Are fast limited-data phylogenetic reconstruction algorithms useful for biological applications?

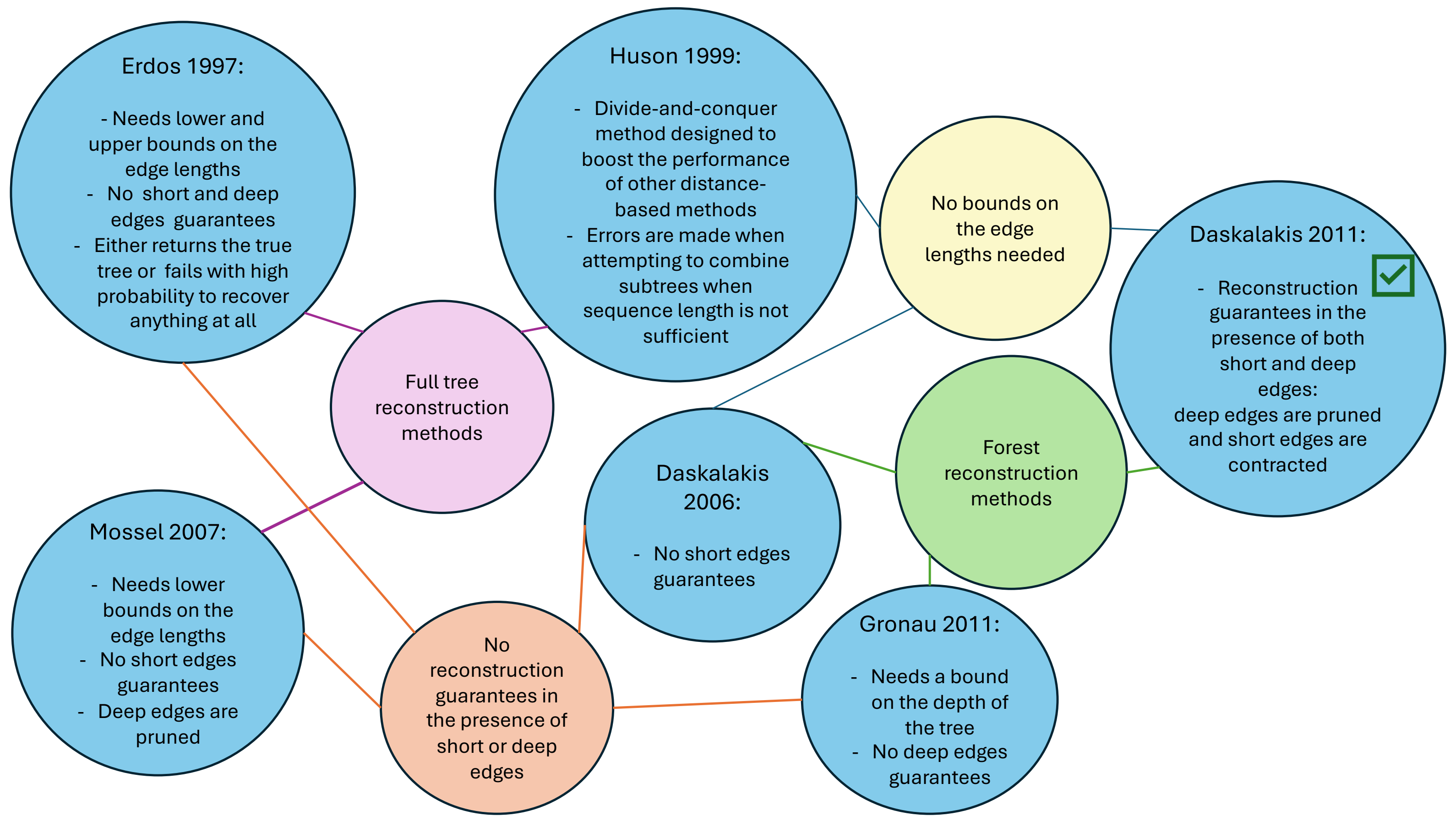
Anastasiia Kim, Ethan Romero-Severson, Andrey Lokhov, Marc Vuffray, Emma Goldberg
Los Alamos National Laboratory

Abstract

- Optimal algorithms are needed to infer phylogenies as global databases provide millions of pathogen genomic samples yet genome size is limited.
- Theoretical computer science community developed polynomial time distance-based methods to return the correct phylogeny from limited amount of data.
- However, it is less known how these methods perform in practice under reasonable biological assumptions for inferring large-scale phylogenies for public health applications.

Fast limited-data methods

- These methods guarantee tree or forest reconstruction (a collection of subtrees) with sequence lengths growing polynomially with the number of taxa, unlike Neighbor-Joining (NJ), which requires exponentially long sequences.
- Such methods also require user input parameters related to the edge lengths bounds, depth of the tree, or forest size.
- Forest reconstruction algorithms construct a forest of subtrees that share multiple properties with the original tree when sequence length is not sufficient.



Erdos 1997:

- Needs lower and upper bounds on the edge lengths
- No short and deep edges guarantees
- Either returns the true tree or fails with high probability to recover anything at all

Huson 1999:

- Divide-and-conquer method designed to boost the performance of other distance-based methods
- Errors are made when attempting to combine subtrees when sequence length is not sufficient

No bounds on the edge lengths needed

Daskalakis 2011: ☑

- Reconstruction guarantees in the presence of both short and deep edges: deep edges are pruned and short edges are contracted

Full tree reconstruction methods

Daskalakis 2006:

- No short edges guarantees

Forest reconstruction methods

Mossel 2007:

- Needs lower bounds on the edge lengths
- No short edges guarantees
- Deep edges are pruned

No reconstruction guarantees in the presence of short or deep edges

Gronau 2011:

- Needs a bound on the depth of the tree
- No deep edges guarantees

Forest reconstruction algorithm

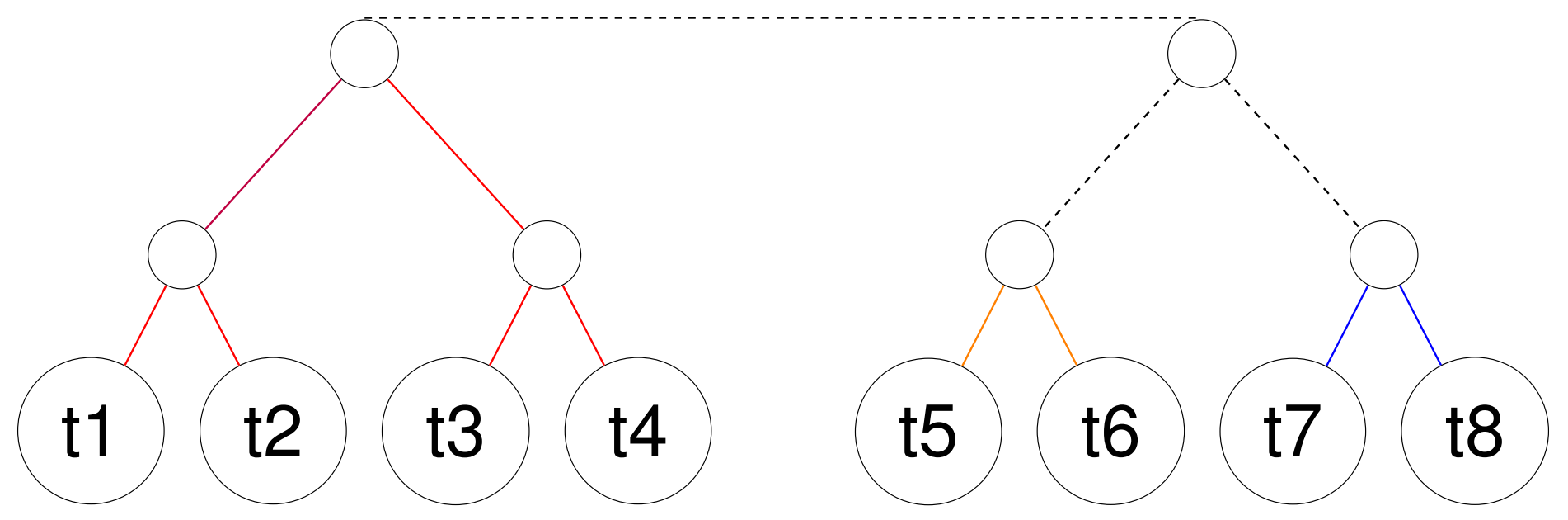
We compare the Daskalakis et. al 2011 forest reconstruction method with NJ.

- No restrictions on the edge lengths and depth of the tree.
- Requires user input parameters (τ, m, M) and $\hat{d}$ (observed distance matrix):
  - $\hat{d}$: a (τ, M)-distorted metric of d: if either d or $\hat{d} < M + \tau$ then $|d - \hat{d}| < \tau$.
  - τ: all edges smaller than τ are contracted, if all edges are larger than f then a tree can be recovered from any (τ, M)-distortion of d if $\tau < f/2$.
  - M: determines which edges are sufficiently long and lie on sufficiently short paths in the forest component.
  - m: determines the forest size, chosen such that $m < \frac{1}{2}(M - 3\tau)$.
- Reconstructs approximately-disjoint forest with chord depth $\approx M/2$ of the true phylogeny where deep edges are pruned and short edges are contracted.

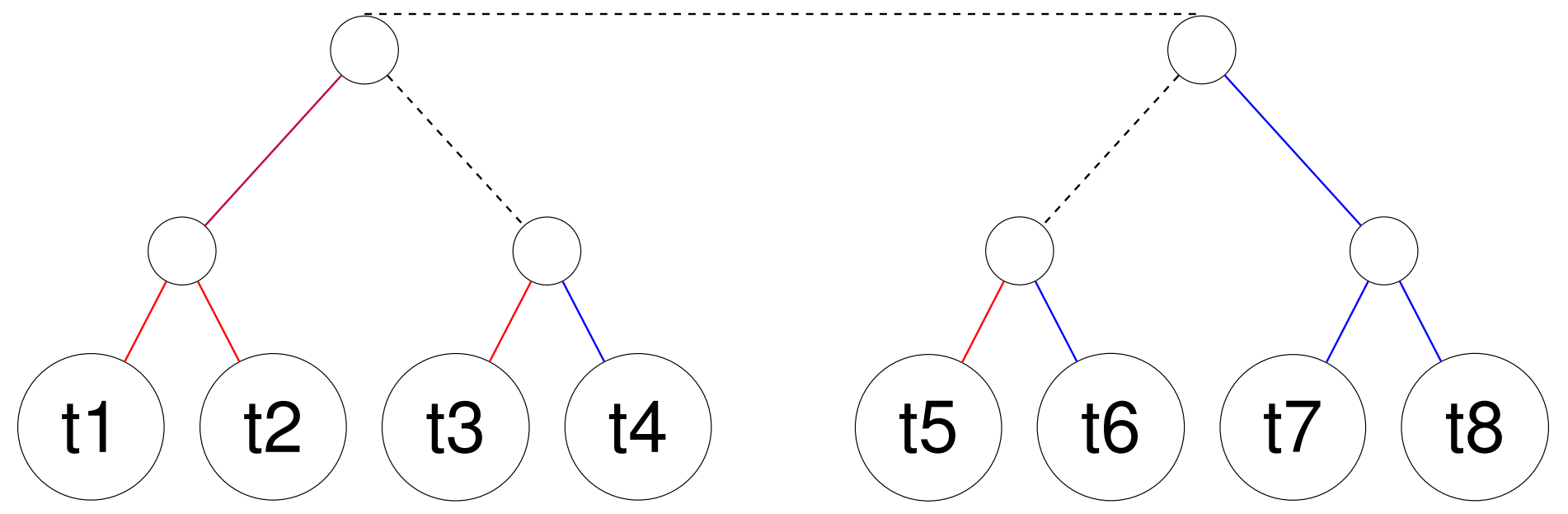
Testing cases under model-match

- Simulate $n = 8, 16, ..., 256$ tips random shaped trees with edge lengths $\lambda \sim$ Uniform(0.05, 0.1).
- Using IQ-TREE 2, simulate alignments under the Jukes-Cantor model with sequence lengths $k = 64, 128, ..., 1024$.
- Explore parameter space of the algorithm to examine the reconstruction:
  - Set τ values close to the half of the minimum edge length.
  - Set m values around the chord depth of the tree.
  - Pick values of M so that $m < \frac{1}{2}(M - 3\tau)$.
- Compare with NJ tree. In the case of the forest, compute the induced RF distance on the same leaf set.

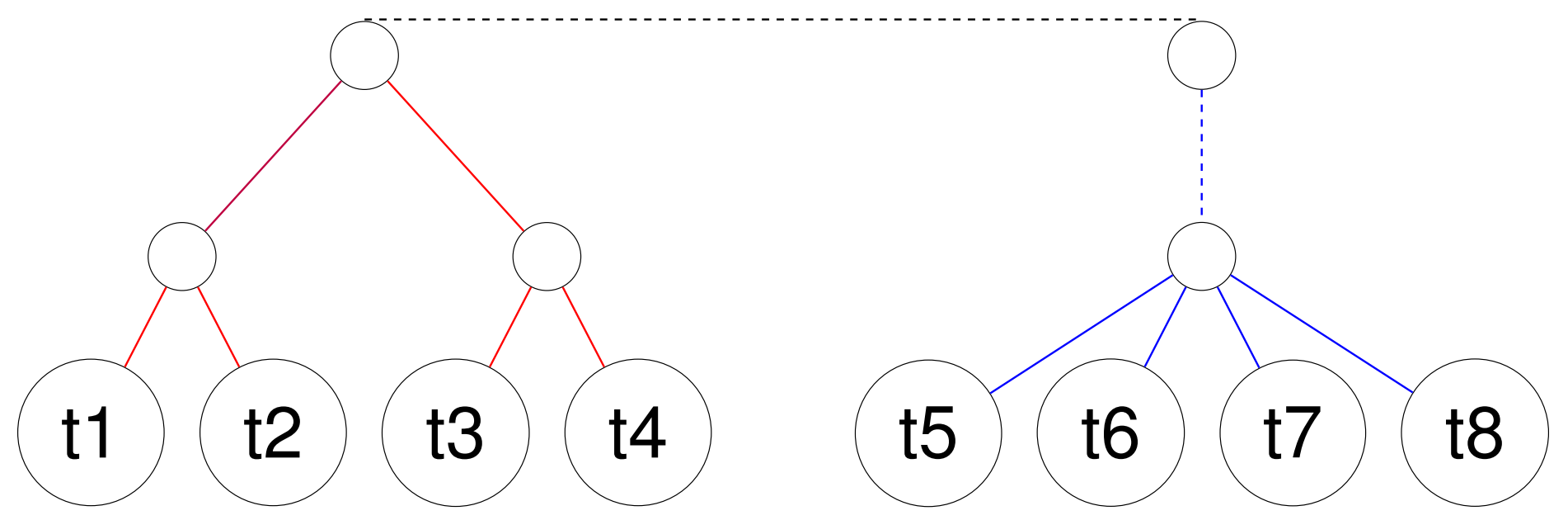
Examples of the algorithm forest output



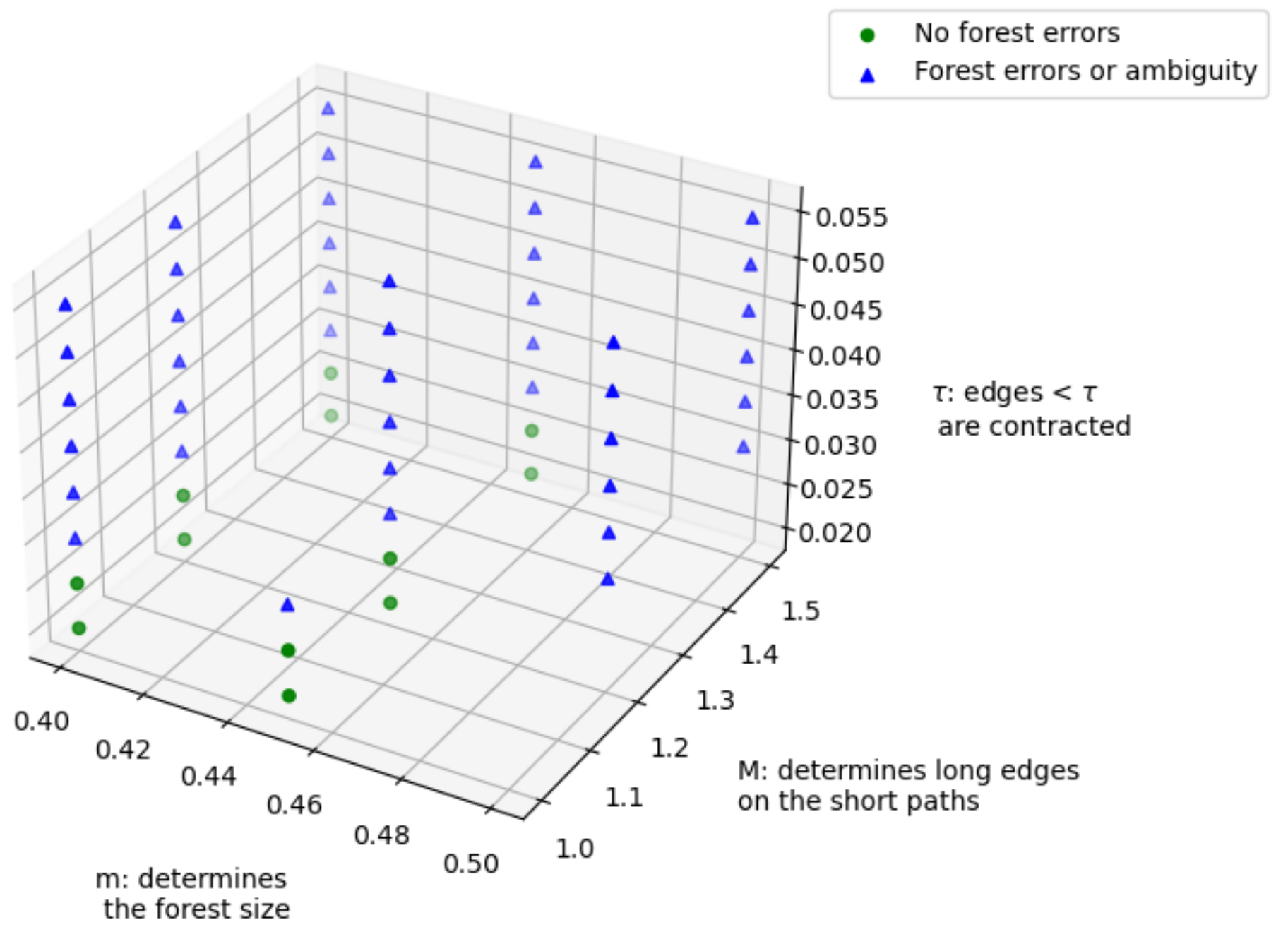
Algorithm returns a correct forest that is obtained from the original tree by removing the dashed edges.



Algorithm fails to split the tree correctly. It is impossible to obtain this forest by cutting some edges in the true phylogeny. Still, the topology within each connected component in this forest is preserved.

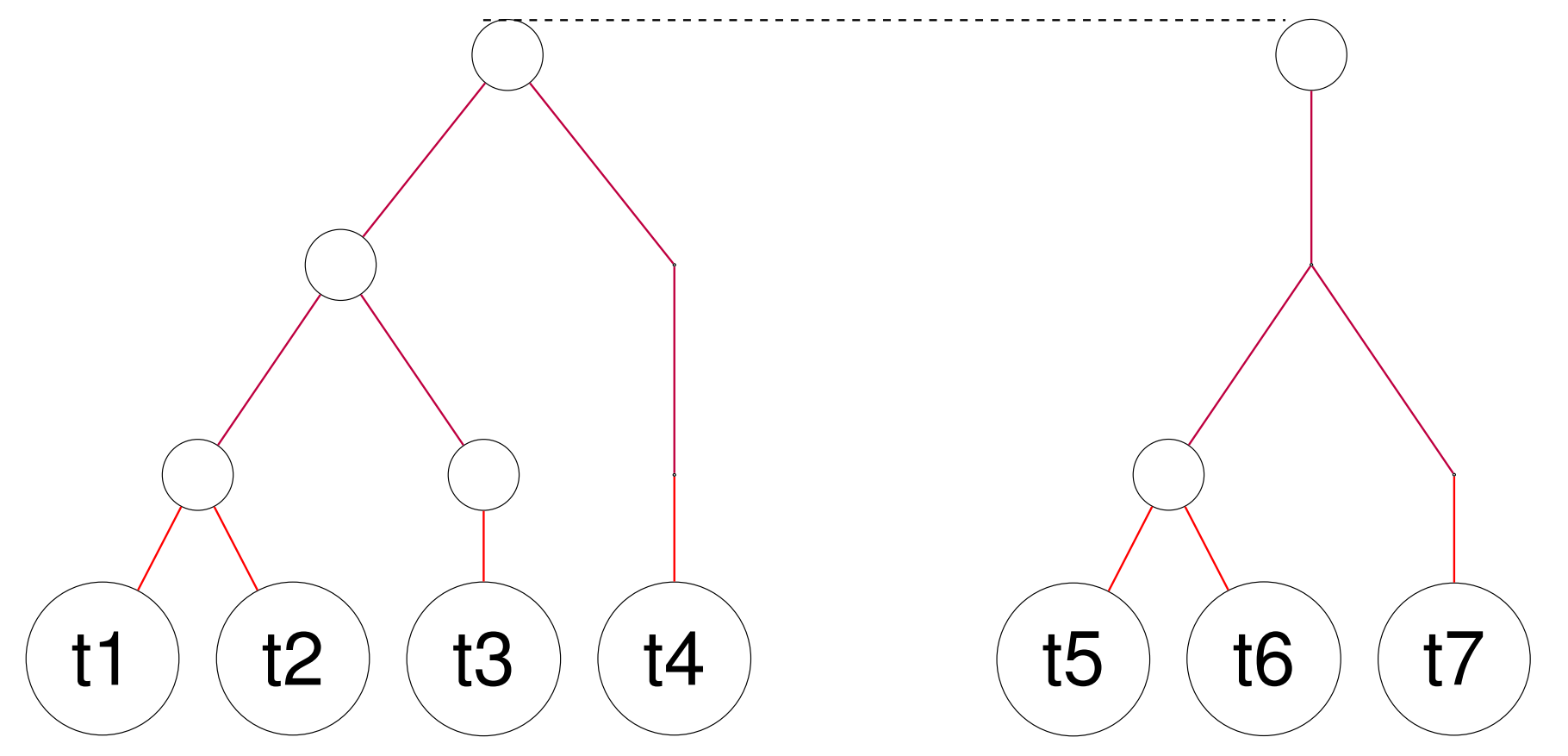


An example of edge contraction leading to multiple topologies the algorithm confused between. The algorithm could not tell the true topology.

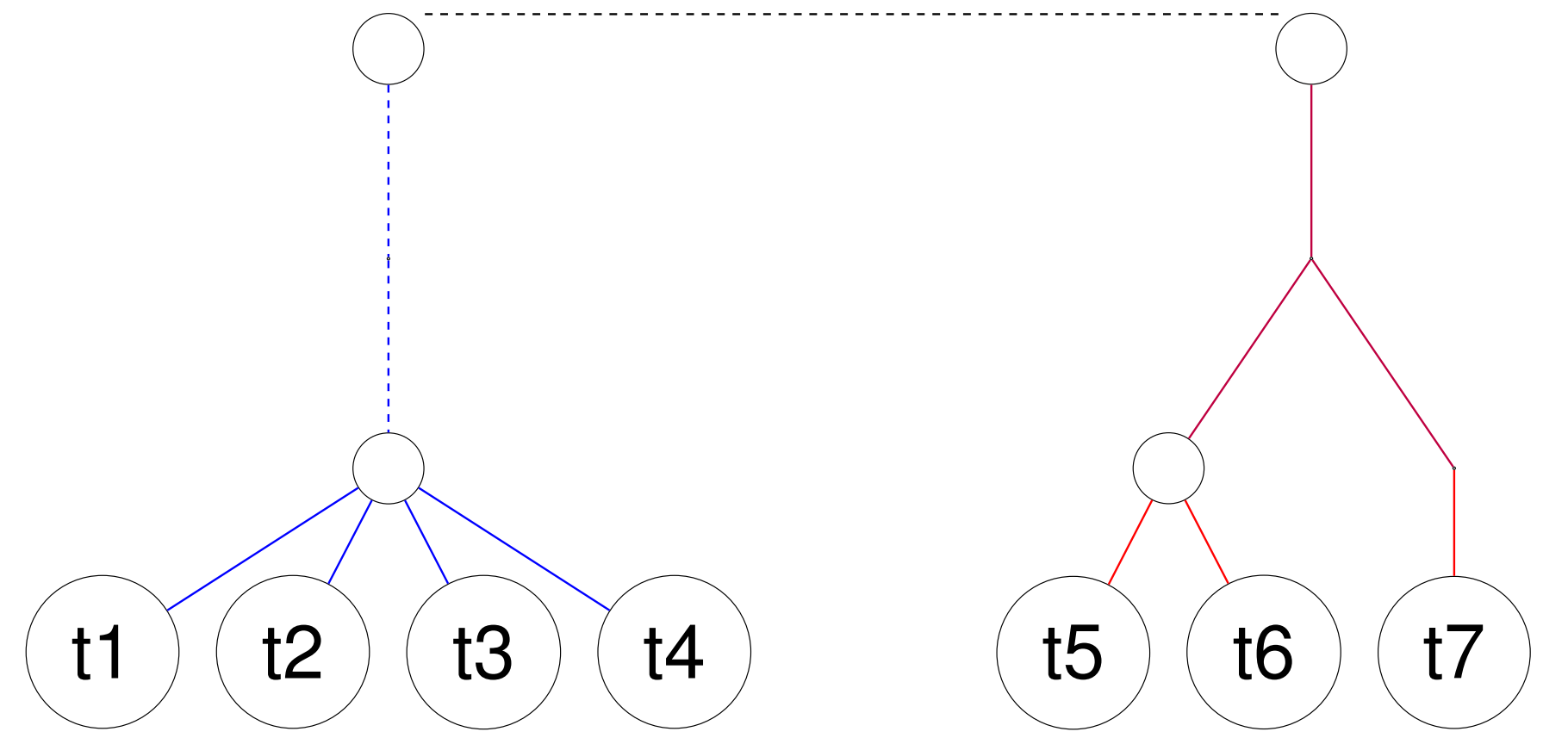


Correctness of the forest reconstruction depends on the input parameters.

Comparison with Neighbor-Joining



Correct forest component of size 7 of the true 10-taxon phylogeny.



Ambiguous forest component produced by the algorithm.

- NJ produced a forest component with the same topology as the true tree, except *t2* was swapped with *t4*.
- The algorithm produced a polytomy in one of the forest components, which was preferable to inferring incorrect taxa relationships, as NJ could do.
- In a simulation study, this algorithm and NJ performed comparably.
  - There were cases when NJ results were incorrect; however, the forest algorithm either reconstructed the correct full tree or the correct forest.
  - When the forest algorithm did not yield subtrees with RF = 0 while NJ produced correct subtrees for the same leaf sets, the forest algorithm still returned ambiguous results with unresolved polytomies.

Discussion

- Practically, the forest algorithm may reconstruct confident subtrees in the forest, which is better than reconstructing an incorrect full tree.
- Depending on the input parameters and sequence length, the algorithm may recover the full tree or the forest.
- One challenge is that the algorithm may not always recover a topologically sensible forest, but the reconstructed subtrees might still be useful:
  - Subtrees are correct for leaf sets, even though the corresponding leaf set is not always a clade in the true tree; still, these subtrees can be used to glue them into one supertree.
  - Given the forest subtrees, topological constraints can be put to use in another algorithm, which would restrict the tree search space.
- Testing this algorithm under model-mismatch conditions may result in more robust reconstruction than NJ.

References

- Daskalakis, C., Mossel, E. and Roch, S., 2011. Phylogenies without branch bounds: Contracting the short, pruning the deep. SIAM Journal on Discrete Mathematics, 25(2), pp.872-893.
- Mossel, E., 2007. Distorted metrics on trees and phylogenetic forests. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 4(1), pp.108-116.
- Erdős, P.L., Steel, M.A., Székely, L.A. and Warnow, T.J., 1999. A few logs suffice to build (almost) all trees (I). Random Structures Algorithms, 14(2), pp.153-184.

Managed by Triad National Security, LLC, for the U.S. Department of Energy's NNSA.
This study was supported by the LDRD program, Grant 20240198ER.