Master Thesis

Phylogenetic decisive and non-decisive taxon sets - analysis and comparison

Janne Pott

April 23rd, 2015

Supervised by Prof. Mareike Fischer
Department for Mathematics and Computer Science
Ernst-Moritz-Arndt University Greifswald

Contents

1. Introduction

Charles Darwin had a big idea, arguably the most powerful idea ever.
And like all the best ideas it is beguilingly simple.
Richard Dawkins, "Why Darwin matters", *The Guardian*, 9 February 2008

Charles Darwin is best known for his work on evolutionary theory: he established that all species of life have descended over time from common ancestors. In a joint publication with Alfred Russel Wallace he introduced the process of natural selection, which results in branching patterns. In the following, Darwin expressed the concept of the branching divergence of varieties and then species in a process of common descent from ancestors with the metaphor **tree of life** ([5]).

This led to the term **phylogenesis**, derivated from the Greek terms *phylé* or *phylon* for "tribe" or "race", and *genesis* for "origin" or "source", meaning the history of the evolutionary development of a species or taxonomic group of organisms ([2]). The term **phylogenetics** is the study of phylogenesis. This can be done by comparing morphological and anatomical data. The result is a phylogenetical tree.

Due to modern sequencing methods a new way of analyzing phylogenesis was established: instead of taking morphological data into account, molecular sequencing data of DNA, RNA and proteins is used.

Ever since Darwin, evolutionary biologists use tree diagrams to depict evolution. There are different types of trees, the most common are the rooted and unrooted type. The rooted phylogenetic tree has one unique node which corresponds to the most recent common ancestor of all the entities at the leaves of the tree. Usually we have no exact knowledge of that ancestor, in which case it is better to use an unrooted tree.

A **supertree** is a phylogenetic tree built out of a combination of smaller phylogenetic trees with incomplete overlapping sets of taxa ([3]). However, the complexity of the construction of a supertree increases exponentially with the number of taxa included. Therefore, Steel and Sanderson [9] stepped back from analyzing trees but looked at the taxa sets. They coined the term **phylogenetically decisive**, which means that a given collection of subsets of the taxon set uniquely determines a supertree (up to isomorphism). They found a combinatorial characterization of the covering subsets to ensure that at most one supertree can be constructed from the smaller trees, called **four-way partition property**.

As the number of these partitions increases exponentially with the number of taxa, Fischer [6] tried to find another characterization for phylogenetic decisiveness.

In this work we will first review the four-way partition property and some of its qualities, as a minimal bound and an equivalence to a possibly NP-hard problem. Then we will introduce Fischers characterization and an algorithm with its limitations and bounds. Finally, we will look at some special cases and minimal phylogenetically decisive sets of taxon sets.

2. Definitions

We begin with some basic definitions from phylogenetic theory and notations used in this work.

Definition 1 (Taxon). In biology, a *taxon* (plural *taxa*) is a group of one or more populations of an organism, which are usually inferred to be phylogenetically related (see [2]).

Definition 2 (Unrooted tree). Following Steel and Sanderson [9], given a set X of taxa, a *binary, unrooted phylogenetic X-tree T* is a connected acyclic graph in which the degree 1 vertices (*leaves* of T) consist of the set X and all the remaining vertices of T are unlabeled and of degree 3.

A *cherry* of a tree is a pair of leaves that are adjacent to the same vertex.

A *quartet tree* is an unrooted phylogenetic X-tree with exactly four taxa. It consists therefore out of two cherries and is denoted as $ab|cd$, if we have cherry ab and cherry cd.

Definition 3 (Displayed trees). Let X be a set of taxa, $X = \{1, \dots, n\}$, $Y \subset X$, and T be an unrooted phylogenetic X-tree. Then, we denote by $T|Y$ the tree which can be derived by T by deleting all elements of X which are not in Y and suppressing all nodes of degree 2. In this case we say that T *displays* $T|Y$

Definition 4 (Supertree and Compatibility). Let $Y_1, \dots, Y_k \subset X$, $k \in \mathbb{N}$, and $T_1, \dots, T_k$ be unrooted binary phylogenetic trees on $Y_1, \dots, Y_k$, respectively. If there is an unrooted binary phylogenetic tree T which displays all trees $T_1, \dots, T_k$, we call T a *supertree* of $T_1, \dots, T_k$.

Moreover, two trees T_1 and T_2 on taxon sets Y_1 and Y_2, respectively, are called *compatible* if there is a supertree T on taxon set $\bar{X} := Y_1 \cup Y_2$ displaying both T_1 and T_2.

Definition 5 (Phylogenetic decisiveness). Let S be a collection of subsets Y of a set X, and let $n = |X|$ throughout. Following Steel and Sanderson [9], we say that S is *phylogenetically decisive* if it satisfies the following property: If T and T' are binary phylogenetic X-trees, with $T|Y = T'|Y$ for all $Y \in S$, then $T = T'$. In other words, for any binary phylogenetic X-tree T, the collection of induced subtrees $\{T|Y : Y \in S\}$ uniquely determines T (up to isomorphism).

Definition 6 (Quadruple). A *quadruple* $Z = \{a, b, c, d\}$ is a subset of the taxa set X, $|X| = n \geq 4$, $a, b, c, d \in X$. A subset Y with $|Y| = m > 4$ taxa can be split into $\binom{m}{4}$ quadruples embedded in Y.

Every quadruple $\{a, b, c, d\}$ can display three quartet trees: $ab|cd$, $ac|bd$, and $ad|bc$.

Let S_n be the set of all quadruples from X that lie in at least one set in S.

$$S_n := \bigcup_{Y \in S} \binom{Y}{4} \tag{1}$$

It can be easily shown that S is phylogenetically decisive if and only if S_n is phylogenetically decisive (see [9]). In the following, we can therefore focus on sets consisting of quadruples only.

Definition 7 (Other annotations)**.** Let X be a set of n taxa, $|X| = n$.

- X_n is the set of all possible quadruples of set X: $|X_n| = \binom{n}{4}$
- Y_n is the set of all possible triples of set X: $|Y_n| = \binom{n}{3}$
- s_n is the number of quadruples of X_n that share one taxon: $s_n = \frac{4|X_n|}{n} = \binom{n-1}{3}$
- Any quadruple {*a*, *b*, *c*, *d*} will be denoted $abcd$ for short
- Any triple {*a*, *b*, *c*} will be denoted abc for short
- Any tuple {*a*, *b*} will be denoted ab for short

Let $S_n = \{Z_1 \ldots, Z_k\}$ be a subset of X_n, $S_n \subseteq X_n$. Then $s_n(i)$ is the number of quadrupels of S_n that include taxon *i*: $s_n(i) = |\{Z_j \in S_n : i \in Z_j\}|$

In the examples we will use two types of notations for the subset S_n:

- $S_{n,i}$, in which n gives the number of used taxa, and i as index for the used examples with that n. If there is only one example, i is ignored.
- $S_{min,n}$, in which n gives the number of used taxa, and *min* to indicate that it is a set with a minimal triple covering.

3. The four-way partition property

3.1. Introduction and Definitions

Steel and Sanderson ([9]) characterized phylogenetic decisiveness for arbitrary sets S of subsets of X. However, we already know that it is sufficient to look at S_n, the set of the embedded quadruples of S. So we can alter the characteristics given by Steel and Sanderson [9] for quadruple sets.

Definition 8 (Four-way partition property)**.** Let $S_n = \{Z_1 \ldots, Z_k\}$ be a set of quadruples of taxa set X, $|X| = n$. Then, S_n satisfies the *four-way partition property* (for X) if, for all partitions of X into four disjoint, nonempty sets A_1, A_2, A_3, and A_4 (with $A_1 \cup A_2 \cup A_3 \cup A_4 = X$) there exists $a_i \in A_i$ for $i = 1, 2, 3, 4$ for which $\{a_1, a_2, a_3, a_4\} \in S_n$.

Theorem 1 (Theorem 2 of [9]). *A collection S_n of quadruples of X is phylogenetically decisive if and only if S_n satisfies the four-way partition property for X.*

In other words every possible partition of the n taxa has to be covered by at least one quadruple. The number of these partitions is simply the Stirling number of the second kind with n objects and 4 boxes, denoted $S(n, 4)$. This number increases rapidly, i.e. for $n = 8$ there are already 1701 different partitions. Therefore, checking all partitions is not the best way for proving decisiveness.

However, the number of kinds of partitions is much smaller, denoted $P(n, 4)$. In the case of eight taxa, there are still only five different types: $a|b|c|defgh$, $a|b|cd|efgh$, $a|b|cde|fgh$, $a|bc|de|fgh$, and $ab|cd|ef|gh$ for taxa set $X = \{a, b, c, d, e, f, g, h\}$.

The first partition shows clearly one condition for decisiveness: every possible triple abc has to be covered by at least one quadruple of S_n. But a set S_n of quadruples that covers all triples exactly once cannot be decisive. The second type of partitions prohibits that: W.l.o.g. $abcd$ is the only quadruple that covers the triples abc, abd, acd and bcd. Then the partition $a|b|cd|efgh$ cannot be covered by any quadruple, because all possible quadruples must have abc or abd, but only $abcd$ is available.

Lemma 1. *A set S_n is not phylogenetically decisive, if it does not cover all triples.*

Proof. Let $|X| = n$, $a, b, c \in X$. Assume the triple abc is not covered by any quadruple of a given set S_n. Then the partition $a|b|c|x_i$, $x_i \in X \setminus \{a, b, c\}$ is not covered. Therefore, the four-way partition property is not fulfilled, and the set S_n is not phylogenetically decisive. □

Lemma 2. *A set S_n is not phylogenetically decisive, if one quadruple covers two or more triples alone.*

Proof. Let $|X| = n$, and $a, b, c, d \in X$. Assume the quadruple $abcd$ covers alone the triples abc and abd. Then the partition $a|b|cd|x_i$, $x_i \in X \setminus \{a, b, c, d\}$ is not covered: all possible quadruples need the triple abc or abd, but the only one available is $abcd$, which has no element in the forth subset. □

3.2. Minimal Triple Covering with Steiner Quadruple Systems

Covering all triples once cannot result in a decisive set, but it gives a lower bound. The question how many quadruples are needed to cover all triples was already answered by Hanani [7]:

Definition 9. Given a set *X* of *n* elements we denote by $S(l, m, n)$, $(l \leq m \leq n)$ a system of subsets of *X*, having *m* elements each, such that every subset of *X* having *l* elements is contained in exactly one set of the system $S(l, m, n)$. The set $S(3, 4, n)$ is called *Steiner Quadruple System*, or SQS(n) for short.

A necessary and sufficient condition for the existence of an $S(3,4,n)$ is that $n \equiv 2(mod 6)$ or $4(mod 6)$ (compare Main Theorem of [7]). In other words, a Steiner Quadruple System exists only for even *n*, which are not completely divisible by six. For these *n* it is true that $min_0(n) := |S(3,4,n)| = \frac{1}{4}\binom{n}{3}$, which is simply the number of all possible triples divided by four, as one quadruple covers four triples. We can reconvert the equation for $min_0(n)$ to get a statement for tuples:

Lemma 3. *To cover all triples once all tuples must appear in at least $\frac{n-2}{2}$ quadruples.*

Proof. For n taxa there are $\binom{n}{2}$ tuples. Every quadruple covers six tuples. So if every tuples appears exactly $\frac{n-2}{2}$ times we have:

$$\begin{aligned} \frac{1}{6}\frac{n-2}{2}\binom{n}{2} &= \frac{(n-2)\cdot n!}{12\cdot 2!\cdot (n-2)!} \\ &= \frac{n!}{4\cdot 3!\cdot (n-3)!} \\ &= \frac{1}{4}\binom{n}{3} = min_0(n) \end{aligned} \tag{2}$$

□

Up to isomorphism, $S(3,4,8)$ is unique.

Example 1. Let $n = 8$, then there are $|X_8| = 70$ possible quadruples and $|Y_8| = 56$ possible triples. They can all be covered exactly once as this resolves into a Steiner Quadruple System. To cover all triples exactly once, one can first simply choose seven quadruples with one taxon (e.g. *a*) to cover all 21 triples with *a*, and then create the complement quadruples (e.g. *abcd* → *efgh*).

Alternatively, one can use the Steiner Triple System for $n = 7$, $S(2,3,7)$, also known as the Fano plane. In this, all tuples appear exactly once. Then you simply add to each of the seven triples the eighth taxon. After having the first seven quadruples you can proceed as above and create the complement quadruples.

The set $S_{8,1}$ is such a Steiner Quadruple System, and it is unique up to isomorphism.

$$\begin{aligned} S_{min,8} = S_{8,1} := \{&1234, 1256, 1278, 1357, 1368, 1458, 1467, \\ &5678, 3478, 3456, 2468, 2457, 2367, 2358\} \end{aligned} \tag{3}$$

Obviously this set is not phylogenetically decisive, as it does not cover the partition $1|2|34|5678$ with any quadruple.

For the other n it is more difficult to find minimal covering sets. The next two examples will show that.

Example 2. Let $n = 6$, then there are $|X_6| = 15$ quadruples and $|Y_6| = 20$ triples, 10 triples containing taxon a. Every a-quadruple covers three a-triples. Using three quadruples with a cannot cover all a-triples, and using four quadruples with a results in two a-triples, which are covered twice. This is true for all six taxa. We need therefore for a minimal triple covering at least $\frac{n \cdot 4}{4} = 6$ quadruples. Note that $min_0(n)$ would have suggested only $\frac{1}{4}\binom{6}{3} = 5$. The set $S_{min,6}$ covers all 20 triples, and all taxa appear four times.

$$S_{min,6} := \{1235, 1246, 1346, 1456, 2345, 2356\} \tag{4}$$

This suggests to add a correcting term to $min_0(n)$: $min_1(n) := \frac{1}{4}\binom{n}{3} + \frac{n}{6}$ for $n = 6m$, $m \in \mathbb{N}$.

Note that for $n = 6m$ it is true that $min_1(n) \in \mathbb{N}$ (see Appendix).

We take a look on how many quadruples with taxon a are in a set $S_{min,n}$, with $|S_{min,n}| = min_1(n)$ and $n = 6m$, $m \in \mathbb{N}$. We can assume that every taxon appears equally as every taxon is in $\binom{n-1}{2}$ triples. So all taxa have to be in the same number of quadruples. To get $s_n(a)$ we simply multiply $min_1(n)$ with the factor $\frac{4}{n}$, as there are four of n taxa in one quadruple:

$$\begin{aligned} s_n(a) = \frac{4}{n} min_1(n) &= \frac{n-2}{3n}\binom{n}{2} + \frac{4}{6} \\ &= \frac{(n-2)n!}{3n2!(n-2)!} + \frac{4}{6} \\ &= \frac{1}{n}\frac{n!}{3!(n-3)!} + \frac{4}{6} \\ &= \frac{1}{n}\binom{n}{3} + \frac{4}{6} \\ &= \frac{1}{3}\underbrace{\binom{n-1}{2}}_{\text{\# triples with taxon } a} + \frac{4}{6} \end{aligned} \tag{5}$$

This means if there is one quadruple less in $S_{min,n}$, $n = 6m$, then there would be at least one taxon whose triples could not be all covered. We cannot guarantee that a set of size $min_1(n)$, $n = 6m$, exists that covers all triples. However, it is consistent with the lower bound of the La Jolla Covering Repository ([1]). This site contains a collection of good $S(l, m, n)$-coverings, including the triple coverings by quadruples, $S(3, 4, n)$, which we treat. Next to the sets it also provides a lower bound for every system.

Example 3. Let $n = 7$, then there are 35 quadruples in X_7, 35 triples in Y_7 and the number of triples with ab is already odd: abc, abd, abe, abf, and abg. One of them must be in two quadruples to cover all five triples, e.g. $abcd$, $abef$, and $abcg$. The triple abc is picked twice.

This applies for all 21 tuples. So all 21 tuples have to appear in at least three quadruples, and every quadruples contains six tuples. A set that covers all triples at least once must be larger than

$$min_2(n) = \frac{1}{6}\frac{n-1}{2}\binom{n}{2} \tag{6}$$

For $n = 7$, $min_2(7) = 10,5$. This means that a set $S'_{min,7}$, $|S'_{min,7}| = 11$ could cover all triples. However it is impossible to construct such a set (see also [1]). But a set $S_{min,7}$ with $|S_{min,7}| = 12$ can cover all triples and is quite easy to find.

$$\begin{aligned} S_{min,7} := \{&1234, 1237, 1256, 1356, 1456, 1457 \\ &1567, 2345, 2346, 2357, 2467, 3467\} \end{aligned} \tag{7}$$

This set cannot be decisive, as there are 25 triples covered only once, but only 12 quadruples in the set. Therefore Lemma 2 applies here. E.g. 1256 covers 125, 126 and 256 alone. The partition $1|2|56|347$ cannot be covered. The partition property is not fulfilled, and the set not decisive.

The function $min_2(n)$, $n = 2m + 1$, is not as exact as $min_1(n)$, $n = 6m$. The difference between $min_2(n)$ and the known size provided by LaJ [1] suggests a correcting term similar to $\frac{n}{6}$ in the case of $min_1(n)$. I.e. one could use $\frac{n}{12}$:

$$min_3(n) = \frac{1}{6}\frac{n-1}{2}\binom{n}{2} + \frac{n}{12} \tag{8}$$

For most uneven n, $min_3(n)$ needs to be rounded up to the next whole number. However, there are critical n for which this function results in a higher number, i.e. $n = 13$: there is a set of size 78, that covers all triples (see [1]). But $min_3(13) = 79,08\bar{3} < 80$. This seems to be always the case if $n - 1$ is completely divisible by 12. For these n the function $min_2(n)$ seems to be good enough.

This gives us a general lower bound for the set size which can cover all triples at least once:

- if $n \equiv 2(mod 6)$ or $4(mod 6)$, then $|S_{min,n}| = \frac{1}{4}\binom{n}{3}$
- if $n = 6m$, $m \in \mathbb{N}$, then $|S_{min,n}| = \frac{1}{4}\binom{n}{3} + \frac{n}{6}$
- if $n = 2m + 1$ and $n - 1$ not divisible by 12, then $|S_{min,n}| \geq \frac{1}{6}\frac{n-1}{2}\binom{n}{2} + \frac{n}{12}$
- if $n = 2m + 1$ and $n - 1$ divisible by 12, then $|S_{min,n}| \geq \frac{1}{6}\frac{n-1}{2}\binom{n}{2}$

n	6	7	8	9	10	11	12
lower bound taken from [1]	6	12	14	25	30	47	57
$min_i(n)$, $i \in \{0, 1, 2, 3\}$	6	$11,08\bar{3}$	14	$24,75$	30	$46,75$	57

n	13	14	15	16	17	97	99
lower bound taken from [1]	78	91	124	140	183	37248	39625
$min_i(n)$, $i \in \{0, 1, 2, 3\}$	78	91	$123,75$	140	$182,75$	37248	$39624,75$

Table 1: Comparison between the known size of a minimal cover and the results of $min_i(n)$, $i \in \{0, 1, 2, 3\}$. Even for larger n $min_i(n)$ is close to the known bound.

We know that sets with this minimal size cannot be decisive. However, doubling them seems to be a fair chance to get a decisive set.

3.3. Decisive set by doubling $S(3, 4, 8)$

Knowing that covering all triples just once is not enough leads to the assumption that for decisiveness both Lemma 1 and Lemma 2 must be fulfilled.

Example 4. (Continuing Example 1) The set $S_{8,1}$ is a SQS(8). It is unique up to isomorphism. Therefore we can use a simple permutation, i.e. (4875) to get another SQS(8), $S_{8,2}$. Note that there is no overlap in quadruples between $S_{8,1}$ and $S_{8,2}$.

$$\begin{aligned} S_{8,2} :=& \{1238, 1246, 1257, 1345, 1367, 1478, 1568, \\ & 4567, 3578, 3468, 2678, 2458, 2356, 2347\} \\ S_8 :=& S_{8,1} \cup S_{8,2} \end{aligned} \tag{9}$$

Theorem 2. *The set S_8 and every equally designed set with eight taxa is phylogenetically decisive.*

Proof. To prove this we can simply show that all partitions are covered.

Case A) $a|b|c|defgh$
All partitions of this kind are covered already by $S_{8,1}$, because all triples *abc* are there.

Case B) $a|b|cd|efgh$
All partitions of this kind are covered, because all triples appear twice in S_8. Therefore, even if $abcd \in S_{8,1}$ cannot cover, there must be a quadruple $abcx \in S_{8,2}$, $x \in \{e, f, g, h\}$, which does cover the partition.

Case C) $a|b|cde|fgh$
By construction there are six quadruples in S_8 with the tuple *ab*. Also by construction,

all other taxa appear twice in these six *ab*-quadruples. Note that in the first set, $S_{8,1}$, every tuple was combined with all six possible taxa: e.g. 1234, 1256, 1278. Therefore, all partitions of this kind must already be covered by $S_{8,1}$, because even if *abcd* or *abfg* could not cover, there must be a quadruple of kind *abeh*, that covers the partition.

Case D) $a|bc|de|fgh$
We focus again on the set $S_{8,1}$. There are three quadruples with ab. If there is a quadruple $abxy$ with $x \in \{d, e\}$ and $y \in \{f, g, h\}$, then the partition is covered. If there is no such quadruple we know that the ab-quadruple must be like $abde$, aby_1y_2, and $abcy_3$, with $y_i \neq y_j$ for $i, j = 1, 2, 3$ and $y_i \in \{f, g, h\}$. With this we know there must be a quadruple $acdy_i \in S_{8,1}$, $i = 1, 2$, covering acd, as every triple has to appear exactly once. So $acde$ and $abcd$ are not in $S_{8,1}$, as ade is already used in $abde$ and abc in $abcy_3$. With $acdy_i$, $i = 1, 2$, the partition is covered.

Case E) $ab|cd|ef|gh$
We focus again on $S_{8,1}$. In this, every tuple appears three times. In this partition type, four tuples are not allowed: ab, cd, ef, and gh. Assuming that none of them appear in one quadruple together, this would exclude 12 quadruples for covering. However, $S_{8,1}$ has 14 quadruples. Therefore there must be two quadruples which can cover the partition. To be precise, these two quadruples are complementing each other: i.e. if $aceg$ is in $S_{8,1}$ then $bdfh$ is also in it, and both can cover the partition.

As all four-way partitions can be covered by quadruples of S_8 or similar sets, they are all phylogenetically decisive. □

Note that we only need $S_{8,2}$ for the second kind of partitions, $a|b|cd|efgh$. All others were already covered by the quadruples of set $S_{8,1}$. To be precise, there are three partitions of this type for every tuple ab, which need covering by $S_{8,2}$. Therefore, there are only $3 \cdot 28 = 84$ partitions without covering. I.e. we have the three quadruples 1234, 1256, and 1278 in $S_{8,1}$ with the tuple 12. With this, the partitions 1|2|34|5678, 1|2|56|3478, and 1|2|78|3456 cannot be covered.

However, the two above mentioned Lemmas are not enough to guarantee phylogenetic decisiveness. This is shown in the next example, again with eight taxa.

Example 5. The following set, $S_{8,3}$, has 36 quadruples and every taxon appears 18 times. There are six triples covered by only one quadruple: 137, 245, 248, 258, 278, and 367. They are covered by different quadruples: 1237, 2457, 2468, 2568, 2678, and 3678. Therefore, the above mentioned rules are fulfilled.

$$\begin{aligned} S_{8,3} := \{&1235, 1237, 1238, 1246, 1247, 1257, 1268, 1345, 1346,\\ &1348, 1356, 1358, 1458, 1467, 1478, 1567, 1578, 1678,\\ &2346, 2347, 2356, 2357, 2368, 2457, 2467, 2468, 2567,\\ &2568, 2678, 3456, 3458, 3478, 3578, 3678, 4568, 4578\} \end{aligned} \tag{10}$$

This set is not phylogenetically decisive, because the partition $137|26|45|8$ is not covered by any quadruple.

3.4. Equivalence to a possibly NP-hard problem

The decision whether a given set is decisive or not can be transferred into a known graph problem: As we can always assume quadruples as input, we can use a 4-uniform hypergraph, with n vertices. Four vertices are connected by a hyperedge, if there is a quadruple in S_n with these four taxa.

The four-way-partition-property implies that a set is not decisive, if there is a partition of X which is not covered by a quadruple of S_n.

There is a known problem called *No-Rainbow-Coloring* problem ([4]). The input is a hypergraph H with vertex set X, and the question is whether there is a nontrivial coloring of the vertex set such that no edge is *rainbow-colored*. This means each edge has at least two vertices of the same color.

This can be specified for a 4-uniform hypergraph with four colors. In other words, is there a partition of X into four non-empty subsets A_1, A_2, A_3, and A_4 of X so that no edge has elements in all subsets? This problem was formulated for 3-uniform graphs and three subsets, and is already expected to be NP hard. Therefore, this one with a 4-uniform hypergraph and 4 colors is expected to be NP hard as well, but has yet to be proven.

However, it is equivalent to the decisiveness problem as a set S_n is not phylogenetically decisive if and only if there is a no rainbow coloring: Not decisive means there is a partition that is not covered by a quadruple, therefore we can color the vertices as the partition suggests, and there cannot be any hyperedge with all colors.

4. Fischer's Algorithm

4.1. Introduction

Following Fischer [6] there is an easy algorithm to figure out whether a set of quadruples is phylogenetically decisive or not. For this algorithm we need the following definitions (all taken from [6]).

Definition 10 (Cross quadruples)**.** Let $S_n = \{Z_1 \dots, Z_k\}$ be a set of quadruples of taxa set X, $|X| = n$. A quadruple $abcd$ such that $a, b, c, d \in X$ and $abcd \notin S_n$ is called *cross quadruple* of S_n or CQ for short.

Definition 11 (Fixing taxon)**.** Let $S_n = \{Z_1 \dots, Z_k\}$ be a set of quadruples of taxa set X, $|X| = n$. Let $abcd$ be a CQ. Then, taxon $x \in X \setminus \{a, b, c, d\}$ is called a *fixing taxon* (or FT for short) of $abcd$, if for each of the four sets $abcx$, $abdx$, $acdx$ and $bcdx$ there exists a $j \in \{1, \dots, k\}$ such that this set is contained in Z_j, respectively.

Definition 12 (Resolved quadruples)**.** Let $S_n = \{Z_1 \dots, Z_k\}$ be a set of quadruples of taxa set X, $|X| = n$ and let $abcd$ be a CQ of S_n.
Then, $abcd$ is called *resolved* if there is a choice of unrooted trees on $Z_i, i = 1, \dots, k$, such that all possible supertrees of these trees display the same of the possible trees on $abcd$.
We call $abcd$ *directly resolved*, if it has a fixing taxon.
We call $abcd$ *indirectly resolved*, if there is a taxon $x \in X$ such that for each of the four sets $abcx$, $abdx$, $acdx$ and $bcdx$ one of the following conditions holds:

1. The set is not a CQ.
2. The set is a CQ but has a FT.
3. The set is a CQ and is itself indirectly resolved.

With these three simple definitions, Fischer [6] suggested the following theorem and algorithm.

Theorem 3 ([6], Proposition 7)**.** *Let X, $|X| = n$, be a set of taxa and $S_n = \{Z_1 \dots Z_k\}$ be a set of quadruples of X. Then, S_n is phylogenetically decisive if and only if all cross quadruples are directly or indirectly resolved.*

Based on this Proposition, Fischer [6] developed an algorithm to check on phylogenetic decisiveness:

First, identify all cross quadruples of the set S_n and mark them red. All quadruples of X_n appearing in S_n are marked green.

Then check the CQs for fixing taxa. If there is a fixing taxon, then change the marking from red to green. All CQs with a change of color are now directly (or later indirectly) resolved and can be used in the next round for checking for FTs.

Repeat this step until either all CQs are colored green, then the set is phylogenetically decisive, or there are red CQs left, but none of them has a fixing taxon, then the set is not phylogenetically decisive (compare with [6], Theorem 4).

4.2. Problem

However, resolving all CQs by FTs is only sufficient but not necessary for phylogenetic decisiveness. This is shown by two counter-examples, one with $n = 6$ and one with $n = 8$. The first one shows that the algorithm gives false negatives if the set size is too small, which implies a lower bound for the algorithm. The second one shows that even above that lower border the algorithm can give false negatives if one taxon appears not often enough.

Example 6. Let $n = 6$, then there are 15 quadruples in X_6 and 20 triples in Y_6.

$$\begin{aligned} S_{6,1} &:= X_6 \setminus \{1234, 1256, 1345, 2456, 3456\} \\ |S_{6,1}| &= 10 \\ S_{6,2} &:= X_6 \setminus \{1234, 1236, 1256, 1345, 2456, 3456\} \\ |S_{6,2}| &= 9 \end{aligned} \tag{11}$$

The first set, $S_{6,1}$, is phylogenetically decisive, as all CQs can be directly or indirectly resolved:

Round 1: 1234 and 1256 change to green with FT 6 and 3, respectively.

Round 2: 1345 and 2456 change to green with FT 2 and 1, respectively.

Round 3: 3456 changes to green with FT 1.

For the second set, $S_{6,2}$, there are no directly resolved CQs. There are six CQs and six triples which are covered only once: namely 123, 126, 134, 256, 345, and 456. Every triple connects two CQs, and every CQ two of these triples. That causes a circle, in which no quadruple can find a possible FT. This can be visualized by a bipartite graph, with one set out of the CQs and the other out of the triples, and an edge if the triple is a subset of the CQ (see Figure 1).

However, this set is phylogenetically decisive, as it satisfies the four-way partition property (see Appendix page 38)

The problem of the algorithm is that it only looks onto the triples and ignores the information input given by the tuples:

The CQ 1234 has no fixing taxon, as the quadruples 1345 (FT 5) and 1236 (FT 6) are unresolved. The other necessary quadruples are already in the set, namely 1235, 1245,

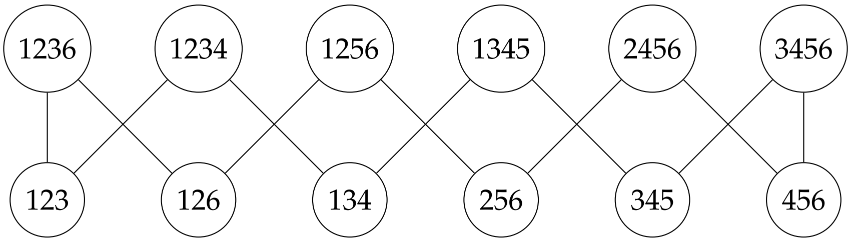

Figure 1: Bipartite graph for the set $S_{6,2}$: The CQs are in one set, and the triples, which are covered only once, in the other. An edge connects a quadruple with a triple, if the triple is in the quadruple.

and 2345 for FT 5, and 1246, 1346, and 2346 for FT 6. The intersection of the first three quadruples is 25, and of the second three 46.

In the worst case described by Fischer [6] we need all four quadruples with the FT to resolve the CQ 1234 uniquely (compare [6], Proposition 5). In this case we would have three trees with five taxa and all three have one cherry in common, e.g. $ax|bc$, $ax|bd$, and $ax|cd$. We need the forth tree with taxa $bcdx$ to decide between $ab|cd$, $ac|bd$ and $ad|bc$.

In this example, this would be the cherry 25 for the quadruples with FT 5, and 46 for FT 6. The algorithm now "sees" only three quadruples with one cherry, which would not be enough to make a decision. It therefore overlooks the second set of three quadruples with a different cherry and does not combine the information.

If we demand the trees to be compatible, the six quadruples with their cherries are enough to resolve the CQ uniquely.

We can choose compatible trees of the taxon sets such that 25 and 46 create cherries on their trees: $25|13$, $25|14$, $25|34$, $46|12$, $46|13$, and $46|23$.

Now we look back at the blocking CQs: 1345 and 1236. For both of them there are three possible trees. We can delete two of them: $14|35$ is incompatible with $46|13$, and $12|36$ is incompatible with $25|13$ (see Fig. 2).

$14|35$ would have resulted in $14|23$, $12|36$ in $12|34$. Therefore the only compatible tree for 1234 with the six input sets is $13|24$, with $13|26$ and $13|45$ (see Table 2).

This can be done with every CQ. Therefore the tree T is defined by the input, and the set is phylogenetically decisive.

This implies that the tuples carry information that the algorithm misses, as it only respects the triples.

So we now know that the algorithm gives false negatives if the set size $|S_n|$ is too small. Later we will see that the algorithm can find decisive sets down to a size of

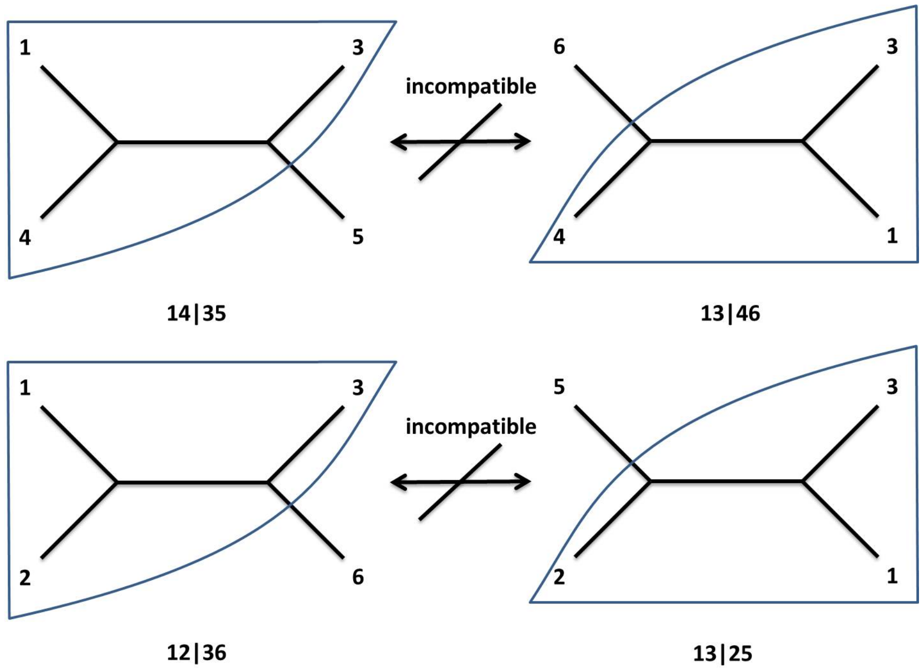

Figure 2: The two trees $14|35$ and $12|36$ can be dismissed as they are incompatible with the input trees $13|46$ and $13|25$.

$|S_n| = \binom{n-1}{3}$. This is exactly the number of all quadruples with taxon a (compare section 5.1).

However, it is not a guarantee to get a true reply. There are phylogenetic decisive sets S_n, with $|S_n| \geqslant \binom{n-1}{3}$, which have CQs that cannot be directly or indirectly be resolved by any FT.

Example 7. We return to Example 4: $n = 8$, and we have a set S_8 with $|S_8| = 28$ and $S_8 = S_{8,1} \cup S_{8,2}$. We already know that this set is phylogenetically decisive. However there are no FTs possible:

W.l.o.g. let $abcd$ be a CQ. To solve it by FT x there need to be $abcx, abdx, acdx, bcdx \in S_8$. We can assume that $abcx \in S_{8,1}$. Then we know, that $abdx, acdx \notin S_{8,1}$, because abx and acx are only covered once in $S_{8,1}$. However, $abdx$ and $acdx$ have adx in common, so they cannot be both in $S_{8,2}$. Therefore, at least one of the ax-quadruples is not in S_8, and no CQ can get resolved by a FT.

Now we can construct a set $S_{8,4}$, $|S_{8,4}| = 49 \geqslant 35 = \binom{7}{3}$.

$$S_{8,4} := S_8 \cup \{abcd : a, b, c, d \in X \setminus \{1\}\} \tag{12}$$

So we have all quadruples without 1, and 14 quadruples with 1 out of S_8. All CQs

cross quadruple	possible trees		
1234	$12\|34$	$13\|24$	$14\|23$
1345	$15\|34$	$13\|45$	$14\|35$
1236	$12\|36$	$13\|26$	$16\|23$

Table 2: All possible trees for the CQs 1234, 1345, and 1236. The compatible trees stand in one column. The first column must be excluded as it is incompatible with $25|13$, the last because of $14|35$. Therefore, only the second is compatible with all trees, which resolves in a unique supertree.

have taxon 1. For these there are still no FTs, as the argumentation above guarantees that one of the three 1-quadruples must be missing. So we still get a false negative answer.

With these two example we showed that resolving all CQs by FTs is only sufficient but not necessary for phylogenetic decisiveness. However, it is still true that a set is phylogentically decisive if all CQs are directly or indirectly resolved by FTs. This leads to the next Lemma.

Lemma 4. *The algorithm cannot give false positives.*

Proof. A false positive reply means that all CQs would be directly or indirectly resolved by FTs but the given set S_n is not phylogenetically decisive.

But this cannot be, following the first part of the proof of Proposition 7 of Fischer [6], showing the sufficiency of resolved CQs by FTs for phylogenetic decisiveness. □

4.3. Bounds of the algorithm

In this section we will introduce an upper and lower bound for the algorithm. Above the upper bound the algorithm will always find FTs for all CQs, and every given set of that size is phylogenetically decisive.

Below the lower bound the algorithm will always give a negative reply, regardless if the set is decisive or not.

Between these bound the algorithm can give true replies, but also false negatives.

4.3.1. Upper border

Moan and Rusinko [8] could prove that there is a minimal number $k(n)$, such that every collection S_n of quadruples with $|S_n| \geq k(n)$ is decisive, with $k(n) = \binom{n}{4} - (n-4)$.

It looks obvious in the first moment, as every triple can be there $(n-3)$ times. If one triple is missing completely, the set cannot be decisive.

However, this upper bound is also true for the algorithm.

Theorem 4. *Let $S_n = \{Z_1 \ldots, Z_k\}$ be a set of quadruples of taxa set X, $|X| = n$. If $|S_n| \geq \binom{n}{4} - (n-4)$ then all $(n-4)$ cross quadruples can be resolved by fixing taxa within two steps.*

Proof. We can prove this by checking all cases for a taxon *d*, with $s_n(d) \geq s_n(x)$, $x \in X \setminus \{d\}$. The cases are simply how often this taxon appears in the CQs.

Case A) Taxon *d* is in no CQ
⇒ Taxon *d* can always be used as FT for all CQs.

Case B) Taxon *d* is in only one CQ
W.l.o.g. *abcd* is a CQ. We know that the triple *abc* must be covered at least once: its maximal coverage is $n-3$, but there are only $n-4$ CQs. So we can assume $abcx \in S_n$, and we know that $abdx, acdx, bcdx \in S_n$, as *d* is missing only once. Choose taxon *x* as FT and resolve the CQ *abcd* directly. Now all quadruples with taxon *d* are in S_n or resolved, so we can use it as FT for all other CQs.

Case C) Taxon *d* is in two CQs: $abcd$ and $a'b'c'd$
Both triples, abc and $a'b'c'$, are covered by at least two different quadruples: following the argumentation of case B) there must be one quadruple, w.l.o.g. $abcx$ covering abc. If it is the only quadruple covering abc, then there would be $n-4$ CQs with abc. However, this forces every taxa to appear once, but taxa d is supposed to be missing twice and to be the taxa missing the least. Therefore, either there is another taxa missing only one time, and we can go to case B, or the triple abc is covered by at least a second quadruple $abcx'$.

Case C1) $abcd, a'b'c'd \notin S_n$, and $|abcd \cap a'b'c'd| \leq 2$ (w.l.o.g. $a = a'$)
Following the argumentation above the triple abc must be covered at least once, therefore there is a quadruple $abcx \in S_n$. So choose again taxon *x* as FT for *abcd*. The other three quadruples, $abdx$, $acdx$, and $bcdx$ must be in S_n, as the intersection of the two CQs is only one or two taxa. Analogously we can choose taxon *y* as FT for *a'b'c'd*, as some quadruple *a'b'c'y* must cover the triple *a'b'c'*. Both CQs with taxon *d* can be directly resolved, so we can use it as FT for all other CQs.

Case C2) $abcd, a'b'c'd \notin S_n$, and $|abcd \cap a'b'c'd| = 3$ (w.l.o.g. $a = a', b = b'$)
The triple *abc* must be covered by at least two different quadruples. Choose taxon $x \neq c'$ as FT for *abcd*. Then $abcx, \underbrace{abdx}_{x \neq c'}, \underbrace{acdx, bcdx}_{\text{as d is missing only twice}} \in S_n$.

The triple *abc'* must be covered by at least two different quadruples. Choose taxon $y \neq c$ as FT for *abc'd*. Then $abc'y, \underbrace{abdy}_{y \neq c}, \underbrace{ac'dy, bc'dy}_{\text{as d is missing only twice}} \in S_n$.

Now all quadruples with taxon *d* are in S_n or resolved, so we can use it as FT for all other CQs.

Case D) Taxon *d* is in three CQs
W.l.o.g. assume that $abcd, a'b'c'd, a''b''c''d \notin S_n$. For all three triples *abc, a′b′c′,* and *a″b″c″* it is true that they are covered by at least three quadruples. Therefore there are three possible FTs for each CQ. For this we can use the same argumentation as above: the triple abc can be covered by $n-3$ quadruples, but one of them is missing for sure, as $abcd$ is a CQ. Therefore, abc can maximal be covered by $n-4$ quadruples. However, there are two CQs without abc for sure, namely $a'b'c'd$ and $a''b''c''d$. Therefore, even if the other $n-7$ CQs would have abc, there must be at least three quadruples with abc in S_n. This is true for all three triples.

Case D1) The cardinality of each pairwise intersections is less than or equal to two
We know that all three triples abc, $a'b'c'$, and $a''b''c''$ are covered by different quadruples: $abcx$, $a'b'c'y$, and $a''b''c''z$. We can choose the taxa *x, y,* and *z* as FTs for the CQs *abcd, a′b′c′d,* and *a″b″c″d*, respectively. The low cardinality of the intersection prohibits problems, as it guarantees that all other quadruples with d are in S_n. So all CQs with taxon *d* can be resolved.
Now all quadruples with taxon *d* are in S_n or resolved, so we can use it as FT for all other CQs.

Case D2) The cardinality of each pairwise intersection is greater than or equal to two
There are three subtypes for this case (see Table 3). Again we have for each of the three triples at least three quadruples covering them. The cardinality of the intersection could cause problems if the FTs are chosen careless: assume as CQs $abcd$, $abed$, and $aced$. We have three possible FTs for $abcd$: x_1, x_2, x_3. One of them could be $e = x_3$. This taxa cannot resolve the CQ $abcd$ as $abed$ and $aced$ are both missing. However, we have two other possible FTs, x_1 and x_2. We can resolve $abcd$ with one of them conflict free. We can argue analogously for the other CQs with d, and the other subtypes.

For all types of this case there is at least one taxon for each CQ which can be used as FT directly. Now all quadruples with taxon *d* are in S_n or resolved, so we can use it as FT for all other CQs.

Case E) Taxon *d* is in four or more CQs
For this case we use the pigeonhole principle.

When $(n-4)$ quadruples are missing, then there are $4 \cdot (n-4)$ taxa used in the CQs.

As d is supposed to be missing the least of all taxa, every other taxon must be missing four times or more: $4 \cdot (n-1)$ taxa must be used in the CQs.

But $4 \cdot (n-1) > 4 \cdot (n-4)$. As the number of CQs is fixed, there must be taxa which are only used three times or less in the CQs.

This is a contradiction to the assumption that taxon d is missing the least of all taxa. This completes the proof. □

$\lvert\cdot\cap\cdot\cap\cdot\rvert = 2$	$\lvert\cdot\cap\cdot\rvert = 3$	Choose as FT
$abcd$		$x \neq e$ for *abcd*
$abed$		$y \neq c$ for *abed*
$aced$		$z \neq b$ for *aced*
$\lvert\cdot\cap\cdot\cap\cdot\rvert = 2$	$\lvert\cdot\cap\cdot\rvert \geq 2$	Choose as FT
$abcd$		$x \neq e, f$ for *abcd*
$abed$		$y \neq c$ for *abed*
$acfd$		$z \neq b$ for *acfd*
$\lvert\cdot\cap\cdot\cap\cdot\rvert = 3$	$\lvert\cdot\cap\cdot\rvert = 3$	Choose as FT
$abcd$		$x \neq e, f$ for *abcd*
$abed$		$y \neq c, f$ for *abed*
$abfd$		$z \neq c, e$ for *abfd*

Table 3: The three types for Case D2. There are always FTs for the CQs with taxon d, but they cannot be chosen freely but with respect to the other CQs. However, as maximal two taxa are blocked as FTs, there must be one left that can be used.

4.3.2. Lower bound

We know that $s_n = \binom{n-1}{3}$. This means simply that there are s_n quadruples with one taxon.

Assume the special case that all quadruples of S_n have taxon x and all quadruples with taxon x are in S_n. Then it is true that $|S_n| = s_n$. We can show that in this special case the algorithm still gives a true reply (see Chapter 5.1).

It seems that below that size the algorithm is not able to resolve all CQs. Therefore we can assume that the lower bound for the algorithm is $\binom{n-1}{3} - 1$.

Proving this for small n is possible by checking all cases. However, the general case is much more difficult to show, as we cannot simply prove it by complete induction.

Proof for $n = 6$

Theorem 5. *Let $n = 6$, X_6 the set of all quadruples with six taxa, and $S_6 \subseteq X_6$. If $|S_6| \leqslant 9$ then there are cross quadruples that cannot be resolved by fixing taxa.*

To prove this theorem we need some new terms.

Definition 13 (Neighbor, Chain, and Circle)**.** Let $Z_1, \ldots, Z_k$ be quadruples. Two quadruples Z_i and Z_j are called *neighbors* if they share a triple *abc*, e.g. $abcd$ is a neighbor of $abce$, depicted as $abcd \leftrightarrow abce$, for $a, b, c, d, e \in X$.

A *chain* is a list of pairwise neighboring quadruples.

A *circle* is a list of pairwise neighboring quadruples in which the starting quadruple is the same as the end quadruple.

Proof. To prove this theorem in general, we have to distinguish five cases: We know that $s_6 = 10$. Let the taxa be *a, b, c, d, e*, and *f*, and six quadruples of X_6 be missing. If it holds for $|S_6| = 9$, it also holds for every smaller set.

A) $s_6(a) = s_6(b) = 4$, the tuple ab missing six times ⇔ ab in all CQs

B) $s_6(a) = 4, s_6(b) = 5$, ab missing five times

C) $s_6(a) = 4, s_6(b) = 6$, ab missing four times

D) $s_6(a) = 5, s_6(b) = 6$, ab missing three times

E) $s_6(a) = s_6(b) = 6$, ab missing two times

Case A) There are exactly six quadruples with *ab* in X_6, all are missing.
⇒ All triples with *ab* are uncovered.
⇒ Following Lemma 1, the set is not phylogenetically decisive.
⇒ Following Lemma 4, the algorithm cannot resolve all CQs.

Case B) Five quadruples with *ab* are missing.
⇒ one quadruple, w.l.o.g. *abef* is in S_6.
⇒ There are two of 12 triple entries with *ab*: *abe* and *abf*. Therefore the triples *abc* and *abd* are not covered by any quadruple.
⇒ Following Lemma 1, the set is not phylogenetically decisive.
⇒ Following Lemma 4, the algorithm cannot resolve all CQs.

Case C) Within the six *ab*-quadruples, every other taxon appears three times. Therefore, there are two options to delete four *ab*-quadruples: one in which a third taxon is missing three times (in taxon frequencies $s_6(a) - s_6(b) - s_6(c) - s_6(d) - s_6(e) - s_6(f) = 6 - 6 - 7 - 8 - 8 - 9$, case C1), the other where the four taxa are all missing two times ($6 - 6 - 8 - 8 - 8 - 8$, case C2).
Case C1) W.l.o.g. all quadruples with *abc* are missing.
⇒ triple *abc* is not covered.
⇒ Following Lemma 1, the set is not phylogenetically decisive.
⇒ Following Lemma 4, the algorithm cannot resolve all CQs.

Case C2) W.l.o.g. the quadruples *abcd* and *abef* are in S_6, and al other quadruples with ab are CQs. Then, the quadruple $abcd$ covers alone both triples abc and abd
⇒ Following Lemma 2, the set is not phylogenetically decisive.
⇒ Following Lemma 4, the algorithm cannot resolve all CQs.

Case D) We focus on the five CQs with taxon a and ignore the last CQ without a.

Three quadruples with *ab* and two others with *a* are missing, so that the other taxa are all missing three time (if not, one taxon x would be missing four times, and there would be four quadruples with *ax* ⇒ case C).

Note that every triple is covered in this case (e.g. if triple *abc* would be uncovered, *abcd, abce* and *abcf* would be missing. To reach then the taxon frequencies $5-7-7-7-7-7$ the quadruple *adef* would have to be deleted twice).

If no triple can be missing, then for the three *ab*-quadruples it is true, that two taxa are missing twice, two taxa are missing once. W.l.o.g let *abcd, abce,* and *abdf* be CQs.

Then *acef* and *adef* must be CQs, too. With this, there are five CQ and five triples which are covered only once (namely *abc, abd, ace, adf* and *aef*). Each of these triples binds to two CQs, and every CQ binds to two triples (e.g. *abcd* to *abc* and *abd*). Therefore, each CQ is neighbored by two other CQs (e.g. $abcd$ has the neighbors $abce$ and *abdf*).

By construction we know that the intersection of the two neighbors of one CQ must be 2 taxa (e.g. ab for the neighbors $abce$ and $abdf$ of $abcd$). But this means, that both neighboring CQs contain the two possible FTs (e.g. possible FTs for $abcd$ are e and f, but $abce$ blocks e and $abdf$ blocks f). Therefore, no CQ has can be resolved by a FT.

Case E) All taxa are missing four times. For the four missing *a*-quadruples it is true, that two taxa must be missing three times, three taxa two times (if not, then at least one taxon must be missing five times after deleting also two more quadruples without *a* ⇒ case D). ⇒ taxon frequencies $6-7-7-8-8-8$.

W.l.o.g. b and c are missing three times. Then, two missing quadruples must have *abc*: *abcd* and *abce*. The other two *a*-quadruples now must both have the f taxon, and *b* or *c*, and *d* or *e*: *abdf* and $acef$ or $abef$ and *acdf*. In both cases the last two missing quadruples must be $bdef$ and $cdef$, as there is no other way to reach taxon frequencies $6-6-6-6-6-6$. No matter what pair is picked in the second step, it always resolves in six triples which are only covered once: *abc, abd/abe, ace/acd, bdf/bef, cef/cdf* and *def*.

Like in case D, each triple connects two CQs, and every CQ binds to two triples. We have analogously neighboring CQs, and again the two neighbors of one CQ have only two taxa in common. Therefore, all possible FTs are blocked for every CQ, and non can be resolved. □

Proof for general n

For *n* taxa there are sets S_n with cardinality $|S_n| = \binom{n-1}{3}$ that are decisive. We can split the binomial coefficient:

$$\binom{n-1}{3} = \binom{n-2}{3} + \binom{n-2}{2} \tag{13}$$

However, using this for a general approach cannot work. Let A be the number of all taxa in a set of size $\binom{n-1}{3}$, as every quadruple has four taxa, and B the number that are necessary if all taxa appear $\binom{n-2}{2}$ times.

$$\begin{aligned} A &:= 4\binom{n-1}{3} = \frac{4(n-1)!}{3!(n-4)!} \\ B &:= n\binom{n-2}{2} = \frac{n(n-2)!}{2!(n-4)!} = \frac{3n(n-2)!}{3!(n-4)!} \end{aligned} \tag{14}$$

The difference $A - B$ is not only positive, but also larger than n.

$$A - B = \frac{(n-2)!(4n-4-3n)}{3!(n-4)!} = \frac{(n-2)!(n-4)}{3!(n-5)!(n-4)} = \binom{n-2}{3} > n \tag{15}$$

Therefore, for a random set S_n of size $\binom{n-1}{3}$ there may not be one taxon x with $s_n(x) = \binom{n-2}{2}$, but all taxa can appear more often and splitting up the binomial coefficient cannot be used in a proof by complete induction.

Conjecture: The algorithm replies always *not phylogenetically decisive* if $|S_n| < \binom{n-1}{3}$

Proof. For the algorithm, not decisive means that there are still CQs which have no FTs.

W.l.o.g. let $|S_n| = \binom{n-1}{3} - 1$. If we can prove the conjecture for this size, it must also hold for every smaller set size.

We know that each taxon x can appear $s_n = \binom{n-1}{3}$ times in X_n, that is if all quadruples with taxon x are in S_n. Therefore, if $|S_n| = \binom{n-1}{3} - 1$, then all taxa must be missing at least once.

W.l.o.g. let $s_n(x) \geq s_n(y), \forall y \in X \setminus x$

$$\begin{aligned} s_n(x) = m &\Rightarrow \exists(\binom{n-1}{3} - m) =: l \text{ CQs with } x \\ &\Rightarrow \exists(l-1) \text{ quadruples without } x \text{ in } S_n \end{aligned} \tag{16}$$

Case A)$\forall$ CQs with x the pairwise intersection is only x: $|xabc \cap xdef| = 1$
All CQs with x need one quadruple without x to be resolved.
Every CQ with x has a different triple abc. To use one quadruple to resolve two CQs the intersection of the two CQ must have cardinality 3.
$\Rightarrow$ The $l-1$ quadruples without x in S_n cannot resolve all CQs with x
$\Rightarrow$ $\exists$ at least one CQ $xabc$, with abc not covered by any quadruple of S_n
$\Rightarrow$ Following Lemma 1, the set S_n is not phylogenetically decisive

$\Rightarrow$ Following Lemma 4, the algorithm cannot resolve all CQs.

Case B)$\exists$ CQs with x: $|xabc \cap xade| = 2$
In this case there must also be one triple missing, as the quadruples without x in S_n cannot cover all triples.
$\Rightarrow$ Following Lemma 1, the set S_n is not phylogenetically decisive
$\Rightarrow$ Following Lemma 4, the algorithm cannot resolve all CQs.

Case C)$\exists$ CQs with x: $|xabc \cap xabd| = 3$
With this we could hope to find quadruples in S_n so that all l CQs with taxon x can be resolved. This would require that two quadruples, namely the one with intersection of a triple xab, $xabc$ and $xabd$, use the same quadruple of S_n.

Case C1) $abce \in S_n, e \neq d$
Then the two CQs cannot use the same quadruple of S_n. Therefore, again one triple missing.
$\Rightarrow$ Following Lemma 1, the set S_n is not phylogenetically decisive
$\Rightarrow$ Following Lemma 4, the algorithm cannot resolve all CQs.

Case C2) $abcd \in S_n$
The quadruple $abcd$ is the only possible quadruple in S_n that can be used to resolve $xabc$ and $xabd$. There are enough quadruples in S_n to resolve all other CQs with x but $xabc$ and $xabd$. But this also means that $abcd$ covers two triples alone, namely abc and abd. $\Rightarrow$ Following Lemma 2, the set S_n is not phylogenetically decisive
$\Rightarrow$ Following Lemma 4, the algorithm cannot resolve all CQs.

Case C3) $abcd, abce \in S_n$
In this case $xabc$ could get resolved with FT e if $xabe$, $xace$ and $xbce$ are in S_n or resolved.

For all of these three it is true that the intersection with $xabc$ has cardinality 3.

Assume all of them are in S_n. Then $xabc$ could be resolved with FT e and in a next step $xabd$ with FT c. But then we have again used two different quadruples for two CQs. Therefore, again one triple missing completely.

Assume one of the three quadruples is not in S_n, w.l.o.g. $xace$. If there is no other ace-quadruple in S_n but $abce$, then $xace$ cannot get resolved as $xabc$ blocks b as FT.

Another ace-quadruple could start a chain reaction, and we can argue as above. But as S_n is a finite set, at some point it must stop. We get a chain of CQs, where the inner CQs are blocked by their neighbors (i.e. $xabc$ blocked by $xabd$ and $xace$) and the ending CQs are blocked by the missing quadruple without taxon x.

If we do this for the other taxa, we can assemble the chains and get circles. □

Example 8. To illustrate this we can look back to Example 6 and the set $S_{6,2}$. Remember, we have six CQs, namely 1234, 1236, 1256, 1345, 2456, and 3456. Case C of the proof applies for all six taxa, as we already know that each CQ can be neighbored to two other CQs by triples, which is nothing but the intersection of size 3.

There are four quadruples with taxon 1. 1234 and 1236 intersect in the triple 123. This means nothing but 6 is blocked as FT for 1234, and 4 is blocked as FT for 1236. Both 2346 and 2345 are in $S_{6,2}$. Therefore we must be in case C3. We cannot use 5 as FT for 1234, as 1345 is not in $S_{6,2}$. So we have a chain $1345 \leftrightarrow 1234 \leftrightarrow 1236$.

If we make now the same consideration for 1236, we can prolong the chain with 1256: $1345 \leftrightarrow 1234 \leftrightarrow 1236 \leftrightarrow 1256$.

Now all CQs with taxon 1 are in one chain and we can check the CQs at the ends: 1345 and 1256. The intersection of 1345 and 3456 is the triple 345, and of 1256 and 2456 the triple 256. Arguing as above, we can prolong the chain again:

$$3456 \leftrightarrow 1345 \leftrightarrow 1234 \leftrightarrow 1236 \leftrightarrow 1256 \leftrightarrow 2456$$

Now, all six CQs of $S_{6,2}$ are used and we need check the CQs at the ends for partners: the intersection of 2456 and 3456 is the triple 456. Therefore, we can connect the end points of the chain to get a circle.

We can depicted that in a graph in which two CQs are connected if they share one triple (see Figure 3). The adjacent CQs of one CQ block both possible FTs.

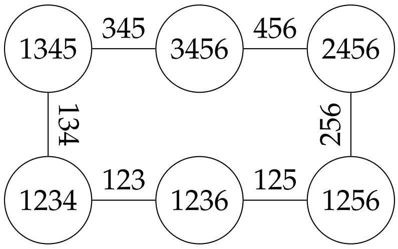

Figure 3: A circle of the six CQs of set $S_{6,2}$. Two CQs are connected if they share a triple.

The larger n gets, the more circles of CQs we need. It seems necessary that every CQ is in at least $\frac{n-4}{2}$ circles to ensure that every taxa is excluded as FT: every CQ has theoretically $n-4$ possible FTs, and each neighbor blocks one. To block all possible FTs a CQ needs at least $n-4$ neighbors. In every circle the CQ picks two different neighbors. Then we need at least $\frac{n-4}{2}$ neighbors.

5. Special cases

5.1. One taxon appears in every quadruple

Given n taxa, there are $\binom{n}{4}$ possible quadruples and $s_n = \binom{n-1}{3}$ quadruples with taxon x. If we choose a set S_n, with $|S_n| = \binom{n-1}{3}$, then we can pick the special case in which taxon x is in all quadruples of S_n.

This set is phylogenetically decisive, as it fulfills the four-way partition property: in any partition x must be in one of the four subsets A_1, A_2, A_3 or A_4. Let $x \in A_1$. Then one can choose freely $a \in A_2$, $b \in A_3$, and $c \in A_4$. The quadruple *abcx* must be in S_n as all x-quadruples are in S_n.

This kind of set is also correctly identified by the algorithm. For all cross quadruples one can easily choose taxon x as fixing taxon. Therefore, all of them can be directly resolved and the set is correctly recognized as phylogenetically decisive.

As soon one quadruple is deleted out of the set, $S'_n = S_n \setminus \{abcx\}$, the set cannot be decisive: the triple *abc* is missing completely. With Lemma 1, this leads directly to none decisiveness.

All $n-3$ quadruples with the triple *abc* are missing. The algorithm can solve all but these $n-3$ quadruples. Therefore, there are CQs without FTs left and the algorithm gives the true reply of no decisiveness.

This case is quite similar to the rooted case: a binary rooted phylogenetic X-tree T is defined as the unrooted one but with the additional property that there is exactly one node of degree 2, namely the so-called root r. As taxon x appears in all quadruples, we can take it as root.

Assume the quadruple $abcx$ is displayed in the unrooted tree $ab|cx$ (see Figure 4). If we delete taxon x in this tree, we also delete the edge connecting x with the internal node r. If we not suppress this node r, we can use it as root for the tree induced by the triple abc.

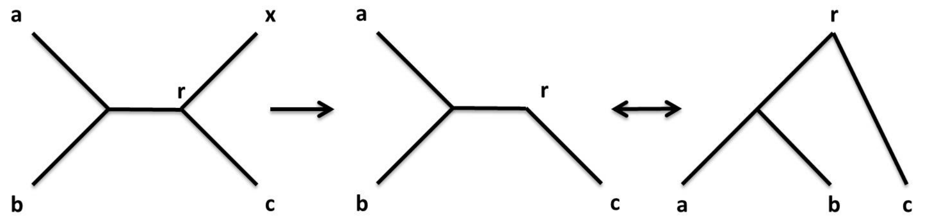

Figure 4: The unrooted tree $ab|cx$ can be transferred to the rooted tree with the cherry ab by deleting taxon x and not suppressing the node of degree 2.

Fischer [6] could show that a given set of taxon sets is phylogenetically decisive in the rooted case if all triples are in the set (see Corollary 1 of [6]). Therefore, if we delete taxon x from all quadruples and use the internal node of degree 2 every time as root, we have all possible triples of $X' = X \setminus \{x\}$. We get a unique supertree T' of the rooted case (up to isomorphism). Now we can add taxon x again and connect it directly to the root r to get the supertree T. Now we have the original set size $|X| = n$, and T displays all $T|Z_i$, with Z_i being the input quadruples.

5.2. One tuple appears too rarely

We already have rules for triples which would make it impossible for a set to be decisive: one triple missing completely and two triples, which are covered only once, are covered by the same quadruple. In both cases the four-way partition property is not fulfilled and the algorithm finds no resolvable CQs.

In this section we will have a look at the special case that one tuple xy appears less than $n - 3$ times. We know that every tuple appears $\binom{n-2}{2}$ times in X_n.

Example 9. Let $n = 9$ and $S_{9,1}$ be a set of quadruples, containing all quadruples without the tuple 12, $|S_{9,1}| = \binom{9}{4} - \binom{9-2}{2}$. The set $S_{9,2}$ contains only quadruples with the tuple 12, and $|S_{9,2}| = 9 - 4 = 5$:

$$\begin{aligned} S_{9,2} &:= \{1234, 1239, 1256, 1257, 1278\} \\ S_9 &:= S_{9,1} \cup S_{9,2} \end{aligned} \tag{17}$$

All 12-triples of Y_9 are covered, and there are four triples covered only once: 124 by 1234, 126 by 1256, 128 by 1278, and 129 by 1239 (see Table 4).

There are $\binom{9-2}{2} - (9 - 4) = 21 - 5 = 16$ CQs, and all of them have the tuple 12 in common. To get resolved by any fixing taxon they need two quadruples of $S_{9,2}$.

The taxon 3 appears twice as possible FT, with the triples 124 and 129. Therefore 3 is a FT for 1249. Analogous, 5 is a FT for 1267, and 7 for 1258. Having resolved these three CQs, one can resolve 1268 with FT 5 or 7 in another round.

There are now 12 CQs left, but none of them can get resolved by a FT. For example 1235 can use 4 or 9 as FT considering the triple 123, but 125 allows only 6, 7, 8. The intersection of the possible FTs is always empty. Therefore there is no FT (see Table 5).

We can illustrate this phenomenon with a graph consisting of two unconnected components, namely the complete graphs K_3 and K_4. The vertices are the triples, and an edge connects two vertices if there is a quadruple in S_9 containing both triples, e.g. 123 and 124 are connected by $1234 \in S_9$ (see Figure 5). To resolve any more quadruple we need at least one more edge that would connect both components. As long as we have two disconnected components, the set is not phylogenetically decisive.

triples	123	124	125	126	127	128	129
possible FTs	$4,9$	3	$6,7$	5	$5,8$	7	3

Table 4: Possible fixing taxa in the original set S_9. The taxa 3, 5, and 7 appear two times.

triples	123	124	125	126	127	128	129
possible FTs	$4,9$	$3,9$	$6,7,8$	$5,7,8$	$5,6,8$	$5,6,7$	$3,4$

Table 5: Possible fixing taxa after two rounds of the algorithm. The taxa 3, 4, and 9 appear two times, and 5, 6, 7, and 8 appear three times.

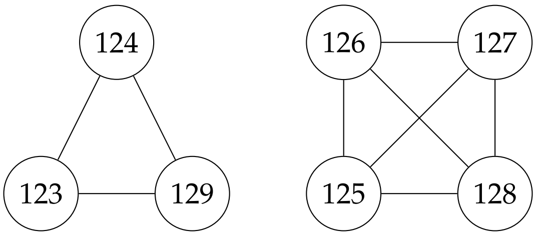

Figure 5: The graph with two components: the complete graphs K_3 and K_4. Two triples are connected if there is a quadruple containing both in S_9 or resolved by a FT. As there is no further connection between the two components no more CQ can be resolved by any FT.

Speaking in terms of the four-way partition property, we can use chains with an additional condition: all quadruples must have a tuple ab in common. In this example we have two disconnected chains for the tuple 12:

$$1239 \leftrightarrow 1234 \text{ and } 1256 \leftrightarrow 1257 \leftrightarrow 1278$$

In other words, if we ignore the tuple taxa, we have two subsets without an overlap: $\{349\}$ and $\{5678\}$. Therefore, we can denote them as A_3 and A_4, and put 1 in A_1 and 2 in A_2 and we get a partition that cannot be covered by any quadruple of S_9. Even if we would include the resolved quadruples of the algorithm this partition cannot be covered, as the resolved ones are simply combinations within the subsets.

Theorem 6. *Let $S_n = \{Z_1 \ldots, Z_k\}$ be a set of quadruples of taxa set X, $|X| = n$. If one tuple appears in less than $n-3$ quadruples of S_n then S_n is not phylogenetically decisive.*

Proof. As we know that the algorithm can only give false negatives but not false positives (see Lemma 4), it is sufficient to prove the theorem for the four-way partition property only.

We have *n* taxa and the tuple *xy* appears only $n-4$ times. There are $n-2$ triples with *xy*.

We focus on the $n-4$ quadruples with xy. Each of them has two more taxa. Therefore there are $2(n-4)$ positions for the $(n-2)$ other taxa. As all triples have to be covered (if not, proving no decisiveness is trivial, see Lemma 1), there must be at least four triples which are covered by only one quadruple.

W.l.o.g. let these triples be *axy, bxy, cxy,* and *dxy*. They must be covered by four different quadruples (if not, it is trivial again to prove no decisiveness, see Lemma 2): *auxy, bvxy, cwxy,* and *dzxy*. There must be quadruples covering the triples *uxy, vxy, wxy,* and *zxy* a second time.

As we know that axy, bxy, cxy, and dxy are covered only once we can use them as start points or end points of a chain: they have one neighbor at most.

Case A) $uvxy, wzxy \in S_n$

These six quadruples form two disconnected chains:

$$auxy \leftrightarrow uvxy \leftrightarrow bvxy \text{ and } cwxy \leftrightarrow wzxy \leftrightarrow dzxy$$

With these two chains we can define two subsets of X: $A_3 := \{a, b, u, v\}$ and $A_4 := \{c, d, w, z\}$. We can denote $x \in A_1$, $y \in A_2$, with $|A_1| = |A_2| = 1$, and $A_5 = X \setminus \{A_1, A_2, A_3, A_4\}$. Then the partitions $A_1|A_2|A_3|A_4A_5$ and $A_1|A_2|A_3A_5|A_4$ cannot be covered by any quadruple.

Therefore the four-way partition property is not fulfilled in this case.

Case B) $uvxy, e_1wxy, e_2zxy \in S_n$
Again we have at least two chains, and the partition $A_1|A_2|A_3|A_4A_5$ of subsets defined above cannot be covered by a quadruple.

Case C) $e_1uxy, e_1vxy, e_2wxy, e_2zxy \in S_n$
In this case we still have at least two disconnected chains, but of length four instead of length three:

$$auxy \leftrightarrow e_1uxy \leftrightarrow e_1vxy \leftrightarrow bvxy \text{ and } cwxy \leftrightarrow e_2wxy \leftrightarrow e_2zxy \leftrightarrow dzxy$$

Speaking in partitions, we have to alter the subsets: $B_3 = A_3 \cup \{e_1\}$, $B_4 = A_4 \cup \{e_2\}$, and $B_5 = X \setminus \{A_1, A_2, B_3, B_4\}$. Then again we can pick any partition that keeps the subsets together, like $A_1|A_2|B_3|B_4B_5$ and we know that it cannot be covered.

Case D) $e_1uxy, e_2vxy, e_3wxy, e_4zxy \in S_n$
As there are four triples covered only once, we have four start or end points of a chain. This implies that there must be two chains of variable length with the quadruples $auxy$, $bvxy$, $cwxy$, and $dzxy$. Therefore we get at least two subsets of $X \setminus \{x, y\}$, namely C_3 and C_4.

If $A_1 \cup A_2 \cup C_3 \cup C_4 = X$, then we know that the partition $A_1|A_2|C_3|C_4$ cannot be covered.

If $A_1 \cup A_2 \cup C_3 \cup C_4 \neq X$, then there are taxa left which form a circle instead of a chain, as we have no start or end points left. We can put all of these taxa in one subset C_5. Then the partition $A_1|A_2|C_3|C_4C_5$ is not covered by any quadruple.

In no case the set fulfills the four-way partition property. Therefore, it cannot be phylogentically decisive.

The algorithm cannot give false positives (see Lemma 4). Therefore it gives in these cases the true reply of none decisiveness. □

This theorem shows that having a tuple only $n - 4$ times in S_n leads directly to no decisiveness. It seems to be necessary for decisiveness to have no unconnected subsets for every tuple. Looking back at Example 9, we could simply add the quadruple 1246 to have one chain instead of two:

$$1239 \leftrightarrow 1234 \leftrightarrow 1246 \leftrightarrow 1256 \leftrightarrow 1257 \leftrightarrow 1278$$

So we do not just need to add any quadruple, but a specific quadruple that connects the chains for the four-way partition property.

Lemma 5 (Chain condition). *Let $S_n = \{Z_1 \ldots, Z_k\}$ be a set of quadruples of taxa set X, $|X| = n$. If there is more than one chain for any tuple ab, $a, b \in X$, then the set is not phylogenetically decisive.*

Proof. Assume there are two chains for the tuple ab. Then we can define two subsets A_3 and A_4 of $X \setminus \{a, b\}$ so that each subset contains one chain but the tuple taxa. So the partition $a|b|A_3|A_4$ is not covered. Then the four-way partition property is not fulfilled and the set is not phylogenetically decisive.

If there are m chains or circles we analogously define m subsets A_i, $i = 1, \ldots, m$ of $X \setminus \{a, b\}$. Any partition $a|b|B_3|B_4$ that does not break up the subsets A_i is not covered by any quadruple. Thus the four-way partition property is not fulfilled and the set is not phylogenetically decisive. □

5.3. Minimal decisive sets for small n

5.3.1. $n = 6$

We already know that a set of nine quadruples can be decisive for six taxa (compare 4.2, $S_{6,2}$, see also Appendix page 38). In this set all taxa appear six times, and all tuples are covered by at least three quadruples.

A set of eight quadruples cannot be decisive. We know that every tuple ab must be there at least $n-3 = 3$ times. So if one taxon a appears in only four quadruples, then we have only 12 ax-tuples, $x \in X \setminus \{a\}$, as every quadruple with a covers three tuples with a. But there are five ax-tuples, which all require at least three covering quadruples. As we have only four, there must be at least 3 tuples with a covered only twice. Such a set cannot be decisive.

So every taxon must appear in at least 5 quadruples. A set of five a-quadruples covering all a-tuples three times can easily be constructed:

$$S_{6,3} = \{abce, abdf, abef, acde, acdf\} \tag{18}$$

If we want a set of eight quadruples and all taxa appearing at least five times, we know there must be three other taxa besides a which also appear only five times. Therefore, we use the permutation (ab) to get a compatible set for taxon b, so that all b-tuples are there three times.

$$S_{6,4} = \{abce, abdf, abef, bcde, bcdf\} \tag{19}$$

The tuple ef is there only once, and we can only choose one more quadruple. Therefore, this tuple cannot be there three times, and any set with eight quadruples cannot be phylogenetically decisive.

5.3.2. $n = 7$

Following the tuple rule that every tuple has to be there $n-3$ times, we know that we need at least 14 quadruples, and every taxon eight times:

$$\begin{aligned} |S_7| &\geq \frac{7-3}{6}\binom{7}{2} = 14 \\ s_7(x) &= \frac{14 \cdot 4}{7} = 8 \forall x \in X \end{aligned} \tag{20}$$

It is possible to construct such a set: We know the tuple *ab* has to be there four times, and there must be a chain that connects the four quadruples, e.g. $abcd$, $abde$, $abef$, and $abfg$. Chain simply means that three of the five ab-triples are covered twice, in this example abd, abe, and abf. If there is no chain, then one quadruple would be "alone", e.g. $abcd$, $abef$, $abfg$, and $abeg$. But then the partition $a|b|cd|efg$ cannot be covered.

Fix the first four quadruples, $abcd$, $abde$, $abef$, and $abfg$. Then we know that there are four more quadruples with a (without b), and four with b (without a). We also know that c and g must be in six of the eight quadruples, otherwise ac, ag, bc, and bg would not appear four times. The other three taxa, d, e, and f appear twice for both a and b. Therefore, the last two quadruples must be $cdef$ and $defg$.

Now we only need to determine the eight quadruples. As c and g have to appear three times in four quadruples there must be two quadruples with acg, and bcg, respectively. To get a chain for the cg tuple there must be one taxon e with both a and b: $acge$ and $bcge$. The other two triples must be filled with different taxa to get a chain and not a circle. A circle of four quadruples would mean that one triple is not covered. I.e. if $acgf$ and $bcgf$ would be both there, then cgd is not covered. Therefore, we need $acgd$ and $bcgf$.

The last four quadruples can easily be determined with the chain condition of Lemma 5: having $abcd$, $acgd$, and $acge$ we need $acef$ (for the tuple ac). Then there is only $agdf$ left, as we know which taxa we need to use. The same way we can fix $bgde$ and $bcdf$.

Now we have 14 quadruples, and all tuples appear exactly four times:

$$\begin{aligned} S_7 := \{&abcd, abde, abef, abfg, acdg, acef, aceg, adfg, \\ &bcdf, bceg, bcfg, bdeg, cdef, defg\} \\ = \{&1234, 1245, 1256, 1267, 1347, 1356, 1357, 1467, \\ &2346, 2357, 2367, 2457, 3456, 4567\} \end{aligned} \tag{21}$$

This set is phylogenetically decisive, as all partitions are covered:

- $a|b|c|defg$: covered, as all triples appear at least once.

- $a|b|cd|efg$: covered, as we have chains for all possible tuples (see also Appendix page 40 and 41)
- $a|bc|de|fg$: covered (see Appendix page 39)

In every set smaller than S_7 there must be at least six tuples appearing only three times, and this leads directly to none decisiveness following Theorem 6.

5.3.3. $n = 8$

We already know that a set of 28 quadruples can be decisive for eight taxa (compare 3.3, S_8). A set $S_{8,5}$ with only 26 quadruples is also decisive, $s_8(x) = 13$.

$$S_{8,5} = S_8 \setminus \{1234, 5678\} \tag{22}$$

We know that S_8 resolves from two SQS(8) and that $S_{8,1}$ covers all partitions but 84 of the kind $a|b|cd|efgh$. As SQS(8) is unique up to isomorphism, $S_{8,2}$ also covers all partitions but 84. The two quadruples 1234 and 5678 were originally in $S_{8,1}$. Therefore we simply have to check the second kind of partition if still all 84 partitions missed by $S_{8,2}$ are covered.

As the two sets $S_{8,1}$ and $S_{8,2}$ had no overlap in quadruples, we know that it is true for all tuples that we have not chains but circles of length six in S_8, i.e.

$$1234 \leftrightarrow 1238 \leftrightarrow 1278 \leftrightarrow 1257 \leftrightarrow 1256 \leftrightarrow 1246 \leftrightarrow 1234$$

By deleting two quadruples we also delete 12 tuples. However, this means that we have a chain for these 12 tuples instead of a circle. But this still allows us to cover all partitions of the kind $a|b|cd|efgh$. Therefore, $S_{8,5}$ is still phylogenetically decisive.

In the set $S_{8,5}$, every taxon appears 13 times, and every possible tuple at least five times. However, it is not possible to delete one more quadruple of $S_{8,5}$ without having at least one tuple appearing only four times. Therefore, the combination of the two Steiner Quadruple Systems cannot be reduced any more.

The tuple rule we used previously for $n = 7$ implies that every tuple needs to be there at least five times. It is possible to construct such a set with only 24 quadruples:

$$\begin{aligned} S_{8,6} := S_{8,2} \cup \{&1234, 1268, 1378, 1458, 1567 \\ &2358, 2367, 2457, 3456, 4678\} \end{aligned} \tag{23}$$

In this set, the tuples 18, 23, 45, and 67 appear six times, all others five times.

However, this set is not phylogenetically decisive: all partition-types but $a|b|cd|efgh$ are completely covered by $S_{8,2}$. But for this one type 84 partitions are not covered by construction. In $S_{8,2}$ every triple is there only once. Therefore all tuples are there three times, i.e. there is 1238, 1345, and 1367 in $S_{8,2}$ covering all 13-triples. But this also means that partitions uniting two of them in one subset, i.e. $1|3|28|4567$ cannot be covered. So for all 28 tuples we have three partitions left which are not covered.

To cover these partitions we need to create a chain to prevent subsets as in the previous chapter (Proof of Theorem 6). In example for tuple 13: In $S_{8,6}$ we have also 1234 and 1378. With this we get the chain:

$$1345 \leftrightarrow 1234 \leftrightarrow 1238 \leftrightarrow 1378 \leftrightarrow 1367$$

All partitions of the tuple 13 are covered.

However this is not possible for all tuples, i.e. 12. Here, we can get a circle of length 4:

$$1234 \leftrightarrow 1238 \leftrightarrow 1268 \leftrightarrow 1246 \leftrightarrow 1234$$

Therefore, there is no connection to the quadruple 1257, and the partition $1|2|57|3468$ cannot be covered. This is true for three more tuples: 38 (3468), 46 (1246) and 57 (3578).

Therefore a random set with all tuples five times is not phylogenetically decisive and a set like $S_{8,6}$ would need at least four more quadruples, as there is no overlap within the four problematic tuples.

Proving that there is no way to construct a set of 24 quadruples in which all tuples are connected in a chain remains unsolved, especially as randomizing the quadruples by not choosing a Steiner Quadruple System as basis requires checking all 1701 partitions again, and not just 84.

6. Outlook

In this work we showed that there is a lower bound for phylogenetic decisiveness: $min_i(n)$, $i \in \{0, 1, 2, 3\}$. A set S_n of size $min_i(n)$ can cover all triples once. We could also show for small n, that a set of size $2 \cdot min_i(n)$ can be decisive.

For all i it is true that we can write $min_i(n)$ regarding to tuples, $\binom{n}{2}$, instead of triples, $\binom{n}{3}$ (see Lemma 3). Therefore, we know that

$$2 \cdot min_i(n) \geq \frac{1}{6}(n-2)\binom{n}{2} > \frac{1}{6}(n-3)\binom{n}{2} \tag{24}$$

So we know that a set S_n with

$$|S_n| \geq 2 \cdot min_i(n) \text{ and } S_n = S_{min,n} \cup S'_{min,n} \text{ with } S_{min,n} \cap S'_{min,n} = \emptyset,$$

does not fall below the tuple rule (Theorem 6), which is that all tuples must appear in at least $n-3$ quadruples. However, that doubling a minimal set is sufficient for decisiveness has yet to be proven for greater n.

Another interesting open question is whether the tuple rule itself is sufficient for greater n. In the case $n=7$, having all tuples $n-3=4$ times was enough. However, this might only work as we have only three types of partitions, and almost all with two subsets with only one element. For greater n, the other partition types might play a bigger role.

Also proving or disproving the NP hardness of the *No-Rainbow-Coloring* problem, and therefore for phylogenetic decisiveness problem would be very interesting.

If the phylogenetic decisiveness problem is NP hard, then the algorithm is a good approximation. However, there is still two important open questions regarding the bounds of the algorithm:

First, we know that the algorithm misses some decisive cases if one taxon x appears not often enough. Is this only true if the other $n-1$ taxa are missing "equally", as in $S_{8,4}$, or is it true in general? Does it depend on the triple covering, which was in $S_{8,4}$ two for all triples with 1? Or is it depending on the fact that $s_n(x) = \binom{n-2}{2} - 1$, which forces all other $n-1$ taxa to be in at least one quadruple?

Second, can the algorithm be fixed so that it does no longer miss the information provided by the tuples? This would lead to a better lower bound, but checking all tuple combination might have its price in complexity, which would decrease the efficiency of the algorithm.

Acknowledgment

I want to thank my supervisor, Prof. Dr. Mareike Fischer, for the support through the crucial steps of this work. I also want to thank Prof. Dr. Leo van Iersel for his help and insight to NP hard problems.

I thank my family who gave me feedback in earlier versions of this work, and who supported me by discussions and suggestions.

References

[1] La jolla covering repository. http://ccrwest.org/cover.html.

[2] Biology-online dictionary. www.biology-online.org/dictionary/Phylogeny.

[3] Burleigh J.G. Eulenstein O. Bansal, M.S. and D. Fernández-Baca. Robinson-foulds supertrees. *Algorithms for Molecular Biology*, 5(18):1–12, 2010.

[4] M. Bodirsky. *Constraint Satisfaction with Infinite Domains*. PhD thesis, Humboldt-Universität zu Berlin, 2004.

[5] C. Darwin. *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. John Murray, 1859.

[6] M. Fischer. Perfect taxon sampling and phylogenetically decisive taxon coverage. arXiv: 1206.3472, 2012.

[7] H. Hanani. On quadruple systems. *Canad. J. Math.*, pages 145–157, 1959.

[8] E. Moan and J. Rusinko. Combinatorics of linked systems of quartet trees. *arXiv preprint*, (arXiv:1405.2464v2):1–7, 2015.

[9] M. Steel and M.J. Sanderson. Characterizing phylogenetically decisive taxon coverage. *Applied Mathematics Letters*, pages 82–86, 2010. doi: 10.1016/j.aml.2009.08.009.

A. Appendix

Proof $min_0(n)$ being a natural number

$min_0(n) \in \mathbb{N}$, for $n = 6m$, $m \in \mathbb{N}$

$$\begin{aligned}\frac{1}{6}\frac{n-2}{2}\binom{n}{2} + \frac{n}{6} &= \frac{n!}{4 \cdot 3! \cdot (n-3)!} + \frac{n}{3!} \\ &= \frac{1}{4! \cdot (n-3)!} \cdot (n! + 4 \cdot n \cdot (n-3)!) \\ &= \frac{(n-3)!}{4! \cdot (n-3)!} \cdot (n \cdot (n-1) \cdot (n-2) + 4n) \\ &= \frac{n}{24} \cdot (n^2 - 3n + 2 + 4) \\ &= \frac{6 \cdot 6 \cdot m}{24} \cdot (6 \cdot m^2 - 3m + 1) \\ &= \frac{1}{2} \cdot (18m^3 - 9m^2 + 3m)\end{aligned} \tag{25}$$

If m is even, both $9m^2$ and $3m$ are even, and therefore the sum. Thus the sum is divisible by 2.

If m is not even, both $9m^2$ and $3m$ are not even, but therefore the sum must be even. Thus the sum is divisible by 2.

$\Rightarrow min_0(n) \in \mathbb{N}$, for $n = 6m$, $m \in \mathbb{N}$

Partitions for $n = 6$ and $n = 7$

In the following are the tables with the partitions. First all partitions for $n = 6$ with quadruple set $S_{6,2}$ (compare chapter 4.2. and 5.3.1.) in Figure 6 (page 38), then two types of partitions for $n = 7$, namely $a|bc|de|fg$ and $a|b|cd|efg$, with quadruple set S_7 (compare chapter 5.3.2:) in Figure 7 - 9 (pages 39 - 41). The third type, $a|b|c|defg$, is not listed. It is not necessary as we know that this type only checks whether all triples are covered. If we know that all triples are covered, then we know that all partitions of that kind are covered as well.

The fist four columns give the partition, the fifth column the quadruple of the set that covers the partition.

S_{6,2} := 1235, 1245, 1246, 1346, 1356,1456, 2345, 2346, 2356
All partitions for n=6

1	2	34	56	1235
1	2	35	46	1245
1	2	36	45	1235
1	3	24	56	1346
1	3	25	46	1356
1	3	26	45	1346
1	4	23	56	1246
1	4	25	36	1246
1	4	26	35	1245
1	5	23	46	1356
1	5	24	36	1456
1	5	26	34	1235
1	6	23	45	1246
1	6	24	35	1346
1	6	25	34	1246

2	3	14	56	2345
2	3	15	46	2345
2	3	16	45	2356
2	4	13	56	2345
2	4	15	36	2345
2	4	16	35	2346
2	5	13	46	2356
2	5	14	36	2345
2	5	16	34	2356
2	6	13	45	2356
2	6	14	35	2346
2	6	15	34	2356

3	4	12	56	2345
3	4	15	26	2345
3	4	16	25	2346
3	5	12	46	1356
3	5	14	26	1356
3	5	16	24	1235
3	6	12	45	1356
3	6	14	25	1356
3	6	15	24	1346

4	5	12	36	1456
4	5	13	26	1456
4	5	16	23	1245
4	6	12	35	1456
4	6	13	25	1456
4	6	15	23	1246

5	6	12	34	1356
5	6	13	24	1456
5	6	14	23	1356

1	2	3	456	1235
1	2	4	356	1245
1	2	5	346	1235
1	2	6	345	1246
1	3	4	256	1346
1	3	5	246	1235
1	3	6	245	1346
1	4	5	236	1245
1	4	6	235	1246
1	5	6	234	1356

2	3	4	156	2345
2	3	5	146	2345
2	3	6	145	2346
2	4	5	136	1245
2	4	6	135	1246
2	5	6	134	2356

3	4	5	126	2345
3	4	6	125	2346
3	5	6	124	1356

4	5	6	123	1456

Figure 6: List of all partitions with 6 taxa. All partitions are covered at least once by one quadruple of $S_{6,2}$

S_7:= 1234, 1245, 1256, 1267, 1347, 1356, 1357, 1467, 2346, 2357, 2367, 2457, 3456, 4567

Partition-Type: a|bc|de|fg

1	23	45	67	1256
1	23	46	57	1256
1	23	47	56	1267
1	24	35	67	1256
1	24	36	57	1256
1	24	37	56	1267
1	25	34	67	1357
1	25	36	47	1357
1	25	37	46	1267
1	26	34	57	1245
1	26	35	47	1245
1	26	37	45	1234
1	27	34	56	1245
1	27	35	46	1245
1	27	36	45	1234

2	13	45	67	1256
2	13	46	57	1256
2	13	47	56	1267
2	14	35	67	1256
2	14	36	57	1256
2	14	37	56	1267
2	15	34	67	2357
2	15	36	47	1234
2	15	37	46	1234
2	16	34	57	1245
2	16	35	47	1245
2	16	37	45	1234
2	17	34	56	1245
2	17	35	46	1245
2	17	36	45	1234

3	12	45	67	1356
3	12	46	57	1356
3	12	47	56	1357
3	14	25	67	1356
3	14	26	57	1356
3	14	27	56	1357
3	15	24	67	1347
3	15	26	47	1234
3	15	27	46	1347
3	16	24	57	1347
3	16	25	47	1234
3	16	27	45	1234
3	17	24	56	2357
3	17	25	46	1356
3	17	26	45	1356

4	12	35	67	1347
4	12	36	57	1347
4	12	37	56	1467
4	13	25	67	2346
4	13	26	57	1245
4	13	27	56	1245
4	15	23	67	1347
4	15	26	37	1234
4	15	27	36	1234
4	16	23	57	1347
4	16	25	37	1234
4	16	27	35	1234
4	17	23	56	1245
4	17	25	36	1234
4	17	26	35	1234

5	12	34	67	1357
5	12	36	47	1357
5	12	37	46	1356
5	13	24	67	1256
5	13	26	47	1245
5	13	27	46	1245
5	14	23	67	1357
5	14	26	37	1356
5	14	27	36	1357
5	16	23	47	1357
5	16	24	37	4567
5	16	27	34	4567
5	17	23	46	1356
5	17	24	36	4567
5	17	26	34	4567

6	12	34	57	1356
6	12	35	47	2367
6	12	37	45	1356
6	13	24	57	1256
6	13	25	47	1267
6	13	27	45	1256
6	14	23	57	1256
6	14	25	37	1267
6	14	27	35	1256
6	15	23	47	1267
6	15	24	37	1267
6	15	27	34	1467
6	17	23	45	1256
6	17	24	35	4567
6	17	25	34	4567

7	12	34	56	1467
7	12	35	46	1347
7	12	36	45	1347
7	13	24	56	1267
7	13	25	46	1267
7	13	26	45	1467
7	14	23	56	1267
7	14	25	36	1267
7	14	26	35	2457
7	15	23	46	2457
7	15	24	36	4567
7	15	26	34	4567
7	16	23	45	1357
7	16	24	35	4567
7	16	25	34	4567

Figure 7: List of partition-type $a|bc|de|fg$ with 7 taxa. All partitions are covered at least once by one quadruple of S_7

S_7:= 1234, 1245, 1256, 1267, 1347, 1356, 1357, 1467, 2346, 2357, 2367, 2457, 3456, 4567

Partition-Type: a|b|cd|efg - Part 1

1	2	34	567	1245
1	2	35	467	1245
1	2	36	457	1234
1	2	37	456	1234
1	2	45	367	1234
1	2	46	357	1234
1	2	47	356	1234
1	2	56	347	1245
1	2	57	346	1245
1	2	67	345	1256

1	3	24	567	1347
1	3	25	467	1234
1	3	26	457	1234
1	3	27	456	1234
1	3	45	267	1234
1	3	46	257	1234
1	3	47	256	1234
1	3	56	247	1357
1	3	57	246	1356
1	3	67	245	1357

1	4	23	567	1245
1	4	25	367	1234
1	4	26	357	1234
1	4	27	356	1234
1	4	35	267	1234
1	4	36	257	1234
1	4	37	256	1234
1	4	56	237	1245
1	4	57	236	1245
1	4	67	235	1347

1	5	23	467	1256
1	5	24	367	1256
1	5	26	347	1245
1	5	27	346	1245
1	5	34	267	1245
1	5	36	247	1256
1	5	37	246	1356
1	5	46	237	1356
1	5	47	236	1245
1	5	67	234	1256

1	6	23	457	1256
1	6	24	357	1256
1	6	25	347	1356
1	6	27	345	1256
1	6	34	257	1356
1	6	35	247	1256
1	6	37	245	1356
1	6	45	237	1256
1	6	47	235	1267
1	6	57	234	1267

1	7	23	456	1267
1	7	24	356	1267
1	7	25	346	1267
1	7	26	345	1467
1	7	34	256	1467
1	7	35	246	1347
1	7	36	245	1347
1	7	45	236	1347
1	7	46	235	1347
1	7	56	234	1267

2	3	14	567	2346
2	3	15	467	1234
2	3	16	457	2346
2	3	17	456	1234
2	3	45	167	2346
2	3	46	157	1234
2	3	47	167	2346
2	3	56	147	2346
2	3	57	146	2367
2	3	67	145	2346

2	4	13	567	1245
2	4	15	367	1234
2	4	16	357	1245
2	4	17	356	1245
2	4	35	167	1245
2	4	36	157	1234
2	4	37	156	1234
2	4	56	137	1245
2	4	57	136	1245
2	4	67	135	2346

2	5	13	467	1245
2	5	14	367	1256
2	5	16	347	1245
2	5	17	346	1245
2	5	34	167	1245
2	5	36	147	1256
2	5	37	146	2457
2	5	46	137	1245
2	5	47	136	1245
2	5	67	134	1256

2	6	13	457	1256
2	6	14	357	1256
2	6	15	347	1267
2	6	17	345	1256
2	6	34	157	2367
2	6	35	147	1256
2	6	37	145	1267
2	6	45	137	1256
2	6	47	135	1267
2	6	57	134	1256

2	7	13	456	1267
2	7	14	356	1267
2	7	15	346	1267
2	7	16	345	2367
2	7	34	156	2367
2	7	35	146	2367
2	7	36	145	1267
2	7	45	136	2357
2	7	46	135	1267
2	7	56	134	1267

Figure 8: Part 1 of the partition-type $a|b|cd|efg$ with 7 taxa. All partitions are covered at least once by one quadruple of S_7

S_7:= 1234, 1245, 1256, 1267, 1347, 1356, 1357, 1467, 2346, 2357, 2367, 2457, 3456, 4567

Partition-Type: a|bc|de|fg - Part 2

3	4	12	567	1347
3	4	15	267	1347
3	4	16	257	1347
3	4	17	256	1234
3	4	25	167	1234
3	4	26	157	1234
3	4	27	156	1347
3	4	56	127	2346
3	4	57	126	1347
3	4	67	125	1347

3	7	12	456	1347
3	7	14	256	1357
3	7	15	246	1347
3	7	16	245	1347
3	7	24	156	1347
3	7	25	146	1357
3	7	26	145	2357
3	7	45	126	1347
3	7	46	125	1347
3	7	56	124	1357

4	5	12	367	2457
4	5	13	267	1245
4	5	16	237	1245
4	5	17	236	2457
4	5	23	167	2457
4	5	26	137	2457
4	5	27	136	1245
4	5	36	127	4567
4	5	37	126	2457
4	5	67	123	2457

5	6	12	347	1356
5	6	13	247	1256
5	6	14	237	1356
5	6	17	234	1356
5	6	23	147	1356
5	6	24	137	1256
5	6	27	134	1256
5	6	34	127	1356
5	6	37	124	1356
5	6	47	123	3456

3	5	12	467	1356
3	5	14	267	1356
3	5	16	247	1357
3	5	17	246	1356
3	5	24	167	2357
3	5	26	147	1356
3	5	27	146	1357
3	5	46	127	1356
3	5	47	126	1357
3	5	67	124	1356

4	6	12	357	1467
4	6	13	257	1467
4	6	15	237	1467
4	6	17	235	4567
4	6	23	157	3456
4	6	25	137	4567
4	6	27	135	1467
4	6	35	127	4567
4	6	37	125	1467
4	6	57	123	1467

5	7	12	346	1357
5	7	13	246	2357
5	7	14	236	1357
5	7	16	234	1357
5	7	23	146	1357
5	7	24	136	2357
5	7	26	134	2357
5	7	34	126	1357
5	7	36	124	1357
5	7	46	123	2457

3	6	12	457	1356
3	6	14	257	1356
3	6	15	247	3456
3	6	17	245	1356
3	6	24	157	3456
3	6	25	147	1356
3	6	27	145	2346
3	6	45	127	1356
3	6	47	125	3456
3	6	57	124	1356

4	7	12	356	1347
4	7	13	256	1467
4	7	15	236	1347
4	7	16	235	1347
4	7	23	156	1347
4	7	25	136	4567
4	7	26	135	1467
4	7	35	126	1347
4	7	36	125	1347
4	7	56	123	1467

6	7	12	345	1467
6	7	13	245	1467
6	7	14	235	1267
6	7	15	234	1467
6	7	23	145	1267
6	7	24	135	1467
6	7	25	134	1267
6	7	34	125	1467
6	7	35	124	2367
6	7	45	123	1467

Figure 9: Part 2 of the partition-type $a|b|cd|efg$ with 7 taxa. All partitions are covered at least once by one quadruple of S_7