The Annals of Applied Probability
2012, Vol. 22, No. 6, 2429–2459
DOI: 10.1214/11-AAP837

PHYLOGENETIC MIXTURES: CONCENTRATION OF MEASURE IN THE LARGE-TREE LIMIT

BY ELCHANAN MOSSEL[1] AND SEBASTIEN ROCH[2]

University of California, Berkeley, and University of California, Los Angeles

The reconstruction of phylogenies from DNA or protein sequences is a major task of computational evolutionary biology. Common phenomena, notably variations in mutation rates across genomes and incongruences between gene lineage histories, often make it necessary to model molecular data as originating from a mixture of phylogenies. Such mixed models play an increasingly important role in practice.

Using concentration of measure techniques, we show that mixtures of large trees are typically identifiable. We also derive sequence-length requirements for high-probability reconstruction.

1. Introduction. Phylogenetics [10, 22] is centered around the reconstruction of evolutionary histories from molecular data extracted from modern species. The assumption is that molecular data consists of aligned sequences and that each position in the sequences evolves independently according to a Markov model on a tree, where the key parameters are (see Section 3 for formal definitions):

- *Rate matrix.* An $r \times r$ mutation rate matrix Q, where r is the alphabet size. A typical alphabet is the set of nucleotides $\{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$, but here we allow more general state spaces. Without loss of generality, we denote the alphabet by $\mathcal{R} = [r] = \{1, \ldots, r\}$. The (i, j)th entry of Q encodes the rate at which state i mutates into state j.
- *Binary tree.* An evolutionary tree T, where the leaves are the modern species and each branching represents a past speciation event. The leaves are labeled with names of species. Without loss of generality, we assume the labels are $X = [n]$.
- *Branch lengths.* For each edge e, we have a scalar branch length w_e which measures the expected total number of substitutions per site along edge e. Roughly speaking, w_e is the amount of mutational change between the end points of e.

The classical problem in phylogenetics can be stated as follows:

Received August 2011; revised December 2011.
[1]Supported by DMS 0548249 (CAREER) award, by DOD ONR Grant N000141110140, by ISF Grant 1300/08 and by ERC Grant PIRG04-GA-2008-239137.
[2]Supported by NSF Grant DMS-10-07144.
MSC2010 subject classifications. Primary 60K35; secondary 92D15.
Key words and phrases. Phylogenetic reconstruction, random trees, concentration of measure.

- *Phylogenetic tree reconstruction* (*PTR*): *Unmixed case.* Given n molecular sequences of length k,

$$\{s_a = (s_a^i)_{i=1}^k\}_{a\in[n]}$$

with $s_a^i \in [r]$, which have evolved according to the process above with independent sites, reconstruct the topology of the evolutionary tree.

There exists a vast theoretical literature on this problem; see, for example, [22] and references therein.

However, various phenomena, notably variations in mutation rates across genomes and incongruences between gene lineage histories, often make it necessary to model molecular data as originating from a *mixture* of different phylogenies.

Here, using concentration of measure techniques, we show that mixtures of large trees are typically identifiable. By typically, we mean informally that our results hold under conditions guaranteeing that the tree topologies present in the mixture are sufficiently distinct. (See Section 2.2 for a careful statement of the theorems.) In particular, we give a broad new class of conditions under which mixtures are identifiable, and we extend, to more general substitution models, previous results on the total variation distance between Markov models on trees. Our proofs are constructive in that we provide a computationally efficient reconstruction algorithm. We also derive sequence-length requirements for high-probability reconstruction.

Our identifiability and reconstruction results represent an important first step toward dealing with more biologically relevant mixture models (such as the ones mentioned above) in which the tree topologies tend to be similar. In particular, in a recent related paper [18], we have used the techniques developed here to reconstruct common rates-across-sites models.

1.1. *Related work.* Most prior theoretical work on mixture models has focused on the question of *identifiability*. A class of phylogenetic models is identifiable if any two models in the class produce different data distributions. It is well known that unmixed phylogenetic models are typically identifiable [6]. *This is not the case in general for mixtures of phylogenies.* For instance, Steel et al. [24] showed that for any two trees one can find a random scaling on each of them, such that their data distributions are identical. Hence it is hopeless, in general, to reconstruct phylogenies under mixture models. See also [9, 13, 14, 23, 26, 27] for further examples of this type.

However, the negative examples constructed in the references above are not necessarily typical. They use special features of the mutation models (and their invariants) and allow themselves quite a bit of flexibility in setting up the topologies and branch lengths. In fact, recently a variety of more standard mixture models have been shown to be identifiable. These include the common GTR + Γ model

[1, 28] and GTR + Γ + I model [5], as well as some covarion models [3], some group-based models [2] and so-called r-component identical tree mixtures [20]. Although these results do not provide practical algorithms for reconstructing the corresponding mixtures, they do give hope that these problems may be tackled successfully.

Beyond the identifiability question, there seems to have been little rigorous work on reconstructing phylogenetic mixture models. One positive result is the case of the molecular clock assumption with across-sites rate variation [24], although no sequence-length requirements are provided. There is a large body of work on practical reconstruction algorithms for various types of mixtures, notably rates-across-sites models and covarion-type models, using mostly likelihood and Bayesian methods; see, for example, [10] for references. But the optimization problems they attempt to solve are likely NP-hard [7, 21]. There also exist many techniques for testing for the presence of a mixture (e.g., for testing for rate heterogeneity), but such tests typically require the knowledge of the phylogeny; see, for example, [11].

Here we give both identifiability and reconstruction results. The proof of our main results relies on the construction of a *clustering statistic* that discriminates between distinct phylogenies. A similar approach was used recently in [18]. There, however, the problem was to distinguish between phylogenies with the *same* topology, but *different* branch lengths. In the current work, a main technical challenge is to analyze the simultaneous behavior of such a clustering statistic on *distinct* topologies. A similar statistic was also used in [25] to prove a special case of Theorem 2 below. However, in contrast to [25], our main result requires that a clustering statistic be constructed based only on data generated by the mixture—that is, *without* prior knowledge of the topologies to be distinguished. Finally, unlike [18] and [25], we consider the more general GTR model.

2. Definitions and results.

2.1. *Basic definitions*.

Phylogenies. A phylogeny is a graphical representation of the speciation history of a group of organisms. The leaves typically correspond to current species. Each branching indicates a speciation event. Moreover we associate to each edge a positive weight. This weight can be thought roughly as the amount of evolutionary change on the edge. More formally, we make the following definitions; see, for example, [22]. Fix a set of leaf labels $X = [n] = \{1, \ldots, n\}$.

DEFINITION 2.1 (Phylogeny). A *weighted binary phylogenetic X-tree* (or *phylogeny*) $T = (V, E; \phi; w)$ is a tree with vertex set V, edge set E, leaf set L with $|L| = n$, and a bijective mapping $\phi : X \to L$ such that:

(1) The degree of all internal vertices $V - L$ is exactly 3.
(2) The edges are assigned weights $w : E \to (0, +\infty)$.

We let $\mathcal{T}_l[T] = (V, E; \phi)$ be the *leaf-labelled topology* of T.

DEFINITION 2.2 (Tree metric). A phylogeny $T = (V, E; \phi; w)$ is naturally equipped with a *tree metric* $d_T : X \times X \to (0, +\infty)$ defined as follows:

$$\forall a, b \in X \qquad d_T(a, b) = \sum_{e \in \mathrm{Path}_T(\phi(a), \phi(b))} w_e,$$

where $\mathrm{Path}_T(u, v)$ is the set of edges on the path between u and v in T. We will refer to $d_T(a, b)$ as the *evolutionary distance* between a and b. In a slight abuse of notation, we also sometimes use $d_T(u, v)$ to denote the evolutionary distance as above between any two vertices u, v of T.

We will restrict ourselves to the following standard special case.

DEFINITION 2.3 (Regular phylogenies). Let $0 < f \leq g < +\infty$. We denote by $\mathbb{Y}_{f,g}^{(n)}$ the set of phylogenies $T = (V, E; \phi; w)$ with n leaves such that $f \leq w_e \leq g$, $\forall e \in E$. We also let $\mathbb{Y}_{f,g} = \bigcup_{n \geq 1} \mathbb{Y}_{f,g}^{(n)}$.

GTR model. A commonly used model of DNA sequence evolution is the following *GTR model*; see, for example, [22]. We first define an appropriate class of rate matrices.

DEFINITION 2.4 (GTR rate matrix). Let $\mathcal{R}$ be a set of character states with $r = |\mathcal{R}|$. Without loss of generality we assume that $\mathcal{R} = [r]$. Let π be a probability distribution on $\mathcal{R}$ satisfying $\pi_x > 0$ for all $x \in \mathcal{R}$. A *general time-reversible* (*GTR*) *rate matrix* on $\mathcal{R}$, with respect to stationary distribution π, is an $r \times r$ real-valued matrix Q such that:

(1) $Q_{xy} > 0$ for all $x \neq y \in \mathcal{R}$.
(2) $\sum_{y \in \mathcal{R}} Q_{xy} = 0$, for all $x \in \mathcal{R}$.
(3) $\pi_x Q_{xy} = \pi_y Q_{yx}$, for all $x, y \in \mathcal{R}$.

By the reversibility assumption, Q has r real eigenvalues

$$0 = \Lambda_1 > \Lambda_2 \geq \cdots \geq \Lambda_r.$$

We normalize Q by fixing $\Lambda_2 = -1$.

DEFINITION 2.5 (GTR model). Consider the following stochastic process. We are given a phylogeny $T = (V, E; \phi; w)$ and a finite set $\mathcal{R}$ with r elements.

Let π be a probability distribution on $\mathcal{R}$ and Q be a GTR rate matrix with respect to π. Associate to each edge $e \in E$ the stochastic matrix

$$M(e) = \exp(w_e Q).$$

The process runs as follows. Choose an arbitrary root $\rho \in V$. Denote by $E_\downarrow$ the set E directed away from the root. Pick a state for the root at random according to π. Moving away from the root toward the leaves, apply the channel $M(e)$ to each edge e independently. Denote the state so obtained $s_V = (s_v)_{v \in V}$. In particular, s_L is the state at the leaves, which we also denote by s_X. More precisely, the joint distribution of s_V is given by

$$\mu_V(s_V) = \pi_\rho(s_\rho) \prod_{e=(u,v)\in E_\downarrow} [M(e)]_{s_u s_v}.$$

For $W \subseteq V$, we denote by μ_W the marginal of μ_V at W. Under this model, the weight w_e is the expected number of substitutions on edge e in the continuous-time process. We denote by $\mathcal{D}[T, Q]$ the probability distribution of s_V. We also let $\mathcal{D}_l[T, Q]$ denote the probability distribution of

$$s_X \equiv \left(s_{\phi(a)}\right)_{a \in X}.$$

More generally, we consider k independent samples $\{s_V^i\}_{i=1}^k$ from the model above, that is, $s_V^1, \ldots, s_V^k$ are i.i.d. $\mathcal{D}[T, Q]$. We think of $(s_v^i)_{i=1}^k$ as the sequence at node $v \in V$. Typically, $\mathcal{R} = \{\mathrm{A}, \mathrm{G}, \mathrm{C}, \mathrm{T}\}$ and the model describes how DNA sequences stochastically evolve by point mutations along an evolutionary tree under the assumption that each site in the sequences evolves independently. When considering many samples $\{s_V^i\}_{i=1}^k$, we drop the subscript to refer to a single sample s_V.

Mixed model. We introduce the basic mixed model which will be the focus of this paper. We will use the following definition. We assume that Q is fixed and known throughout.

REMARK 2.1 (Unknown rate matrix). See the concluding remarks for an extension of our techniques when Q is unknown.

DEFINITION 2.6 (Θ-mixture). Let Θ be a positive integer. In the Θ-*mixture model*, we consider a finite set of phylogenies

$$\mathbb{T} = \{T_\theta = (V_\theta, E_\theta; \phi_\theta; w_\theta)\}_{\theta=1}^\Theta$$

on the same set of leaf labels $X = [n]$ and a positive probability distribution $\nu = (\nu_\theta)_{\theta=1}^\Theta$ on $[\Theta]$. Consider k i.i.d. random variables $N^1, \ldots, N^k$ with distribution ν. Then, conditioned on $N^1, \ldots, N^k$, the samples $\{s_X^i\}_{i=1}^k$ generated under the Θ-mixture model $(\mathbb{T}, \nu, Q)$ are independent with conditional distribution $s_X^j \sim \mathcal{D}_l[T_{N^j}, Q]$, $j = 1, \ldots, k$. We denote by $\mathcal{D}_l[(\mathbb{T}, \nu, Q)]$ the probability distribution of s_X^1. We will refer to T_θ as the θ-*component* of the mixture $(\mathbb{T}, \nu, Q)$.

We assume that Θ is fixed and known throughout. As above, we drop the superscript to refer to a single sample s_X with corresponding component indicator N. To simplify notation, we let

$$d_{T_\theta} = d_\theta \qquad \forall \theta \in [\Theta].$$

Some notation. We will use the notation $[n]^2 = \{(a,b) \in [n] \times [n] : a \leq b\}$, $[n]^2_= = \{(a,a)\}_{a\in[n]}$ and $[n]^2_{\neq} = [n]^2 - [n]^2_=$. We also denote by $[n]^4_{\neq}$ the set of pairs $(a_1,b_1),(a_2,b_2) \in [n]^2_{\neq}$ such that $(a_1,b_1) \neq (a_2,b_2)$ (as pairs). We use the notation $\mathrm{poly}(n)$ to denote the growth condition usually written $\Theta(n^C)$ for some $C > 0$.

2.2. *Main results.* We make the following assumptions on the mutation model.

ASSUMPTION 1. Let $0 < f \leq g < +\infty$, and $\underline{\nu} > 0$. We will use the following set of assumptions on a Θ-mixture model $(\mathbb{T}, \nu, Q)$:

(1) *Regular phylogenies*: $T_\theta \in \mathbb{Y}_{f,g}, \forall \theta \in [\Theta]$.
(2) *Minimum frequency*: $\nu_\theta \geq \underline{\nu}, \forall \theta \in [\Theta]$.

We denote by $\Theta\text{-M}[f,g,\underline{\nu},n]$ the set of Θ-mixture models on n leaves satisfying these conditions.

REMARK 2.2 (No minimum frequency). See the concluding remarks for an extension of our techniques when the minimum frequency assumption is not satisfied.

Tree identifiability. Our first result states that, under Assumption 1, Θ-mixture models are identifiable—except for an "asymptotically negligible fraction." To formalize this notion, we use the following definition. Note that $\Theta\text{-M}[f,g,\underline{\nu},n]$ is a compact subset of a finite product of metric spaces [4] which we equip with its Borel σ-algebra.

DEFINITION 2.7 (Permutation-invariant measure). Let

$$A \subseteq \Theta\text{-M}(f,g,\underline{\nu},n)$$

be a Borel set. Given Θ permutations $\Pi = \{\Pi_\theta\}_{\theta\in[\Theta]}$ of X, we let

$$\Pi[\mathbb{T}] \equiv \{\Pi_\theta[T_\theta]\}_{\theta\in[\Theta]} \equiv \{(V_\theta, E_\theta; \phi_\theta \circ \Pi_\theta; w_\theta)\}_{\theta\in[\Theta]},$$

where $\circ$ indicates composition, and

$$A_\Pi = \{(\mathbb{T},\nu,Q) \in \Theta\text{-M}(f,g,\underline{\nu},n) : (\Pi[\mathbb{T}],\nu,Q) \in A\}.$$

A probability measure λ on $\Theta\text{-M}(f,g,\underline{\nu},n)$ is *permutation-invariant* if for all A and Π as above, we have the following:

$$\lambda[A] = \lambda[A_\Pi].$$

REMARK 2.3. Alternatively one can think of a permutation-invariant measure as first picking unlabeled trees, branch weights and mixture frequencies according to a specified joint distribution, and then labeling the leaves of each tree in the mixture independently, uniformly at random. Note that the *independent* labeling of the trees is needed for our proof. It ensures that the phylogenies in the mixture are typically, "sufficiently distinct." Generalizing our results, possibly in a weaker form, to mixtures of "similar" phylogenies is an important open problem. See [18] for recent progress in this direction.

For two Θ-mixture models $(\mathbb{T}, \nu, Q)$ and $(\mathbb{T}' = \{T'_\theta\}_{\theta\in[\Theta]}, \nu', Q)$, we write

$$(\mathbb{T}, \nu, Q) \nsim (\mathbb{T}', \nu', Q),$$

if there is no bijective mapping h of $[\Theta]$ such that

$$\mathcal{T}_l[T_\theta] = \mathcal{T}_l\big[T'_{h(\theta)}\big] \qquad \forall\theta \in [\Theta].$$

In words, $(\mathbb{T}, \nu, Q)$ and $(\mathbb{T}', \nu', Q)$ are *not* equivalent up to component re-labeling.

THEOREM 1 (Tree identifiability). *Fix* $0 < f \le g < +\infty$, *and* $\underline{\nu} > 0$. *Then, there exists a sequence of Borel subsets*

$$A_n \subseteq \Theta\text{-M}(f, g, \underline{\nu}, n), \qquad n \ge 1,$$

such that the following hold:

(1) *For any sequence of permutation-invariant measures* λ_n, $n \ge 1$, *respectively, on* Θ-M$(f, g, \underline{\nu}, n)$, $n \ge 1$, *we have*

$$\lambda_n[A_n] = 1 - o_n(\underline{\nu}, f, g)$$

as $n \to \infty$. *Here* $o_n(\underline{\nu}, f, g)$ *indicates convergence to* 0 *as* $n \to \infty$ *for fixed* $\underline{\nu}, f, g$.

(2) *For all*

$$(\mathbb{T}, \nu, Q) \nsim (\mathbb{T}', \nu', Q) \in \bigcup_{n\ge1} A_n,$$

we have

$$\mathcal{D}_l[(\mathbb{T}, \nu, Q)] \ne \mathcal{D}_l[(\mathbb{T}', \nu', Q)].$$

REMARK 2.4. As remarked above, our proof requires that the phylogenies in the mixture are "sufficiently different." This is typically the case under a permutation-invariant measure. Roughly speaking, the complements of the sets A_n in the previous theorem contain those exceptional instances where the phylogenies are too "similar." See the proof for a formal definition of A_n.

Tree distance. We also generalize to GTR models a result of Steel and Székely: phylogenies are typically far away in variational distance [25]. The techniques in [25] apply only to group-based models and other highly symmetric models; see [25] for details. Let $\|\cdot\|_{\mathrm{TV}}$ denote total variation distance; that is, for two probability measures $\mathcal{D}, \mathcal{D}'$ on a measure space $(\Omega, \mathcal{F})$ define

$$\|\mathcal{D} - \mathcal{D}'\|_{\mathrm{TV}} = \sup_{B \in \mathcal{F}} |\mathcal{D}(B) - \mathcal{D}'(B)|.$$

THEOREM 2 (Tree distance). *Let $\{A_n\}_n$ be as in Theorem 1 where $\Theta = 2$ and $\underline{\nu} = 1/2$ [in which case we necessarily have $\nu = (1/2, 1/2)$]. Then for all*

$$(\mathbb{T}, \nu, Q) \in \bigcup_{n \geq 1} A_n,$$

we have

$$\|\mathcal{D}_l[T_1, Q] - \mathcal{D}_l[T_2, Q]\|_{\mathrm{TV}} = 1 - o_n(1).$$

REMARK 2.5. Note that ν plays no substantive role in the previous theorem other than to determine A_n.

Tree reconstruction. The proof of Theorems 1 and 2 rely on the following reconstruction result of independent interest. We show that the topologies can be reconstructed efficiently with high confidence using polynomial length sequences. Recall that k denotes the sequence length.

THEOREM 3 (Tree reconstruction). *Fix $0 < f \leq g < +\infty$, and $\underline{\nu} > 0$. Then, there exists a sequence of Borel subsets*

$$A_n \subseteq \Theta\text{-M}(f, g, \underline{\nu}, n), \qquad n \geq 1,$$

such that the following hold:

(1) *For any sequence of permutation-invariant measures λ_n, $n \geq 1$, respectively, on Θ-M$(f, g, \underline{\nu}, n)$, $n \geq 1$, we have*

$$\lambda_n[A_n] = 1 - o_n(\underline{\nu}, f, g)$$

as $n \to \infty$.

(2) *For all*

$$(\mathbb{T}, \nu, Q) \in \bigcup_{n \geq 1} A_n,$$

the topologies of $(\mathbb{T}, \nu, Q)$ can be reconstructed in time polynomial in n and k using polynomially many samples (i.e., k is polynomial in n) with probability $1 - o_n(\underline{\nu}, f, g)$ under the samples and the randomness of the algorithm.

REMARK 2.6. The subsets $\{A_n\}_n$ in Theorems 1 and 3 are in fact the same.

The rest of the paper is devoted to the proof of Theorem 3 which implies Theorems 1 and 2.

2.3. *Proof overview.* The proof of Theorem 3 relies on the construction of a *clustering statistic* that discriminates between distinct phylogenies.

Clustering statistic. Fix $0 < f \le g < +\infty$ and $\underline{\nu} > 0$. Suppose for now that $\Theta = 2$, and let λ be a permutation-invariant probability measure on Θ-M$[f, g, \underline{\nu}, n]$. It will be useful to think of λ as a two-step procedure: first pick unlabeled, weighted topologies; and second, assign a uniformly random labeling to the leaves of each tree. Pick a Θ-mixture model $(\mathbb{T}, \nu, Q)$ according to λ. We will denote by $\mathbb{P}_\lambda$ and $\mathbb{E}_\lambda$ probability and expectation under λ. Similarly, we denote by $\mathbb{P}_l$ and $\mathbb{E}_l$ (resp., $\mathbb{P}_\mathcal{A}$ and $\mathbb{E}_\mathcal{A}$) probability and expectation under $(\mathbb{T}, \nu, Q)$ (resp., under the randomness of our algorithm), as well as combinations such as $\mathbb{P}_{\mathcal{A},\lambda}$ with the obvious meaning.

Let $\mathbf{z} = (z_x)_{x=1}^r$ be a (real-valued) right eigenvector of Q corresponding to eigenvalue $\Lambda_2 = -1$ and normalize $\mathbf{z}$ so that

$$\sum_{x=1}^{r} \pi_x z_x^2 = 1.$$

(Any negative eigenvalue could be used instead.) Consider the following one-dimensional mapping of the samples ([17], Lemma 5.3): for all $i = 1, \ldots, k$ and $a \in X$,

$$\sigma_a^i = z_{s_a^i}. \tag{1}$$

Recall that we drop the superscript when referring to a single sample. It holds that

$$\mathbb{E}_l[\sigma_a | N = \theta] = 0. \tag{2}$$

Moreover, following a computation in [17], Lemma 5.3, letting $a \wedge b$ be the most recent common ancestor of a and b (under the arbitrary choice of root ρ) one has

$$\begin{aligned} q_\theta(a,b) &= \mathbb{E}_l[\sigma_a\sigma_b | N = \theta] - \mathbb{E}_l[\sigma_a | N = \theta]\mathbb{E}[\sigma_b | N = \theta] \\ &= \mathbb{E}_l[\sigma_a\sigma_b | N = \theta] \\ &= \sum_{x=1}^{r} \pi_x \mathbb{E}_l[\sigma_a\sigma_b | N = \theta, s_{a\wedge b} = x] \\ &= \sum_{x=1}^{r} \pi_x \mathbb{E}_l[\sigma_a | N = \theta, s_{a\wedge b} = x]\mathbb{E}_l[\sigma_b | N = \theta, s_{a\wedge b} = x] \\ &= \sum_{x=1}^{r} \pi_x \big(e^{-d_\theta(a\wedge b, a)} z_x\big)\big(e^{-d_\theta(a\wedge b, b)} z_x\big) \\ &= e^{-d_\theta(a,b)} \end{aligned} \tag{3}$$

and

$$q(a,b) = \mathbb{E}_l[\sigma_a\sigma_b] - \mathbb{E}_l[\sigma_a]\mathbb{E}_l[\sigma_b] = \mathbb{E}_l[\sigma_a\sigma_b] = \sum_{\theta=1}^{\Theta} \nu_\theta e^{-d_\theta(a,b)}. \tag{4}$$

We use a statistic of the form

$$\mathcal{U} = \frac{1}{|\Upsilon|} \sum_{(a,b)\in\Upsilon} \sigma_a\sigma_b, \tag{5}$$

where $\Upsilon \subseteq [n]^2_{\neq}$. For $\mathcal{U}$ to be effective in discriminating between T_1 and T_2, we require the following (informal) conditions:

(C1) The difference in conditional expectations

$$\Delta = \bigl|\mathbb{E}_l[\mathcal{U}|N=1] - \mathbb{E}_l[\mathcal{U}|N=2]\bigr|$$

is large.

(C2) The statistic $\mathcal{U}$ is concentrated around its mean under both $\mathcal{D}_l[T_1, Q]$ and $\mathcal{D}_l[T_2, Q]$.

(C3) The set Υ can be constructed from data generated by the mixture $(\mathbb{T}, \nu, Q)$.

A $\mathcal{U}$ satisfying C1–C3 could be used to infer the *hidden* variables $N^1, \ldots, N^k$ and, thereby, to cluster the samples in their respective component.

Prior work. In [18], it was shown in a related context that taking $\Upsilon = [n]^2_{\neq}$ is not in general an appropriate choice, as it may lead to a large variance. Instead, the following lemma was used.

CLAIM (Disjoint close pairs [25]; see also [18]). *For any $T \in \mathbb{Y}^{(n)}_{f,g}$, there exists a subset $\Upsilon \subseteq [n]^2_{\neq}$ such that the following hold*:

(1) $|\Upsilon| = \Omega(n)$;
(2) $\forall (a,b) \in \Upsilon$, $d_T(a,b) \leq 3g$;
(3) $\forall (a_1,b_1) \neq (a_2,b_2) \in \Upsilon$, *the paths* $\mathrm{Path}_T(a_1,b_1)$ *and* $\mathrm{Path}_T(a_2,b_2)$ *are edge-disjoint. We will say that such pairs are T-disjoint.*

For special Q matrices, it was shown in [25] and [18] that such a Υ for $T = T_1$, say, can be used to construct a clustering statistic [similar to (5)] concentrated under $\mathcal{D}_l[T_1, Q]$. In particular, the T_1-disjointness assumption above implies the independence of the variables $\sigma_{a_1}\sigma_{b_1}$ and $\sigma_{a_2}\sigma_{b_2}$ under the Q matrices considered in [18, 25]. Moreover, Steel and Székely [25] proved the existence of a further subset that is also T_2-disjoint, but their construction requires the knowledge of T_2. Here we show how to satisfy conditions C1–C3 under GTR models.

High-level construction. We give a sketch of our techniques. Formal statements and full proofs can be found in Sections 3, 4 and 5. For $\alpha > 0$, let

$$\Upsilon_{\alpha,\theta} = \{(a,b) \in [n]^2_{\neq} : d_\theta(a,b) \leq \alpha\}$$

and

$$\Upsilon_\alpha = \bigcup_{\theta \in [\Theta]} \Upsilon_{\alpha,\theta}.$$

Because the variables $N^1, \ldots, N^k$ are hidden, we cannot infer $\Upsilon_{\alpha,\theta}$ directly from the samples, for instance, using (3). Instead:

(*Step* 1) Using (4) and the estimator

$$\hat{q}(a,b) = \frac{1}{k} \sum_{i=1}^{k} \sigma_a^i \sigma_b^i,$$

we construct a set with size linear in n satisfying

$$\Upsilon_{4g} \subseteq \Upsilon' \subseteq \Upsilon_{C_c}$$

for an appropriate constant C_c; see Lemma 4.1.

Define

$$\Upsilon'_\theta = \Upsilon' \cap \Upsilon_{C_c,\theta}.$$

For general GTR rate matrices, T_θ-disjointness of $(a_1,b_1), (a_2,b_2) \in \Upsilon'_\theta$ does not guarantee independence of $\sigma_{a_1}\sigma_{b_1}$ and $\sigma_{a_2}\sigma_{b_2}$ under $\mathcal{D}_l[T_\theta, Q]$. Instead, we choose pairs that are far enough from each other by picking a sufficiently sparse random subset of Υ'; see Lemma 3.8. We say that $(a_1,b_1), (a_2,b_2) \in \Upsilon'_\theta$ are T_θ-*far* if the smallest evolutionary distance between $\{a_1,b_1\}$ and $\{a_2,b_2\}$ is at least $C_f \log\log n$ for a constant $C_f > 0$ to be determined.

(*Step* 2) We take a random subset Υ'' of Υ' with

$$|\Upsilon''| = \Theta(\log n);$$

see Lemma 4.2.

Denoting

$$\Upsilon''_\theta = \Upsilon'' \cap \Upsilon_{C_c,\theta},$$

we show that all $(a_1,b_1) \neq (a_2,b_2) \in \Upsilon''_\theta$ are T_θ-far. Under a permutation-invariant λ, a pair $(a,b) \in \Upsilon_{\alpha,1}$ is unlikely to be in $\Upsilon_{\alpha,2}$. In particular, we show that, under λ, the intersection of Υ''_1 and Υ''_2 is empty. In fact, a pair $(a,b) \in \Upsilon_{\alpha,1}$ is likely to be such that $d_2(a,b)$ is large. We say that $(a,b) \in [n]^2_{\neq}$ is T_θ-*stretched* if $d_\theta(a,b) \geq C_{st} \log\log n$ for a constant $C_{st} > 0$ to be determined. We show that all $(a,b) \in \Upsilon''_1$ are T_2-stretched; see Lemma 3.7.

To infer Υ''_θ, we consider the quantity

$$\hat{r}(c_1, c_2) = \frac{1}{k}\sum_{i=1}^{k}[\sigma^i_{a_1}\sigma^i_{b_1}\sigma^i_{a_2}\sigma^i_{b_2} - \hat{q}(a_1, b_1)\hat{q}(a_2, b_2)]$$

for $c_1 = (a_1, b_1) \neq c_2 = (a_2, b_2) \in [n]^2_{\neq}$. We note that if $(a, b) \in \Upsilon''$ is T_2-stretched, then

$$\mathbb{E}_l[\sigma_a\sigma_b|N = 2] \approx \mathbb{E}_l[\sigma_a|N = 2]\mathbb{E}_l[\sigma_b|N = 2] = 0$$

and

$$q(a, b) \approx \nu_1 q_1(a, b).$$

There are then two cases:

(I) If $c_1 = (a_1, b_1) \neq c_2 = (a_2, b_2) \in \Upsilon''_1$ (and similarly for Υ''_2), they are T_1-far and each is T_2-stretched. Moreover we show that (c_1, c_2) is T_2-far. Therefore,

$$q(a_1, b_1) \approx \nu_1 q_1(a_1, b_1), q(a_2, b_2) \approx \nu_1 q_1(a_2, b_2),$$

and we show further that

$$\begin{aligned}
&\mathbb{E}_l[\sigma_{a_1}\sigma_{b_1}\sigma_{a_2}\sigma_{b_2}] \\
&\quad\approx \nu_1\mathbb{E}_l[\sigma_{a_1}\sigma_{b_1}|N = 1]\mathbb{E}_l[\sigma_{a_2}\sigma_{b_2}|N = 1] \\
&\qquad + \nu_2\mathbb{E}_l[\sigma_{a_1}|N = 2]\mathbb{E}_l[\sigma_{b_1}|N = 2]\mathbb{E}_l[\sigma_{a_2}|N = 2]\mathbb{E}_l[\sigma_{b_2}|N = 2] \\
&\quad\approx \nu_1 q_1(a_1, b_1)q_1(a_2, b_2).
\end{aligned}$$

So

$$\hat{r}(c_1, c_2) \approx \nu_1(1 - \nu_1)q_1(a_1, b_1)q_1(a_2, b_2) > 0.$$

(II) On the other hand, if $c_1 = (a_1, b_1) \in \Upsilon''_1$ and $c_2 = (a_2, b_2) \in \Upsilon''_2$, then c_1 is T_2-stretched, and c_2 is T_1-stretched. Moreover we show that (c_1, c_2) is both T_1-far and T_2-far. Therefore,

$$q(a_1, b_1) \approx \nu_1 q_1(a_1, b_1), \qquad q(a_2, b_2) \approx \nu_2 q_2(a_2, b_2),$$

and we show that

$$\begin{aligned}
\mathbb{E}_l[\sigma_{a_1}\sigma_{b_1}\sigma_{a_2}\sigma_{b_2}] &\approx \nu_1\mathbb{E}_l[\sigma_{a_1}\sigma_{b_1}|N = 1]\mathbb{E}_l[\sigma_{a_2}|N = 1]\mathbb{E}_l[\sigma_{b_2}|N = 1] \\
&\quad + \nu_2\mathbb{E}_l[\sigma_{a_1}|N = 2]\mathbb{E}_l[\sigma_{b_1}|N = 2]\mathbb{E}_l[\sigma_{a_2}\sigma_{b_2}|N = 2] \\
&\approx 0.
\end{aligned}$$

So

$$\hat{r}(c_1, c_2) \approx -\nu_1 q_1(a_1, b_1)\nu_2 q_2(a_2, b_2) < 0;$$

see Lemma 3.9.

The argument above leads to the following step.

(*Step* 3) For all pairs $c_1 = (a_1, b_1)$ and $c_2 = (a_2, b_2)$ in Υ'', we compute $\hat{r}(c_1, c_2)$. Using cases I and II, we then infer the sets Υ''_1 and Υ''_2. We form the clustering statistics

$$\mathcal{U}^i_\theta = \frac{1}{|\Upsilon''_\theta|} \sum_{(a,b)\in\Upsilon''_\theta} \sigma^i_a \sigma^i_b$$

for $\theta = 1, 2$ and $i = 1, \ldots, k$; see Lemma 4.3.

By the arguments in cases I and II above, we get that for $(a, b) \in \Upsilon''_1$,

$$\mathbb{E}_l[\sigma_a \sigma_b | N = 1] \approx \nu_1 q_1(a, b),$$

whereas

$$\mathbb{E}_l[\sigma_a \sigma_b | N = 2] \approx \mathbb{E}_l[\sigma_a | N = 2]\mathbb{E}_l[\sigma_b | N = 2] \approx 0,$$

so that (dropping the superscript to refer to a single sample)

$$\mathbb{E}_l[\mathcal{U}_1 | N = 1] > C_\Delta,$$

whereas

$$\mathbb{E}_l[\mathcal{U}_1 | N = 2] < C_\Delta$$

for a constant $C_\Delta > 0$ to be determined later; see Lemma 3.10. Moreover, the properties of Υ''_θ discussed in cases I and II allow us to prove further that $\mathcal{U}_\theta$ is concentrated around its mean; see Lemma 3.11. This leads to the following step.

(*Step* 4) Divide the samples $i = 1, \ldots, k$ into two clusters K_1 and K_2, according to whether

$$\mathcal{U}^i_1 > C_\Delta \quad \text{or} \quad \mathcal{U}^i_2 > C_\Delta,$$

respectively; see Lemma 5.1.

Once the samples are divided into pure components, we apply standard reconstruction techniques to infer each topology.

(*Step* 5) For $\theta = 1, 2$, reconstruct the topology $\mathcal{T}_l[T_\theta]$ from the samples in K_θ; see Lemma 5.3.

General Θ. When $\Theta > 2$, we proceed as above and construct a clustering statistic for each component.

3. Main lemmas. In this section, we derive a number of preliminary results. These results are also described informally in Section 2.3.

Fix a GTR matrix Q and constants $\Theta \geq 2$, $0 < f \leq g < +\infty$ and $\underline{\nu} > 0$. Let λ be a permutation-invariant probability measure on Θ-M$[f, g, \underline{\nu}, n]$. Pick a Θ-mixture

model $(\mathbb{T}, \nu, Q)$ according to λ, and generate k independent samples $\{s_X^i\}_{i=1}^k$ from $\mathcal{D}_l[(\mathbb{T}, \nu, Q)]$. We work with the mapping $\{\sigma_X\}_{i=1}^k$ defined in (1).

Throughout we assume that the number of samples is $k = n^{C_k}$ for some $C_k > 0$ to be fixed later.

3.1. *Useful lemmas.* We will need the following standard concentration inequalities; see, for example, [19]:

LEMMA 3.1 (Azuma–Hoeffding inequality). *Suppose* $\mathbf{Z} = (Z_1, \ldots, Z_m)$ *are independent random variables taking values in a set* S, *and* $h : S^m \to \mathbb{R}$ *is any* t*-Lipschitz function*: $|h(\mathbf{z}) - h(\mathbf{z}')| \le t$ *whenever* $\mathbf{z}, \mathbf{z}' \in S^m$ *differ at just one coordinate. Then*, $\forall \zeta > 0$,

$$\mathbb{P}[|h(\mathbf{Z}) - \mathbb{E}[h(\mathbf{Z})]| \ge \zeta] \le 2\exp\left(-\frac{\zeta^2}{2t^2 m}\right).$$

LEMMA 3.2 (Chernoff bounds). *Let* $Z_1, \ldots, Z_m$ *be independent Poisson trials such that, for* $1 \le i \le m$, $\mathbb{P}[Z_i = 1] = p_i$ *where* $0 < p_i < 1$. *Then, for* $Z = \sum_{i=1}^m Z_i$, $M = \mathbb{E}[Z] = \sum_{i=1}^m p_i$, $0 < \delta_- \le 1$, *and* $\delta_+ > 2e - 1$,

$$\mathbb{P}[Z < (1 - \delta_-)M] < e^{-M\delta_-^2/2}$$

and

$$\mathbb{P}[Z > (1 + \delta_+)M] < 2^{-(1+\delta_+)M}.$$

3.2. *Large-sample asymptotics.* Denoting $\mathcal{K} = [k]$, let $\mathcal{K}_\theta \subseteq \mathcal{K}$ be those samples coming from component θ, that is,

$$\mathcal{K}_\theta = \{i \in \mathcal{K} : N^i = \theta\}.$$

LEMMA 3.3 (Size of $\mathcal{K}_\theta$). *Under* $\mathbb{P}_l$, *for any* $C_s > 1$, *we have*

$$C_s^{-1} \le \frac{|\mathcal{K}_\theta|}{\nu_\theta k} \le C_s$$

for all $\theta \in [\Theta]$, *except with probability* $\exp(-\Omega(n^{C_k}))$.

PROOF. Recall that $\underline{\nu} \le \nu_\theta \le 1 - \underline{\nu}$. Using Lemma 3.1 with $m = k$ and

$$\zeta = \nu_\theta k \max\{1 - C_s^{-1}, C_s - 1\} = \nu_\theta k(C_s - 1)$$

gives the result. □

Consider the estimators

$$\hat{q}_\theta(a, b) = \frac{1}{|\mathcal{K}_\theta|} \sum_{i \in \mathcal{K}_\theta} \sigma_a^i \sigma_b^i$$

and

$$\hat{q}(a,b) = \frac{1}{k}\sum_{i=1}^{k} \sigma_a^i \sigma_b^i.$$

Let

$$q_\theta(a,b) = e^{-d_\theta(a,b)}$$

and

$$q(a,b) = \sum_{\theta=1}^{\Theta} \nu_\theta q_\theta(a,b).$$

LEMMA 3.4 (Accuracy of $\hat{q}$). *Fix $0 < C_q < C_k/2$. Under $\mathbb{P}_l$, we have*

$$|\hat{q}(a,b) - q(a,b)| \le n^{-C_q}$$

and

$$|\hat{q}_\theta(a,b) - q_\theta(a,b)| \le n^{-C_q}$$

for all $\theta \in [\Theta]$ and all $(a,b) \in [n]^2_{\ne}$ except with probability $\exp(-\mathrm{poly}(n))$.

PROOF. For each $(a,b) \in [n]^2_{\ne}$, $\hat{q}(a,b)$ is a sum of k independent variables. By Lemma 3.1, taking $m = k$, $t = k^{-1}\max_i |z_i|^2$, $\zeta = n^{-C_q}$, we have

$$|\hat{q}(a,b) - q(a,b)| \le n^{-C_q},$$

except with probability $2\exp(-\Omega(n^{C_k - 2C_q}))$. Note that there are at most n^2 elements in $[n]^2_{\ne}$ so that the probability of failure is at most

$$2n^2 \exp(-\Omega(n^{C_k - 2C_q})) = \exp(-\Omega(n^{C_k - 2C_q})).$$

Using Lemma 3.3, the same holds for each θ. The overall probability of failure under $\mathbb{P}_l$ is $\exp(-\Omega(n^{C_k - 2C_q}))$. □

Following the same argument, a similar result holds for

$$\hat{r}(c_1,c_2) = \frac{1}{k}\sum_{i=1}^{k} [\sigma_{a_1}^i \sigma_{b_1}^i \sigma_{a_2}^i \sigma_{b_2}^i - \hat{q}(a_1,b_1)\hat{q}(a_2,b_2)]$$

for $c_1 = (a_1,b_1) \ne c_2 = (a_2,b_2) \in [n]^2_{\ne}$. Let

$$r(c_1,c_2) = \mathbb{E}_l[\hat{r}(c_1,c_2)].$$

LEMMA 3.5 (Accuracy of $\hat{r}$). *Under $\mathbb{P}_l$, we have*

$$|\hat{r}(c_1,c_2) - r(c_1,c_2)| \le n^{-C_q}$$

for all $c_1 = (a_1,b_1) \ne c_2 = (a_2,b_2) \in [n]^2_{\ne}$ except with probability $\exp(-\mathrm{poly}(n))$.

3.3. *Combinatorial properties*. For $\alpha > 0$, let

$$\Upsilon_{\alpha,\theta} = \{(a,b) \in [n]^2_{\neq} : d_\theta(a,b) \leq \alpha\} \tag{6}$$

and

$$\Upsilon_\alpha = \bigcup_{\theta \in [\Theta]} \Upsilon_{\alpha,\theta}. \tag{7}$$

The lower bound below follows from a (stronger) lemma in [25]; see also [18].

LEMMA 3.6 (Size of $\Upsilon_{\alpha,\theta}$). *For all $\alpha > 0$ and $\theta \in [\Theta]$,*

$$\tfrac{1}{4}n \leq |\Upsilon_{\alpha,\theta}| \leq 2^{\lfloor \alpha/f \rfloor} n.$$

In particular,

$$\tfrac{1}{4}n \leq |\Upsilon_\alpha| \leq \Theta 2^{\lfloor \alpha/f \rfloor} n.$$

PROOF. For $a \in X$ and $\alpha \geq 4g$, let

$$\mathcal{B}_\alpha(a) = \{v \in V : d_\theta(\phi_\theta(a), v) \leq \alpha\}.$$

Since T_θ is binary, there are at most $2^{\lfloor \alpha/f \rfloor}$ vertices within evolutionary distance α, that is,

$$|\mathcal{B}_\alpha(a)| \leq 2^{\lfloor \alpha/f \rfloor}.$$

Restricting to leaves gives the upper bound.

Let

$$\Gamma_\alpha = \big\{a \in [n] : d_\theta(a,b) > \alpha, \forall b \in [n] - \{a\}\big\},$$

that is, Γ_α is the set of leaves with no other leaf at evolutionary distance α in T_θ. We will bound the size of Γ_α. Note that for all $a, b \in \Gamma_\alpha$ with $a \neq b$, we have $\mathcal{B}_{\alpha/2}(a) \cap \mathcal{B}\alpha/2(b) = \varnothing$ by the triangle inequality. Moreover, it holds that for all $a \in \Gamma_\alpha$

$$|\mathcal{B}_{\alpha/2}(a)| \geq 2^{\lfloor \alpha/(2g) \rfloor},$$

since T_θ is binary, and there is no leaf other than a in $\mathcal{B}_{\alpha/2}(a)$. Hence, we must have

$$|\Gamma_\alpha| \leq \frac{2n-2}{2^{\lfloor \alpha/(2g) \rfloor}} \leq \left(\frac{1}{2^{\lfloor \alpha/(2g) \rfloor - 1}}\right) n$$

as there are $2n-2$ nodes in T_θ. Now, for all $a \notin \Gamma_\alpha$ assign an arbitrary leaf at evolutionary distance at most α. Then

$$\begin{aligned} |\Upsilon_{\alpha,\theta}| &\geq \frac{1}{2}(n - |\Gamma_\alpha|) \\ &\geq \frac{1}{2}\left(1 - \frac{1}{2^{\lfloor \alpha/(2g) \rfloor - 1}}\right) n, \end{aligned}$$

where we divided by 2 to avoid double-counting. The result follows from the assumption $\alpha \geq 4g$. $\square$

Let $C_c > 4g$, $C_f > 0$, and $C_{st} > C_f$ to be fixed later.

DEFINITION 3.1 (T_θ-quasicherry). We say that $(a, b) \in [n]^2_{\neq}$ is a *T_θ-quasicherry* if $(a, b) \in \Upsilon_{C_c,\theta}$.

DEFINITION 3.2 (T_θ-stretched). We say that $(a, b) \in [n]^2_{\neq}$ is *T_θ-stretched* if $d_\theta(a, b) \geq C_{st} \log\log n$.

DEFINITION 3.3 (T_θ-far). We say that $c_1 = (a_1, b_1) \neq c_2 = (a_2, b_2) \in [n]^2_{\neq}$ are *T_θ-far* if

$$d_\theta(c_1, c_2) \equiv \min\{d_\theta(x_1, x_2) : x_1 \in \{a_1, b_1\}, x_2 \in \{a_2, b_2\}\} \geq C_f \log\log n.$$

Let Υ' be any subset satisfying

$$\Upsilon_{4g} \subseteq \Upsilon' \subseteq \Upsilon_{C_c} \tag{8}$$

and let

$$\Upsilon'_\theta = \Upsilon' \cap \Upsilon_{C_c,\theta}. \tag{9}$$

Let $C^p_{sp} > 0$ to be fixed later. Keep each $(a, b) \in \Upsilon_{C_c}$ independently with probability

$$p_{sp} = \frac{C^p_{sp} \log n}{n}$$

to form the set Υ''_{C_c}, and let

$$\Upsilon'' = \Upsilon' \cap \Upsilon''_{C_c}.$$

Let $0 < C^-_{sp} < C^+_{sp} < +\infty$ be constants (to be determined).

DEFINITION 3.4 (Properly sparse). A subset $\Upsilon_{4g} \subseteq \Upsilon'' \subseteq \Upsilon_{C_c}$ with

$$\Upsilon''_\theta = \Upsilon'' \cap \Upsilon_{C_c,\theta}, \qquad \theta \in [\Theta],$$

is *properly sparse* if it satisfies the following properties: For all $\theta \in [\Theta]$:

(1) We have $C^-_{sp} \log n \leq |\Upsilon''_\theta| \leq C^+_{sp} \log n$.
(2) All $c_1 = (a_1, b_1) \neq c_2 = (a_2, b_2) \in \Upsilon''$ are T_θ-far.
(3) All pairs in Υ''_θ are $T_{\theta'}$-stretched for $\theta' \neq \theta$.

Let

$$\Upsilon''_{C_c,\theta} = \Upsilon''_{C_c} \cap \Upsilon_{C_c,\theta}, \qquad \theta \in [\Theta],$$

and

$$\Upsilon''_{4g,\theta} = \Upsilon_{4g} \cap \Upsilon''_{C_c,\theta}, \qquad \theta \in [\Theta].$$

LEMMA 3.7 (Sparsification). *There exist constants* $0 < C^-_{sp} < C^+_{sp} < +\infty$ *such that, under* $\mathbb{P}_{\mathcal{A},\lambda}$, *the set* Υ''_{C_c} *as above satisfies the following properties, except with probability* $1/\operatorname{poly}(n)$: *for all* $\theta \in [\Theta]$:

(1) *We have* $C^-_{sp} \log n \le |\Upsilon''_{4g,\theta}|$ *and* $|\Upsilon''_{C_c,\theta}| \le C^+_{sp} \log n$.
(2) *All* $c_1 = (a_1, b_1) \ne c_2 = (a_2, b_2) \in \Upsilon''_{C_c}$ *are* T_θ*-far.*
(3) *All pairs in* $\Upsilon''_{C_c,\theta}$ *are* $T_{\theta'}$*-stretched for* $\theta' \ne \theta$.

In particular, the set Υ'' *as above is properly sparse. Moreover, the claim holds for any* $C^-_{sp} > 0$ *by taking* $C^p_{sp} > 0$ *large enough.*

Intuitively, part (2) follows from the sparsification step whereas part (3) is a consequence of the permutation-invariance of λ. We give a formal proof next.

PROOF OF LEMMA 3.7. For part (1), we use Lemma 3.2. Take

$$\frac{1}{4} C^p_{sp} \log n \le M_{4g} \equiv \frac{C^p_{sp} \log n}{n} |\Upsilon_{4g,\theta}|$$

and

$$M_{C_c} \equiv \frac{C^p_{sp} \log n}{n} |\Upsilon_{C_c,\theta}| \le 2^{\lfloor C_c/f \rfloor} C^p_{sp} \log n.$$

With $\delta_- = 1/2$, $\delta_+ = 5$, we have

$$\mathbb{P}_{\mathcal{A}}[|\Upsilon''_{4g,\theta}| < (1-\delta_-) M_{4g}] < e^{-M_{4g}\delta_-^2/2} = \frac{1}{\operatorname{poly}(n)}$$

and

$$\mathbb{P}_{\mathcal{A}}[|\Upsilon''_{C_c,\theta}| > (1+\delta_+) M_{C_c}] < 2^{-(1+\delta_+)M_{C_c}} = \frac{1}{\operatorname{poly}(n)}.$$

The first part follows from the choice

$$C^-_{sp} = \frac{C^p_{sp}}{8}$$

and

$$C^+_{sp} = 6 C^p_{sp} 2^{\lfloor C_c/f \rfloor}.$$

For the second part, let $c_1 = (a_1, b_1)$ be a pair in Υ''_{C_c}. Let $\mathcal{S}$ be the collection of pairs $c_2 = (a_2, b_2) \neq c_1$ in the original set Υ_{C_c} that are within evolutionary distance $C_f \log\log n$ of c_1 in T_θ, that is,

$$d(c_1, c_2) \leq C_f \log\log n.$$

Note that the number of leaves within evolutionary distance $C_f \log\log n$ from a_1 or b_1 is at most $2 \cdot 2^{\lfloor C_f \log\log n/f \rfloor}$. Moreover, each such leaf can be involved in at most $\Theta 2^{\lfloor C_c/f \rfloor}$ pairs, since any pair in Υ_{C_c} must be a $T_{\theta'}$-quasicherry for some $\theta' \in [\Theta]$ and the number of leaves at evolutionary distance C_c from a vertex in a tree in $\mathbb{Y}_{f,g}$ is at most $2^{\lfloor C_c/f \rfloor}$. Hence

$$|\mathcal{S}| \leq 2 \cdot 2^{\lfloor C_f \log\log n/f \rfloor} \cdot \Theta 2^{\lfloor C_c/f \rfloor} = O(\log n).$$

Therefore the probability that any $c_2 \in \mathcal{S}$ remains in Υ''_{C_c} is at most $O(\log^2 n/n)$. Assuming part (1) holds, summing over Υ''_{C_c}, and applying Markov's inequality, we get

$$\mathbb{P}_{\mathcal{A}}[|c_1 \neq c_2 \in \Upsilon''_{C_c} : c_1, c_2 \text{ are not } T_\theta\text{-far}| \geq 1] = O\left(\frac{\log^3 n}{n}\right) + \frac{1}{\text{poly}(n)}.$$

This gives the second part.

For the third part, consider a T_θ-quasicherry (a, b). Thinking of λ as assigning leaf labels in $T_{\theta'}$ uniformly at random, the probability that b is within evolutionary distance $C_{st} \log\log n$ of a in $T_{\theta'}$ is at most

$$\mathbb{P}_\lambda[(a, b) \text{ is not } T_{\theta'}\text{-stretched}] \leq \frac{2^{\lfloor C_{st} \log\log n/f \rfloor}}{n} = O\left(\frac{\log n}{n}\right),$$

where the numerator in the second expression is an upper bound on the number of vertices at evolutionary distance $C_{st} \log\log n$ of a in $T_{\theta'}$. Summing over all pairs in $\Upsilon''_{C_c,\theta}$ and assuming the bound in part (1) holds, the expected number of pairs in $\Upsilon''_{C_c,\theta}$ that are not $T_{\theta'}$-stretched is $O(\log^2 n/n)$. By Markov's inequality,

$$\mathbb{P}_{\mathcal{A},\lambda}[|\{(a, b) \in \Upsilon''_{C_c,\theta} : (a, b) \text{ is not } T_{\theta'}\text{-stretched}\}| \geq 1] \leq O\left(\frac{\log^2 n}{n}\right) + \frac{1}{\text{poly}(n)}.$$

This gives the third part. □

3.4. *Mixing*. We use a mixing argument similar to [15]. Let

$$Q_{\min} = \min_{x \neq y} Q_{xy},$$

which is positive by assumption. We think of Q as acting as follows. From a state x, we have two type of transitions to $y \neq x$:

(i) We jump to state y at rate $Q_{\min} > 0$.
(ii) We jump to state y at rate $Q_{xy} - Q_{\min} \geq 0$.

Note that a transition of type (i) does not depend on the starting state. Hence if $\mathcal{P}$ is a path from u to v in T_θ, $N=\theta$, and a transition of type (i) occurs along $\mathcal{P}$, then σ_u is *independent* of σ_v. The probability, conditioned on $N=\theta$, that such a transition does not occur, is $e^{-d_\theta(u,v)(r-1)Q_{\min}}$.

Let $\Upsilon'' \subseteq [n]^2_{\neq}$ be a properly sparse set. We show next that pairs in Υ'' are independent with high probability. We proceed by considering the paths joining them and arguing that transitions of type (i) are likely to occur on them by the combinatorial properties in Definition 3.4. Formally, fix $\theta \in [\Theta]$, and consider two pairs $c_1=(a_1,b_1) \neq c_2=(a_2,b_2) \in \Upsilon''$. By Definition 3.4, c_1 and c_2 are T_θ-far. There are three cases without loss of generality:

(1) c_1, c_2 *are* T_θ*-quasicherries.* In the subtree of T_θ connecting $\{a_1,b_1,a_2,b_2\}$, called a *quartet*, the paths $\mathrm{Path}_{T_\theta}(a_1,b_1)$ and $\mathrm{Path}_{T_\theta}(a_2,b_2)$ are disjoint. This is denoted by the *quartet split* $a_1b_1|a_2b_2$. Let $\mathcal{P}^\theta[c_1,c_2]$ be the internal path of the quartet. Note that by Definition 3.4 the length of $\mathcal{P}^\theta[c_1,c_2]$ is at least $C_f \log\log n - 2C_c$. Denote by $\mathcal{P}^\theta_{c_1}[c_1,c_2]$ the subpath of $\mathcal{P}^\theta[c_1,c_2]$ within evolutionary distance $\frac{1}{3}C_f \log\log n$ of c_1.

(2) c_1 *is a* T_θ*-quasicherry, and* c_2 *is* T_θ*-stretched.* Consider the subtree of T_θ connecting $\{a_1,b_1,a_2\}$, called a *triplet*, and let u be the central vertex of it. Let $\mathcal{P}^\theta[c_1,a_2]$ be the path connecting u and a_2. Note that by Definition 3.4, the length of $\mathcal{P}^\theta[c_1,a_2]$ is at least $C_f \log\log n - C_c$. Denote by $\mathcal{P}^\theta_{c_1}[c_1,a_2]$ the subpath of $\mathcal{P}^\theta[c_1,a_2]$ within evolutionary distance $\frac{1}{3}C_f \log\log n$ of c_1. Similarly, denote by $\mathcal{P}^\theta_{a_2}[c_1,a_2]$ the subpath of $\mathcal{P}^\theta[c_1,a_2]$ within evolutionary distance $\frac{1}{3}C_f \log\log n$ of a_2.

(3) c_1, c_2 *are* T_θ*-stretched.* Let $\mathcal{P}^\theta[a_1,a_2]$ be the path connecting a_1 and a_2. Note that by Definition 3.4 the length of $\mathcal{P}^\theta[a_1,a_2]$ is at least $C_f \log\log n$. Denote by $\mathcal{P}^\theta_{a_1}[a_1,a_2]$ the subpath of $\mathcal{P}^\theta[a_1,a_2]$ within evolutionary distance $\frac{1}{3}C_f \log\log n$ of a_1. Similarly, let $\mathcal{P}^\theta[a_1,b_1]$ be the path joining a_1 and b_1, and let $\mathcal{P}^\theta_{a_1}[a_1,b_1]$ be the subpath of $\mathcal{P}^\theta[a_1,b_1]$ within evolutionary distance $\frac{1}{3}C_{st} \log\log n > \frac{1}{3}C_f \log\log n$ of a_1.

Condition on $N=\theta$. For each $c_1=(a_1,b_1) \in \Upsilon''_\theta$, let $\mathcal{E}^\theta_{c_1}$ be the following event:

> Each subpath $\mathcal{P}^\theta_{c_1}[c_1,c_2]$, $c_2 \neq c_1 \in \Upsilon''_\theta$, and each subpath $\mathcal{P}^\theta_{c_1}[c_1,a_2]$, $c_2=(a_2,b_2) \in \Upsilon'' - \Upsilon''_\theta$, undergo a transition of type (i) during the generation of sample σ_X.

Similarly, for each $c_1=(a_1,b_1) \in \Upsilon'' - \Upsilon''_\theta$, let $\mathcal{E}^\theta_{c_1} = \mathcal{E}^\theta_{a_1} \cap \mathcal{E}^\theta_{b_1}$ where $\mathcal{E}^\theta_{a_1}$ is the following event (and similarly for $\mathcal{E}^\theta_{b_1}$):

> Each subpath $\mathcal{P}^\theta_{a_1}[c_2,a_1]$, $c_2 \in \Upsilon''_\theta$, each subpath $\mathcal{P}^\theta_{a_1}[a_1,a_2]$, $c_2=(a_2,b_2) \in \Upsilon'' - \Upsilon''_\theta$ with $c_1 \neq c_2$, as well as subpath $\mathcal{P}^\theta_{a_1}[a_1,b_1]$ undergo a transition of type (i) during the generation of sample σ_X.

Note that, under $\mathcal{E}^\theta_{c_1}$, the random variable $\sigma_{a_1}\sigma_{b_1}$ is independent of every other such random variable in Υ''. Moreover, in the case $c_1 \in \Upsilon'' - \Upsilon''_\theta$, then further

σ_{a_1} is independent of σ_{b_1}. The next lemma shows that most of the events above occur with high probability implying that a large fraction of $\sigma_{a_1}\sigma_{b_1}$'s are mutually independent.

LEMMA 3.8 (Pair independence). *Let $\Upsilon'' \subseteq [n]^2_{\neq}$ be a properly sparse set. Conditioned on $N = \theta$, let*

$$\mathcal{I} = \{c_1 \in \Upsilon'' : \mathcal{E}^{\theta}_{c_1} \text{ holds}\}.$$

For any $0 < \varepsilon_{\mathcal{I}} < 1$ and $C_{\mathcal{I}} > 0$, there exist C_f, $C_{st} > C_f$ and $C^-_{sp} > 0$ large enough so that the following holds except with probability $n^{-C_{\mathcal{I}}}$ under $\mathbb{P}_l$:

$$|\mathcal{I}| \geq (1 - \varepsilon_{\mathcal{I}})|\Upsilon''|.$$

PROOF. Condition on $N = \theta$. Note that the $\mathcal{E}^{\theta}_{c_1}$'s are mutually independent because the corresponding paths are disjoint by construction. By a union bound over Υ'', for all $c_1 \in \Upsilon''$,

$$\begin{aligned}\mathbb{P}_l[(\mathcal{E}^{\theta}_{c_1})^c | N = \theta] &\leq 2C^+_{sp} \log n \cdot e^{-((1/3)C_f \log\log n - 2C_c)(r-1)Q_{\min}} \\ &= \frac{1}{\text{poly}(\log n)}\end{aligned} \tag{10}$$

for C_f large enough. Applying Lemma 3.2 with

$$M = |\Upsilon''| \cdot \mathbb{P}_l[(\mathcal{E}^{\theta}_{c_1})^c | N = \theta]$$

and $\delta_+ > 2e$ such that

$$(1 + \delta_+)M = \varepsilon_{\mathcal{I}}|\Upsilon''| \geq \varepsilon_{\mathcal{I}} C^-_{sp} \log n,$$

we get

$$\mathbb{P}_l[|\Upsilon'' - \mathcal{I}| > \varepsilon_{\mathcal{I}}|\Upsilon''|] \leq 2^{-\varepsilon_{\mathcal{I}}|\Upsilon''|} = \frac{1}{n^{C_{\mathcal{I}}}}$$

by taking C^-_{sp} large enough in Definition 3.4. □

We use the independence claims above to simplify expectation computations.

LEMMA 3.9 (Expectation computations). *Let $\Upsilon'' \subseteq [n]^2_{\neq}$ be a properly sparse set. The following hold. For all $\theta \neq \theta' \in [\Theta]$:*

(1) $\forall (a, b) \in \Upsilon''_{\theta}$,

$$q_{\theta}(a, b) \geq e^{-C_c}.$$

(2) $\forall (a, b) \in \Upsilon'' - \Upsilon''_{\theta}$,

$$q_{\theta}(a, b) = \frac{1}{\text{poly}(\log n)}.$$

(3) $\forall (a,b) \in \Upsilon''_\theta$,

$$q(a,b) = \nu_\theta q_\theta(a,b) + \frac{1}{\mathrm{poly}(\log n)}.$$

(4) $\forall c_1 = (a_1,b_1) \neq c_2 = (a_2,b_2) \in \Upsilon''_\theta$,

$$\begin{aligned} r(c_1,c_2) &= \nu_\theta(1-\nu_\theta) q_\theta(a_1,b_1) q_\theta(a_2,b_2) + \frac{1}{\mathrm{poly}(\log n)} \\ &\geq \frac{1}{2}\underline{\nu}(1-\underline{\nu}) e^{-2C_c} > 0. \end{aligned}$$

(5) $\forall c_1 = (a_1,b_1) \in \Upsilon''_\theta, c_2 = (a_2,b_2) \in \Upsilon''_{\theta'}$,

$$\begin{aligned} r(c_1,c_2) &= -\nu_\theta q_\theta(a_1,b_1) \nu_{\theta'} q_{\theta'}(a_2,b_2) + \frac{1}{\mathrm{poly}(\log n)} \\ &\leq -\frac{1}{2}\underline{\nu} e^{-2C_c} < 0. \end{aligned}$$

PROOF. Parts (1) and (2) follow from the fact that $q_\theta(a,b) = e^{-d_\theta(a,b)}$, $d_\theta(a,b) \leq C_c$ for all $(a,b) \in \Upsilon''_\theta$ and $d_\theta(a,b) \geq C_{st} \log\log n$ for all $(a,b) \in \Upsilon'' - \Upsilon''_\theta$ from Definition 3.4. Part (3) follows from parts (1) and (2).

For part (4), let $c_1 = (a_1,b_1) \neq c_2 = (a_2,b_2) \in \Upsilon''_\theta$. Note that

$$\begin{aligned} \mathbb{E}_l[\sigma_{a_1}\sigma_{b_1}\sigma_{a_2}\sigma_{b_2} | N = \theta, \mathcal{E}^\theta_{c_1}, \mathcal{E}^\theta_{c_2}] &= \mathbb{E}_l[\sigma_{a_1}\sigma_{b_1} | N = \theta] \mathbb{E}_l[\sigma_{a_2}\sigma_{b_2} | N = \theta] \\ &= q_\theta(a_1,b_1) q_\theta(a_2,b_2) \end{aligned}$$

and

$$\begin{aligned} \mathbb{E}_l[\sigma_{a_1}\sigma_{b_1}\sigma_{a_2}\sigma_{b_2} | N = \theta', \mathcal{E}^{\theta'}_{c_1}, \mathcal{E}^{\theta'}_{c_2}] &= \mathbb{E}_l[\sigma_{a_1} | N = \theta'] \mathbb{E}_l[\sigma_{b_1} | N = \theta'] \\ &\quad \times \mathbb{E}_l[\sigma_{a_2} | N = \theta'] \mathbb{E}_l[\sigma_{b_2} | N = \theta'] \\ &= 0 \end{aligned}$$

by (2), so that

$$\mathbb{E}_l[\sigma_{a_1}\sigma_{b_1}\sigma_{a_2}\sigma_{b_2}] = \nu_\theta q_\theta(a_1,b_1) q_\theta(a_2,b_2) + \frac{1}{\mathrm{poly}(\log n)}$$

from (10). Then part (4) follows from Lemma 3.4 and part (3).

For part (5), let $c_1 = (a_1,b_1) \in \Upsilon''_\theta, c_2 = (a_2,b_2) \in \Upsilon''_{\theta'}$. Let $\theta'' \neq \theta, \theta'$. Note that

$$\begin{aligned} \mathbb{E}_l[\sigma_{a_1}\sigma_{b_1}\sigma_{a_2}\sigma_{b_2} | N = \theta, \mathcal{E}^\theta_{c_1}, \mathcal{E}^\theta_{c_2}] &= \mathbb{E}_l[\sigma_{a_1}\sigma_{b_1} | N = \theta] \\ &\quad \times \mathbb{E}_l[\sigma_{a_2} | N = \theta] \mathbb{E}_l[\sigma_{b_2} | N = \theta] \\ &= 0 \end{aligned}$$

and

$$\begin{aligned}\mathbb{E}_l[\sigma_{a_1}\sigma_{b_1}\sigma_{a_2}\sigma_{b_2}|N=\theta',\mathcal{E}_{c_1}^{\theta'},\mathcal{E}_{c_2}^{\theta'}] &= \mathbb{E}_l[\sigma_{a_1}\sigma_{b_1}|N=\theta'] \\ &\quad\times \mathbb{E}_l[\sigma_{a_2}|N=\theta']\mathbb{E}_l[\sigma_{b_2}|N=\theta'] \\ &= 0.\end{aligned}$$

Moreover, since $c_1, c_2 \notin \Upsilon''_{\theta''}$,

$$\begin{aligned}\mathbb{E}_l[\sigma_{a_1}\sigma_{b_1}\sigma_{a_2}\sigma_{b_2}|N=\theta'',\mathcal{E}_{c_1}^{\theta''},\mathcal{E}_{c_2}^{\theta''}] &= \mathbb{E}_l[\sigma_{a_1}|N=\theta'']\mathbb{E}_l[\sigma_{b_1}|N=\theta''] \\ &\quad\times \mathbb{E}_l[\sigma_{a_2}|N=\theta'']\mathbb{E}_l[\sigma_{b_2}|N=\theta''] \\ &= 0.\end{aligned}$$

Hence

$$\mathbb{E}_l[\sigma_{a_1}\sigma_{b_1}\sigma_{a_2}\sigma_{b_2}] = 0 + \frac{1}{\operatorname{poly}(\log n)}$$

from (10). Then part (5) follows from Lemma 3.4 and part (3). □

3.5. *Large-tree concentration.* Let $\Upsilon'' \subseteq [n]^2_{\neq}$ be a properly sparse set. Consider the clustering statistic

$$\mathcal{U}_\theta = \frac{1}{|\Upsilon''_\theta|}\sum_{(a,b)\in\Upsilon''_\theta}\sigma_a\sigma_b.$$

We show that $\mathcal{U}_\theta$ is concentrated and separates the θ-component from all other components.

LEMMA 3.10 (Separation). *There exists $C_\Delta > 0$ such that for $\theta' \neq \theta$*

$$\mathbb{E}_l[\mathcal{U}_\theta|N=\theta] > C_\Delta$$

and

$$\mathbb{E}_l[\mathcal{U}_\theta|N=\theta'] < C_\Delta.$$

PROOF. By Definition 3.4, all $(a,b) \in \Upsilon''_\theta$ are $T_{\theta'}$-stretched. Hence

$$\mathbb{E}_l[\mathcal{U}_\theta|N=\theta] \geq e^{-C_c}$$

and

$$\mathbb{E}_l[\mathcal{U}_\theta|N=\theta'] = \frac{1}{\operatorname{poly}(\log n)}$$

by Lemma 3.9. Taking $C_\Delta = \frac{1}{2}e^{-C_c}$ gives the result. □

LEMMA 3.11 (Concentration of $\mathcal{U}_\theta$). *For all $\varepsilon_{\mathcal{U}} > 0$ and $C_{\mathcal{U}} > 0$, there are $C_f > 0$, $C_{st} > C_f$ and $C_{sp}^- > 0$ large enough such that for all θ, θ' (possibly equal)*

$$\mathbb{P}_l\big[\big|\mathcal{U}_{\theta'} - \mathbb{E}_l[\mathcal{U}_{\theta'}|N=\theta]\big| \geq \varepsilon_{\mathcal{U}}|N=\theta\big] \leq \frac{1}{n^{C_{\mathcal{U}}}}.$$

PROOF. Let $\mathcal{I}$ be as in Lemma 3.8, and let $\mathcal{U}_\theta^{\mathcal{I}}$ be the same as $\mathcal{U}_\theta$ with the sum restricted to $\mathcal{I}$. From Lemmas 3.7 and 3.8, conditioned on $\mathcal{I}$, $\mathcal{U}_\theta^{\mathcal{I}}$ is a normalized sum of $\Theta(\log n)$ independent bounded variables. Concentration of $\mathcal{U}_\theta^{\mathcal{I}}$ therefore follows from Lemma 3.1 using $m = \Omega(\log n)$, $t = O(1/\log n)$ and $\zeta = \frac{1}{2}\varepsilon_{\mathcal{U}}$. Taking $\varepsilon_{\mathcal{I}} = \frac{1}{2}\varepsilon_{\mathcal{U}} \max_i z_i^2$ and $C_{\mathcal{I}} > C_{\mathcal{U}}$ in Lemma 3.8 as well as $C_{sp}^- > 0$ large enough gives the result. □

4. Constructing the clustering statistic from data. In this section, we provide details on the plan laid out in Section 2.3.

Fix a GTR matrix Q and constants $\Theta \geq 2$, $0 < f \leq g < +\infty$ and $\underline{\nu} > 0$. Let λ be a permutation-invariant probability measure on Θ-M$[f, g, \underline{\nu}, n]$. In this section, we work directly with samples $\{\sigma_X^i\}_{i=1}^k$ generated from an *unknown* Θ-mixture model $(\mathbb{T}, \nu, Q)$ picked according to λ.

Our goal is to construct the clustering statistics $\{\mathcal{U}_\theta\}_{\theta=1}^\Theta$ from $\{\sigma_X^i\}_{i=1}^k$. These statistics will be used in the next section to reconstruct the topologies of the model $(\mathbb{T}, \nu, Q)$.

4.1. *Clustering algorithm.* We proceed in three steps. Let

$$C_c = -\ln\left(\frac{1}{3\Theta(1-\underline{\nu})}\underline{\nu}e^{-4g}\right)$$

and

$$\omega = \tfrac{2}{3}\underline{\nu}e^{-4g}.$$

The algorithm is the following:

(1) (*Finding quasicherries*) For all pairs of leaves $a, b \in [n]$, compute $\hat{q}(a, b)$, and set

$$\hat{\Upsilon}' = \{(a, b) \in [n]_{\neq}^2 : \hat{q}(a, b) \geq \omega\}.$$

(2) (*Sparsification*) Construct $\hat{\Upsilon}''$ by keeping each $(a, b) \in \hat{\Upsilon}'$ independently with probability

$$p_{sp} = \frac{C_{sp}^p \log n}{n}.$$

(3) (*Inferring clusters*) For all $c_1 \neq c_2 \in \hat{\Upsilon}'$, compute $\hat{r}(c_1, c_2)$, and set $c_1 \sim c_2$ if

$$\hat{r}(c_1, c_2) > 0.$$

Let $\hat{\Upsilon}''_\theta$, $\theta = 1, \ldots, \hat{\Theta}$, be the equivalence classes of the transitive closure of $\sim$.

(4) (*Final sets*) Return $\hat{\Upsilon}''_\theta$, $\theta \in [\Theta]$.

4.2. *Analysis of the clustering algorithm.* We show that each step of the previous algorithm succeeds with high probability.

LEMMA 4.1 (Finding quasicherries). *The set* $\hat{\Upsilon}'$ *satisfies the following, except with probability at most* $\exp(-\text{poly}(n))$ *under* $\mathbb{P}_l$:

$$\Upsilon_{4g} \subseteq \hat{\Upsilon}' \subseteq \Upsilon_{C_c}.$$

PROOF. We prove both inclusions. For all $\theta \in [\Theta]$ and $(a, b) \in \Upsilon_{4g,\theta}$,

$$q_\theta(a, b) \geq e^{-4g}$$

and

$$q(a, b) \geq \underline{\nu} e^{-4g} > \tfrac{2}{3}\underline{\nu} e^{-4g} = \omega.$$

By Lemma 3.4,

$$\hat{q}(a, b) \geq \omega,$$

except with probability $\exp(-\text{poly}(n))$.

Similarly for any $(a, b) \in \hat{\Upsilon}'$, by Lemma 3.4, if

$$\hat{q}(a, b) \geq \omega = \tfrac{2}{3}\underline{\nu} e^{-4g},$$

then

$$q(a, b) \geq \tfrac{1}{3}\underline{\nu} e^{-4g},$$

so that there is $\theta \in [\Theta]$ with

$$\nu_\theta q_\theta(a, b) \geq \frac{1}{3\Theta}\underline{\nu} e^{-4g}.$$

That is,

$$q_\theta(a, b) \geq \frac{1}{3\Theta(1 - \underline{\nu})}\underline{\nu} e^{-4g}$$

and

$$d_\theta(a, b) \leq -\ln\left(\frac{1}{3\Theta(1 - \underline{\nu})}\underline{\nu} e^{-4g}\right) = C_c.$$

Hence $(a, b) \in \Upsilon_{C_c,\theta}$. $\square$

LEMMA 4.2 (Sparsification). *Assuming that the conclusions of Lemma 4.1 hold, $\hat{\Upsilon}''$ is properly sparse, except with probability* $1/\operatorname{poly}(n)$.

PROOF. This follows from Lemma 4.1 and the choice of p_{sp}. □

LEMMA 4.3 (Inferring clusters). *Assuming that the conclusions of Lemmas 4.1 and 4.2 hold, we have $\hat{\Theta} = \Theta$, and there is a bijective mapping h of $[\Theta]$ such that*

$$\hat{\Upsilon}''_{h(\theta)} = \Upsilon''_\theta$$

with the choice $\Upsilon' = \hat{\Upsilon}'$ in Section 3.3, except with probability $\exp(-\operatorname{poly}(n))$.

PROOF. It follows from Lemmas 3.5 and 3.9 that $\sim$ is an equivalence relation with equivalence classes Υ''_θ, $\theta = 1, \ldots, \Theta$, except with probability $\exp(-\operatorname{poly}(n))$. □

5. Tree reconstruction. We now show how to use the clustering statistics to build the topologies. The algorithm is composed of two steps: we first bin the sites according to the value of the clustering statistics; we then use the sites in one of those bins and apply a standard distance-based reconstruction method. We show that the content of the bins is made of sites from the same component—thus reducing the situation to the unmixed case.

Let

$$C_\Delta = \tfrac{1}{2} e^{-C_c},$$
$$\varepsilon_{\mathcal{U}} = \tfrac{1}{3} e^{-C_c}$$

and

$$\varepsilon_{\mathcal{I}} = \frac{1}{2} \varepsilon_{\mathcal{U}} \max_i z_i^2.$$

Moreover take C_f, C_{st}, C_{sp}^p and C_{sp}^- so that the lemmas in Section 3 hold.

To simplify notation, we rename the components so that h is the *identity*.

5.1. *Site binning.* Let $\hat{\Upsilon}''_\theta$, $\theta \in [\Theta]$, be the sets returned by the algorithm in Section 4. Assume that the conclusions of Lemmas 4.1, 4.2 and 4.3 hold. We bin the sites with the following procedure:

(1) (*Clustering statistics*) For all $i = 1, \ldots, k$ and all $\theta = 1, \ldots, \Theta$, compute

$$\hat{\mathcal{U}}_\theta^i = \frac{1}{|\hat{\Upsilon}''_\theta|} \sum_{(a,b) \in \hat{\Upsilon}''_\theta} \sigma_a^i \sigma_b^i.$$

(2) (*Binning sites*) For all $\theta = 1, \ldots, \Theta$, set

$$\hat{\mathcal{K}}_\theta = \{i \in [k] : \hat{\mathcal{U}}^i_\theta > C_\Delta\}.$$

We show that the binning is successful with high probability.

LEMMA 5.1 (Binning the sites). *Assume that the conclusions of Lemmas 4.1, 4.2 and 4.3 hold. For any C_k, there exists $C_{\mathcal{U}}$ large enough so that, for all $\theta \in [\Theta]$,*

$$\hat{\mathcal{K}}_\theta = \mathcal{K}_\theta,$$

except with probability $1/\operatorname{poly}(n)$.

PROOF. This follows from Lemmas 3.10 and 3.11 by a union bound over all samples. □

5.2. *Estimating a distorted metric.*

Estimating evolutionary distances. We estimate evolutionary distances on each component. For all $\theta \in [\Theta]$, let $\hat{\mathcal{K}}_\theta$ be as above and assume the conclusions of Lemma 5.1 hold.

(1) (*Estimating distances*) For all $\theta = 1, \ldots, \Theta$ and $a \neq b \in [n]$, compute

$$\hat{q}_\theta(a, b) = \frac{1}{|\hat{\mathcal{K}}_\theta|} \sum_{i \in \hat{\mathcal{K}}_\theta} \sigma^i_a \sigma^i_b.$$

LEMMA 5.2 (Estimating distances). *Assume the conclusions of Lemma 5.1 hold. The following hold except with probability* $\exp(-\operatorname{poly}(n))$: *for all $\theta \in [\Theta]$ and all $a \neq b \in [n]$,*

$$|\hat{q}_\theta(a, b) - q_\theta(a, b)| \leq \frac{1}{n^{C_q}}.$$

PROOF. The result follows from Lemma 3.4. □

Tree construction. To reconstruct the tree, we use a distance-based method of [8]. We require the following definition.

DEFINITION 5.1 (Distorted metric [12, 16]). Let $T = (V, E; \phi; w)$ be a phylogeny with corresponding tree metric d, and let $\tau, \Psi > 0$. We say that $\hat{d} : X \times X \to (0, +\infty]$ is a (τ, Ψ)-*distorted metric* for T or a (τ, Ψ)-*distortion* of d if:

(1) (*Symmetry*) For all $a, b \in X$, $\hat{d}$ is symmetric, that is,

$$\hat{d}(a, b) = \hat{d}(b, a);$$

(2) (*Distortion*) $\hat{d}$ is accurate on "short" distances; that is, for all $a, b \in X$, if either $d(a,b) < \Psi + \tau$ or $\hat{d}(a,b) < \Psi + \tau$, then

$$|d(a,b) - \hat{d}(a,b)| < \tau.$$

An immediate consequence of [8], Theorem 1, is the following.

CLAIM (Reconstruction from distorted metrics [8]). *Let $T = (V, E; \phi; w)$ be a phylogeny in $\mathbb{Y}_{f,g}$. Then the topology of T can be recovered in polynomial time from a (τ, Ψ)-distortion $\hat{d}$ of d as long as*

$$\tau \leq \frac{f}{5}$$

and

$$\Psi \geq 5g \log n.$$

REMARK 5.1. The constants above are not optimal but will suffice for our purposes.

See [8] for the details of the reconstruction algorithm.

We now show how to obtain a $(f/5, 5g \log n)$-distortion with high probability for each component.

LEMMA 5.3 (Distortion estimation). *There exist $C_q, C_k > 0$ so that, given that the conclusions of Lemma 5.2 hold, for all $\theta \in [\Theta]$,*

$$\hat{d}_\theta(a,b) = -\ln(\hat{q}_\theta(a,b)_+), \qquad (a,b) \in X \times X,$$

is a $(f/5, 5g \log n)$-distortion of d_θ.

PROOF. Fix $\theta \in [\Theta]$. Define

$$Ł_2^- = \{(a,b) \in X \times X : d_\theta(a,b) \leq 15g \log n\}$$

and

$$Ł_2^+ = \{(a,b) \in X \times X : d_\theta(a,b) > 12g \log n\}.$$

Let $(a,b) \in Ł_2^-$. Note that

$$e^{-d_\theta(a,b)} \geq \exp(-15g \log n) \equiv \frac{1}{n^{C_q'}},$$

where the last equality is a definition. Then, taking C_q (and hence C_k) large enough, from Lemma 5.2, we have

$$|\hat{d}_\theta(a,b) - d_\theta(a,b)| \leq \frac{f}{5}.$$

Similarly, let $(a, b) \in Ł_2^+$. Note that

$$e^{-d_\theta(a,b)} < \exp(-12g \log n) \equiv \frac{1}{n^{C_q''}},$$

where the last equality is a definition. Then, taking C_q large enough, from Lemma 5.2 we have

$$\hat{d}_\theta(a, b) \geq 5g \log n + \frac{f}{5}. \qquad \square$$

6. Proof of main theorems. We are now ready to prove the main theorems.

PROOF OF THEOREM 3. Let $C_1, C_2 > 0$. Let A_n be the subset of those Θ-mixture models $(\mathbb{T}, \nu, Q)$ in Θ-M$[f, g, \underline{\nu}, n]$ for which part (3) of Lemma 3.7 holds with probability at least $1 - n^{-C_1}$ under the random choices of the algorithm. By the proof of Lemma 3.7, for small enough $C_1, C_2 > 0$, we have $\lambda_n[A_n^c] \leq n^{-C_2}$. On A_n, the lemmas in Sections 3, 4 and 5 hold with probability $1 - 1/\operatorname{poly}(n)$. Then the topologies are correctly reconstructed by the claim in Section 5.2. $\square$

PROOF OF THEOREM 1. Let

$$(\mathbb{T}, \nu, Q) \nsim (\mathbb{T}', \nu', Q) \in \bigcup_{n \geq 1} A_n.$$

Then, by Theorem 3, the algorithm correctly reconstructs the topologies in $(\mathbb{T}, \nu, Q)$ with probability $1 - 1/\operatorname{poly}(n)$ on sequences of length $k = \operatorname{poly}(n)$. Repeating the reconstruction on independent sequences and taking a majority vote, we get almost sure convergence to the correct topologies. The same holds for $(\mathbb{T}', \nu', Q)$. Hence,

$$\mathcal{D}_l[(\mathbb{T}, \nu, Q)] \neq \mathcal{D}_l[(\mathbb{T}', \nu', Q)]. \qquad \square$$

PROOF OF THEOREM 2. Let

$$(\mathbb{T}, \nu, Q) \in \bigcup_{n \geq 1} A_n$$

with $\Theta = 2$ and $\nu = (1/2, 1/2)$. Then, from the proof of Lemma 5.1, there exists a clustering statistic such that samples from T_1 and T_2 are correctly distinguished with probability $1 - 1/\operatorname{poly}(n)$. Recall that

$$\|\mathcal{D} - \mathcal{D}'\|_{\mathrm{TV}} = \sup_{B \in \mathcal{F}} |\mathcal{D}(B) - \mathcal{D}'(B)|.$$

Taking B to be the event that a site is recognized as belonging to component 1 by the clustering statistic above, we get

$$\|\mathcal{D}_l[T_1, Q] - \mathcal{D}_l[T_2, Q]\|_{\mathrm{TV}} = 1 - o_n(1). \qquad \square$$

7. Concluding remarks. Our techniques also admit the following extensions:

- When Q is unknown, one can still apply our technique by using the following idea. Note that all we need is an eigenvector of Q with negative eigenvalue. Choose a pair (a, b) of close leaves using, for instance, the classical log-det distance [22]. Under a permutation-invariant measure, (a, b) is stretched in all but one component, with high probability. One can then compute an eigenvector decomposition of the transition matrix between a and b. We leave out the details.
- The minimum frequency assumption is not necessary as long as one has an upper bound on the number of components and that one requires only that frequent enough components be detected and reconstructed. We leave out the details.

REFERENCES

[1] ALLMAN, E. S., ANÉ, C. and RHODES, J. A. (2008). Identifiability of a Markovian model of molecular evolution with gamma-distributed rates. *Adv. in Appl. Probab.* **40** 229–249. MR2411822

[2] ALLMAN, E. S., PETROVIC, S., RHODES, J. A. and SULLIVANT, S. (2011). Identifiability of two-tree mixtures for group-based models. *IEEE/ACM Trans. Comput. Biology Bioinform.* **8** 710–722.

[3] ALLMAN, E. S. and RHODES, J. A. (2006). The identifiability of tree topology for phylogenetic models, including covarion and mixture models. *J. Comput. Biol.* **13** 1101–1113 (electronic). MR2255411

[4] BILLERA, L. J., HOLMES, S. P. and VOGTMANN, K. (2001). Geometry of the space of phylogenetic trees. *Adv. in Appl. Math.* **27** 733–767. MR1867931

[5] CHAI, J. and HOUSWORTH, E. A. (2011). On Rogers' proof of identifiability for the GTR + Gamma + I model. Available at http://sysbio.oxfordjournals.org/content/early/2011/03/27/sysbio.syr023.short.

[6] CHANG, J. T. (1996). Full reconstruction of Markov models on evolutionary trees: Identifiability and consistency. *Math. Biosci.* **137** 51–73. MR1410044

[7] CHOR, B. and TULLER, T. (2006). Finding a maximum likelihood tree is hard. *J. ACM* **53** 722–744 (electronic). MR2263067

[8] DASKALAKIS, C., MOSSEL, E. and ROCH, S. (2009). Phylogenies without branch bounds: Contracting the short, pruning the deep. In *RECOMB* (S. Batzoglou, ed.). *Lecture Notes in Computer Science* **5541** 451–465. Springer, New York.

[9] EVANS, S. N. and WARNOW, T. (2004). Unidentifiable divergence times in rates-across-sites models. *IEEE/ACM Trans. Comput. Biology Bioinform.* **1** 130–134.

[10] FELSENSTEIN, J. (2004). *Inferring Phylogenies*. Sinauer, Sunderland, MA.

[11] HUELSENBECK, J. P. and RANNALA, B. (1997). Phylogenetic methods come of age: Testing hypotheses in an evolutionary context. *Science* **276** 227–232.

[12] KING, V., ZHANG, L. and ZHOU, Y. (2003). On the complexity of distance-based evolutionary tree reconstruction. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms* (*Baltimore*, *MD*, 2003) 444–453. ACM, New York. MR1974948

[13] MATSEN, F. A., MOSSEL, E. and STEEL, M. (2008). Mixed-up trees: The structure of phylogenetic mixtures. *Bull. Math. Biol.* **70** 1115–1139. MR2391182

[14] MATSEN, F. A. and STEEL, M. (2007). Phylogenetic mixtures on a single tree can mimic a tree of another topology. *Syst. Biol.* **56** 767–775.

[15] MOSSEL, E. (2003). On the impossibility of reconstructing ancestral data and phylogenies. *J. Comput. Biol.* **10** 669–678.

[16] MOSSEL, E. (2007). Distorted metrics on trees and phylogenetic forests. *IEEE/ACM Trans. Comput. Biology Bioinform.* **4** 108–116.
[17] MOSSEL, E. and PERES, Y. (2003). Information flow on trees. *Ann. Appl. Probab.* **13** 817–844. MR1994038
[18] MOSSEL, E. and ROCH, S. (2011). Identifiability and inference of non parametric rates-across-sites models on large-scale phylogenies. Preprint.
[19] MOTWANI, R. and RAGHAVAN, P. (1995). *Randomized Algorithms*. Cambridge Univ. Press, Cambridge. MR1344451
[20] RHODES, J. and SULLIVANT, S. (2010). Identifiability of large phylogenetic mixture models. Preprint.
[21] ROCH, S. (2006). A short proof that phylogenetic tree reconstruction by maximum likelihood is hard. *IEEE/ACM Trans. Comput. Biology Bioinform.* **3** 92–94.
[22] SEMPLE, C. and STEEL, M. (2003). *Phylogenetics*. *Oxford Lecture Series in Mathematics and Its Applications* **24**. Oxford Univ. Press, Oxford. MR2060009
[23] STEEL, M. (2009). A basic limitation on inferring phylogenies by pairwise sequence comparisons. *J. Theoret. Biol.* **256** 467–472.
[24] STEEL, M., SZÉKELY, L. A. and HENDY, M. D. (1994). Reconstructing trees when sequence sites evolve at variable rates. *J. Comput. Biol.* **1** 153–163.
[25] STEEL, M. A. and SZÉKELY, L. A. (2006). On the variational distance of two trees. *Ann. Appl. Probab.* **16** 1563–1575. MR2260073
[26] ŠTEFANKOVIČ, D. and VIGODA, E. (2007). Phylogeny of mixture models: Robustness of maximum likelihood and non-identifiable distributions. *J. Comput. Biol.* **14** 156–189 (electronic). MR2299868
[27] STEFANKOVIC, D. and VIGODA, E. (2007). Pitfalls of heterogeneous processes for phylogenetic reconstruction. *Syst. Biol.* **56** 113–124.
[28] WU, J. and SUSKO, E. (2010). Rate-variation need not defeat phylogenetic inference through pairwise sequence comparisons. *J. Theoret. Biol.* **263** 587–589.

DEPARTMENTS OF STATISTICS
AND COMPUTER SCIENCE
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720
USA
E-MAIL: mossel@stat.berkeley.edu

DEPARTMENT OF MATHEMATICS
AND BIOINFORMATICS PROGRAM
UNIVERSITY OF CALIFORNIA
LOS ANGELES, CALIFORNIA 90095
USA
E-MAIL: roch@math.ucla.edu