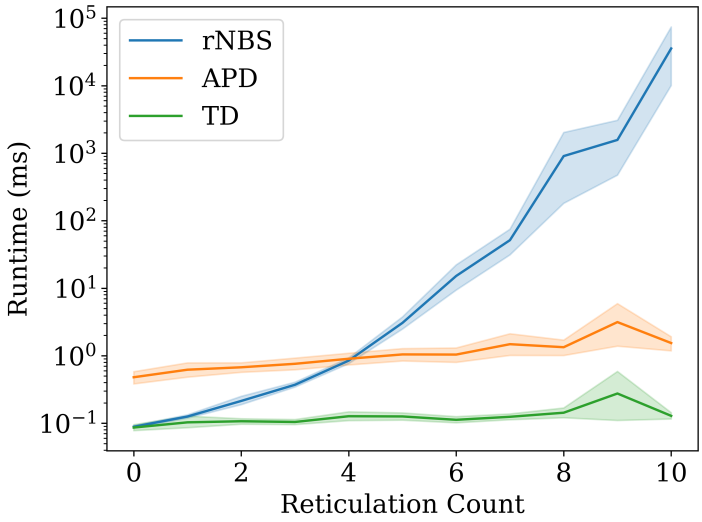



Fig. 7: Runtimes of rNBS, APD, and topological dissimilarity (TD) with respect to number of reticulation nodes in the networks. Time is measured in milliseconds.

## 4 Conclusions and Future Work

By taking branch lengths into account, our two novel dissimilarity measures of phylogenetic networks will address an important deficit in the ability to analyse phylogenetic network space and its exploration. By taking different approaches to measuring dissimilarity, and through our analysis of scaled, perturbed, and MCMC sampled phylogenetic networks, we have shown that the path distance approach has more immediate promise than edge-covering of displayed trees. We implemented both measures in PhyloNet [26, 28] and studied their properties on pairs of perturbed networks. Furthermore, we illustrated the use of the dissimilarity measures for clustering and summarizing the phylogenetic networks in a posterior sample of the Bayesian inference method MCMC_SEQ.

There are many directions for future work, two of which we will discuss here.

*Incorporating inheritance probabilities in the dissimilarity measures.* Statistical inference of phylogenetic networks results not only in estimates of the topologies and branch lengths, but also of inheritance probabilities, which annotate the reticulation edges. The measures we presented above ignore inheritance probabilities. In particular, the APD measure treats all paths equally. One possible extension that accounts for inheritance probabilities is to weight each path that is counted by a combination of the inheritance probabilities used by the reticulation edges on that path. Thus, a further extension to average path distance can incorporate inheritance probabilities by weighting the path distance between two leaves $i$ and $j$ at each MRCA. Thus, for each MRCA, we multiply the path

distance from the MRCA to leaf $i$ with the inheritance probabilities along the path, perform the same for the path from the MRCA to leaf $j$, and sum them up. Finally, the weighted path distance (WPD) becomes the sum of weighted distances at each MRCA of $i$ and $j$. For example, for the network of Fig. 4, the weighted path distance between $x_2$ and $x_3$ in the network is the sum of weighted path distances at the three MRCAs: $[(\lambda_{10}+\lambda_6+\lambda_2+\lambda_1)\cdot\gamma_2+(\lambda_{12}+\lambda_8+\lambda_7+\lambda_4)\cdot(1-\gamma_1)]+[(\lambda_{10}+\lambda_6+\lambda_2)\cdot\lambda_2+(\lambda_{12}+\lambda_8+\lambda_5)\cdot\lambda_1]+[(\lambda_{10}+\lambda_9)\cdot(1-\lambda_2)+\lambda_{12}]$. Similarly, for the rNBS measure, we can weight each tree by the product of the inheritance probabilities of the reticulation edges used to display the tree. However, it is important to note here that these measures could be very sensitive to inaccuracies of the inheritance probability estimates. For example, two networks that are identical in terms of topologies and branch lengths but vary significantly in the inheritance probabilities of even a small number of reticulation edges could result in a very large dissimilarity value, unless the inheritance probabilities are weighted carefully. We will explore these directions.

*Tree-based dissimilarity in the presence of incomplete lineage sorting.* Zhu *et al.* [31] showed that when incomplete lineage sorting occurs, inference and analysis of phylogenetic networks are more adequately done with respect to the set of parental trees of the network, rather than the set of displayed trees. Thus, a different approach to defining the rooted network branch score could involve building a complete bipartite graph on the sets of parental trees of the two networks. The major challenge here is computational; as we demonstrated in our analysis of rNBS, runtime may dramatically increase with the number of reticulations $k$ for a bipartite graph approach. Worse, while a phylogenetic network has up to $2^k$ displayed trees, the number of parental trees could be significantly much larger to the extent that $2^k$ is feasible in the case of a small $k$ whereas computing the set of all parental trees explicitly for the same value of $k$ can be infeasible. Therefore, faster computations and/or heuristics for computing such a dissimilarity measure would be needed.

**Acknowledgements** We thank Zhen Cao for contributing the MCMC_SEQ posterior sample files. This work was supported in part by NSF grants CCF-1514177, CCF-1800723 and DBI-2030604 to L.N.

## References

1. Allen, B.L., Steel, M.: Subtree transfer operations and their induced metrics on evolutionary trees. Annals of combinatorics **5**(1), 1–15 (2001)
2. Avino, M., Ng, G.T., He, Y., Renaud, M.S., Jones, B.R., Poon, A.F.Y.: Tree shape-based approaches for the comparative study of cophylogeny. Ecology and Evolution **9**(12), 6756–6771 (2019). https://doi.org/https://doi.org/10.1002/ece3.5185, https://onlinelibrary.wiley.com/doi/abs/10.1002/ece3.5185
3. Billera, L.J., Holmes, S.P., Vogtmann, K.: Geometry of the space of phylogenetic trees. Advances in Applied Mathematics **27**(4), 733–767 (2001)
4. Bouckaert, R., Vaughan, T.G., Barido-Sottani, J., Duchêne, S., Fourment, M., Gavryushkina, A., Heled, J., Jones, G., Kühnert, D., De Maio, N.,

Matschiner, M., Mendes, F.K., Müller, N.F., Ogilvie, H.A., du Plessis, L., Popinga, A., Rambaut, A., Rasmussen, D., Siveroni, I., Suchard, M.A., Wu, C.H., Xie, D., Zhang, C., Stadler, T., Drummond, A.J.: Beast 2.5: An advanced software platform for bayesian evolutionary analysis. PLOS Computational Biology **15**(4), 1–28 (04 2019). https://doi.org/10.1371/journal.pcbi.1006650, https://doi.org/10.1371/journal.pcbi.1006650

5. Bouvel, M., Gambette, P., Mansouri, M.: Counting phylogenetic networks of level 1 and 2. Journal of Mathematical Biology **81**(6), 1357–1395 (2020)
6. Cardona, G., Llabrés, M., Rosselló, F., Valiente, G.: Metrics for phylogenetic networks i: Generalizations of the robinson-foulds metric. IEEE/ACM Transactions on Computational Biology and Bioinformatics **6**(1), 46–61 (2008)
7. Cardona, G., Llabrés, M., Rosselló, F., Valiente, G.: Metrics for phylogenetic networks ii: Nodal and triplets metrics. IEEE/ACM Transactions on Computational Biology and Bioinformatics **6**(3), 454–469 (2008)
8. Elworth, R., Ogilvie, H.A., Zhu, J., Nakhleh, L.: Advances in computational methods for phylogenetic networks in the presence of hybridization. In: Bioinformatics and Phylogenetics, pp. 317–360. Springer (2019)
9. Felsenstein, J.: The number of evolutionary trees. Systematic Zoology **27**(1), 27–33 (1978)
10. Gavryushkin, A., Drummond, A.J.: The space of ultrametric phylogenetic trees. Journal of Theoretical Biology **403**, 197–208 (2016). https://doi.org/https://doi.org/10.1016/j.jtbi.2016.05.001, https://www.sciencedirect.com/science/article/pii/S002251931630073X
11. Heled, J., Drummond, A.J.: Bayesian Inference of Species Trees from Multilocus Data. Molecular Biology and Evolution **27**(3), 570–580 (11 2009). https://doi.org/10.1093/molbev/msp274, https://doi.org/10.1093/molbev/msp274
12. Hoffer, B.L.: Language borrowing and the indices of adaptability and receptivity. Intercultural Communication Studies **14**(2), 53 (2005)
13. Jain, C., Rodriguez-R, L.M., Phillippy, A.M., Konstantinidis, K.T., Aluru, S.: High throughput ani analysis of 90k prokaryotic genomes reveals clear species boundaries. Nature Communications **9**(1), 5114 (Nov 2018). https://doi.org/10.1038/s41467-018-07641-9, https://doi.org/10.1038/s41467-018-07641-9
14. Kannan, L., Wheeler, W.: Maximum parsimony on phylogenetic networks. Algorithms for molecular biology : AMB **7**, 9 (05 2012). https://doi.org/10.1186/1748-7188-7-9
15. Koskela, J.: Zig-zag sampling for discrete structures and non-reversible phylogenetic mcmc. Journal of Computational and Graphical Statistics (2022). https://doi.org/10.1080/10618600.2022.2032722, https://doi.org/10.1080/10618600.2022.2032722, just accepted
16. Kuhner, M.K., Felsenstein, J.: A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Molecular biology and evolution **11**(3), 459–468 (1994)
17. Maddison, D.R.: The discovery and importance of multiple islands of most-parsimonious trees. Systematic Biology **40**(3), 315–328 (1991)
18. Michael, T.P., VanBuren, R.: Building near-complete plant genomes. Current Opinion in Plant Biology **54**, 26–33 (2020). https://doi.org/https://doi.org/10.1016/j.pbi.2019.12.009, https://www.sciencedirect.com/science/article/pii/S1369526619301244, genome studies and molecular genetics

19. Nakhleh, L.: Evolutionary phylogenetic networks: models and issues. In: Problem Solving Handbook in Computational Biology and Bioinformatics, pp. 125–158. Springer (2010)
20. Nakhleh, L.: A metric on the space of reduced phylogenetic networks. IEEE/ACM Transactions on Computational Biology and Bioinformatics **7**(2), 218–222 (2010). https://doi.org/10.1109/TCBB.2009.2
21. Pardi, F., Scornavacca, C.: Reconstructible phylogenetic networks: Do not distinguish the indistinguishable. PLOS Computational Biology **11**(4), 1–23 (04 2015). https://doi.org/10.1371/journal.pcbi.1004135, https://doi.org/10.1371/journal.pcbi.1004135
22. Rhie, A., McCarthy, S.A., Fedrigo, O., Damas, J., Formenti, G., Koren, S., Uliano-Silva, M., Chow, W., Fungtammasan, A., Kim, J., Lee, C., Ko, B.J., Chaisson, M., Gedman, G.L., Cantin, L.J., Thibaud-Nissen, F., Haggerty, L., Bista, I., Smith, M., Haase, B., Mountcastle, J., Winkler, S., Paez, S., Howard, J., Vernes, S.C., Lama, T.M., Grutzner, F., Warren, W.C., Balakrishnan, C.N., Burt, D., George, J.M., Biegler, M.T., Iorns, D., Digby, A., Eason, D., Robertson, B., Edwards, T., Wilkinson, M., Turner, G., Meyer, A., Kautt, A.F., Franchini, P., Detrich, H.W., Svardal, H., Wagner, M., Naylor, G.J.P., Pippel, M., Malinsky, M., Mooney, M., Simbirsky, M., Hannigan, B.T., Pesout, T., Houck, M., Misuraca, A., Kingan, S.B., Hall, R., Kronenberg, Z., Sović, I., Dunn, C., Ning, Z., Hastie, A., Lee, J., Selvaraj, S., Green, R.E., Putnam, N.H., Gut, I., Ghurye, J., Garrison, E., Sims, Y., Collins, J., Pelan, S., Torrance, J., Tracey, A., Wood, J., Dagnew, R.E., Guan, D., London, S.E., Clayton, D.F., Mello, C.V., Friedrich, S.R., Lovell, P.V., Osipova, E., Al-Ajli, F.O., Secomandi, S., Kim, H., Theofanopoulou, C., Hiller, M., Zhou, Y., Harris, R.S., Makova, K.D., Medvedev, P., Hoffman, J., Masterson, P., Clark, K., Martin, F., Howe, K., Flicek, P., Walenz, B.P., Kwak, W., Clawson, H., Diekhans, M., Nassar, L., Paten, B., Kraus, R.H.S., Crawford, A.J., Gilbert, M.T.P., Zhang, G., Venkatesh, B., Murphy, R.W., Koepfli, K.P., Shapiro, B., Johnson, W.E., Di Palma, F., Marques-Bonet, T., Teeling, E.C., Warnow, T., Graves, J.M., Ryder, O.A., Haussler, D., O'Brien, S.J., Korlach, J., Lewin, H.A., Howe, K., Myers, E.W., Durbin, R., Phillippy, A.M., Jarvis, E.D.: Towards complete and error-free genome assemblies of all vertebrate species. Nature **592**(7856), 737–746 (Apr 2021). https://doi.org/10.1038/s41586-021-03451-0, https://doi.org/10.1038/s41586-021-03451-0
23. Robinson, D., Foulds, L.: Comparison of phylogenetic trees. Mathematical Biosciences **53**(1), 131–147 (1981). https://doi.org/https://doi.org/10.1016/0025-5564(81)90043-2, https://www.sciencedirect.com/science/article/pii/0025556481900432
24. Stockham, C., Wang, L.S., Warnow, T.: Statistically based postprocessing of phylogenetic analysis by clustering. Bioinformatics **18**(suppl_1), S285–S293 (2002)
25. Swofford, D., Olson, G., Waddell, P., Hillis, D.: Molecular systematics, chapter entitled "phylogenetic inference" (1996)
26. Than, C., Ruths, D., Nakhleh, L.: Phylonet: a software package for analyzing and reconstructing reticulate evolutionary relationships. BMC bioinformatics **9**(1), 1–16 (2008)
27. Wen, D., Nakhleh, L.: Coestimating Reticulate Phylogenies and Gene Trees from Multilocus Sequence Data. Systematic Biology **67**(3), 439–457 (10 2017). https://doi.org/10.1093/sysbio/syx085, https://doi.org/10.1093/sysbio/syx085
28. Wen, D., Yu, Y., Zhu, J., Nakhleh, L.: Inferring phylogenetic networks using phylonet. Systematic biology **67**(4), 735–740 (2018)

29. Whidden, C., Matsen, Frederick A., I.: Quantifying MCMC Exploration of Phylogenetic Tree Space. Systematic Biology **64**(3), 472–491 (01 2015). https://doi.org/10.1093/sysbio/syv006, https://doi.org/10.1093/sysbio/syv006
30. Zhang, C., Ogilvie, H.A., Drummond, A.J., Stadler, T.: Bayesian Inference of Species Networks from Multilocus Sequence Data. Molecular Biology and Evolution **35**(2), 504–517 (12 2017). https://doi.org/10.1093/molbev/msx307, https://doi.org/10.1093/molbev/msx307
31. Zhu, J., Yu, Y., Nakhleh, L.: In the light of deep coalescence: revisiting trees within networks. BMC Bioinformatics **17**(14), 415 (Nov 2016). https://doi.org/10.1186/s12859-016-1269-1, https://doi.org/10.1186/s12859-016-1269-1