SIAM J. DISCRETE MATH.
Vol. 28, No. 3, pp. 1493–1507

ON THE COMPATIBILITY OF QUARTET TREES*

NOGA ALON†, SAGI SNIR‡, AND RAPHAEL YUSTER§

Abstract. Phylogenetic tree reconstruction is a fundamental biological problem. Quartet trees, trees over four species, are the minimal informational unit for phylogenetic classification. While every phylogenetic tree over n species defines $\binom{n}{4}$ quartets, not every set of quartets is compatible with some phylogenetic tree. Here we focus on the compatibility of quartet sets. We provide several results addressing the question of what can be inferred about the compatibility of a set from its subsets. Most of our results use probabilistic arguments to prove the sought characteristics. In particular we show that there are quartet sets Q of size $m = cn \log n$ in which every subset of cardinality $c'n/\log n$ is compatible, and yet no fraction of more than $1/3 + \epsilon$ of Q is compatible. On the other hand, in contrast to the classical result stating when Q is the densest, i.e., $m = \binom{n}{4}$ and the compatibility of any set of three quartets implies full compatibility, we show that even for $m = \Theta\left(\binom{n}{4}\right)$ there are (very) incompatible sets for which every subset of large constant cardinality is compatible. Our final result relates to the conjecture of Bandelt and Dress regarding the maximum quartet distance between trees. We provide asymptotic upper and lower bounds for this value.

Key words. phylogenetic reconstruction, tree compatibility, quartet amalgamation, quartet fit

AMS subject classification. 92B05

DOI. 10.1137/130941043

1. Introduction.

The study of evolution and the construction of phylogenetic (evolutionary) trees are classical subjects in biology. A phylogeny, the evolutionary history of a set of species, is normally represented by a tree where the species under study are mapped to the leaves of the tree and the tree structure represents evolutionary relationships. Frequently, it is desirable to combine several trees over overlapping sets of species. This task is called the *supertree* task [6, 5], where the goal is to find a tree over the full set of species that satisfies most of the partial input trees. We distinguish between *rooted* and *unrooted* settings. In the rooted setting a rooted triplet tree (Figure 1(a)) is the basic unit of information. We denote the triplet over leaves $\{a, b, c\}$ as $[ab|c]$ if the lowest common ancestor (LCA) for a, b is a (proper) descendant of the LCA for all $\{a, b, c\}$. However, the more prevailing setting is the unrooted setting. In the unrooted setting, the notion of LCA is meaningless, and therefore the basic unit of information is a quartet tree (Figure 1(b)). We denote a quartet over leaves $\{a, b, c, d\}$ as $[ab|cd]$, meaning that there is an edge in the underlying tree separating a and b from c and d.

For its fundamental role in phylogenetics, a vast research effort has been devoted to phylogenetic tree reconstruction from quartets, a field denoted as *quartet-based reconstruction* (see, e.g., [3, 4, 23, 27, 34, 31, 36], among many others). Here, accurate

*Received by the editors October 14, 2013; accepted for publication (in revised form) May 19, 2014; published electronically September 25, 2014. A preliminary version of this paper appeared in *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms*, 2014.
http://www.siam.org/journals/sidma/28-3/94104.html

†Tel Aviv University, Tel Aviv 69978, Israel, and Institute for Advanced Study, Princeton, NJ 08540 (nogaa@tau.ac.il). This author's research was supported in part by an ERC Advanced grant, a USA-Israeli BSF grant, an ISF grant, the Israeli I-Core program, and the Simonyi Fund.

‡Department of Evolutionary Biology, University of Haifa, Haifa 31905, Israel (ssagi@research.haifa.ac.il). This author's research was supported in part by the Israeli ISF and the USA-Israel BSF grants.

§Department of Mathematics, University of Haifa, Haifa 31905, Israel (raphy@math.haifa.ac.il).

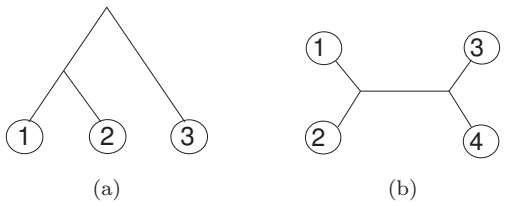


FIG. 1. *In a* rooted *setting, a rooted triplet is the smallest informative tree versus an unrooted quartet in the* unrooted *setting.* (a) *A rooted triplet tree* [12|3]. (b) *An unrooted quartet tree* [12|34].

trees over four species are constructed, normally using some *quartet oracle* [15, 16, 20, 23], from a (possibly partial) set of 4-sets. These trees are subsequently combined together into the big tree on the full species set. The latter problem, however, deciding whether there exists a tree satisfying all the quartets in an arbitrary given set, is NP-complete [35]. Moreover, the ideal case in which all quartets agree on a single tree is very rare. This raises the problem of finding a tree maximizing the number of compatible quartets: *maximum quartet compatibility* (or *consistency*) (MQC) [29]. MQC is obviously NP-hard, and therefore several approximation algorithms have been suggested [4, 8, 26, 32, 33], but the best approximation to the general problem is still obtained by a naive "random tree" with an expected approximation ratio of 1/3 [33].

A possible direction to follow in this respect is to check compatibility of relatively small subsets of the ground set as "witnesses" for compatibility, under the trivial observation that if the ground set is compatible (agreeing on some tree), then necessarily every subset of it is compatible. A related approach was also studied in this context [12], motivated by the investigation of property testers for quartet compatibility. Of course the number of such subsets may be exponentially large, but one might suspect that a sampling approach can lead to a randomized algorithm for set compatibility. In this paper we address the question of how much we can infer about a quartet set just by examining its constituting subsets. We examine this question in several settings, all of which are abstract and asymptotically large. We use probabilistic arguments to base our claims, allowing us to show large incompatibility in (relatively sparse) sets of quartets, despite the compatibility of all their (rather large) subsets. One consequence of the results here is that no sampling algorithm, even one based on $m/polylog(m)$ quartets among m given ones, can be used to determine the compatibility of the given quartet set or estimate the maximum cardinality of a compatible subset of it.

More concretely, one might expect that the larger the subsets with respect to the ground quartet set, the more information they endow us about the set. However, our first result shows, perhaps surprisingly, that even when the subsets are only a small polylogarithmic factor smaller than the ground set, the compatibility of all subsets of this size cannot guarantee even relative compatibility (compatibility of a large fraction) of the ground set. We note that previous works give constructions of incompatible sets of $\Theta(n)$ quartets [35] and recently even $\Theta(n^2)$ quartets [30] for arbitrary n, for which every proper subset is compatible. However, the sets we analyze here are inherently incompatible in the sense that a constant fraction of the elements (specifically 2/3) should be removed in order to achieve compatibility. In terms of property testing, our results imply that any (one-sided, nonadaptive) tester for the property of quartet compatibility for a given subset of quartets of an n-element set must perform at least $\tilde{\Omega}(n)$ queries. See [12] for an upper bound for the number of

queries when the input consists of all quartets.

At the other extreme, however, a classical result [13] shows that when the ground set is the densest, that is, the full $\binom{n}{4}$ quartets on n species, compatibility of every subset of at least three quartets implies set compatibility. There is a big gap between these two results. In the first result, even a polylog factor between the sizes of the ground set and the subsets cannot guarantee compatibility of the set, whereas in the second example, even a constant size, in fact the minimal meaningful set size, is enough to warrant compatibility. This raises the natural question of whether it is the density of the ground set that determines whether subset compatibility implies set compatibility.

Our next result tackles this latter question. Specifically, we show that for any $0 < \varepsilon < 1$ there is a quartet set of size $\varepsilon\binom{n}{4}$ that is (very) incompatible, yet every subset of it of size $\tilde{\Theta}\left((1/\varepsilon)^{1/3}\right)$ is compatible. This means that for every given (arbitrarily large) constant C, we can take a quartet set on n species which is dense (namely, of constant density) that is very incompatible, and yet every subset of size C is compatible. We use a blowup argument for this result, where a constant size set is expanded to arbitrary, asymptotic size. The latter is also related to a result of [25] showing an example of a constant size incompatible quartet set, all of whose subsets are compatible (we note that the subsets in [25] satisfy a much stronger property of closeness that we describe below). We can use the same blowup technique we use here on that example, to get a big incompatible set $\left($of magnitude a constant fraction of $\binom{n}{4}\right)$, but the size of the compatible subsets this way will only be limited to 5.

Compatibility is inherently linked to the subject of closure operations on a quartet set [7, 10, 18, 19, 25]. These operations rely on inference rules allowing a recursive augmentation of a set of quartets with quartets that are implied by the existing set. A quartet set Q is closed when it contains all the quartets induced by all the trees satisfying Q. Part of our work is also related and relies on these inference rules, as discussed in section 6.

Our last result touches compatibility from a different angle. We tackle the question of how much a compatible set can be violated. In particular we are interested in answering what fraction of the largest, unambiguous, compatible set—the full set of $\binom{n}{4}$ compatible quartets—can be violated. This question has significance when measuring similarity between trees based on quartet similarity, *qFit* [22]. It was presented by Bandelt and Dress [2], where they conjectured that this value tends to 2/3. We give an upper and lower bound on the fraction of violated quartets of the (compatible) quartet set of any tree T. For every fixed n, our lower bound is strictly greater than 2/3, but as n tends to infinity, it tends to 2/3.

We end the paper with open problems and further research questions arising as a result of this work.

2. Preliminaries. A binary rooted tree is a tree whose edges are directed, every internal vertex has two children, and every vertex but one distinguished vertex, the *root*, has a single ancestor. In a binary undirected (also called unrooted) tree, edges are undirected (and hence parenthood does not exists), and all internal vertices have three neighbors each. Such a tree is also called a trivalent tree. Throughout this paper, unless stated otherwise, all trees are assumed to be unrooted binary trees, with leaves labeled bijectively by a taxa (species) set $\mathcal{X}$ of size n. For brevity, we will refer to such trees as *phylogenetic trees*, or more briefly, as just *trees*. For a tree $T = (V, E)$, the set of leaves of T is denoted by $\mathcal{L}(T)$.

The removal of an edge e in a tree splits the tree into two subtrees and therefore

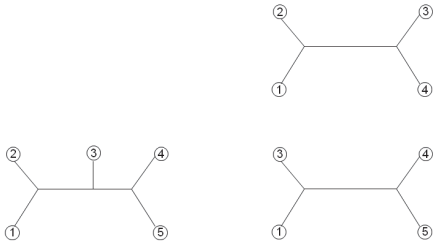


FIG. 2. (a) *A phylogenetic tree over five leaves.* (b) *Two quartets compatible with the tree on the left.*

induces a *split* among the leaves of the tree. We identify an edge e by the split $(U, \mathcal{L}(T) \setminus U)$ it generates on the set of leaves and denote the split by U_e. As external edges (edges adjacent to leaves) induce trivial splits, we refer only to internal edges. Let T be a tree and $A \subseteq \mathcal{L}(T)$ be a subset of the leaves of T. We denote by $T|_A$ the topological subtree of T induced by A where all leaves in $\mathcal{L}(T) \setminus A$ and paths leading exclusively to them are removed, and subsequently internal vertices with degree two are contracted.

For two trees T and T', we say that T *satisfies* T' (or, equivalently, that T' is *satisfied* by T) if $\mathcal{L}(T') \subseteq \mathcal{L}(T)$ and $T|_{\mathcal{L}(T')} = T'$. Otherwise, T' is *violated* by T. Let $\mathcal{T} = \{T_1, \ldots, T_k\}$ be a set of trees with possibly overlapping leaves, and denote by $\mathcal{L}(\mathcal{T}) = \bigcup_i \mathcal{L}(T_i)$ the union of the set of leaves of all trees $T_i \in \mathcal{T}$. Then, for a tree T with $\mathcal{L}(T) = \mathcal{L}(\mathcal{T})$, we denote by $\mathcal{T}_s(T)$ the set of trees in $\mathcal{T}$ that are satisfied by T. We say that $\mathcal{T}$ is *compatible* (or *consistent*) if there exists a tree T^* over the set of leaves $\mathcal{L}(\mathcal{T})$ that satisfies every tree $T_i \in \mathcal{T}$, i.e., $\mathcal{T}_s(T^*) = \mathcal{T}$ (see Figure 2). We denote by $co(\mathcal{T})$ the set of trees that satisfy $\mathcal{T}$, $co(\mathcal{T}) = \{T : \mathcal{T}_s(T) = \mathcal{T}\}$.

Further, we say that T^* is *defined* by $\mathcal{T}$ if $co(\mathcal{T})$ is the singleton $\{T^*\}$. If there is no such compatible tree T^* (i.e., $co(\mathcal{T}) = \emptyset$), we say that $\mathcal{T}$ is *incompatible* (or *inconsistent*).

A *quartet* tree (or just a quartet for short) is a phylogenetic tree over four leaves $\{a, b, c, d\}$. We denote a quartet over $\{a, b, c, d\}$ as $[ab|cd]$ if there exists an edge e whose corresponding split U_e satisfies $a, b \in U$ and $c, d \notin U$. Quartets are the most elementary informational unit in an unrooted phylogenetic tree, since trees on quartets of leaves (in contrast to pairs or triplets of leaves) can violate other phylogenetic trees. Every phylogenetic tree T with n leaves uniquely defines $\binom{n}{4}$ quartets, one for each set of four leaves. Let $\mathcal{Q}(T)$ denote this full quartet set of T. Obviously, $\mathcal{Q}(T)$ uniquely defines T.

We mention two basic problems and results regarding quartet-based reconstruction. When a set of trees $\mathcal{T}$ is incompatible, it is desirable to find a tree T^* over $\bigcup_i \mathcal{L}(T_i)$ that maximizes some objective function. This meta-problem is commonly known as the *supertree* problem, and a corresponding solution T^* is a *supertree*. An important case of the supertree problem is when the set of input trees is a set of quartet trees $\mathcal{Q}$ and the task is to find a tree T such that $|\mathcal{Q}_s(T)|$, the subset of $\mathcal{Q}$ satisfied by T (see more precisely below), is maximized. This problem is called *maximum quartet compatibility* (MQC).

PROBLEM 2.1 (MQC). *Given a set of quartets $\mathcal{Q}$ with leaves labeled by a taxa set $\mathcal{X}$, find a tree T with leaves labeled bijectively by $\mathcal{X}$ that satisfies the maximum*

number of elements of $\mathcal{Q}$.

We note that MQC is NP-hard. In fact, the decision problem of whether $\mathcal{Q}$ is compatible is NP-complete [35].

Let T be any tree with n leaves labeled by a taxa set $\mathcal{X}$. Consider a random bijection π between a taxa set $\mathcal{X}$ of size n and the leaves of T. The corresponding labeled tree is denoted by T^π. As each of the $n!$ possible bijections is equally likely, we notice that a quartet $[ab|cd]$ with labels from $\mathcal{X}$ is satisfied by T^π with probability $1/3$. We therefore have, by linearity of expectation, the following lemma.

LEMMA 2.2. *Let* $\mathcal{Q}$ *be an arbitrary set of quartets over a taxa set* $\mathcal{X}$ *of size* n, *and let* T^π *be a random bijection between the leaves of a tree* T *and* $\mathcal{X}$. *Then the expected number of elements in* $\mathcal{Q}$ *satisfied by* T *is* $|\mathcal{Q}|/3$.

Lemma 2.2 immediately yields an efficient (randomized) $1/3$ approximation algorithm for MQC. This naive solution is presently the best known approximation ratio for general instances of MQC that can be achieved by a polynomial time algorithm.

3. Subset compatibility in very incompatible sets. Our first result shows that even when all relatively large subsets of a given set are compatible, the whole set can still be very far from compatible.

THEOREM 3.1. *There exists a set* Q *of* $\Theta(n \log n)$ *quartets such that every subset* $Q' \subseteq Q$ *of size* $\Theta(\frac{n}{\log n})$ *is compatible, yet* Q *is (very) incompatible, that is, no subset consisting of more than a* $1/3 + \epsilon$ *fraction of its quartets is compatible.*

In the following we prove the existence of Q in Theorem 3.1 via probabilistic arguments.

DEFINITION 3.2 (random quartet set). *A set of* m *quartets* Q *is a* random quartet set *if it is constructed by the following process. A quartet* q *is composed by first sampling four distinct taxa* $a, b, c, d \in \{1, \ldots, n\}$ *uniformly at random. Then* q *is set to be one out of the three possible quartets over* $\{a, b, c, d\}$ *(namely, one of* $[ab|cd], [ac|bd], [ad|bc]$*) with equal probability. Finally,* q *is added to* Q. *The process repeats* m *times (observe that* Q *is constructed with replacement, so it may be a multiset; however, the probability of having a repeated element in* Q *tends to zero as* n *grows, assuming* $m = o(n^2)$*).*

LEMMA 3.3. *Fix* $\epsilon > 0$. *Let* Q *be a random quartet set of size* $|Q| = m \geq \frac{4}{\epsilon^2} n \log n$. *Then with probability of at least* $2/3$ *there is no phylogenetic tree satisfying more than a fraction of* $1/3 + \epsilon$ *quartets of* Q.

Proof. Let T be any phylogenetic tree over the taxa set $\mathcal{X} = \{1, \ldots, n\}$, and let $Q_s(T)$ be the quartets of Q satisfied by T. Clearly, by the construction of Q, a quartet q is satisfied with probability $1/3$, and by linearity of expectation, $E[m - |Q_s(T)|] = 2m/3$. By a Chernoff bound (cf. [1, Theorem A.1.13]),

$$\Pr\left[|Q_s(T)| - \frac{m}{3} > \epsilon m\right] = \Pr\left[(m - |Q_s(T)|) - \frac{2m}{3} < -\epsilon m\right] < e^{-\frac{\epsilon^2 m^2}{4m/3}} \leq \frac{1}{n^{3n}}.$$

Now, a phylogenetic tree with n leaves has $2n - 2$ vertices and it is well known [11] that there are exactly $(2n-2)^{2n-4}$ (not necessarily phylogenetic) labeled trees with $2n - 2$ vertices. Hence, using the union bound, we obtain that the probability that *any* phylogenetic tree satisfies more than $m/3 + \epsilon m$ quartets is less than

$$(2n-2)^{2n-4} \cdot \frac{1}{n^{3n}} \ll \frac{1}{3}. \qquad \square$$

We now turn to prove the existence of a set Q as above, but for which all its

subsets of size $O(n/\log n)$ are compatible. Again, we use probabilistic arguments to establish the desired properties.

DEFINITION 3.4 (quartet cover). *Let Q' be a set of quartets. Let $Cover(Q')$ denote the set of taxa contained in some quartet of Q'. If $\mathcal{X}' \subseteq Cover(Q')$, then we say that Q'* covers *$\mathcal{X}'$.*

LEMMA 3.5. *Let Q' be a random quartet set of size m'. For $k \leq n$,*

$$\Pr[|Cover(Q')| \leq k] \leq \left(\frac{k}{n}\right)^{4m'-k} e^k \,. \tag{3.1}$$

Proof. For a given subset $\mathcal{X}'$ of k taxa, we consider the event that $Cover(Q') \subseteq \mathcal{X}'$. The probability that a random quartet has all its taxa in $\mathcal{X}'$ is

$$\frac{k(k-1)(k-2)(k-3)}{n(n-1)(n-2)(n-3)} \leq \left(\frac{k}{n}\right)^4 .$$

Hence,

$$\Pr[Cover(Q') \subseteq \mathcal{X}'] \leq \left(\frac{k}{n}\right)^{4m'} .$$

As there are $\binom{n}{k}$ possible choices for $\mathcal{X}'$, we have (by the union bound)

$$\Pr[|Cover(Q')| \leq k] \leq \binom{n}{k}\left(\frac{k}{n}\right)^{4m'} \leq \left(\frac{ne}{k}\right)^k \left(\frac{k}{n}\right)^{4m'} \leq \left(\frac{k}{n}\right)^{4m'-k} e^k \,,$$

where the second inequality is derived by the bound on binomial coefficients $\binom{n}{k} \leq \left(\frac{ne}{k}\right)^k$(see, e.g., [14, p. 1097]). It can easily be observed that the lemma holds even when Q' is a multiset. ∎

Equation (3.1) bounds the probability that a *given* random quartet set Q' satisfies $|Cover(Q')| \leq k$. As our ground set of m quartets Q is also random, so is every subset of Q. We want to bound the probability of having a small value of $|Cover(Q')|$ for *any* subset Q' of Q of size m'. As there are $\binom{m}{m'}$ subsets to consider, we immediately obtain the following corollary from Lemma 3.5.

COROLLARY 3.6. *Let Q be a random quartet set of size m. For $m' \leq m$, the probability that there exists some subset Q' of Q of size m' with $|Cover(Q')| \leq k$ is at most*

$$\binom{m}{m'}\left(\frac{k}{n}\right)^{4m'-k} e^k \leq \left(\frac{me}{m'}\right)^{m'}\left(\frac{k}{n}\right)^{4m'-k} e^k = \left(\frac{m}{m'}\right)^{m'}\left(\frac{k}{n}\right)^{4m'-k} e^{k+m'} \,,$$

where the first inequality again is due to the bound that is bound on binomial coefficients.

On the one hand, the larger the k, the stronger the result in Theorem 3.1. On the other hand, making k too large makes it more difficult to maintain compatibility. The next few claims establish a value of k which is relatively large, while guaranteeing compatibility.

DEFINITION 3.7. *A quartet set $\hat{Q}$ satisfies* twofold coverage *(TFC) if $|Cover(\hat{Q})| > 2|\hat{Q}|$.*

We now show that when the TFC property applies to all nonempty subsets of a quartet set, this latter set must be compatible. We note that a weaker constraint on

$\hat{Q}$, in which $|Cover(\hat{Q}))| > |\hat{Q}|+3$, was shown in [24] to be sufficient for compatibility. However, the proof below is simpler and suffices for our needs.

LEMMA 3.8. *Let Q' be a set of quartets with the property that every nonempty subset $\hat{Q} \subseteq Q'$ satisfies TFC; then Q' is compatible.*

Proof. Suppose that every nonempty subset $\hat{Q} \subseteq Q'$ satisfies TFC. We first note that any such $\hat{Q}$ has at least one taxon covered by a single quartet of $\hat{Q}$. Indeed, the average number of quartets containing a specific taxon is precisely

$$\frac{4|\hat{Q}|}{|Cover(\hat{Q})|} < \frac{4|\hat{Q}|}{2|\hat{Q}|} = 2,$$

and so at least one taxon is contained in less than two quartets of $\hat{Q}$, namely in one.

OBSERVATION 3.9. *Let T be a tree over a set $\mathcal{X}$, and let q be a quartet over $\{a,b,c,d\}$ such that $|\{a,b,c,d\} \setminus \mathcal{X}| \geq 1$. Then it is possible to construct a tree (over $\{a,b,c,d\} \cup \mathcal{X}$) satisfying both q and T.*

Proof. We consider the case $|\{a,b,c,d\} \setminus \mathcal{X}| = 1$; if this cardinality is bigger, the argument is simpler. Without loss of generality (w.l.o.g.), let $q = [ab|cd]$ and assume $d \notin \mathcal{X}$. Let T' be constructed from T as follows. Split the edge incident with c in T by adding a new internal vertex v, and attach a new edge connecting v and d. Hence, T' satisfies both T and q. ∎

As any nonempty $\hat{Q} \subseteq Q'$ satisfies TFC and thus has a taxon x covered by a single quartet of $\hat{Q}$, the last observation shows that if we already have a tree over the rest of the taxa $\mathcal{X} \setminus x$ that satisfies all $Q' \setminus q$, we can construct a tree over the whole set $\mathcal{X}$ by adding x such that q is satisfied, and hence all Q' are satisfied.

We now construct a tree satisfying Q'. The proof follows by induction on the size of Q'. The basis of the induction is trivial: Q' is a single quartet, and so is the tree. In order to prove the induction step, we need some auxiliary observation.

OBSERVATION 3.10. *For a quartet set Q' such that every nonempty subset $\hat{Q} \subseteq Q'$ satisfies TFC, it is possible to order Q' as (q_i) such that every q_i contains a new taxon, i.e., a taxon not contained in any quartet q_j for $j < i$.*

Proof. We build the ordering backward, starting from the complete set Q'. Since Q' satisfies TFC, there must be a quartet q' with a taxon covered only by q'. Then we take out q' to be the last in the ordering. Since $Q' \setminus \{q'\}$ also satisfies TFC, we can repeat this process all the way until Q' is depleted. ∎

We can now prove the induction step. Since every q_i in the ordering (q_i) contains a taxon not covered by any q_j, for $j < i$, by Observation 3.9, it can be added to the existing tree. ∎

It now remains to establish the probability of the event described by the conditions of Lemma 3.8 for the values of interest.

LEMMA 3.11. *Let $C \geq 1$ be a constant, and let n be sufficiently large as a function of C. Let Q be a random quartet set of size $m \leq Cn \log n$. Then the probability that every subset $Q' \subseteq Q$ of size at most $n^2/(4e^4 m)$ is compatible is at least $2/3$.*

Proof. By Corollary 3.6, the probability that there exists a quartet subset of size m' (of a set of m quartets) that covers at most k taxa is at most

$$\left(\frac{m}{m'}\right)^{m'} \left(\frac{k}{n}\right)^{4m'-k} e^{k+m'}.$$

As we wish to use the properties of TFC , we set $k = 2m'$. The probability in the last

equation becomes

$$\left(\frac{m}{m'}\right)^{m'}\left(\frac{2m'}{n}\right)^{4m'-2m'}e^{3m'}=\left(\frac{4e^3mm'}{n^2}\right)^{m'}. \tag{3.2}$$

However, we need to bound the above probability not for a single value of m', but for all values $m'=1,\ldots,m^*$, where $m^*=n^2/(4e^4m)$. Thus, by (3.2) and the union bound, we must prove

$$\sum_{m'=1}^{m^*}\left(\frac{4e^3mm'}{n^2}\right)^{m'}\le\frac{1}{3}.$$

Indeed, such a bound implies that with probability at least 2/3, every subset of quartets of size $m'\le m^*$ covers more than $k=2m'$ quartets and hence satisfies TFC. Consequently, by Lemma 3.8, every such subset is compatible.

To bound the left-hand side of the last inequality, note that for $m'=1$ we have $4e^3mm'/n^2=O(\log n/n)$ and for each $2\le m'\le\log n$ we have $(4e^3mm'/n^2)^{m'}\le O(\frac{\log^2 n}{n^2})$. For bigger values of m' the assumption that $m'\le n^2/(4e^4m)$ implies $(4e^3mm'/n^2)^{m'}\le(1/e)^{m'}\le\frac{1}{n}$. Therefore the sum is at most

$$O\left(\frac{\log n}{n}\right)+\sum_{2\le i\le\log n}O\left(\frac{\log^2 n}{n^2}\right)+\sum_{i>\log n}\left(\frac{1}{e}\right)^i=O\left(\frac{\log n}{n}\right),$$

which is smaller than 1/3 provided n is sufficiently large. ∎

Proof of Theorem 3.1. Let $\epsilon>0$, and let $C=4/\epsilon^2$. Let n be sufficiently large as a function of ϵ (and hence C). Let Q be a random quartet set of size $|Q|=m=\frac{4}{\epsilon^2}n\log n$. Then, by Lemma 3.3, with probability of at least 2/3, there is no phylogenetic tree satisfying more than a fraction of $1/3+\epsilon$ quartets of Q. By Lemma 3.11, with probability at least 2/3, every subset of Q of size at most $\frac{\epsilon^2 n}{16e^4\log n}$ is compatible. Since with probability at least 1/3 (namely, positive probability) both lemmata hold for a randomly chosen Q, the theorem follows. ∎

4. Extension to dense inputs. In section 3 we have shown that even when all quartet subsets of a fairly large magnitude are compatible, the whole quartet set Q may still be very far from compatible. In this section we first mention a result of an opposite case: when Q is very dense, this cannot happen. We use a classic result of Colonius and Schulze [13] dealing with quartet sets of size three, or for short, quartet triplets. The motivation to study quartet triplets stems from the following reason. Trivially, a single quartet is compatible, and by our assumption (that every four taxa induce a single quartet) and Observation 3.9, also a pair of quartets is compatible. This, however, changes when considering three or more quartets. For example, the set $[ab|cd]$, $[ac|ed]$, and $[ae|cd]$ is easily seen to be incompatible. Figure 3(a) shows two compatible quartets inducing a tree over five taxa, whereas Figure 3(b) shows a quartet triplet that is incompatible—every pair of quartets induces a different tree. We denote trees over five taxa as *quintets* and usually mark quintets by r (as q is reserved for quartets). Observe that all unlabeled phylogenetic trees on five leaves are isomorphic. Precisely two pairs of leaves have a common neighbor (such a pair is called a *cherry*). We therefore have the following definition.

DEFINITION 4.1. *A* quintet *is an unrooted phylogenetic tree over five taxa. We mark a quintet* $r=[12|3|45]$ *denoting that* $\{1,2\}$ *and* $\{4,5\}$ *are the two cherries and* 3 *is the remaining leaf.*

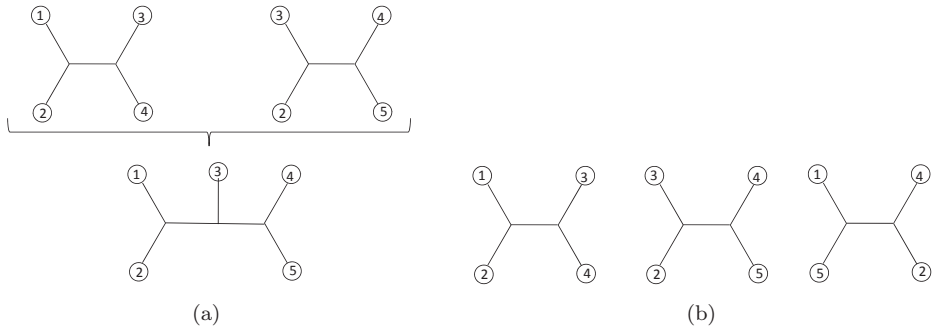


FIG. 3. (a) *Two quartets inducing a tree over five taxa.* (b) *Three incompatible quartets.*

In [13], Colonius and Schulze defined the following quartet condition over a full set Q.

CONDITION 4.2. *If* $[1,5|3,4] \in Q$ *or* $[1,5|2,3] \in Q$, *then* $[1,2|3,4] \in Q$ *implies* $[2,5|3,4] \in Q$.

They then stated the following theorem (see [13, Theorem 2]).

THEOREM 4.3. *If Q is the full quartet set and for every $\{a,b,c,d,e\} \in \mathcal{X}$, Condition* 4.2 *holds, then Q is compatible.*

By a simple observation [29, Chapter 6, exercise 19], it can be shown that Condition 4.2 implies that every quartet triplet is compatible, and hence, by Theorem 4.3, the following corollary follows.

COROLLARY 4.4. *If Q is the full quartet set and every quartet triplet in Q is compatible, then Q is compatible.*

In the previous section we have shown that when the quartet set is rather sparse, specifically $O(n \log n)$ quartets, it is possible that even if every relatively large subset of it is compatible, the entire set may still be very far from compatible. On the other hand, by Corollary 4.4, if the set is the full $\binom{n}{4}$ quartet set, then it is enough that every quartet triplet is compatible, in order to force compatibility of the entire set. A natural question to ask is how far we can push up the size of the ground set, that is, increasing the number of quartets towards $\binom{n}{4}$, and still obtain a negative result while requiring that all nonnegligible size subsets be compatible. In this section we show that even for size $\Theta\left(\binom{n}{4}\right)$ there are (very) incompatible sets of this size in which every subset of large constant cardinality is compatible.

THEOREM 4.5. *For any $\gamma > 0$, there exists $0 < \delta \leq \gamma$ and there exist quartet sets of size δn^4 that are incompatible, and yet they have the property that every subset of size $\tilde{\Theta}\left((1/\delta)^{1/3}\right)$ is compatible.*[1]

We give a short overview of the proof. We start with a small constant size basic set $\mathcal{X}^*$ of N taxa that is big enough for Theorem 3.1 to hold. We therefore have a quartet set of size $\Theta(N \log N)$ for which we know that it is not compatible, and yet every subset of size $\Theta(N/\log N)$ is compatible. Next, we use a blowup argument to expand the basic set to a set of taxa $\mathcal{X}$ of size n and $\Theta\left(\binom{n}{4}\right)$ quartets. We prove that the expanded quartet set is still very far from being compatible, yet any appropriate large constant size subset of it is compatible. Adjusting the parameter γ and hence δ controls the size of N and the blowup scale, and the theorem follows. We proceed with the details.

[1]The notation $\tilde{\Theta}(.)$ is used to suppress polylogarithmic factors. Namely, $\tilde{\Theta}(t) = \Theta(t \log^k t)$.

Proof. By the proof of Theorem 3.1, we know that for any $\epsilon > 0$ there exists $N_0 = N_0(\epsilon)$ such that for all $N > N_0$ there exists a quartet set Q over a taxa set of size N with $|Q| = \frac{4}{\epsilon^2} N \log N$ with the following properties: No tree satisfies more than a fraction of $1/3 + \epsilon$ elements of Q, and yet every subset of Q with at most $\varepsilon^2 N/(16e^4 \log N)$ elements is compatible. For $\gamma > 0$, let $\epsilon > 0$ be the largest constant for which $N_0 = N_0(\epsilon)$ satisfies

$$\gamma \geq \frac{4 \log N_0}{\epsilon^2 N_0^3} \ .$$

Let $N \geq N_0$ be the smallest constant which satisfies

$$\left(\frac{1}{\delta (\log \frac{1}{\delta})^4} \right)^{1/3} \leq \frac{\epsilon^2 N}{16 e^4 \log N}, \tag{4.1}$$

where

$$\delta = \frac{4 \log N}{\epsilon^2 N^3} \ .$$

Observe that indeed $\delta \leq \gamma$.

Let $\mathcal{X}^*$ be a ground set of N taxa, and let Q^* be a set of $|Q^*| = \frac{4}{\epsilon^2} N \log N$ quartets over $\mathcal{X}^*$ with the following properties: No tree satisfies more than a fraction of $1/3 + \epsilon$ elements of Q^*, and yet every subset of Q^* with at most $\varepsilon^2 N/(16e^4 \log N)$ elements is compatible.

Let $\mathcal{X}$ be obtained from $\mathcal{X}^*$ by replacing each taxon $x \in \mathcal{X}^*$ with k copies denoted by $x_1, \ldots, x_k$, where $k = n/N$. Observe that $|\mathcal{X}| = n$. Correspondingly, construct a quartet set $Q = (\mathcal{Q}^*)^k$ by constructing, for each $q = [ab|cd] \in \mathcal{Q}^*$, a set of k^4 copies of the form $q_{i,j,\ell,p} = [a_i b_j | c_\ell d_p]$ for $i, j, \ell, p = 1, \ldots, k$. Notice that

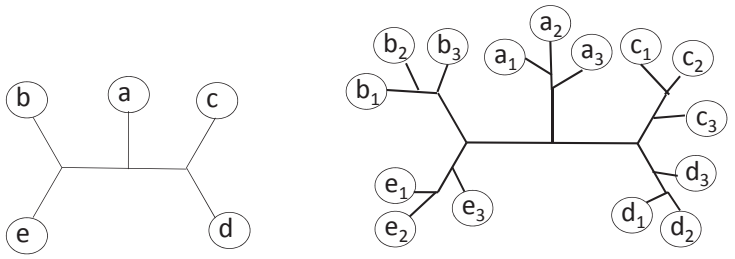
$$|Q| = k^4 |Q^*| = \left(\frac{n}{N} \right)^4 \frac{4}{\epsilon^2} N \log N = \delta n^4 \ .$$

We next show that Q is (very) incompatible. Let $Q' \subseteq Q$ be any set of size at least $(1/3 + \epsilon)|Q|$. We show that Q' is incompatible. We can partition Q' into k^4 equivalence classes $Q'(i, j, \ell, p)$ according to the indices of the quartets in it. A quartet $[a_i b_j | c_\ell d_p]$ belongs to the class $Q'(i, j, \ell, p)$. By averaging, there is some equivalence class, say w.l.o.g. $Q'(1, 1, 1, 1)$, of size at least

$$\frac{|Q'|}{k^4} \geq \frac{(1/3 + \epsilon)|Q|}{k^4} = \frac{(1/3 + \epsilon)\delta n^4}{(n/N)^4} = (1/3 + \epsilon)|Q^*| \ .$$

But $Q'(1, 1, 1, 1)$ is isomorphic to a subset of quartets of Q^* (simply ignore the indices), and by the properties of Q^* we have that no subset of more than a fraction of $\frac{1}{3} + \epsilon$ elements of it is satisfied, and hence $Q'(1, 1, 1, 1)$ is incompatible. Thus, Q' is incompatible.

It remains to prove that every subset of Q of size $\left(\frac{1}{\delta (\log \frac{1}{\delta})^4} \right)^{1/3}$ is compatible. Indeed, let Q' be such a subset. If we ignore the indices of the taxa, we get that the index-free Q' is a (multi)set of Q^* of size at most $\left(\frac{1}{\delta (\log \frac{1}{\delta})^4} \right)^{1/3}$, and hence, by (4.1) and the property of Q^*, it is compatible. Let T' be a tree satisfying the index-free Q' over $\mathcal{X}^*$. Replace every leaf (taxa) in T' with a tree on its k copies, as shown in Figure 4, to obtain a tree over $\mathcal{X}$ satisfying Q'. ∎



(a) (b)

FIG. 4. (a) *A tree T' on taxa set $\mathcal{X}^* = \{a, b, c, d, e\}$.* (b) *A tree T' on taxa set $\mathcal{X}$ after replacing each leaf with its k copies ($k = 3$ in this example).*

5. Maximum violation of compatible quartets. The issue of tree similarity is central in phylogenetics. The most common tree similarity measure is the Robinson–Foulds (RF) [28] symmetric difference, which is actually a distance measure. Removing an edge in the tree induces two subtrees. Hence, an edge in a tree induces a bipartition on the taxa set, where each part corresponds to the taxa set in an induced subtree. A tree is therefore the set of its bipartitions. The RF distance between two trees counts how many edges (bipartitions) are exhibited by exactly one tree.

Another common measure that involves quartet compatibility is *quartet fit* (qFit), or *quartet distance* [22]. Let T be a tree over n leaves, and let $Q(T)$ be its full quartet set of $\binom{n}{4}$ quartets. For another tree T' with n leaves, let $qfit_T(T')$ denote the number of quartets out of $Q(T)$ satisfied by T'. Clearly, $qfit_T(T) = \binom{n}{4}$ and $qfit_T(T') = qfit_{T'}(T)$. A natural question to ask is how small $qfit_T(T'$ can be. It is clear that a tree over four taxa, i.e., a quartet q, can easily be violated by another quartet q', simply by rewiring (permuting) the taxa of q, yielding $qfit_q(q') = 0$. However, as we show next, for larger trees, the problem gets significantly more involved. The problem was presented by Bandelt and Dress [2] (there $qfit$ is denoted by δ), and they conjecture that it goes to $2/3$ as n goes to infinity. We here present upper and lower bounds for this value. In particular, we show that for every fixed n it is strictly larger than $2/3$, but we believe that the conjecture is likely to be true. We now define the problem rigorously using, for convenience, the following notation. Let $f(n)$ denote the maximum cardinality of a set of compatible quartets Q over n taxa such that there is some tree which violates all of Q. Trivially, $f(n) \leq \binom{n}{4}$ and $f(4) = 1$.

The simplest nontrivial tree is a quintet. There is a single unlabeled structure for a quintet, so for $n = 5$, we can only rewrite the labels on the leaves. Observe that the quintet $r' = [15|3|24]$ violates all the quartets defined by the quintet $r = [12|3|45]$, thereby showing that $f(5) = 5$ (or, equivalently, that $qfit_r(r') = 0$). However, as we shall see in Lemma 5.2, already for $n = 6$ we no longer have total violation, as $f(6) = 14 < \binom{6}{4}$.

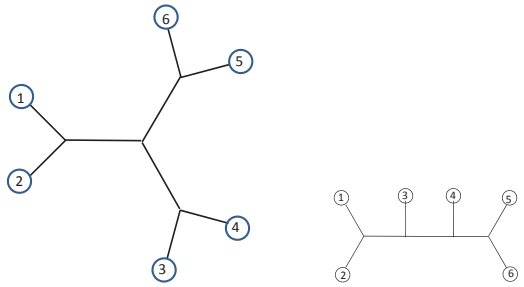
The main result of this section provides upper and lower bounds for $f(n)$.

THEOREM 5.1. $\frac{9}{10}\binom{n}{4}(1 + o(1)) \geq f(n) > \frac{2}{3}\binom{n}{4}$.

We note that this result (specifically the lower bound) does not refute the conjecture of Bandelt and Dress, as in the limit $f(n)/\binom{n}{4}$ may still equal $2/3$.

We first need to establish the following lemma (the same lemma was also proved in [2]).

LEMMA 5.2. $f(6) = 14$.



(a) (b)

FIG. 5. (a) *The* 6*-flower tree.* (b) *The* 6*-caterpillar tree* [12|3|4|56].

Proof. There are two nonisomorphic unlabeled phylogenetic trees on six leaves. One has three cherries and another has two cherries, which we denote as the *flower* and the *caterpillar*, respectively (see Figure 5). Let T and T' be two labeled trees on the six taxa $\{1,2,3,4,5,6\}$. We need to show that at least one quartet is satisfied by both of them.

Assume first that at least one of T or T' is a flower. W.l.o.g., T is a flower. Observe that the flower has the property that for any two of its taxa x, y, there is at least one quartet of the form $[xy|ab]$. Indeed, if $\{a,b\}$ is a cherry not involving x or y, then the quartet on $\{x,y,a,b\}$ has this form. Now, for any cherry $\{x,y\}$ of T', and for any cherry $\{a,b\}$ of T which does not involve x, y, the quartet $[xy|ab]$ is satisfied by both T' and T.

We may now assume that both T and T' are caterpillars. Assume w.l.o.g. that $\{1,2\},\{5,6\}$ are the cherries of T and that 3 is the "noncherry" closer to $\{1,2\}$. For convenience, we can denote T by $[12|3|4|56]$.

Now, if 1 and 2 are not in opposite cherries in T' (namely, T' is *not* of the form $[1a|b|c|2d]$), then there is at least one quartet in T' of the form $[12|ab]$, as we can take a, b to be a cherry of T' not involving $1, 2$.

So, we may now assume that 1 and 2 are in opposite cherries in T', and, symmetrically, we can assume that 5 and 6 are in opposite cherries in T'. So, T' is equivalent to either $[15|3|4|26]$ or $[15|4|3|26]$. In the first case, $[13|46]$ is the only quartet satisfied by both trees, and in the second case, $[23|45]$ is the only quartet satisfied by both trees, proving that, in fact, $f(6) = 14$. ∎

Proof of Theorem 5.1. We first prove that $f(n) > \frac{2}{3}\binom{n}{4}$.

Let T be any tree on the n-taxa set $\mathcal{X}$. Let $Q = Q(T)$ be the full quartet set of T. In particular, any subset of Q is a compatible set. Let T^π be obtained from T by randomly permuting the labels on the leaves. Then, for any quartet in Q, the probability that it is satisfied by T^π is $1/3$. By linearity of expectation, the expected number of quartets of Q satisfied by T^π is $\binom{n}{4}/3$. As the random variable (that counts the number of quartets of Q satisfied by T^π) is nonconstant, we get that there is a permutation on the leaves such that the resulting T^π does not satisfy more than $\frac{2}{3}\binom{n}{4}$ quartets of Q.

We next prove that $f(n) \le \frac{9}{10}\binom{n}{4}(1+o(1))$. Recall that an r-uniform hypergraph on n vertices is a subset of r-sets of $\{1,\dots,n\}$ (the r-sets are the edges of the hypergraph). An r-uniform hypergraph is *complete* if it contains all $\binom{n}{r}$ possible edges. Denote by K_k^r the complete r-uniform hypergraph on k vertices. For positive integers $r < k < n$, let $T(n,k,r)$ denote the maximum possible number of edges in an

r-uniform hypergraph that does not contain K_k^r as a subgraph. A result of de Caen [17] asserts that $T(n,6,4) \leq \frac{9}{10}\binom{n}{4}(1+o(1))$.

Given two trees T and T^* on the n taxa set $\{1,\ldots,n\}$, construct the following 4-uniform hypergraph H, whose vertices are the taxa. For any 4-set $\{a,b,c,d\}$, we make $\{a,b,c,d\}$ an edge of H if and only if the quartet on $\{a,b,c,d\}$ in T^* is different from the quartet on $\{a,b,c,d\}$ in T. By Lemma 5.2, we have that H does not contain K_6^4 as a subgraph. Hence, the number of edges of H is at most $T(n,6,4) \leq \frac{9}{10}\binom{n}{4}(1+o(1))$. □

6. Further research and open problems. In this work we presented several results regarding quartet subset compatibility. We drew a direct linkage between the density of the incompatible ground quartet set and the gap to its compatible constituting subsets. Our results rely on probabilistic arguments and hence do not necessarily provide explicit examples for the given proofs of existence.

As mentioned in the introduction, one consequence of the results here is that no sampling algorithm can effectively be used to determine the compatibility of the given quartet set or estimate the maximum cardinality of a compatible subset of it.

The results presented here have a tight link to quartet set closure as studied extensively in [9, 19, 25]. Recall that $co(Q)$ is the set of trees that satisfy Q. Then the closure of Q, $cl(Q)$ is defined as follows.

DEFINITION 6.1.

$$cl(Q) = \bigcap_{T \in co(Q)} Q(T).$$

That is, $cl(Q)$ is the quartet set that exists in every tree compatible with Q. There is no polynomial time algorithm known to compute $cl(Q)$.

Condition 4.2 gives rise to two immediate quartet rules, denoted as the *dyadic closure*, introduced first by [18] and used further in [20, 21].

DEFINITION 6.2. *The dyadic closure of a quartet set Q, denoted as $cl_2(Q)$, is a minimal set of quartets that contains Q and satisfies the following two rules:*

1. dc1: $[ab|cd], [ab|ce] \in cl_2(Q) \Rightarrow [ab|de] \in cl_2(Q)$.
2. dc2: $[ab|cd], [bc|de] \in cl_2(Q) \Rightarrow [ab|ce], [ab|de], [ac|de] \in cl_2(Q)$.

Observe that, indeed, any tree which satisfies the quartets on the left-hand side of a dyadic rule must also satisfy the right-hand side of the rule.

An implied property of $cl(Q)$ is that it cannot be extended by any quartet inference rule, such as the dyadic closure rules presented above. In [25] an explicit set of six quartets over $n = 8$ taxa is given. This set has the property that it is incompatible, yet every subset of it is compatible *and closed*. The example demonstrates that there are cases in which not any quartet rule can be applied on any compatible subset, in order to arrive at a conflict at the ground set. However, it is not shown how this example can be extended to any n.

In this respect, it is interesting to find whether we can extend our compatible subsets by applying some quartet rules. We give here a partial answer to this question, which extends our result from Theorem 3.1. We say that a set Q is 2-closed if it cannot be extended by any of the two dyadic rules (see Definition 6.2), i.e., $Q = cl_2(Q)$. Below is a restatement of Theorem 3.1 with the restriction to 2-closed sets.

THEOREM 6.3. *There exists a set Q of m quartets such that $m = \Theta(n \log n)$ and every subset $Q' \subseteq Q$ of size $m' = O\left(\frac{n}{\log n}\right)$ is compatible and 2-closed, yet Q is incompatible.*

Proof. We need to show that in no subset of Q can we apply one of the dyadic rules. Note that to apply each of these rules we need to jointly have three taxa in both quartets. We need to show that there exists such a set that satisfies the conditions of Theorem 3.1 and also satisfies that no two quartets share the same three taxa.

OBSERVATION 6.4. *Let Q be a set of $\Theta(n \log n)$ random quartets. Then the probability of two quartets sharing the same three taxa is $o(1)$.*

Proof. For three taxa a, b, c, the probability of being selected in a quartet is $\frac{1}{n(n-1)(n-2)}\binom{4}{3}3! = \Theta\left(\frac{1}{n^3}\right)$ and for being selected in two quartets q_1 and q_2 is $\Theta\left(\frac{1}{n^6}\right)$. Since we have $\binom{n}{3}$ triplets of taxa, the probability of any three taxa appearing in two quartets is $\Theta(\frac{n^3}{n^6})$. Finally, since the size of Q is $\Theta(n \log n)$, there are $\Theta(n^2 \log^2 n)$ pairs of quartets, so the total probability is at most $O(\frac{\log^2 n}{n}) = o(1)$. ∎

Since the probability of finding such a pair of quartets is so low, it is clear that this constraint does not affect the existence of such a set Q, as we showed in Theorem 3.1. ∎

The *order* of a closure inference rule is the minimal number of quartets necessary to derive the inference (i.e., the number of quartets on the left side of the rule). Theorem 6.3 answers only partially the question of under what order our subsets are closed, that is, what the closure inference rules are that cannot be applied to them. In particular, if no inference rule can be applied to these subsets, they are fully closed. While using arguments similar to those used here can push the limit a bit upward, it seems that a different approach is necessary to prove/disprove full closeness.

As the majority of this work provides negative results regarding the inability to infer set compatibility based on subset compatibility, it would be beneficial to expand the positive result of [13] to other cases, e.g., when some quartets are missing. It is noteworthy that compatibility of the full quartet set can be checked in polynomial time [2], and therefore the result of Corollary 4.4 is not of direct practical algorithmic use and is mostly a structural result.

In the realm of quartet fit measure studied in section 5, the task of closing the gap between the two bounds remains open. This is essential for providing a normalization means for this measure, representing the *relative* (as opposed to the *absolute*) quartet violation between trees.

Acknowledgments. We thank the referees of an earlier version of this manuscript for their useful comments and for pointing us to the results of Colonius and Schulze [13], the conjecture from Bandelt and Dress [2], and the result of Grünewald [24].

REFERENCES

[1] N. ALON AND J.H. SPENCER, *The Probabilistic Method*, 2nd ed., Wiley-Interscience, New York, 2000.

[2] H. BANDELT AND A. DRESS, *Reconstructing the shape of a tree from observed dissimilarity data.*, Adv. in Appl. Math., 7 (1986), pp. 309–343.

[3] V. BERRY AND O. GASCUEL, *Inferring evolutionary trees with strong combinatorial evidence*, Theoret. Comput. Sci., 240 (2001), pp. 271–298.

[4] V. BERRY, T. JIANG, P. KEARNEY, M. LI, AND T. WAREHAM, *Quartet cleaning: Improved algorithms and simulations*, in Algorithms—ESA '99 (Prague), Springer, Berlin, 1999, pp. 313–324.

[5] O. BININDA-EMONDS, ED., *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, Comput. Biol. 4, Kluwer Academic Publishers, Dordrecht, The Netherlands, 2004.

[6] O. BININDA-EMONDS, J. GITTLEMAN, AND M. STEEL, *The (super)tree of life: Procedures, problems, and prospects*, Annu. Rev. Ecol. Systemat., 33 (2002), pp. 265–289.

[7] S. Böcker, D. Bryant, A. Dress, and M. Steel, *Algorithmic aspects of tree amalgamation*, J. Algorithms, 37 (2000), pp. 522–537.
[8] D. Bryant and J. Lagergren, *Compatibility of unrooted phylogenetic trees is FPT*, Theoret. Comput. Sci., 351 (2006), pp. 296–302.
[9] D. Bryant and M. Steel, *Extension operations on sets of leaf-labelled trees*, Adv. in Appl. Math., 16 (1995), pp. 425–453.
[10] D. Bryant and M. Steel, *Constructing optimal trees from quartets*, J. Algorithms, 38 (2001), pp. 237–259.
[11] A. Cayley, *A theorem on trees*, Quart. J. Math, 23 (1889), pp. 376–378.
[12] M. S. Chang, C. C. Lin, and P. Rossmanith, *A property tester for tree-likeness of quartet topologies*, Theory Comput. Syst., 49 (2011), pp. 576–587.
[13] H. Colonius and H. Schulze, *Tree structures for proximity data*, British J. Math. Statist. Psych., 34 (1981), pp. 167–180.
[14] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms.*, 2nd ed., MIT Press, Cambridge, MA, 2001.
[15] M. Csürös, *Fast recovery of evolutionary trees with thousands of nodes*, J. Comput. Biol., 9 (2002), pp. 277–297.
[16] C. Daskalakis, E. Mossel, and S. Roch, *Phylogenies without branch bounds: Contracting the short, pruning the deep*, SIAM J. Discrete Math., 25 (2011), pp. 872–893.
[17] D. de Caen, *Extension of a theorem of Moon and Moser on complete subgraphs*, Ars Combin., 16 (1983), pp. 5–10.
[18] M. Dekker, *Reconstruction Methods for Derivation Trees*, Masters thesis, Vrije Universiteit, Amsterdam, 1986.
[19] T. Dezulian and M. Steel, *Phylogenetic closure operations and homoplasy-free evolution*, in Classification, Clustering, and Data Mining Applications, Studies in Classification, Data Analysis, and Knowledge Organization, Springer, Berlin, Heidelberg, 2004, pp. 395–416.
[20] P. Erdös, M. Steel, L. Szekely, and T. Warnow, *A few logs suffice to build (almost) all trees* (i), Random Structures Algorithms, 14 (1999), pp. 153–184.
[21] P. Erdös, M. Steel, L. Szekely, and T. Warnow, *A few logs suffice to build (almost) all trees* (ii), Theoret. Comput. Sci., 221 (1999), pp. 77–118.
[22] J. Estabrook, *Comparison of undirected phylogenetic trees based on subtrees of four evolutionary units*, Syst. Biol., 34 (1985), pp. 193–200.
[23] I. Gronau, S. Moran, and S. Snir, *Fast and reliable reconstruction of phylogenetic trees with indistinguishable edges*, Random Structures Algorithms, 40 (2012), pp. 350–384.
[24] S. Grünewald, *Slim sets of binary trees*, J. Combin. Theory Ser. A, 119 (2012), pp. 323–330.
[25] S. Grünewald, M. Steel, and M. Shel Swenson, *Closure operations in phylogenetics*, Math. Biosci., 208 (2007), pp. 521–537.
[26] T. Jiang, P. Kearney, and M. Li, *Orchestrating quartets: Approximation and data correction*, in Proceedings of the 39th Annual IEEE Symposium on Foundations of Computer Science, Palo Alto, CA, 1998, pp. 416–425.
[27] T. Jiang, P. Kearney, and M. Li, *A polynomial time approximation scheme for inferring evolutionary trees from quartet topologies and its application*, SIAM J. Comput., 30 (2001), pp. 1942–1961.
[28] D. R. Robinson and L. R Foulds, *Comparison of phylogenetic trees*, Math. Biosci., 53 (1981), pp. 131–147.
[29] C. Semple and M. A. Steel, *Phylogenetics*, Oxford University Press, Oxford, UK, 2003.
[30] B. Shutters, S. Vakati, and D. Fernandez-Baca, *Incompatible quartets, triplets, and characters*, Algorithm Mol. Biol., 8 (2013), 11.
[31] S. Snir and S. Rao, *Quartets maxcut: A divide and conquer quartets algorithm*, IEEE/ACM Trans. Comput. Biol. Bioinformatics (TCBB), 7 (2010), pp. 714–718.
[32] S. Snir and R. Yuster, *A linear time approximation scheme for maximum quartet consistency on sparse sampled inputs*, SIAM J. Discrete Math., 25 (2011), pp. 1722–1736.
[33] S. Snir and R. Yuster, *Reconstructing approximate phylogenetic trees from quartet samples*, SIAM J. Comput., 41 (2012), pp. 1466–1480.
[34] K. St. John, T. Warnow, B. Moret, and L. Vawter, *Performance study of phylogenetic methods: (Unweighted) quartet methods and neighbor-joining*, in Proceedings of the Twelfth Annual ACM-SIAM Symposium on Discrete Algorithms, 2001, pp. 196–206.
[35] M. Steel, *The complexity of reconstructing trees from qualitative characters and subtrees*, J. Classification, 9 (1992), pp. 91–116.
[36] K. Strimmer and A. von Haeseler, *Quartet puzzling: A quartet maximum-likelihood method for reconstructing tree topologies*, Mol. Biol. Evol., 13 (1996), pp. 964–969. Software available online from ftp://ftp.ebi.ac.uk/pub/software/unix/puzzle/.