Supplementary - Material for the work Maximum Likelihood of Phylogenetic Networks

Guohua Jin[1], Luay Nakhleh[1], Sagi Snir[2], and Tamir Tuller[3]

[1] Dept. of Computer Science Rice University Houston, TX, USA
[2] Dept. of Mathematics University of California Berkeley, CA, USA
[3] School of Computer Science. Tel-Aviv University, Israel.

1 ML on Networks is NP-hard

Problem 1. 3-Satisfiability (3SAT) [1]
Input: A formula F over a set U of variables, collection C of clauses over U such that each clause $c \in C$ has $|c| = 3$.
Question: Is there a truth assignment for U that simultaneously satisfies all clauses in C?

Given such a formula F we construct a network $N(F)$ as follows. In the sequel, we use x is *connected* to y to indicate that x is a child of y.

1. A root R with a 1-leaf child.
2. For every variable we crate a node connected to a *diamond loop* as depicted in Fig. 1.

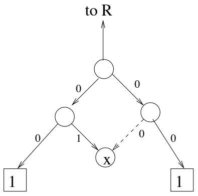


Fig. 1. For every variable we create a diamond gadget.

3. For every clause $c_i = (i \vee j \vee k)$ we generate a 1-leaf, called *clause leaf*, and connect it by a tree edge to variable node i and by a reticulation edge to an *intermediate node* c'_i. Intermediate node c'_i is connected by a tree edge to variable node j and by a reticulation edge to an intermediate node c''_i. Finally, intermediate node c''_i is connected by a tree edge to variable node k (see Fig. 2). The length of a (tree) edge emanating from a variable node is 1 if the literal is negated and 0 otherwise.
4. Finally, we create a complementing 1-leaf child connected by a 0 long edge to every internal node with out-degree 1.

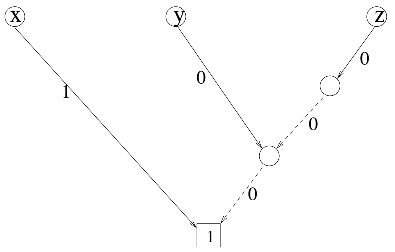


Fig. 2. A clause gadget for the clause $(\bar{x} \vee y \vee z)$. (complementing 1-leaves removed).

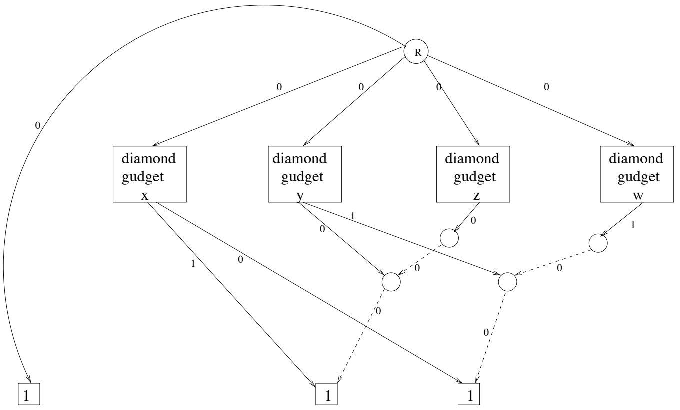


Fig. 3. The reduction from 3-SAT to the tiny best ancestral likelihood. Each variable has a node to which all literals of that variable are connected. In the figure we see the sub network representing the formula $(\bar{x} \vee y \vee z) \wedge (x \vee \bar{y} \vee \bar{w})$.

A complete network representing the formula $(\bar{x} \vee y \vee z) \wedge (x \vee \bar{y} \vee \bar{w})$ (without the complementing 1-leaves) is shown in Fig. 3.

Observation 1 *The network $N(F)$ is a valid phylogenetic network.*

Proof. We observe the following:

- There is a single internal node with in-degree 0.
- Every internal node has in-degree > 1.
- Every node with in-degree > 1 has exactly one entering tree edge and the rest are 0-length reticulation edges.
- The temporal constraint property holds.

Let S denote the leaf assignment under the reduction. Then we get the following claim:

Claim. F has a satisfying assignment if and only if there is a tree $T \in \mathbf{T}(N)$ and internal assignment $\mathbf{a} \in \{0,1\}^r$ s.t. $L(S|T, \mathbf{a}, p) > 0$.

Proof. We begin with some auxiliary observations.

Observation 2 *For any tree T in the network and internal assignment $\mathbf{a}$, $L(S|T, \mathbf{a}, p) > 0$ if and only if the root R and all non variable nodes (nodes which are not assigned to variables) must have internal assignment 1.*

We first show that if F has a satisfying assignment, then there is a tree $T \in \mathbf{T}(N)$ and internal assignment $a \in \{0,1\}^r$ s.t. $L(S|T,p) > 0$. For every variable we assign its value from the satisfying assignment. For every clause c_i, we choose the path from the 1-leaf representing c_i to the literal satisfying it. Note that if that literal is negated, then the edge emanating from it has substitution probability 1. Now, every variable is connected to the root by either the $0-0$ path in the diamond if the variable is assigned 1 or the $1-0$ path otherwise. Finally, if we set the probability of every reticulation edge to be positive, we get the desired result.
To prove the other direction, assume there exists a tree $T \in \mathbf{T}(N)$ and internal assignment $a \in \{0,1\}^r$ s.t. $L(S|T,p) > 0$. Then by Observation 2, all non-variable internal nodes are forced to the value 1.

Observation 3 *For every variable v in T, all emanating edges from v in T have the same probability which is either 0 if v has assignment 1 or 1 otherwise.*

Since every leaf must be connected to the root we get that every clause is satisfied.

We comment that in the final tree, few internal nodes may remain with out-degree 0 and hence disappear in the resulting tree. In addition, other internal nodes can remain with out-degree 1 and contracted with the resulting obvious probability on the new edge.

Theorem 4. *The tiny best tree ancestral sequences of phylogenetic networks is NP-hard.*

Proof. It is easy to see that the problem is in NP, since given a tree T it is easy to check if $T \in \mathbf{T}(N)$ and subsequently calculate its likelihood. Additionally, the reduction of Claim 1 is performed in time polynomial in the size of F.

Corollary 1. *The tiny best tree average likelihood of phylogenetic networks is NP-hard.*

Proof. Using the same reduction as in Claim 1, and Observation 2, we see that for a tree to obtain likelihood greater than zero, all internal nodes are forced to 1. Therefore, the only freedom is in the variable nodes and by Observation 3 each such assignment defines a truth assignment to the variables in F.

By using similar arguments to Corollary 1 we also obtain the two following hardness results:

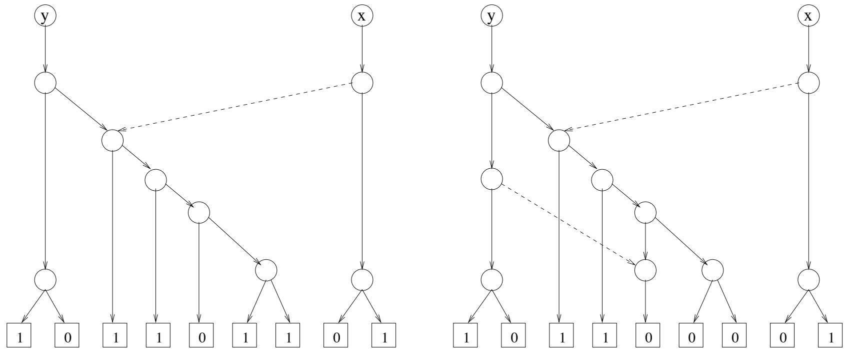
Corollary 2. *The tiny all trees ancestral likelihood and the tiny all trees average likelihood of phylogenetic networks are NP-hard.*

All the above results set the complexity of the tiny versions of the network likelihood problems. The next reduction deals with the "small" problem where the network is given but we seek to find the tree **and** the edge probabilities that maximize the likelihood over the whole trees of the network. For the reduction we use the same construct we used in [2]. We also restrict the edge probabilities to some interval $[0, p]$ for $p < 1$.

We prove the hardness of the small best tree ancestral ML problem by a reduction from the Maximum 2-Satisfiability (max-2-sat) problem [1], which is formally defined as follows.

Problem 2. Maximum 2-Satisfiability (max-2-sat)
Input: Set U of variables, collection C of clauses over U such that each clause $c \in C$ has $|c| = 2$, and a positive integer $K \leq |C|$.
Question: Is there a truth assignment for U that simultaneously satisfies at least K of the clauses in C?

We start with a lemma which will be used in our main proof. Let a "True-True" denote a clause that has no negated literals, "True-False" denote a clause that has exactly one negated literal, and "False-False" denote a clause in which both literals are negated. For the "True-True" clause we generate the subnetwork shown in Fig. 4 on the right, For the "True-False" clause we generate the subnetwork shown in Fig. 4 on the left and for the "False-False" we generate the same network as for "True-True" but flip the value of the leaves.



$X \vee Y$ $\bar{X} \vee Y$

Fig. 4. Part of the reduction from max-2-sat to small best tree ancestral likelihood.

Lemma 1. *[2]*
(1) The minimal number of substitutions is 3 *for a "True-True" network is obtained by labelling* $x = 1$*,* $y = 1$*, or both. Otherwise,it is* 4*.*

(2) The minimal number of substitutions is 3 *for a "True-False" network is obtained by labelling* $x = 0$, $y = 1$, *or both. Otherwise, it is* 4.
(3) The minimal number of substitutions is 3 *for a "False-False" network is obtained by labelling* $x = 0$, $y = 0$, *or both. Otherwise, it is* 4.

Given a formula F as input to max-2-sat, we create a node for every variable and connect every variables to two ancestral nodes and these two to the root. In addition, for every clause c_i we generate the appropriate subnetwork as in Fig. 4 and connect it to the variables involved. Fig. 5 shows a complete network generated for a specific clause.

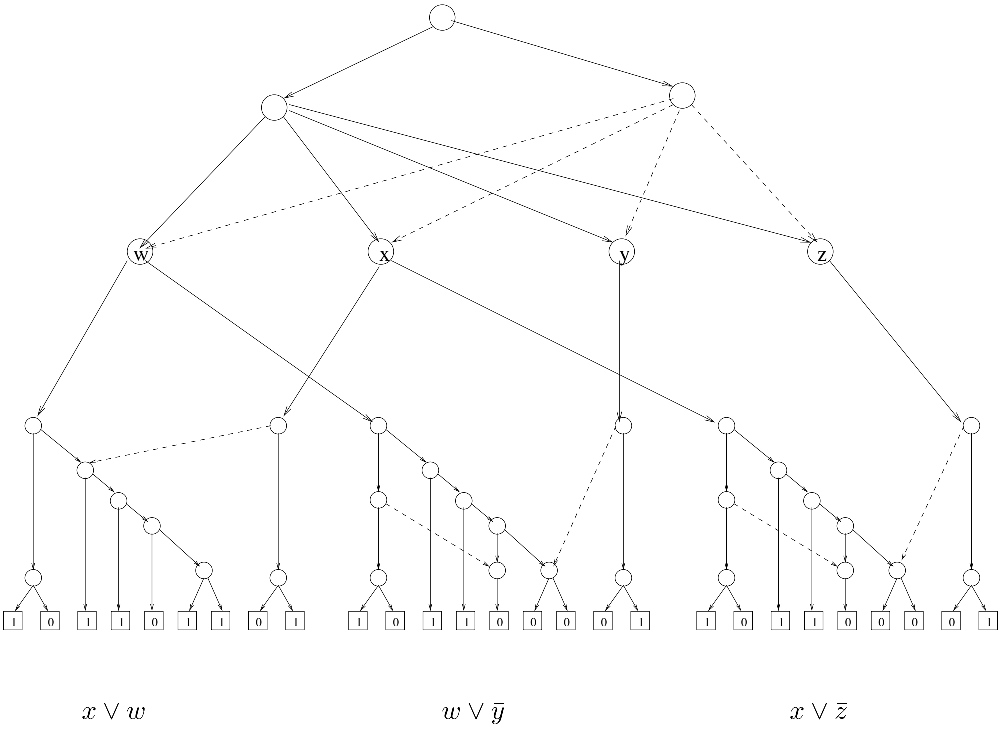


Fig. 5. The complete network generated for the clauses $(x \vee w), (w \vee \bar{y}), (x \vee \bar{z})$.

Since this version of problem deals with the ancestral version where every internal node is assigned a value of either 0 or 1, we get the following observation:

Observation 5 *At any optimal tree, the probability of every edge will be set to either zero or* p.

Observation 6 *Unless all variables have the same value (0 or 1) there is exactly one substitution on the subnetwork above the variable nodes.*

Therefore, WLOG, we will assume the optimal assignment has both values.

Claim. Given a set of clauses C over variables X, as input to max-2-Sat, X has an assignment which k clauses are satisfied, if and only if the network constructed has a tree with likelihood $p^{4|C|-k+1}$.

Proof. $\Rightarrow$ By Lemma 1 there are $4|C|-k$ substitutions at the networks below the variable nodes and by Observation 6 exactly one more above the variable nodes. Now, by Observation 5 the edges on which a substitution occurs get probability p and the others 0. This yields the desired result.
The proof of the other direction proceeds similarly.

Theorem 7. *The small best ancestral tree is NP-hard.*

References

1. M. R. Garey and D. S. Johnson. *Computer and Intractability.* Bell Telephone Laboratories, incorporated, 1979.
2. G. Jin, L. Nakhleh, S. Snir, and T. Tuller. Efficient parsimony-based methods for phylogenetic network reconstruction. *ECCB*, 2006.