Noname manuscript No.
(will be inserted by the editor)

Perfect taxon sampling and phylogenetically decisive taxon coverage

Mareike Fischer

the date of receipt and acceptance should be inserted later

arXiv:1206.3472v1 [q-bio.PE] 15 Jun 2012

Abstract In a recent study, Steel and Sanderson defined and characterized phylogenetically decisive sets of taxon sets. A set is called phylogenetically decisive if regardless of the trees chosen for each of its taxon sets, as long as these trees are compatible with one another, their supertree is always unique. This implies that the sampled taxon sets always lead to a unique supertree, regardless what tree they support (as long as the trees of all the taxon sets are compatible) – which is why this setting can be referred to as 'perfect taxon sampling'. However, the complexity of the decision problem to determine whether a set of taxon sets is phylogenetically decisive remained unknown. This problem was one of the 'Penny Ante' prize questions of the Annual New Zealand Phylogenetics Meeting in 2012. In this paper, we explain phylogenetic decisiveness and demonstrate a new characterization, which then leads to a polynomial time algorithm for the case where the number of taxon sets under consideration is polynomial in the number of taxa - both for the (simpler) rooted tree case as well as for the (more complicated) unrooted tree case.

Mathematics Subject Classification (2010) 92B05; 94C15

Keywords phylogenetic tree, grove, supertree, compatibility

Mareike Fischer
Ernst-Moritz-Arndt University Greifswald
Department for Mathematics and Computer Science
Walther-Rathenau-Str. 47
17487 Greifswald
Germany
E-mail: email@mareikefischer.de

1 Introduction

Reconstructing the Tree of Life, i.e. the phylogenetic tree displaying all living species on earth is one of the main challenges of biological sciences to-date. The genetic sequence data on some clusters of species are already available in databases like GenBank or SwissProt, and there are algorithms available to reconstruct the tree of each cluster. In many studies, data from different loci are combined by building trees from each locus and combining trees to a so-called 'supertree'. In this setting, it is common that the supertree contains all taxa whereas the input trees for each individual locus oftentimes do not contain all taxa under consideration. While it is even possible that these input trees are incompatible with one another (which then makes it impossible to find a perfect supertree, i.e. a supertree displaying all the input trees); even in the case of compatibility, it is not always clear which supertree is best as there may be more than one.

In a recent study, Steel and Sanderson (2010) mathematically characterized so-called phylogenetically decisive sets of taxon sets. These sets consist of input taxon sets which have the property that all possible compatible input trees chosen for the input sets lead to a unique supertree. This is an interesting setting, as it shows that the decision which taxa to sample for each input tree may already ensure that the supertree of all input trees is unique. In this context, a set of taxon set which is phylogenetically decisive can also be referred to as a set of perfect taxon samples. However, the complexity of deciding whether or not a given set of taxon sets is phylogenetically decisive remained unknown.

Problem 1 (Phylogenetic decisiveness decision problem) Given a taxon set X with $|X| = n$ and a set $\mathcal{S} = \{Y_1, \dots, Y_k\}$ of subsets of X, what is the complexity of determining whether $\mathcal{S}$ is phylogenetically decisive or not?

Problem 1 was listed on the 2012 'Penny Ante' prize list of the Annual New Zealand Phylogenetics Meeting (Steel (2011)), and it is the aim of this paper to present a solution to the problem.

In their study, Steel and Sanderson showed that a set of taxon sets is phylogenetically decisive if and only if it satisfies the so-called four-way partition property. However, this characterization does unfortunately not provide an answer to Problem 1, as checking the four-way partition property is not a priori efficient. In our paper, we extend the work of Steel and Sanderson. First, while the work of Steel and Sanderson focuses on the unrooted tree case, i.e. the case where all input trees and the supertree are unrooted phylogenetic trees, we additionally consider the rooted case. Moreover, we show that in both cases, for a given set of taxon sets it can be decided in polynomial time if this set is phylogenetically decisive as long as the number of taxon sets under consideration is polynomial in the number of taxa. We present algorithms for both cases, which make explicit use of some characterizing properties we derive in this work, as well as some worst-case runtime bounds for these algorithms.

More importantly, our algorithm for the unrooted tree case is not only useful for checking whether or not a set of taxon sets is phylogenetically decisive – which would imply that all possible quartets of four taxa get uniquely resolved when the input trees are combined into a supertree. This would be the perfect case which may not occur very often in practice. However, oftentimes it is not necessary to have a unique supertree but rather to find out the relationship (and thus the unique subtree common to all supertrees, if it exists) of a particular quartet which is not already resolved by any of the input trees. In terms of our algorithm, this would mean that only this particular quartet would need to be examined – i.e. one iteration would be needed. This makes our approach useful for phylogeneticists, as even if the set of taxon sets they are investigating is not phylogenetically decisive, it may still bear some useful new information which can be determined this way.

2 Notation and Model Assumptions

We first introduce some notations and definitions required for presenting our results. We start with the standard notion of phylogenetic trees both for the rooted and the unrooted case.

Definition 1 A *rooted phylogenetic X-tree* $\mathcal{T}$ is a connected, acyclic graph with one node of degree 2, namely the so-called *root*, all other internal nodes of degree 3 and $|X| = n$ nodes of degree 1, namely the so-called *leaves* or *taxa*. Whenever there is no ambiguity, we refer to a rooted phylogenetic X-tree as *rooted tree* for short.

Definition 2 An *unrooted phylogenetic X-tree* $\mathcal{T}$ is a connected, acyclic graph where all internal nodes are of degree 3 where there are and $|X| = n$ nodes of degree 1, namely the so-called *leaves* or *taxa*. Whenever there is no ambiguity, we refer to an unrooted phylogenetic X-tree as *unrooted tree* for short.

We are now in a position to define phylogenetic decisiveness.

Definition 3 A collection $\mathcal{S} = \{Y_1, \ldots, Y_k\}$ of subsets of X (i.e. $Y_i \subseteq X \;\; \forall i = 1, \ldots, k$) is said to be *phylogenetically decisive* for rooted or unrooted trees, respectively, if for every rooted or unrooted binary phylogenetic X–tree $\mathcal{T}$, the collection $\mathcal{T}|_{Y_i} : Y_i \in \mathcal{S}$ defines $\mathcal{T}$ (i.e. $\mathcal{T}$ is the only tree that displays these trees).

In order to work with phylogenetic decisiveness, we require some concepts which are similarly used in the context of phylogenetic groves, e.g. by Ané et al (2009) and Fischer (2011).

Definition 4 Let $\pi = S_1|S_2|\dots|S_m$ be a partition of $\mathcal{S} = \{Y_1,\dots,Y_k\}$ (where $Y_i \subseteq X \;\; \forall i = 1,\dots,k$), for some $m \leq k$. A set $\{x,y,z\}$ such that $x,y,z \in X$ and $\{x,y,z\} \not\subseteq \mathcal{L}(S_i)$ for all $i = 1,\dots,m$, where $\mathcal{L}(S_i) = \bigcup_{j:Y_j \in S_i} Y_j$, is called a *cross triple of $\mathcal{S}$ with respect to π* or *CT wrt π* for short.

Definition 5 Let $\pi = S_1|S_2|\dots|S_m$ be a partition of $\mathcal{S} = \{Y_1,\dots,Y_k\}$ (where $Y_i \subseteq X \;\; \forall i = 1,\dots,k$), for some $m \leq k$. A set $\{a,b,c,d\}$ such that $a,b,c,d \in X$ and $\{a,b,c,d\} \not\subseteq \mathcal{L}(S_i)$ for all $i = 1,\dots,m$, where $\mathcal{L}(S_i) = \bigcup_{j:Y_j \in S_i} Y_j$, is called a *cross quadruple of $\mathcal{S}$ with respect to π* or *CQ wrt π* for short.

Definition 6 Let $\mathcal{S} = Y_1,\dots,Y_k$ be a set of taxon sets and let $\pi = S_1|\dots|S_m$ be a partition of $\mathcal{S}$. Let $\{x,y,z\}$ be a cross triple of $\mathcal{S}$ with respect to π (or $\{a,b,c,d\}$ a cross quadruple of $\mathcal{S}$ with respect to π). Then, $\{x,y,z\}$ (or $\{a,b,c,d\}$, respectively) is called *resolved* if there is a choice of rooted (or unrooted, respectively) trees on Y_i, $i = 1,...,k$, such that all possible supertrees of these trees display the same of the three possible trees on $\{x,y,z\}$ (or $\{a,b,c,d\}$, respectively).

These few notions suffice to derive the desired results.

3 Results

3.1 Rooted tree case

First, we present the results for the case where all trees are thought of as being rooted. It turns out that this case is much simpler than the unrooted case, which is analyzed lateron. Moreover, the rooted case can be characterized in such a simple way that the decision whether a particular set of taxon sets is phylogenetically decisive can be reduced to a search for cross triples. This also leads to a new characterization of phylogenetically decisiveness.

We first state the main result, which then is proven subsequently.

Theorem 1 (Main theorem 1) *Given a set $\mathcal{S} = \{Y_1,\dots,Y_k\}$ of subsets of X, where $|X| = n$, the question whether $\mathcal{S}$ is phylogenetically decisive can be answered in at most $\mathcal{O}(k \cdot n^3)$ steps. In particular, as long as k is polynomial in n, the question can be answered in polynomial time.*

In order to prove the above theorem, we need to prove some useful properties concerning cross triples first.

Proposition 1 *Let $\mathcal{S} = \{Y_1, \ldots, Y_k\}$ be phylogenetically decisive set of subsets of X. Let π be a partition of $\mathcal{S}$. Then, there are no cross triples of $\mathcal{S}$ with respect to π.*

Proof Assume there is a CT $\{x, y, z\}$ of $\mathcal{S}$ wrt π. We now construct k trees $T_1, \ldots, T_k$ as follows: If Y_i contains only one of the elements x, y or z, we choose T_i such that this element is directly connected to the root, say on the right-hand side of T_i and all other taxa of Y_i are contained in the left-hand side of T_i. If Y_i contains two of the elements x, y or z, we choose T_i such that these two elements form a cherry which is directly connected to the root, say on the right-hand side of T_i and all other taxa of Y_i are contained in the left-hand side of T_i. Note that no Y_i can contain all three elements x, y and z as $\{x, y, z\}$ is a CT. The construction of T_i is depicted in Figures 1 and 2.

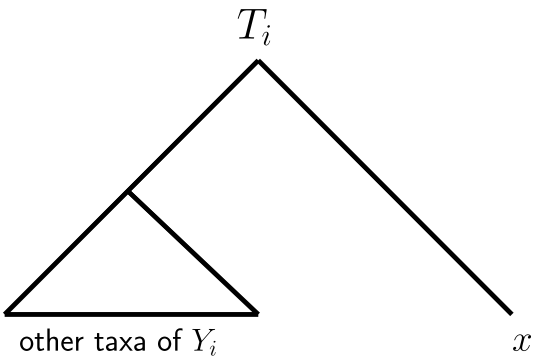

Fig. 1: Construction of T_i for the case where $x \in Y_i$ and $y, z \notin Y_i$.

Note that we do not specify the rest of the left-hand side subtrees of T_i or the trees T_i for those subsets Y_i which do not contain x, y or z. They can all be chosen arbitrarily as long as one makes sure that all T_i are compatible.

Now since the T_i are all compatible, we can combine them into a supertree $\mathcal{T}$. However, by construction, $\mathcal{T}$ cannot be unique as the relationship of x, y and z cannot be resolved: no T_i contains all of them, and those T_i which contain up two of them just bear the information that they are equally unrelated to the other taxa. Thus, $\{x, y, z\}$ is not resolved and therefore there is no unique supertree of the T_i. Therefore, $\mathcal{T}$ is not the only tree that displays the T_i. An example of two trees $\mathcal{T}$ and $\mathcal{T}'$ displaying the T_i as constructed above is depicted in Figure 3.

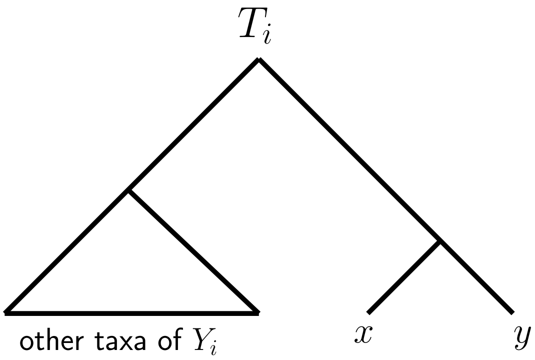

Fig. 2: Construction of T_i for the case where $x, y \in Y_i$ and $z \notin Y_i$.

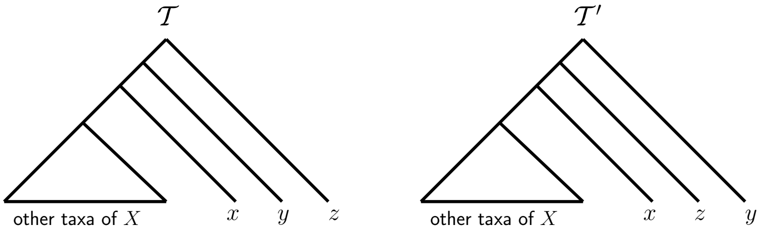

Fig. 3: Two possible supertrees for the T_i showing two different resolutions of the cross triple $\{x, y, z\}$.

The fact that $\mathcal{T}$ is not the only tree displaying the T_i contradicts the fact that $\mathcal{S}$ is phylogenetically decisive. This completes the proof.

□

Next, we state a sufficient criterion for phylogenetic decisiveness.

Proposition 2 *Let $\mathcal{S} = \{Y_1, \ldots, Y_k\}$ be a set of subsets of X such that there is no cross triple with respect to any partition of $\mathcal{S}$. Then, $\mathcal{S}$ is phylogenetically decisive.*

Proof Let $\mathcal{S} = \{Y_1, \ldots, Y_k\}$ be such that there is no cross triple with respect to any partition of $\mathcal{S}$. Assume $\mathcal{S}$ is not phylogenetically decisive. Then there exists a choice of trees $T_1, \ldots, T_k$ on $Y_1, \ldots, Y_k$, such that these trees are represented by at least two different supertrees $\mathcal{T}_1$ and $\mathcal{T}_2$. Since these two trees are different, they resolve at least one triple $\{x, y, z\}$ of taxa in X differently, and thus in particular this triple $\{x, y, z\}$ is not resolved by any of the input trees $T_1, \ldots, T_k$. But as there is no CT wrt $\pi = Y_1|Y_2|\ldots|Y_k$, i.e. the partition which splits all elements of $\mathcal{S}$ apart, by Definition 4, all triples are subset of at least one Y_i. So $\{x, y, z\} \subseteq Y_i$ for some $i \in \{1, \ldots, k\}$, and therefore $\{x, y, z\}$ is resolved by T_i. This is a contradiction and thus completes the proof.

□

Now we show a useful property of cross triples, which is needed in the following.

Proposition 3 *Let $\mathcal{S} = \{Y_1, \ldots, Y_k\}$ be a set of taxon sets with $Y_i \subseteq X$ for all $i = 1, \ldots, k$. Let $\pi = Y_1|Y_2|\ldots|Y_k$ be the partition of $\mathcal{S}$ which splits all elements of $\mathcal{S}$ apart. Then, if there is no cross triple of $\mathcal{S}$ with respect to π, there are no cross triples with respect to any other partitions of $\mathcal{S}$, either.*

Proof Assume there is no cross triple of $\mathcal{S}$ with respect to $\pi = Y_1|Y_2|\ldots|Y_k$. Let $\phi = S_1|S_2|\ldots|S_m$ be a partition of $\mathcal{S}$ with $m < k$, i.e. each Y_j belongs to exactly one S_i, but each S_i may contain more than one Y_j. Now suppose there is a CT $\{x, y, z\}$ of $\mathcal{S}$ wrt ϕ. By Definition 4, this implies that no S_i contains x, y and z together. Thus, in particular no Y_j can contain x, y and z together. Then, again by Definition 4, if there is no Y_j which contains x, y and z together, this implies that $\{x, y, z\}$ is a CT with respect to $\pi = Y_1|Y_2|\ldots|Y_k$. This contradicts the assumption and thus completes the proof.

□

We are now in a position to state a novel characterization of phylogenetic decisiveness in terms of the following corollary.

Corollary 1 *A set $\mathcal{S} = \{Y_1, \ldots, Y_k\}$ of taxon sets with $Y_i \subseteq X$ for all $i = 1, \ldots, k$ is phylogenetically decisive if and only if there is no cross triple of $\mathcal{S}$ with respect to $\pi = Y_1|Y_2|\ldots|Y_k$.*

Proof

1. Let $\mathcal{S} = \{Y_1, \ldots, Y_k\}$ be a set of taxon sets with $Y_i \subseteq X$ for all $i = 1, \ldots, k$ which is phylogenetically decisive. Then, by Proposition 1, there are no cross triples of $\mathcal{S}$ with respect to any partition of $\mathcal{S}$. This implies in particular that there is no cross triple with respect to the partition $\pi = Y_1|Y_2|\ldots|Y_k$.
2. Let $\mathcal{S} = \{Y_1, \ldots, Y_k\}$ be a set of taxon sets with $Y_i \subseteq X$ for all $i = 1, \ldots, k$ such that there is no cross triple of $\mathcal{S}$ with respect to the partition $\pi = Y_1|Y_2|\ldots|Y_k$ which splits all elements of $\mathcal{S}$ apart. By Proposition 3, this implies that there is no cross triple with respect to any partition of $\mathcal{S}$. By Proposition 2, it follows that $\mathcal{S}$ is phylogenetically decisive. This completes the proof.

□

Now we can prove the main theorem of this section. In the proof, we also provide an algorithm for solving Problem 1.

Proof (Main Theorem 1) Let $\mathcal{S} = \{Y_1, \ldots, Y_k\}$ be a set of subsets of X, where $|X| = n$. We want to decide if $\mathcal{S}$ is phylogenetically decisive. By Corollary 1 this means we have to check if there is a cross triple of $\mathcal{S}$ with respect to $\pi = Y_1|Y_2|\ldots|Y_k$. We now provide an $\mathcal{O}(k \cdot n^3)$ algorithm.

We assume an arbitrary ordering of $\binom{X}{3} = \{Y : Y \subseteq X \text{ and } |Y| = 3\}$ and we continue as follows:

```
BEGIN
    Initialization: Marker($i$) := 0 for all $i = 1, \ldots, k$.
    For $i = 1, \ldots, |\binom{X}{3}| = \frac{1}{6}(n^3 - 3n^2 + 2n)$
        For $j = 1, \ldots, k$
            Check if element $i$ of $\binom{X}{3}$ is a subset of $Y_j$. If yes: Marker($i$) := 1.
        If Marker($i$) == 0: STOP. $\mathcal{S}$ is *not* phylogenetically decisive.
    If Marker($i$) == 1 for all $i = 1, \ldots, k$: $\mathcal{S}$ is phylogenetically decisive.
END
```

So the algorithm checks for all subtriples of X if it is a cross triple (in this case, its marker remains 0) or not (in this case, the marker is changed to 1). If a cross triple is found, the algorithm stops immediatly as then $\mathcal{S}$ is not phylogenetically decisive. If, on the other hand, after checking all triples no cross triple is found, $\mathcal{S}$ is phylogenetically decisive. Note that the checking step can be done with optimal hash tables in $\mathcal{O}(1)$, i.e. constant time (where the hash table build-up has complexity $\mathcal{O}(|Y_j|)$). Now as this is repeated $k \cdot \frac{1}{6}(n^3 - 3n^2 + 2n)$ times, the overall time needed is bounded by $\mathcal{O}(k \cdot n^3)$. This completes the proof.

□

Note that Theorem 1 can also be proven in a different way by using a result on the unrooted tree case by Steel and Sanderson **?** and regarding the root as an outgroup taxon. This idea is further explained in Remark 1

Now we provide a simple example for a phylogenetically decisive set of taxon sets in the rooted sense.

Example 1 Let $X = \{1,2,3,4\}$. Then, by Main Theorem 1, the set $\mathcal{S} := \{\{1,2,3\},\{1,2,4\},\{1,3,4\},\{2,3,4\}\}$ is phylogenetically decisive, because it contains all possible triplet subsets of X and thus there is no cross triple with respect to any partition.

Note that Example 1 shows that if there is no cross triple with respect to any partition, this does not imply that $\mathcal{S}$ has to contain X. The latter case would be a trivial case, because then also a supertree of the trees for all $Y_i \in \mathcal{S}$ would equal the tree provided for X. In this respect, Example 1 shows that there are non-trivial ways to keep cross triples out of taxon sample set and thus to ensure that all collections compatible input trees will lead to a unique supertree.

In the following, we consider the case of unrooted trees, which is a bit more complex than the rooted case.

3.2 Unrooted tree case

By considering Example 1, one can easily see that a set $\mathcal{S}$ that is phylogenetically decisive for the case where the input tree topologies chosen are rooted need not be phylogenetically decisive for unrooted input tree topologies. In the above example, all sets in $\mathcal{S}$ contain only three taxa, which implies for the unrooted case that all chosen input tree topologies can only be star trees and thus cannot lead to a unique supertree (as star trees are compatible with all possible supertrees). So the concept of phylogenetic decisiveness cannot be directly transferred from rooted to unrooted trees. However, we now show that Proposition 2 can be generalized to the unrooted case when regarding quadruples instead of triples.

Proposition 4 *Let $\mathcal{S} = \{Y_1, \ldots, Y_k\}$ be a set of subsets of X such that there is no cross quadruple with respect to any partition of $\mathcal{S}$. Then, $\mathcal{S}$ is phylogenetically decisive.*

Proof Let $\mathcal{S} = \{Y_1, \ldots, Y_k\}$ be such that there is no cross quadruple with respect to any partition of $\mathcal{S}$. Assume $\mathcal{S}$ is not phylogenetically decisive. Then there exists a choice of trees $T_1, \ldots, T_k$ on $Y_1, \ldots, Y_k$, such that these trees are represented by at least two different supertrees $\mathcal{T}_1$ and $\mathcal{T}_2$. Since these two trees are different, they resolve at least one quadruple $\{a,b,c,d\}$ of taxa in X differently, and thus in particular this quadruple $\{a,b,c,d\}$ is not resolved by any of the input trees $T_1, \ldots, T_k$. But as there is no CQ wrt the partition $\pi = Y_1|Y_2|\ldots|Y_k$, which splits all elements of $\mathcal{S}$ apart, by Definition 5, all quadruples are subset of at least one Y_i. So $\{a,b,c,d\} \subseteq Y_i$ for some

$i \in \{1, \ldots, k\}$, and therefore $\{a, b, c, d\}$ is resolved by T_i. This is a contradiction and thus completes the proof.

□

Note that for the unrooted case, a characterization of phylogenetic decisiveness was already given by Steel and Sanderson (2010) in terms of the following theorem.

Theorem 2 (Steel and Sanderson (2010), 'Four-way partition property') *A collection $\mathcal{S} = \{Y_1, \ldots, Y_k\}$ of subsets of a taxon set X is phylogenetically decisive (in the unrooted sense) if and only if it satisfies the four-way partition property, i.e. if for all partitions $\pi = X_1|X_2|X_3|X_4$ of X into four non-empty subsets, there exist taxa $x_i \in X_i$ (for $i = 1, 2, 3, 4$) such that the quadruple $\{x_1, x_2, x_3, x_4\}$ is contained in Y_j for some $j \in \{1, \ldots, k\}$.*

However, if one wants to decide if a particular set of taxon sets is phylogenetically decisive, checking all partitions of the set into four subsets is computationally expensive. Therefore, we subsequently develop another characterization.

Remark 1 Theorem 2 gives rise to a different view on the rooted tree case. Given a set of taxon sets $\mathcal{S} = \{Y_1, \ldots, Y_k\}$, one can attach to the root of each input tree T_i on Y_i a pending edge and label the resulting leaf, say, x_ρ. Accordingly, for all $i = 1, \ldots, k$, we define $\tilde{Y}_i := Y_i \cup \{x_\rho\}$ and $\tilde{X} := X \cup \{x_\rho\}$. Then a rooted triple $\{a, b, c\} \subset X$ can be regarded as unrooted quadruple $\{x_\rho, a, b, c\} \subset \tilde{X}$. By Theorem 2, one can find an alternative approach to prove Theorem1: It can be easily checked that $\tilde{\mathcal{S}} := \{\tilde{Y}_1, \ldots, \tilde{Y}_k\}$ fulfills the four-way partition property if and only if there is no cross triple of $\mathcal{S}$ with respect to $Y_1|Y_2|\ldots|Y_k$.

Next we state an additional definition to simplify the terminology for the subsequent proofs.

Definition 7 Let X be a taxon set and $\mathcal{S} = \{Y_1, \ldots, Y_k\}$ be a set of subsets of X. Then, a quadruple $\{a, b, c, d\} \subseteq X$ is called a *global cross quadruple* or just *cross quadruple* for short if it is a cross quadruple with respect to $\pi = Y_1|\ldots|Y_k$, i.e. if $\{a, b, c, d\} \nsubseteq Y_i$ for all $i = 1, \ldots, k$.

Before deriving a new characterization of phylogenetic decisiveness for the unrooted case, we show that Theorem 1 cannot be generalized to this case by just translating triples to quadruples. For this example, we use the characterization provided by Theorem 2.

Example 2 Let $X = \{1,2,3,4,5\}$. Then, the set $\mathcal{S} := \{\{1,2,3,4\},\{1,2,4,5\},\{1,3,4,5\},\{2,3,4,5\}\}$ is phylogenetically decisive. This can be verified by Theorem 2 as follows: four out of the five possible quadruples form elements of $\mathcal{S}$ and are therefore resolved by all possible input tree collections. So the four-way partition property is fulfilled for all partitions $\pi = X_1|X_2|X_3|X_4$ of X which split any of these quadruples apart. However, the only CQ, namely $\{1,2,3,5\}$ could be critical: what if this quadruple gets split apart by a 4-partition? For this case, we need to verify the 4-way partition property. Wlog. $1 \in X_1, 2 \in X_2, 3 \in X_3, 5 \in X_4$. It can be easily verified that for all $i = 1,2,3,4$, if taxon 4 is in X_i, the 4-way partition property holds and therefore, $\mathcal{S}$ is phylogenetically decisive even though there is a cross quadruple. Thus, in the unrooted case there may be CQs even in phylogenetically decisive sets of taxon sets.

The above example shows that the unrooted case is more complicated than the rooted case. In the rooted case, phylogenetic decisiveness is equivalent to the absence of cross triples, but in the unrooted case, cross quadruples do not necessarily destroy the phylogenetic decisiveness. However, this is only true if the cross quadruples have a certain property. Informally speaking, a cross quadruple is a quadruple unresolved by the input trees – so in order for it to be resolved uniquely by all possible supertrees, there must be additional information on the quadruple somewhere in the input sets. Therefore, we now introduce the notion of so-called fixing taxa.

Definition 8 Let X be a set of taxa and $\mathcal{S} = \{Y_1,\ldots,Y_k\}$ be a set of subsets of X. Let $\{a,b,c,d\}$ be a global CQ. Then, taxon $x \in X \setminus \{a,b,c,d\}$ is called a *fixing taxon* of $\{a,b,c,d\}$, if for each of the four sets $\{a,b,c,x\}$, $\{a,b,d,x\}$, $\{a,c,d,x\}$ and $\{b,c,d,x\}$ there exists a $j \in \{1,\ldots,k\}$ such that this set is contained in Y_j, respectively.

We now show the role of fixing taxa in resolving cross quadruples.

Proposition 5 *Let X be a set of taxa and $\mathcal{S} = \{Y_1,\ldots,Y_k\}$ be a set of subsets of X. Let $\{a,b,c,d\}$ be a global CQ of $\mathcal{S}$ with fixing taxon $x \in X$. Then, any assignment of compatible trees $T_1,\ldots,T_k$ on the taxon sets $Y_1,\ldots,Y_k$ resolves $\{a,b,c,d\}$ in a unique way, i.e. for all pairs of supertrees $\mathcal{T}$, $\tilde{\mathcal{T}}$of $T_1,\ldots,T_k$, we have: $\mathcal{T}|_{\{a,b,c,d\}} = \tilde{\mathcal{T}}|_{\{a,b,c,d\}}$.*

Proof Let $\{a,b,c,d\}$ be a global CQ of $\mathcal{S}$ with fixing taxon $x \in X$. Then, for any assignment of compatible trees $T_1,\ldots,T_k$ on the taxon sets $Y_1,\ldots,Y_k$, the sets $\{a,b,c,x\}$, $\{a,b,d,x\}$, $\{a,c,d,x\}$ and $\{b,c,d,x\}$ are all resolved by the definition of a fixing taxon. In the following, a quartet tree on taxa $\{w,x,y,z\}$ inducing the split $wx|yz$ is identified with this split in order to simplify the notation. Then, it can then be easily checked that given a compatible collection $\mathcal{Q}$ of quartet splits, another quartet split $wx|yz \notin \mathcal{Q}$ may be implied (cf. quartet rules in Semple and Steel (2003), p. 129 as well as in Steel (1992)). We now examine each possible input quartet split of the resolved sets $\{a,b,c,x\}$, $\{a,b,d,x\}$, $\{a,c,d,x\}$ and $\{b,c,d,x\}$ to see a) if the respective set is compatible and b) if yes, if and how it resolves the CQ $\{a,b,c,d\}$:

1. $\{ab|cx, ab|dx\} \Rightarrow ab|cd$
2. $\{ab|cx, ad|bx\} \Rightarrow ad|bc$
3. $\{ab|cx, ax|bd\} \Rightarrow ac|bd$
4. $\{ac|bx, ab|dx\} \Rightarrow ac|bd$
5. $\{ac|bx, ad|bx, ac|dx\} \Rightarrow ac|bd$
6. $\{ac|bx, ad|bx, ad|cx\} \Rightarrow ad|bc$
7. $\{ac|bx, ad|bx, ax|cd\}$ incompatible
8. $\{ac|bx, ax|bd\} \Rightarrow ac|bd$
9. $\{ax|bc, ab|dx\} \Rightarrow ad|bc$
10. $\{ax|bc, ad|bx\} \Rightarrow ad|bc$
11. $\{ax|bc, ax|bd, ac|dx\}$ incompatible
12. $\{ax|bc, ax|bd, ad|cx\}$ incompatible
13. $\{ax|bc, ax|bd, ax|cd, bc|dx\} \Rightarrow ad|bc$
14. $\{ax|bc, ax|bd, ax|cd, bd|cx\} \Rightarrow ac|bd$
15. $\{ax|bc, ax|bd, ax|cd, bx|cd\} \Rightarrow ab|cd$

So in all cases where the input quartets are compatible, the CQ $\{a,b,c,d\}$ is uniquely resolved. This completes the proof.

□

Remark 2 Note that in 13 – 15, the first three quartets are not enough to resolve the CQ $\{a,b,c,d\}$ uniquely, so the fixing taxon of all four subtriples of the CQ is needed.

Example 3 (Example 2 continued.) In Example 2, taxon 4 is a fixing taxon of the only CQ $\{1,2,3,5\}$. Thus, by Proposition 3.2, this CQ is resolved in the same way in all possible supertrees of the trees corresponding to the taxon sets in $\mathcal{S}$.

However, the existence of a fixing taxon is sufficient but not necessary for a cross quadruple to be resolved. This is demonstrated by the following example.

Example 4 Let $X = \{1,2,3,4,5,6\}$ and $\mathcal{S} := \{\{1,2,3,5\},\{1,2,4,5\},\{1,2,4,6\},$ $\{1,2,5,6\},\{1,2,3,6\},\{1,3,4,6\},\{1,3,5,6\},\{1,4,5,6\},\{2,3,4,5\},\{2,3,4,6\},$

$\{2,3,5,6\}\}$. In this case, $\mathcal{S}$ has four CQs: $\{1,2,3,4\}$, $\{1,3,4,5\}$, $\{2,4,5,6\}$ and $\{3,4,5,6\}$. It can be checked (calculation not shown) that $\mathcal{S}$ is phylogenetically decisive by verifying the four-way partition property (see Theorem 2) explicitely for all possible partitions of X. However, while the CQs $\{1,2,3,4\}$ and $\{2,4,5,6\}$ have fixing taxa 6 and 1, respectively, the other two CQs do not have any fixing taxa.

Note that cross quadruples are by definition sets of four taxa which are not resolved by any input tree, but by Proposition a cross quadruple is resolved by a supertree of the input trees if it has a fixing taxon. The main idea we present below in Theorem 3 is that cross quadruples can be iteratively resolved: if a cross quadruple has no fixing taxon, but, say, taxon x would be a fixing taxon if another quadruple was resolved, which in turn does have a fixing taxon, then the original cross quadruple can 'inherit' the resolvedness by resolving the second quadruple first. Thus, we need to distinguish directly and indirectly resolved cross quadruples.

Definition 9 Let X be a set of taxa and $\mathcal{S} = \{Y_1, \dots, Y_k\}$ be a set of subsets of X. Let $\{a,b,c,d\}$ be a global CQ of $\mathcal{S}$. Then, we call $\{a,b,c,d\}$ *directly resolved*, if it has a fixing taxon. We call $\{a,b,c,d\}$ *indirectly resolved*, if there is a taxon $x \in X$ such that for each the four sets $\{a,b,c,x\}$, $\{a,b,d,x\}$, $\{a,c,d,x\}$ and $\{b,c,d,x\}$ one of the following conditions holds:

1. The set is not a cross quadruple.
2. The set is a cross quadruple but has a fixing point.
3. The set is a cross quadruple and is itself indirectly resolved.

We now state some helpful characteristics of CQs of phylogenetically decisive sets of taxon sets in order to derive a characterization of phylogenetically decisive sets in the unrooted case.

Proposition 6 *Let $X = \{1, \dots, n\}$ be a set of at least four taxa and $\mathcal{S} = \{Y_1, \dots, Y_k\}$ be a phylogenetically decisive set of subsets of X. Then each triple $\{x,y,z\} \subset X$ is contained in at least one Y_j with $|Y_j| \geq 4$.*

Proof Let $\mathcal{S} = \{Y_1, \dots, Y_k\}$ be phylogenetically decisive. Assume there exist three taxa $x,y,z \in X$ such that for all j with $|Y_j| \geq 4$: $\{x,y,z\} \not\subseteq Y_j$. Let $X_1 := \{x\}$, $X_2 := \{y\}$, $X_3 := \{z\}$ and $X_4 := X \setminus \{x,y,z\}$. Then, $\pi := X_1|X_2|X_3|X_4$ partitions X into four non-empty subsets. Now assume there is no taxon a in X – and thus neither in X_4 – such that $\{x,y,z,a\} \subseteq Y_j$. Then, the four-way partition property fails for π. Thus, $\mathcal{S}$ is not phylogenetically decisive. This is a contradiction and thus completes the proof.

□

Definition 10 Let X be a set of taxa and $\mathcal{S} = \{Y_1, \dots, Y_k\}$ be a set of subsets of X. Then, the *colored 3-overlap graph* of $\mathcal{S}$ is a graph where the nodes are all possible quadruples in $\binom{X}{4} = \{Y : Y \subseteq X \text{ and } |Y| = 4\}$, and where two nodes are connected with an edge if the corresponding quadruples share three taxa. Moreover, all nodes are either colored red or green: they are green if they are contained in at least one Y_j and red otherwise (note: the red ones are the global cross quadruples).

Example 5 We continue Example 4. $X = \{1,2,3,4,5,6\}$ and $\mathcal{S} := \{\{1,2,3,5\}, \{1,2,4,5\}, \{1,2,4,6\}, \{1,2,5,6\}, \{1,2,3,6\}, \{1,3,4,6\}, \{1,3,5,6\}, \{1,4,5,6\}, \{2,3,4,5\}, \{2,3,4,6\}, \{2,3,5,6\}\}$. As stated before, $\mathcal{S}$ has four CQs: $\{1,2,3,4\}$, $\{1,3,4,5\}$, $\{2,4,5,6\}$ and $\{3,4,5,6\}$. We construct the colored 3-overlap graph as depicted by Figure 4. Note that the CQs are all red.

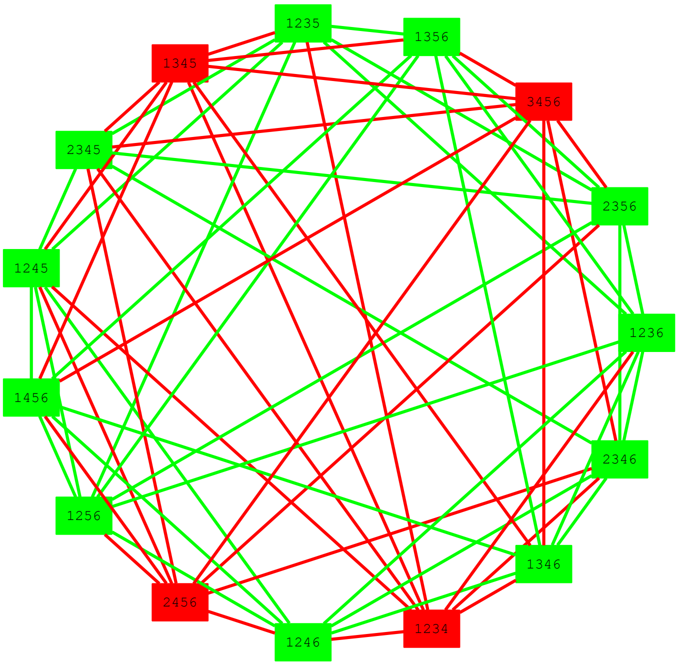

Fig. 4: The colored 3-overlap graph.

We are now in a position to prove a characterization for resolved global cross quadruples.

Theorem 3 *Let X be a set of taxa and $\mathcal{S} = \{Y_1, \dots, Y_k\}$ be a set of subsets of X. Let $\{a, b, c, d\}$ be a global CQ of $\mathcal{S}$. Then, $\{a, b, c, d\}$ is uniquely resolved by all supertrees of compatible input trees $T_1, \dots, T_k$ on $Y_1, \dots, Y_k$ if and only if it is directly or indirectly resolved.*

Proof

- Assume $\{a, b, c, d\}$ is directly or indirectly resolved. If it is directly resolved, it has a fixing taxon and thus, by Proposition 3.2, there remains nothing to show. If it is indirectly resolved, one can apply Proposition 3.2 iteratively in order to derive a chain of resolved quadruples (starting with one that has a fixing taxon and contains a taxon which acts as a fixing taxon for another cross quadruple, and so on) until $\{a, b, c, d\}$ is resolved.
- Assume $\{a, b, c, d\}$ is a global CQ of $\mathcal{S}$ but is uniquely resolved by all supertrees for all possible combinations of compatible input trees $T_1, \dots, T_k$ on the taxon sets $Y_1, \dots, Y_k$. Assume $\{a, b, c, d\}$ is not directly or indirectly resolved. Then there is no fixing taxon resolving $\{a, b, c, d\}$ or other cross quadruples which would in turn deliver a fixing taxon for $\{a, b, c, d\}$. Therefore, for all $x \in X$, at most three of the sets $\{a, b, c, x\}$, $\{a, b, d, x\}$, $\{a, c, d, x\}$ and $\{b, c, d, x\}$ in the colored 3-overlap graph are either green or red but directly or indirectly resolved. This implies that for all $x \in X$, at least one of the sets $\{a, b, c, x\}$, $\{a, b, d, x\}$, $\{a, c, d, x\}$ and $\{b, c, d, x\}$ is unresolved.
  Then, all of these four quadruples which contain three taxa of $\{a, b, c, d\}$ and some taxon $x \in X$ also contain at least one more taxon which lies in all of them: e.g. if $\{b, c, d, x\}$ is not resolved, but $\{a, b, d, x\}$, $\{a, c, d, x\}$ and $\{a, b, d, x\}$ are resolved, the intersection of the resolved sets contains both x and a. In other words, considering the four quadruples $\{a, b, c, x\}$, $\{a, b, d, x\}$, $\{a, c, d, x\}$ and $\{b, c, d, x\}$ and assuming that at least one of them is not directly or indirectly resolved, the intersection of the resolved quadruples has at least cardinality 2. So assume for some $x \in X$ without loss of generality that the intersection of the three sets contains x and a. We now assume for some $x \in X$ without loss of generality that only the sets $\{a, b, c, x\}$, $\{a, b, d, x\}$ and $\{a, c, d, x\}$ (or just one or two of them) are either green or red but directly or indirectly resolved. Note that if they are green, they are contained in one of the input sets Y_j, and if they are red but directly or indirectly resolved, all possible compatible choices of $T_1, \dots, T_k$ lead to a unique resolution of them by the first part of the proof. Thus, choose $T_1, \dots, T_k$ such that they are compatible trees on the taxon sets $Y_1, \dots, Y_k$ and such that all their supertrees display the mentioned sets as follows: $ax|bc, ax|bd$ and $ax|cd$. For example, we can construct the input trees such that x and a are put on a cherry. Then, the resolutions $ab|cd, ac|bd$ and $ad|bc$ are all possible in supertrees of $T_1, \dots, T_k$. Therefore, if $\mathcal{T}$ is a supertree of $T_1, \dots, T_k$ such that $\mathcal{T}|_{\{a,b,c,d\}} = ab|cd$, then also $\tilde{\mathcal{T}}$ is a supertree of $T_1, \dots, T_k$, where we define $\tilde{\mathcal{T}}$ as follows: $\tilde{\mathcal{T}}|_{\{a,b,c,d\}} = ac|bd$

and $\tilde{\mathcal{T}}|_{X\setminus\{a,b,c,d\}} = \mathcal{T}|_{X\setminus\{a,b,c,d\}}$. Thus, $\{a,b,c,d\}$ is not resolved for this choice of $T_1,\ldots,T_k$. This is a contradiction and thus completes the proof.

□

Next we state a straightforward characterization of phylogenetic decisiveness, before we combine this characterization with the previous theorem to obtain Main Theorem 2.

Proposition 7 *Let X be a set of taxa and $\mathcal{S} = \{Y_1,\ldots,Y_k\}$ be a set of subsets of X. Then, $\mathcal{S}$ is phylogenetically decisive if and only if all global cross quadruples are directly or indirectly resolved.*

Proof

1. Assume all global cross quadruples of $\mathcal{S}$ are resolved. Assume $\mathcal{S}$ is not phylogenetically decisive. Then there exist compatible trees $T_1,\ldots,T_k$ on $Y_1,\ldots,Y_k$, such that there are two different trees $\mathcal{T}_1$ and $\mathcal{T}_2$ that display $T_1,\ldots,T_k$. As $\mathcal{T}_1$ and $\mathcal{T}_2$ are different, there is at least one quadruple $\{a,b,c,d\}$ which is resolved differently in $\mathcal{T}_1$ and $\mathcal{T}_2$. Then, in particular $\{a,b,c,d\} \not\subseteq Y_j$ for all $j = 1,\ldots,k$, and thus, by definition, $\{a,b,c,d\}$ is a global cross quadruple. Thus, $\{a,b,c,d\}$ must be resolved for all choices of input trees $T_1,\ldots,T_k$. This contradicts the fact that $\mathcal{T}_1$ and $\mathcal{T}_2$ are supertrees of $T_1,\ldots,T_k$ but display conflicting resolutions of $\{a,b,c,d\}$. This completes the proof.
2. Let X be a set of taxa and $\mathcal{S} = \{Y_1,\ldots,Y_k\}$ be a phylogenetically decisive set of subsets of X. Then, if there are no unresolved cross quadruples, there is nothing to show. So assume there exists a global cross quadruple $\{a,b,c,d\}$ which is not resolved. Then, there exists a choice of trees $T_1,\ldots,T_k$ on taxon sets $Y_1,\ldots,Y_k$ such that there are at least two supertrees $\mathcal{T}_1$ and $\mathcal{T}_2$ displaying two different resolutions of the quadruple $\{a,b,c,d\}$, wlog. $\mathcal{T}_1|_{\{a,b,c,d\}} = ab|cd$ and $\mathcal{T}_2|_{\{a,b,c,d\}} = ac|bd$. So clearly, $\mathcal{T}_1 \neq \mathcal{T}_2$, but both are supertrees of $T_1,\ldots,T_k$. This contradicts the phylogenetic decisiveness of $\mathcal{S}$ and thus completes the proof.

□

We are now in a position to state Main Theorem 2, which shows that even if the setting for unrooted trees is more complicated than the one for rooted trees, the decision whether or not a given set of taxon sets is phylogenetically decisive can still be made in polynomial time.

Theorem 4 (Main theorem 2) *Given a set $\mathcal{S} = \{Y_1, \dots, Y_k\}$ of subsets of X, where $|X| = n$, the question whether $\mathcal{S}$ is phylogenetically decisive can be answered in at most $\mathcal{O}(k \cdot n^{16})$ steps. In particular, as long as k is polynomial in n, the question can be answered in polynomial time.*

Proof (Main Theorem 2) Let $\mathcal{S} = \{Y_1, \dots, Y_k\}$ be a set of subsets of X, where $|X| = n$. We want to decide if $\mathcal{S}$ is phylogenetically decisive.

1. We start by generating the colored 3-overlap graph of the quadruples as follows: For each quadruple $\{a,b,c,d\} \in \binom{X}{4}$, we draw a node and color it green if there is a $j \in \{1, \dots, k\}$ such that $\{a,b,c,d\} \subseteq Y_j$ and red else (Note: This step is bounded by $k \cdot \binom{|X|}{4} < k \cdot n^4$). We then draw the edges of the graph as follows: For any two nodes, we connect them if and only if they share three taxa (Note: This step is bounded by $\binom{|X|}{4} \cdot (\binom{|X|}{4} - 1) < n^8$).
2. All red nodes are global CQs of $\mathcal{S}$ (Note: Their number is bounded by $\binom{|X|}{4} < n^4$). They need to be checked for fixing taxa. So as long as there are red nodes, check for each CQ $\{a,b,c,d\}$ if it has a fixing taxon, i.e. if there is a taxon $x \in X$ such that $\{a,b,c,x\}$, $\{a,b,d,x\}$, $\{a,c,d,x\}$ and $\{b,c,d,x\}$ are all green (Note: This step cannot take more than $\binom{|X|}{4} \cdot 4 < 4n^4$ steps, so the entire order of this step is bounded by $n^4 \cdot 4n^4$, so that this step can be done in $\mathcal{O}(n^{16})$). If there are red nodes but none of them has a fixing taxon, STOP: $\mathcal{S}$ is not phylogenetically decisive. For all red nodes with a fixing taxon, change their color to green. Repeat until there are only green nodes left. In this case, $\mathcal{S}$ is phylogenetically decisive.

Clearly, this algorithm has a polynomial running time as long as k is not exponential in n and it terminates once all nodes are colored green or no more red node has a fixing taxon. The correctness of the algorithm follows from Proposition 7 and Theorem 3.

□

Informally speaking, the algorithm presented in the proof of Main Theorem 2 works as follows: If there are no cross quadruples, all nodes in the colored 3-overlap graph are green and $\mathcal{S}$ is phylogenetically decisive. However, if there are red nodes, i.e. there are cross quadruples, the algorithm checks if they can be resolved. Each cross quadruple with a fixing taxon can be resolved and thus is colored green. These newly green-colored nodes might help to deliver a fixing taxon to a previously unresolved node in the graph. So the red neighbors of a newly colored node need to be checked again for being resolvable. Once nothing can be recolored, the algorithm stops. If then everything is green, $\mathcal{S}$ is phylogenetically decisive, else not. We now demonstrate the algorithm with a short example.

Example 6 We continue Examples 4 and 5. The colored 3-overlap graph is depicted by Figure 4. Now in the first iteration, we check all red nodes for fixing taxa. It turns out that the CQ $\{1,2,3,4\}$ has fixing taxon 6 and $\{2,4,5,6\}$ has fixing taxon 1. Therefore, these two nodes change their color from red to green, see Figure 5. Now for the second iteration, there are still two red nodes, namely $\{1,3,4,5\}$ and $\{3,4,5,6\}$. We check the graph and find that since now $\{1,2,3,4\}$ is green and $\{2,3,4,5\}$, $\{1,2,3,5\}$ and $\{1,2,4,5\}$ were green right from the beginning, now taxon 2 acts as a fixing taxon for $\{1,3,4,5\}$. Thus, we can change the color from $\{1,3,4,5\}$ from red to green. Similarly, 2 now also acts as a fixing taxon for $\{3,4,5,6\}$, as in the previous iteration $\{2,4,5,6\}$ turned green. So we can now change the color of $\{3,4,5,6\}$ from red to green. Now there are only green nodes left as shown in Figure 6, which means that $\mathcal{S}$ is phylogenetically decisive. This verifies the above result derived with the help of the four-way partition property by Steel and Sanderson (2010).

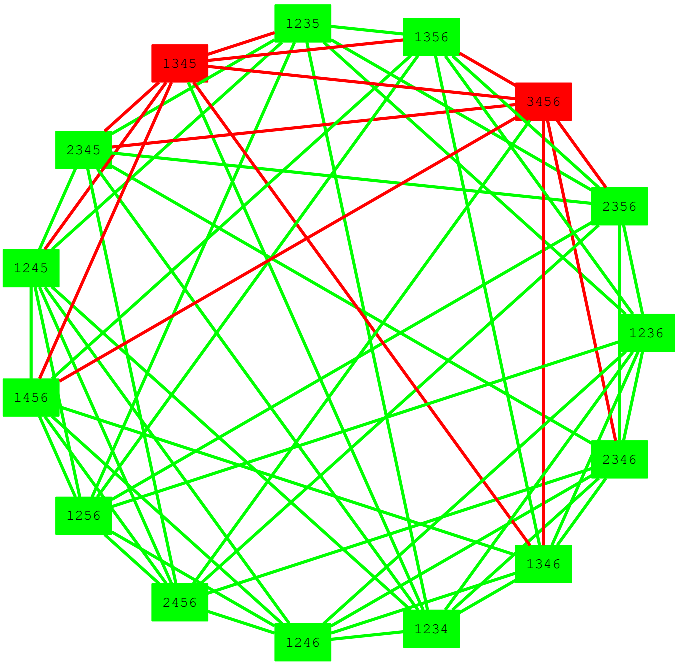

Fig. 5: The colored 3-overlap graph after the first iteration.

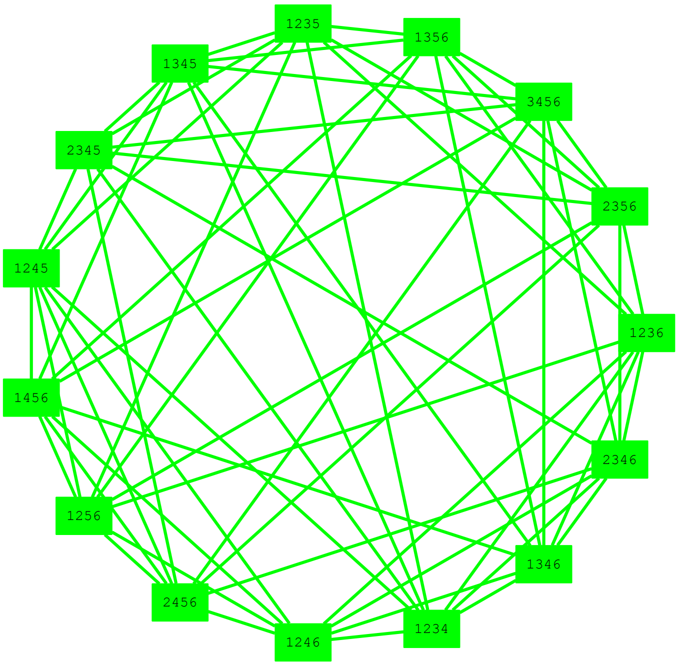

Fig. 6: The colored 3-overlap graph after the second iteration: all nodes are green, so $\mathcal{S}$ is phylogenetically decisive.

4 Discussion

Deriving new information on taxa by combining various compatible trees on different taxon sets into one common supertree is a challenging task, even if the input taxon sets are compatible with one another. It is therefore understandable that phylogeneticists seek to understand a priori if a particular set of input trees has the potential to deliver new information – in the perfect case even to provide a unique supertree. In the perfect case, the set of input taxon sets is called phylogenetically decisive.

In our paper, we provide algorithms (both for the rooted and unrooted tree case) to determine whether a set of taxon sets is phylogenetically decisive, and the algorithms are polynomial as long as the number of taxon sets under investigation are polynomial in the total number of taxa. Our algorithms are quite intuitive: They check for the smalles possible information unit of each setting (i.e. triples in the rooted case and quadruples in the unrooted case), if all such instances are uniquely resolved by all supertrees. If all such small units are uniquely resolved by all supertrees, the supertree itself must be unique. However, this stepwise approach in both algorithms makes them useful even if the underlying set of taxon sets turns out not to be phylogenetically decisive. For instance, particularly relevant quadruples can be investigated on their own – a quadruple can be uniquely resolved by all supertrees, even if the supertree is not unique, and this can be checked with our algorithm. Or, on the other hand, using our approach, one can determine the quadruples which cause a set of taxon sets to not be phylogenetically decisive and thus decide whether these taxa should be excluded from the study. Thus, our approach is not only suitable to determine whether a set of taxa is perfectly sampled, but also which taxa are problematic. However, we are aware that the algorithms as stated here are not runtime optimized and thus can probably be improved. This would be interesting for future work.

input taxon sets which have the property that all possible compatible input trees chosen for the input sets lead to a unique supertree. This is an interesting setting, as it shows that the decision which taxa to sample for each input tree may already ensure that the supertree of all input trees is unique. In this context, a set of taxon set which is phylogenetically decisive can also be referred to as a set of perfect taxon samples. However, the complexity of deciding whether or not a given set of taxon sets is phylogenetically decisive remained unknown.

Acknowledgements I want to thank Mike Sanderson and Mike Steel for helpful discussions on the topic, including in particular the idea presented in Remark 1. I also wish to thank the Ernst-Moritz-Arndt University Greifswald for funding.

References

Ané C, Eulenstein O, Piaggio-Talice R, Sanderson M (2009) Groves of phylogenetic trees. Ann Comb 13:139 – 167

Fischer M (2011) Mathematical aspects of phylogenetic groves. under revision

Semple C, Steel M (2003) Phylogenetics. Oxford University Press

Steel M (1992) The complexity of reconstructing trees from qualitative characters and subtrees. Journal of Classification 9:91–116

Steel M (2011) The penny ante 2012. URL `http://www.math.canterbury.ac.nz/bio/events/south2012/files/\penny_ante_problems.pdf`

Steel M, Sanderson M (2010) Characterizing phylogenetically decisive taxon coverage. Appl Math Letters 23:82–86