Phylogenies without Branch Bounds: Contracting the Short, Pruning the Deep

Extended Abstract

Constantinos Daskalakis[1], Elchanan Mossel[2,⋆], and Sebastien Roch[1]

[1] Microsoft Research
[2] UC Berkeley and Weizman Institute

Abstract. We introduce a new phylogenetic reconstruction algorithm which, unlike most previous rigorous inference techniques, does not rely on assumptions regarding the branch lengths or the depth of the tree. The algorithm returns a forest which is guaranteed to contain all edges that are: 1) sufficiently long and 2) sufficiently close to the leaves. How much of the true tree is recovered depends on the sequence length provided. The algorithm is distance-based and runs in polynomial time.

1 Introduction

In Evolutionary Biology, the speciation history of a family of related organisms is generally represented graphically by a *phylogeny*, that is, a tree where the leaves are the observed (extant) species and the branchings indicate speciation events. Traditional approaches for reconstructing phylogenies from homologous molecular sequences extracted from the observed species [1,2] are typically computationally intractable [3,4,5,6,7], statistically inconsistent [8], or they require impractical sequence lengths [9,10,11,12]. Nevertheless, over the past decade, much progress has been made in the design of efficient, fast-converging reconstruction techniques, starting with the seminal work of Erdös et al. [13]. The algorithm in [13], often dubbed the Short Quartet Method (SQM), is based on well-known distance-matrix techniques, that is, it relies on estimates of the evolutionary distance between each pair of species (roughly the time elapsed since their most recent common ancestor). However, unlike other popular distance methods such as Neighbor-Joining [14], the key behind SQM's performance is that it discards long evolutionary distances, whose estimates from sequence comparisons are known to be statistically unreliable. The algorithm works by first building subtrees of small diameter and, in a second stage, glueing the pieces back together.

The Short Quartet Method is in fact guaranteed to return the correct topology from polynomial-length sequences in polynomial time with high probability. But this appealing theoretical performance comes at a price. The results of [13] rely

⋆ E.M. is supported by an Alfred Sloan fellowship in Mathematics and by NSF grants DMS-0528488, and DMS-0548249 (CAREER) and by ONR grant N0014-07-1-05-06.

S. Batzoglou (Ed.): RECOMB 2009, LNCS 5541, pp. 525–539, 2009.

critically on biological assumptions which, although reasonable, are often not met in practice (see Section 1.3 for a formal statement):

a) *[Dense Sampling of Species]* The observed species are "closely related." In particular, there are no exceptionally long branches in the phylogeny.
b) *[Absence of Polytomies]* The phylogeny is bifurcating. In fact, Erdös et al. assume that speciation events are sufficiently far apart to be easily distinguished.

The point of a) is that it implies a natural bound on the depth of the tree which in turn ensures that enough information about the deep parts of the tree diffuses to the leaves. As for Assumption b), it guarantees that a clear signal can be extracted from each branch of the phylogeny. It is obvious—at least intuitively—that assumptions such as a) and b) are necessary to secure the type of results Erdös et al. obtain: *the guaranteed reconstruction of the full phylogeny.* Hence, to improve over SQM and obtain strong guarantees under more general conditions, one has to relax this last requirement.

In this paper, we design an algorithm which provides strong reconstruction guarantees without Assumptions a) and b). We show that our algorithm is guaranteed to recover a *forest* containing all edges that are "sufficiently long" and "sufficiently close" to the leaves. In fact, we allow a trade-off between the resolution of short branches and the depth of the reconstructed forest, a feature of potential practical interest. Also, we guarantee that our reconstructed forest has the desirable property of being *disjoint* (although the presence of short edges leads us to allow deep intersections of very short branches between the subtrees). Moreover, our algorithm does not require the knowledge of a priori bounds on branch lengths or tree depth. Finally if Assumptions a) and b) are satisfied, we recover the whole phylogeny and provide an alternative to the algorithm of Erdös et al.

Precise statements are given in Section 1.2. For a full comparison to related work see also Section 1.3. For a lack of space, the proofs of our results are not included in this extended abstract. But they can be found in the full version of the paper [15].

1.1 What Can We Hope to Reconstruct?

Well-known identifiability results [16] guarantee that phylogenies—or at least their idealized stochastic models—can be fully reconstructed given enough data at the leaves. However, molecular data gathered from current species are in essence limited, which begs the question: *How much of tree can we really hope to reconstruct?* We pointed out above two important sources of difficulties: short branches produce a low signal that may be hard to detect; similarly, untangling the deep parts of the tree presents challenges that are well documented (see, e.g., [17,18]). Note that these issues are fundamentally "information-theoretic" and that they affect all reconstruction methods.

To avoid these difficulties, most *rigorous* methods impose restrictions on the length of the branches and/or the depth of the tree, which may be unsatisfactory from a practical perspective. On the other hand, commonly used methods

in *practice*, such as likelihood and bayesian methods, typically produce several candidate trees as well as confidence estimates. But theoretical guarantees on the quality of such outputs are hard to obtain.

Here, we seek to give strong reconstruction guarantees without any assumption on the true phylogeny. Our goal is to recover, for any given amount of data, as much of the tree as can rigorously be reconstructed with high confidence. Since the full phylogeny may not always be recoverable, we are led to a more flexible solution concept: we output a *contracted subforest* of the true phylogeny. That is, we output a forest containing all branches that are "sufficiently long" and "sufficiently recent"; note that "sufficiently" here is determined (information-theoretically) by the size of the data (usually in terms of sequence length). In the remainder of this section we formalize this notion.

The input. Formally, a phylogeny is a *weighted, multifurcating tree* on a set of leaves L, which we identify with the labels $[n] = \{1, \ldots, n\}$. We denote a phylogeny by $T = (V, E; L, \lambda)$. Here V and E are respectively the vertex and edge set of the tree, and $\lambda : E \to (0, +\infty)$ assigns a weight to each edge (the branch length). We assume that all internal vertices $V - L$ have degree at least 3.

A phylogeny is naturally equipped with a so-called additive metric on the leaves $d : L \times L \to (0, +\infty)$ defined as follows

$$\forall u, v \in L,\ d(u, v) = \sum_{e \in \mathrm{P}_T(u,v)} \lambda_e,$$

where $\mathrm{P}_T(u, v)$ is the set of edges on the path between u and v in T. Often $d(u, v)$ is referred to as the "evolutionary distance" between species u and v. Since under the assumptions above there is a one-to-one correspondence between d and λ, we write either $T = (V, E; L, d)$ or $T = (V, E; L, \lambda)$. We also sometimes use the natural extension of d to the internal vertices of T. We denote by $\mathcal{T}$ the set of all phylogenies on any number of leaves.

It is well-known that given an additive metric d one can reconstruct the corresponding phylogeny T. However, in practice, one can only derive an *estimate* $\hat{d}$ of d, the accuracy of which depends on the sequence length. (This estimate is known in the literature as the "distance matrix".) Our goal in this paper is to reconstruct a phylogeny—or as much of it as possible—from this "distorted" version of its additive metric. A well-known property of $\hat{d}$ is that estimates of long distances are unreliable. The following definition formalizes this phenomenon. See Figure 1 for an illustration.

Definition 1 (Distorted Metric [19,20]). *Let $T = (V, E; L, d)$ be a phylogeny and let $\tau, M > 0$. We say that $\hat{d} : L \times L \to (0, +\infty]$ is a* (τ, M)-distorted metric *for T or a* (τ, M)-distortion *of d if:*

1. [Symmetry] *For all $u, v \in L$, $\hat{d}$ is symmetric, that is,*

$$\hat{d}(u, v) = \hat{d}(v, u);$$

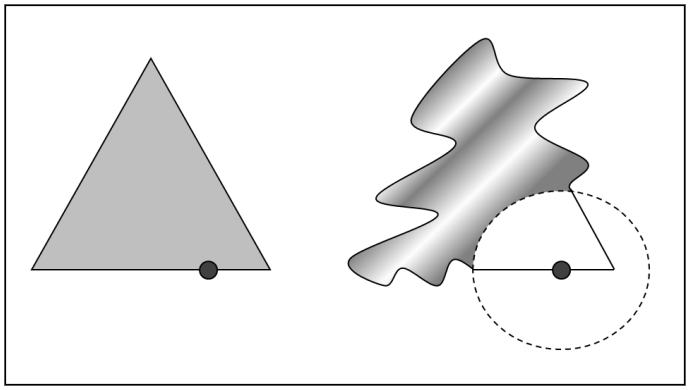

Fig. 1. The effect of distance distortion from the perspective of a leaf. On the left hand side is the true phylogeny. On the right hand side, only distances within a certain radius represent accurately the metric underlying the phylogeny.

2. [Distortion] *$\hat{d}$ is accurate on "short" distances, that is, for all $u, v \in L$, if either $d(u,v) < M + \tau$ or $\hat{d}(u,v) < M + \tau$ then*

$$\left| d(u,v) - \hat{d}(u,v) \right| < \tau.$$

In phylogenetic reconstruction, a distorted metric is naturally derived from samples of a Markov model on a tree—a common model of DNA sequence evolution used in Biology. (See [15] for details.) In the remainder of this paper, we assume that we are given a (τ, M)-distortion $\hat{d}$ of an additive metric d and we seek to recover the underlying phylogeny T.

Contraction and pruning. Given only a (τ, M)-distorted metric, it is clear that the best we can hope for in general is to reconstruct a forest containing those edges of T that are "sufficiently close" to the leaves. Indeed, note that two phylogenies that are identical up to depth M from the leaves, but are otherwise different, can give rise to the same distorted metric. Moreover, since we do not assume that edges are longer than the accuracy τ, some edges may be too short to be reconstructed and, as we mentioned before, we allow ourselves to instead contract them. Hence, we are led to consider subforests of the true phylogeny where deep edges are *pruned* and short edges are *contracted*.

To formalize this idea we need a few definitions. Let us first describe what we mean by a *subforest* of a phylogeny $T = (V, E; L, d)$. Given a set of vertices $V' \subseteq V$, the *subtree of T restricted to V'* is the tree obtained 1) by keeping only nodes and edges on paths between vertices in V' and then 2) by contracting all paths composed of vertices of degree 2, except the nodes in V'. See Figure 2 for an example. We denote this tree by $T|_{V'}$. We typically take $V' \subseteq L$. A *subforest* of T is defined to be a collection of restricted subtrees of T.

We also need a notion of depth. Given an edge $e \in E$, the *chord depth* of e is the length of the shortest path between two leaves on which e lies. That is,

$$\Delta_{\mathrm{c}}(e) = \min \left\{ d(u,v) \ : \ u, v \in L, e \in \mathrm{P}_T(u,v) \right\}.$$

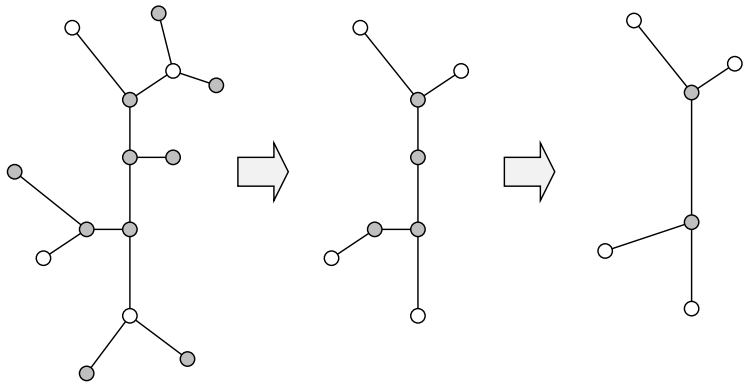

Fig. 2. Restricting the top tree to its white nodes

We define the *chord depth of a tree* T to be the maximum chord depth in T

$$\Delta_{\mathrm{c}}(T) = \max\{\Delta_{\mathrm{c}}(e) \ : \ e \in E\}.$$

Definition 2 (Contracted Subforest). *Let $T = (V, E; L, d)$ be a phylogeny. Fix $M > 0$. Let $\{L_1, \ldots, L_q\}$ be the natural partition of the leaf set L obtained by removing all edges $e \in E$ such that $\Delta_{\mathrm{c}}(e) \geq M$. We define the* M-pruned subforest *of T to be the forest $F_M(T) = (V_M, E_M)$ consisting of the trees $\{T|_{L_1}, \ldots, T|_{L_q}\}$. The metric d is extended as follows for all $u, v \in L$,*

$$d_M(u,v) = \begin{cases} d(u,v), \text{ if } u,v \text{ are in the same subtree of } F_M(T), \\ +\infty, \quad \text{o.w.} \end{cases}$$

Similarly, we define an extension λ_M of λ.

Now, given also $\tau > 0$, the τ-contracted M-pruned subforest *of T is the forest $F_{\tau,M}(T) = (V_{\tau,M}, E_{\tau,M})$ obtained from $F_M(T)$ by contracting edges $e \in E_M$ of weight $\lambda_M(e) < \tau$.*

Path-disjointness. We require that the trees of our reconstructed forest are "not intersecting". This is a natural condition to impose in order to obtain a meaningful reconstruction: we want to avoid as much as possible that the same branches appear in many subtrees. In fact, we can only guarantee approximate *path-disjointness* as defined below.

We first need a notion of depth for vertices. For a phylogeny $T = (V, E; L, d)$ and a vertex $x \in V$, the *vertex depth* of x is the length of the shortest path between x and the set of leaves. That is,

$$\Delta_{\mathrm{v}}(x) = \min\{d(u,x) \ : \ u \in L\}.$$

Given two leaves u, v of T, we denote by $\widetilde{\mathrm{P}}_T(u,v)$ the set of vertices on the path between u and v in T.

We say that two trees are (τ, M)-path disjoint if they are "almost disjoint" in the sense that they only share edges (if any) that are "deep" (endpoints have vertex depth at least $M/2$) and "short" (length at most τ). More formally:

Definition 3 (Approximate Path-Disjointness). *Let $T = (V, E; L, d)$ be a phylogeny. Two subtrees T_1, T_2 of T restricted respectively to $L_1, L_2 \subseteq L$ are* (τ, M)-path-disjoint *if $L_1 \cap L_2 = \emptyset$ and for all pairs of leaves $u_1, v_1 \in L_1$ and $u_2, v_2 \in L_2$ such that*

$$\widetilde{\mathrm{P}}_T(u_1, v_1) \cap \widetilde{\mathrm{P}}_T(u_2, v_2) \neq \emptyset,$$

we have:

$$\min\{\Delta_{\mathrm{v}}(x) \ : \ x \in \widetilde{\mathrm{P}}_T(u_1, v_1) \cap \widetilde{\mathrm{P}}_T(u_2, v_2)\} \geq \frac{1}{2}M,$$

and, if further $\mathrm{P}_T(u_1, v_1) \cap \mathrm{P}_T(u_2, v_2) \neq \emptyset$,

$$\max\{\lambda_e \ : \ e \in \mathrm{P}_T(u_1, v_1) \cap \mathrm{P}_T(u_2, v_2)\} \leq \tau.$$

More generally, a collection of restricted subtrees $T_1, \ldots, T_q$ of T are (τ, M)-path-disjoint *if they are pairwise (τ, M)-path-disjoint. In the case $\tau = 0$, we simply say that the subtrees are* path-disjoint.

1.2 Main Result and Corollaries

Main result. Our main result is an algorithm which, given a (τ, M)-distorted metric, reconstructs a contracted subforest (of the true phylogeny) whose trees are approximately path-disjoint. Typically, M is much larger than τ. In that case, we reconstruct a subforest of T with chord depth $\approx \frac{1}{2}M$ which includes all edges of length at least 4τ. The reconstructed subtrees may "overlap" on edges of length at most 2τ at vertex depth $\approx \frac{1}{4}M$. In an upcoming journal version of the paper, we show that these parameters are essentially optimal. The algorithm runs in polynomial time. An implementation of the algorithm with low running time will be given in the journal version.

More precisely, we show:

Theorem 1 (Main Result). *Let τ and M be monotone functions of n with $M > 3\tau$. Let $m > 3\tau$ be such that*

$$m < \frac{1}{2}[M - 3\tau],$$

for all n. Then, there is an algorithm $\mathcal{A}$ such that, for all phylogenies $T = (V, E; L, d)$ in $\mathcal{T}$ and all (τ, M)-distortions $\hat{d}$ of d, $\mathcal{A}$ applied to $\hat{d}$ satisfies the following:

1. [Approximate Path Disjointness] *$\mathcal{A}$ returns a $(2\tau, m-3\tau)$-path-disjoint subforest $\widehat{F}$ of T;*
2. [Depth Guarantee] *The forest $\widehat{F}$ is a refinement of $F_{4\tau, m-\tau}(T)$;*
3. [Polynomial Time] *$\mathcal{A}$ runs in time polynomial in $n, \log M, \log \tau$.*

We give below a few important special cases of Theorem 1. The proof of Theorem 1 can be found in the full version of the paper [15].

Tree case. When the amount of data is sufficient to produce a distorted metric with $M = \Omega(\Delta_c(T))$, we get a single component, that is, the full tree (up to those edges that are contracted).

Corollary 1 (Tree Case). *Let $\tau > 0$ and $M > 2\Delta_c(T) + 5\tau$. Then, choosing $m > \Delta_c(T) + \tau$ guarantees that the reconstructed forest is composed of only a tree.*

In the case of "dense" phylogenies, $M = \Omega(\log n)$ is sufficient to reconstruct the full tree.

Definition 4 (Dense Phylogenies (see e.g. [13])). *We say that a collection of phylogenies $\mathcal{T}'$ is* dense *if there is a $0 < g < +\infty$ (independent of n) such that for all $T = (V, E; L, \lambda) \in \mathcal{T}'$ we have*

$$\forall e \in E,\ \lambda_e \leq g. \tag{1}$$

We denote by $\mathcal{T}_g$ the set of phylogenies satisfying (1).

Corollary 2 (Dense Case). *In the case of dense phylogenies, $M = \Omega(\log n)$ suffices to guarantee the reconstruction of the full tree, up to contracted edges.*

Absolute variant. All rigorous algorithms prior to our work (see Section 1.3) require knowledge of either the tree depth or bounds on the edge lengths to give strong reconstruction guarantees. This is not satisfactory from a practical point of view. Here given only the sequence length we provide explicit guarantees. The following result assumes that the distorted metric is derived from a Markov model on a tree. (See [15] for details.)

Corollary 3 (Absolute Variant). *Given a number of samples $k = \Omega(\log n)$ from a Markov model on a tree and a chosen level of contraction $\varepsilon > 0$ (small), one can choose τ, M, m so that $\mathcal{A}$ is guaranteed to return a (contracted) subforest of T containing $F_{\varepsilon,M'}(T)$ with probability $1 - o(1)$, where $M' = \Omega_\varepsilon(\log k - \log\log n)$.*

Complete resolution. Finally we remark that, if we further assume that all branch lengths are bounded *from below* by a constant, then by choosing τ accordingly a non-contracted forest is returned. In particular, we also recover the results of [13].

1.3 Related Work

Under a Markov model of evolution, the Short Quartet Method (SQM) of Erdös et al. [13] is guaranteed to recover the full phylogeny as long as the number of samples k satisfies

$$k > cf^{-2}e^{c'g\Delta_c(T)}\log n,$$

for constants $c, c' > 0$, where f and g are respectively lower and upper bounds on the branch lengths possibly depending on n. For instance, if f and g are constants

the sequence length needed for complete reconstruction depends polynomially in the number of species.

Mossel [19] developed a framework that allows the reconstruction of a well-behaved *forest* when sequences are too short to guarantee a complete reconstruction. More precisely, edges which are too deep (in the sense of appearing only on paths between species whose distances are not accurately known) are *pruned* from the final reconstruction. At a high level, Mossel's Distorted Metric Method (DMM) (implicit in [19]), works in a fashion similar to SQM—except for a preprocessing phase that clusters together sufficiently related species. However, for DMM to work, lower bounds on the branch lengths are required and, moreover, these must be known by the algorithm. Following up on [19], Daskalakis et al. [21] gave a variant of DMM that runs without knowledge of a priori bounds on the branch lengths or the tree depth—making their variant somewhat more practical. However, like DMM, the algorithm in [21] does not deal properly with short edges: any part of the tree containing a short edge cannot be reconstructed by the algorithm (even though there may be adjacent edges that are in fact reconstructible). Therefore, in the presence of short edges no guarantee can be given about the depth of the reconstructed forest.

Recently Gronau et al. [22] eliminated the need for a lower bound on the branch length by *contracting* edges whose length is below a user-defined threshold. Their solution uses a Directional Oracle (DO) which closes in on the location of a leaf to be added and, in the process, contracts regions that do not provide a reliable directional signal. Although the DO algorithm does not use an explicit bound on the depth of the tree, their *reconstruction guarantee* requires such a bound, similarly to [13]. In particular, Gronau et al. leave open the question of giving a forest-building version of their algorithm. Moreover, the sequence length in [22] depends exponentially on what the authors call the ε-diameter of the tree—essentially, the maximum diameter of the contracted regions. It is natural to conjecture that an optimal result should not depend on this parameter.

For further related work on efficient phylogeny reconstruction, see also [23,24,25,26,20,27,28].

1.4 Discussion of the Results

The following table summarizes the current status as discussed in the previous sections.

As the table emphasizes, our overarching goal is to design an algorithm with good reconstruction guarantees in the presence of both short and deep edges, whose execution does not rely on a priori bounds on branch lengths. Unfortunately, given the combinatorial complexity of Mossel's forest-building algorithm, it is not straightforward to provide the extra flexibility of edge contraction in this framework. The novelty in our work is twofold:

– *Solution Concept:* A basic complication is that, in some sense, contraction and pruning interfere with each other. Indeed, the presence of unresolved

	Execution	*Guarantees*	
	No branch bound needed	Short edges OK	Deep edges OK
[13]			
[19]			✓
[21]	✓		✓
[22]	✓	✓	
Our method	✓	✓	✓

branches at the boundary of partially reconstructed subtrees creates the possibility of deep "undetectable" intersections. This pitfall seems to be unavoidable. One of our main contributions is to introduce the notion of approximate disjointness, which allows short and deep intersections between subtrees of the reconstructed forest. This suitable solution concept leads to a quite simple algorithm with reasonable guarantees. Moreover, the flexibility in our definition allows us to recover all previously known results as special cases.

– *Algorithmic Technique:* A natural approach to forest building used in [19,21] proceeds along the following three steps:

  1. first, leaves are grouped into clusters for which all pairwise distances are accurately known (the *small* clusters);
  2. by definition, the local topologies on the small clusters can be trivially reconstructed [29];
  3. finally, the local topologies that intersect in the true tree are "glued" together to get a forest (the resulting forest partitions the leaves into *large* clusters).

  This last step involves non-trivial combinatorial considerations. We have found that further allowing contracted edges makes this process somewhat unmanageable. Instead we use a different approach relying on simple metric arguments. In particular, we *directly* partition the leaves into large clusters, whose underlying subtrees are approximately disjoint, and provide a new straightforward method to reconstruct these subtrees.

In addition, we obtain as special cases the results discussed in Section 1.3. In particular, if there are no short edges, we recover the results of [19] and [21], where a path-disjoint forest is returned (by taking τ equal to half the lower bound on the branch lengths in Theorem 1). If furthermore there is an upper bound on the branch lengths, we recover the results of [13] (Corollary 2). Finally, if we keep the upper bound on the edge lengths, but drop the lower bound, we recover the results of [22] (Corollary 1). In fact, we eliminate the dependence on

the ε-diameter. [1] Further, unlike [22], we allow an arbitrary number of states, an extension—it should be noted—that follows easily from [23] and [19].

2 Algorithm

The outline of the algorithm follows. There are three main phases, which are explained in detail after the outline. The input to the algorithm is a (τ, M)-distorted metric $\hat{d}$ on n leaves. In particular, we assume that the values τ and M are known to the algorithm (but see also Corollary 3). Let m be as in Theorem 1. We denote the true tree by $T = (V, E; L, d)$. The details of the subroutines Mini Contractor and Extender are detailed in Figures 4 and 6 (see also their high level description below). For lack of space, the proof of correctness of the algorithm can be found in [15].

- **Pre-Processing: Leaf Clustering.** Build the distorted clustering graph $\widehat{H}_m = (\widehat{V}_m, \widehat{E}_m)$ where $\widehat{V}_m = [n]$ and $(u, v) \in \widehat{E}_m \iff \hat{d}(u, v) < m$; compute the connected components $\{\hat{h}_m^{(i)} = (\hat{v}_m^{(i)}, \hat{e}_m^{(i)})\}_{i=1}^{q}$ of $\widehat{H}_m$;
- **Main Loop.** For all components $i = 1, \ldots, q$:
  - For all pairs of leaves $u, v \in \hat{v}_m^{(i)}$ such that $(u, v) \in \widehat{E}_m$:
    - ∗ **Mini Reconstruction.** Compute

$$\{\psi_j(u, v)\}_{j=1}^{r(u,v)} := \text{Mini Contractor}(\hat{h}_m^{(i)}; u, v);$$

    - ∗ **Bipartition Extension.** Compute

$$\{\bar{\psi}_j(u, v)\}_{j=1}^{r(u,v)} := \text{Extender}(\hat{h}_m^{(i)}, \{\psi_j(u, v)\}_{j=1}^{r(u,v)}; u, v);$$

  - Deduce the tree $\widehat{T}^{(i)}$ from $\{\bar{\psi}_j(u, v)\}_{j=1}^{r(u,v)}$;
- **Output.** Return the resulting forest $\widehat{F}$.

Pre-processing: Leaf clustering. As mentioned before, given a (τ, M)-distortion we cannot hope to reconstruct edges that are too deep inside the tree. This results in the reconstruction of a *forest*. Therefore, the first phase of the algorithm is to determine the "support" of this forest. We proceed as follows. Consider the following graph on L.

Definition 5 (Clustering Graph). *Let $\tau \leq M' \leq M - \tau$. The* distorted clustering graph with parameter M', *denoted $\widehat{H}_{M'} = (\widehat{V}_{M'}, \widehat{E}_{M'})$, is the following*

[1] After the results of the current paper were posted on the arXiv, we were informed by S. Moran that, in parallel to our work, the authors of [22] have improved on their previous results: the dependence on the ε-diameter has been removed. A preprint of this work is currently available on the authors' website. Note however that this new, independent work does not deal with deep edges and still makes assumptions similar to [13] restricting the depth of the generating tree.

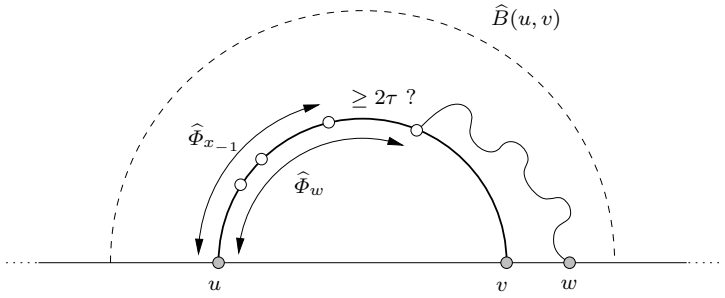

Fig. 3. Illustration of routine MINI CONTRACTOR. See Figure 4 for notation.

graph: the vertices $\widehat{V}_{M'}$ *are the leaves* L *of* T*; two leaves* $u, v \in L$ *are connected by an edge* $e \in \widehat{E}_{M'}$ *if*

$$\hat{d}(u,v) < M'. \tag{2}$$

Note that this is an undirected graph because $\hat{d}$ *is symmetric. Similarly, we define the* clustering graph with parameter M', $H_{M'} = (V_{M'}, E_{M'})$, *where we use* d *instead* $\hat{d}$ *in (2).*

The first phase of the algorithm consists in building the graph $\widehat{H}_m$ from $\hat{d}$. We then compute the connected components of $\widehat{H}_m$ which we denote $\{\hat{h}_m^{(i)}\}_{i=1}^q$. In the next two phases, we build a tree on each of these components.

Building the components I: Mini-reconstruction problem. Fix a component $\hat{h}_m^{(i)}$ of $\widehat{H}_m$. In this and the next phase, we seek to reconstruct a contracted tree on $\hat{h}_m^{(i)}$. Denote by $T^{(i)}$ the true tree T restricted to the leaves in $\hat{h}_m^{(i)}$. First, we find all edges of $T^{(i)}$ that are "sufficiently long" and lie on "sufficiently short" paths. More precisely, we consider all pairs of leaves u, v connected by an edge in $\hat{h}_m^{(i)}$, that is, leaves within distorted distance m. For each such pair, say u, v, the *mini reconstruction problem* consists in finding all edges in $\mathrm{P}_{T^{(i)}}(u,v)$ that have length longer than $\lambda_e \geq 4\tau$. To do this using the distortion $\hat{d}$, we first consider a *ball* $\widehat{B}(u,v)$ of all nodes within distorted distance M of u and v, that is,

$$\widehat{B}(u,v) = \left\{ w \in \hat{h}_m^{(i)} \ : \ \hat{d}(u,w) \vee \hat{d}(v,w) < M \right\},$$

where $a \vee b$ is the maximum of a and b.— The point of using this ball is that we can then guarantee that each edge in $\mathrm{P}_{T^{(i)}}(u,v)$ is "witnessed" by a quartet (i.e., a 4-tuple of leaves) in $\widehat{B}(u,v)$ in the following sense: let (x_1, x_2) be an edge in $\mathrm{P}_{T^{(i)}}(u,v)$ and let (x_j, y_j), $j = 1, 2$, be an edge adjacent to x_j that is *not* in $\mathrm{P}_{T^{(i)}}(u,v)$; for $j = 1, 2$ let $L^{(i)}_{x_j \to y_j}$ be the leaves reachable from y_j using paths not including x_j; then we will show that $L^{(i)}_{x_j \to y_j} \cap \widehat{B}(u,v) \neq \emptyset$ for $j = 1, 2$. In other words, there is enough information in $\widehat{B}(u,v)$ to reconstruct all edges in

Algorithm MINI CONTRACTOR
Input: Component $\hat{h}_m^{(i)}$; Leaves u, v;
Output: Bipartitions $\{\psi_j(u,v)\}_{j=1}^{r(u,v)}$;

- **Ball.** Let
$$\widehat{B}(u,v) := \left\{ w \in \hat{h}_m^{(i)} \ : \ \hat{d}(u,w) \vee \hat{d}(v,w) < M \right\};$$
- **Intersection Points.** For all $w \in \widehat{B}(u,v)$, estimate the point of intersection between u, v, w (distance from u), that is,
$$\widehat{\Phi}_w := \frac{1}{2}\left(\hat{d}(u,v) + \hat{d}(u,w) - \hat{d}(v,w)\right);$$
- **Long Edges.** Set $S := \widehat{B}(u,v) - \{u\}$, $x_{-1} = u$, $j := 0$;
  - Until $S = \emptyset$:
    - Let $x_0 = \arg\min\{\widehat{\Phi}_w \ : \ w \in S\}$ (break ties arbitrarily);
    - If $\widehat{\Phi}_{x_0} - \widehat{\Phi}_{x_{-1}} \geq 2\tau$, create a new edge by setting $\psi_{j+1}(u,v) := \{\widehat{B}(u,v) - S, S\}$ and let $C_{j+1} := \{x_0\}$, $j := j+1$;
    - Else, set $C_j := C_j \cup \{x_0\}$;
    - Set $S := S - \{x_0\}$, $x_{-1} := x_0$;
- **Output.** Return the bipartitions $\{\psi_j(u,v)\}_{j=1}^{r(u,v)}$ (where $r(u,v)$ is the number of bipartitions generated in the previous step).

Fig. 4. Algorithm MINI CONTRACTOR. See Figure 3 for illustration.

$\mathrm{P}_{T^{(i)}}(u,v)$—at least those that are "sufficiently long." This phase is detailed in Figure 4. An illustration is given in Figure 3.

Building the components II: Extending the bipartitions. The previous step reconstructs "sufficiently long" edges on balls of the form $\widehat{B}(u,v)$. By *reconstructing an edge on* $\widehat{B}(u,v)$, we mean *finding the bipartition of* $\widehat{B}(u,v)$ *to which the edge corresponds.* More precisely:

Definition 6 (Bipartitions). *Let $T = (V,E)$ be a multifurcating tree with no vertex of degree 2. Each edge e in T induces a* bipartition *of the leaves L of T as follows: if one removes the edge e from T, then one is left with two connected components; take the partition of the leaves corresponding to those components. Denote by $b_T(e)$ the bipartition of e on T. It is easy to see that given the bipartitions $\{b_T(e)\}_{e\in E}$ one can reconstruct the tree T efficiently [29,30,31]. (Proceed by sequentially "splitting" clusters.)*

The goal of the second phase in the main loop of our reconstruction algorithm is to *extend* the bipartitions previously built from $\widehat{B}(u,v)$ to the full component $\hat{h}_m^{(i)}$. To perform this task, we use the following observation: suppose we want to deduce the bipartition corresponding to edge e; if we take the ball $\widehat{B}(u,v)$ to be much larger than m (yet small enough that it remains within our radius of precision M), we can make sure that a path *from* a leaf in $\hat{h}_m^{(i)}$ that is outside $\widehat{B}(u,v)$ *to* a leaf on the other side of the bipartition is "long." Therefore, we can

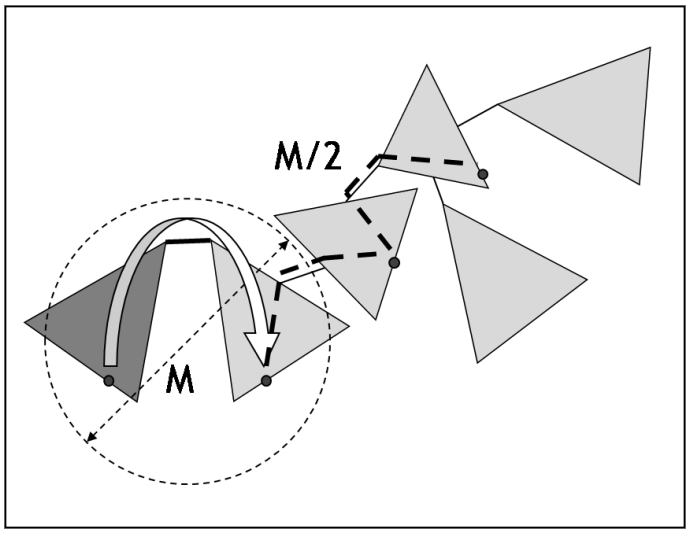

Fig. 5. Illustration of routine EXTENDER. See also Figure 6.

Algorithm EXTENDER
Input: Component $\hat{h}_m^{(i)}$; Bipartitions $\{\psi_j(u,v)\}_{j=1}^{r(u,v)}$; Leaves u, v;
Output: Bipartitions $\{\bar{\psi}_j(u,v)\}_{j=1}^{r(u,v)}$;

- For $j = 1, \ldots, r(u,v)$ (unless $r(u,v) = 0$):
  - **Initialization.** Denote by $\psi_j^{(u)}(u,v)$ the vertex set containing u in the bipartition $\psi_j(u,v)$, and similarly for v; Initialize the extended partition $\bar{\psi}_j^{(u)}(u,v) := \psi_j^{(u)}(u,v)$, $\bar{\psi}_j^{(v)}(u,v) := \psi_j^{(v)}(u,v)$;
  - **Modified Graph.** Let K be $\hat{h}_m^{(i)}$ where all edges between $\psi_j^{(u)}(u,v)$ and $\psi_j^{(v)}(u,v)$ have been removed;
  - **Extension.** For all $w \in \hat{v}_m^{(i)} - (\psi_j^{(u)}(u,v) \cup \psi_j^{(v)}(u,v))$, add w to the side of the partition it is connected to in K (by definition of K, each w as above is connected to exactly one side);
- Return the bipartitions $\{\bar{\psi}_j(u,v)\}_{j=1}^{r(u,v)}$.

Fig. 6. Algorithm EXTENDER. See Figure 5 for an illustration.

easily determine what side of the partition each leaf in $\hat{h}_m^{(i)}$ lies on. For details, see Figure 6. An illustration is given in Figure 5.

3 Concluding Remarks

An interesting question for future work is whether the *approximate* disjointness in our results can be avoided. Since we guarantee that any shared edge lies deep

inside the forest, it is tempting to simply remove all deep edges (say beyond $m/4$) from the output forest. Unfortunately, many of these edges may in fact be contracted and moreover they may be clustered in "supernodes" including both deep and not-so-deep edges. It does not seem to be a trivial task to break these deep supernodes apart and preserve strong reconstruction guarantees.

References

1. Felsenstein, J.: Inferring Phylogenies. Sinauer, Sunderland (2004)
2. Semple, C., Steel, M.: Phylogenetics. Mathematics and its Applications series, vol. 22. Oxford University Press, Oxford (2003)
3. Graham, R.L., Foulds, L.R.: Unlikelihood that minimal phylogenies for a realistic biological study can be constructed in reasonable computational time. Math. Biosci. 60, 133–142 (1982)
4. Day, W.H.E., Sankoff, D.: Computational complexity of inferring phylogenies by compatibility. Syst. Zool. 35(2), 224–229 (1986)
5. Day, W.H.E.: Computational complexity of inferring phylogenies from dissimilarity matrices. Bull. Math. Biol. 49(4), 461–467 (1987)
6. Chor, B., Tuller, T.: Finding a maximum likelihood tree is hard. J. ACM 53(5), 722–744 (2006)
7. Roch, S.: A short proof that phylogenetic tree reconstruction by maximum likelihood is hard. IEEE/ACM Trans. Comput. Biology Bioinform. 3(1), 92–94 (2006)
8. Felsenstein, J.: Cases in which parsimony or compatibility methods will be positively misleading. Syst. Biol., 401–410 (1978)
9. Atteson, K.: The performance of neighbor-joining methods of phylogenetic reconstruction. Algorithmica 25(2-3), 251–278 (1999)
10. Lacey, M.R., Chang, J.T.: A signal-to-noise analysis of phylogeny estimation by neighbor-joining: insufficiency of polynomial length sequences. Math. Biosci. 199(2), 188–215 (2006)
11. Steel, M.A., Székely, L.A.: Inverting random functions. Ann. Comb. 3(1), 103–113 (1999); 3 Combinatorics and biology (Los Alamos, NM, 1998)
12. Steel, M.A., Székely, L.A.: Inverting random functions. II. Explicit bounds for discrete maximum likelihood estimation, with applications. SIAM J. Discrete Math. 15(4), 562–575 (electronic 2002)
13. Erdös, P.L., Steel, M.A., Székely, L.A., Warnow, T.A.: A few logs suffice to build (almost) all trees (part 1). Random Struct. Algor. 14(2), 153–184 (1999)
14. Saitou, N., Nei, M.: The neighbor-joining method: A new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4(4), 406–425 (1987)
15. Daskalakis, C., Mossel, E., Roch, S.: Phylogenies without branch bounds: Contracting the short, pruning the deep (2008) (preprint), `http://arxiv.org/abs/0801.4190`
16. Chang, J.T.: Full reconstruction of Markov models on evolutionary trees: identifiability and consistency. Math. Biosci. 137(1), 51–73 (1996)
17. Philippe, H., Laurent, J.: How good are deep phylogenetic trees? Current Opinion in Genetics & Development 8(8), 616–623 (1998)
18. Ciccarelli, F.D., Doerks, T., von Mering, C., Creevey, C.J., Snel, B., Bork, P.: Toward Automatic Reconstruction of a Highly Resolved Tree of Life. Science 311(5765), 1283–1287 (2006)

19. Mossel, E.: Distorted metrics on trees and phylogenetic forests. IEEE/ACM Trans. Comput. Bio. Bioinform. 4(1), 108–116 (2007)
20. King, V., Zhang, L., Zhou, Y.: On the complexity of distance-based evolutionary tree reconstruction. In: SODA 2003: Proceedings of the fourteenth annual ACM-SIAM symposium on Discrete algorithms, Philadelphia, PA, USA, Society for Industrial and Applied Mathematics, pp. 444–453 (2003)
21. Daskalakis, C., Hill, C., Jaffe, A., Mihaescu, R., Mossel, E., Rao, S.: Maximal accurate forests from distance matrices. In: Apostolico, A., Guerra, C., Istrail, S., Pevzner, P.A., Waterman, M. (eds.) RECOMB 2006. LNCS (LNBI), vol. 3909, pp. 281–295. Springer, Heidelberg (2006)
22. Gronau, I., Moran, S., Snir, S.: Fast and reliable reconstruction of phylogenetic trees with very short edges. To appear in SODA (2008)
23. Erdös, P.L., Steel, M.A., Székely, L.A., Warnow, T.A.: A few logs suffice to build (almost) all trees (part 2). Theor. Comput. Sci. 221, 77–118 (1999)
24. Huson, D.H., Nettles, S.H., Warnow, T.J.: Disk-covering, a fast-converging method for phylogenetic tree reconstruction. J. Comput. Biol. 6(3-4) (1999)
25. Csurös, M., Kao, M.Y.: Provably fast and accurate recovery of evolutionary trees through harmonic greedy triplets. SIAM Journal on Computing 31(1), 306–322 (2001)
26. Csurös, M.: Fast recovery of evolutionary trees with thousands of nodes. J. Comput. Biol. 9(2), 277–297 (2002)
27. Mossel, E., Roch, S.: Learning nonsingular phylogenies and hidden Markov models. Ann. Appl. Probab. 16(2), 583–614 (2006)
28. Daskalakis, C., Mossel, E., Roch, S.: Optimal phylogenetic reconstruction. In: STOC'06: Proceedings of the 38th Annual ACM Symposium on Theory of Computing, pp. 159–168. ACM Press, New York (2006)
29. Buneman, P.: The recovery of trees from measures of dissimilarity. In: Mathematics in the Archaelogical and Historical Sciences, pp. 187–395. Edinburgh University Press, Edinburgh (1971)
30. Meacham, C.A.: A manual method for character compatibility analysis. Taxon 30, 591–600 (1981)
31. Bandelt, H.J., Dress, A.: Reconstructing the shape of a trea from observed dissimilarity data. Adv. Appl. Math. 7(3), 309–343 (1986)