Quartet Compatibility and the Quartet Graph

Stefan Grünewald[1], Peter J. Humphries[2], and Charles Semple[2*]

[1] CAS-MPG Partner Institute for Computational Biology
Shanghai Institutes for Biological Sciences
Shanghai, China
and
Max Planck Institute for Mathematics in the Sciences
Leipzig, Germany
stefan@picb.ac.cn

[2] Department of Mathematics and Statistics
University of Canterbury
Christchurch, New Zealand
p.humphries@math.canterbury.ac.nz, c.semple@math.canterbury.ac.nz

Submitted: Oct 13, 2005; Accepted: Aug 1, 2008; Published: Aug 11, 2008
Mathematics Subject Classification: 05C05, 92B10

Abstract

A collection $\mathcal{P}$ of phylogenetic trees is compatible if there exists a single phylogenetic tree that displays each of the trees in $\mathcal{P}$. Despite its computational difficulty, determining the compatibility of $\mathcal{P}$ is a fundamental task in evolutionary biology. Characterizations in terms of chordal graphs have been previously given for this problem as well as for the closely-related problems of (i) determining if $\mathcal{P}$ is definitive and (ii) determining if $\mathcal{P}$ identifies a phylogenetic tree. In this paper, we describe new characterizations of each of these problems in terms of edge colourings. Furthermore, making use of the tools that underlie these new characterizations, we also determine the minimum number of quartets required to identify an arbitrary phylogenetic tree, thus correcting a previously published result.

1 Introduction

Unrooted phylogenetic (evolutionary) trees are used in computational biology to represent the evolutionary relationships of a set X of extant species. A fundamental way in which such trees are inferred is by amalgamating a collection $\mathcal{P}$ of smaller phylogenetic trees on overlapping subsets of X into a single parent tree. Collectively, such amalgamation

*The first author was supported by the Allan Wilson Centre for Molecular Ecology and Evolution. The second and third authors were supported by the New Zealand Marsden Fund.

methods are known as *supertree methods* and the resulting parent tree is called a *supertree*. The popularity of supertree methods is highlighted in [1, 2].

If the amalgamating collection $\mathcal{P}$ contains no conflicting information, then $\mathcal{P}$ is said to be *compatible*. Furthermore, $\mathcal{P}$ is *definitive* if $\mathcal{P}$ is compatible and there is exactly one supertree that 'displays' all of the ancestral relationships displayed by the trees in $\mathcal{P}$. Precise definitions of these concepts are given in the next section. Within the context of supertree methods, two natural mathematical problems arise:

(i) is $\mathcal{P}$ compatible and, if so,

(ii) is $\mathcal{P}$ definitive?

As computational problems, (i) is known to be NP-complete [3, 11], while the complexity of the second problem continues to remain open. Nevertheless, there are attractive characterizations of these problems in terms of chordal graphs [5, 8, 9, 11].

In practice, while a collection $\mathcal{P}$ of phylogenetic trees might be compatible, it is unlikely to be definitive. A closely related notion, and one that is essentially as good, is the following: $\mathcal{P}$ *identifies* a supertree $\mathcal{T}$ if $\mathcal{T}$ displays $\mathcal{P}$ and all other supertrees that display $\mathcal{P}$ are 'refinements' of $\mathcal{T}$. This means that if $\mathcal{P}$ identifies a supertree, then the collection of supertrees that display $\mathcal{P}$ is well understood. This gives rise to a third mathematical problem:

(iii) does $\mathcal{P}$ identify a supertree?

Like problems (i) and (ii), a characterization of this problem has also been given in terms of chordal graphs [6].

Each of problems (i), (ii), and (iii) are typically stated in terms of collections of *quartets*—that is, binary phylogenetics trees with four leaves—rather than an arbitrary collection of phylogenetic trees. The reason for this is that a phylogenetic tree is completely determined by its collection of induced quartets (see, for example, [10]). Consequently, for the rest of the paper, we will view $\mathcal{P}$ as a collection of quartets.

In this paper, we introduce the 'quartet graph' and show that, in addition to the chordal graph characterizations, these problems can also be characterized in terms of edge colourings via this graph. One of the main motivations for the paper is that it is hoped that the quartet graph may provide new insights not only on the complexity of (ii) but also on other quartet problems in phylogenetics. Indeed, in the second half of the paper, we make use of the quartet graph and its associated concepts to determine, for a given phylogenetic tree $\mathcal{T}$, the size of a minimum-sized set of quartets that identifies $\mathcal{T}$. The resulting theorem corrects a previously published result [10].

The paper is organized as follows. The next section consists of preliminaries and formal statements of the main results of the paper. For completeness, Section 3 contains the chordal graph characterizations of problems (i)-(iii). Section 4 contains the proofs of the characterizations of (i)-(iii) in terms of quartet graphs. The proof of the compatibility characterization is algorithmic and thus provides a phylogenetic tree that displays the original collection of quartets if this collection is compatible. Section 5 contains the proof

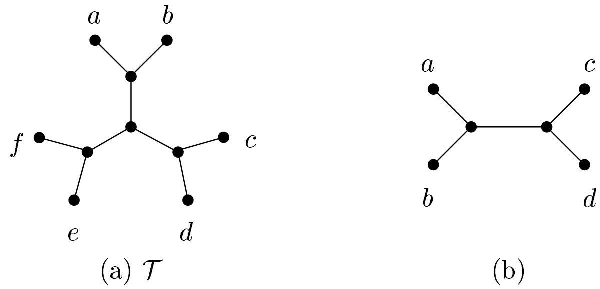

Figure 1: Two phylogenetic trees.

of the minimum number of quartets needed to identify a given phylogenetic tree as well as two closely-related optimality results. Throughout the paper, X will always denote a finite set, and notation and terminology follows [10].

2 Main Results

2.1 Phylogenetic Trees and Compatibility

A *phylogenetic X-tree* $\mathcal{T}$ is an unrooted tree in which every interior vertex has degree at least three and whose leaf set is X. In addition, if all interior vertices of $\mathcal{T}$ have degree three, then $\mathcal{T}$ is *binary*. The set X is called the *label set* of $\mathcal{T}$. A *quartet* is a binary phylogenetic tree whose label set has size 4. To illustrate, two phylogenetic trees are shown in Fig. 1, with the tree on the right being a quartet.

Let $\mathcal{T}$ and $\mathcal{T}'$ be two phylogenetic trees with label sets X and X', respectively, where $X \subseteq X'$. The *restriction* of $\mathcal{T}'$ to X, denoted $\mathcal{T}'|X$, is the phylogenetic tree that is obtained from the minimal subtree of $\mathcal{T}'$ connecting the elements in X by contracting degree-2 vertices. We say that $\mathcal{T}'$ *displays* $\mathcal{T}$ if $\mathcal{T}'|X$ is isomorphic to $\mathcal{T}$. For example, in Fig. 1, $\mathcal{T}$ displays the quartet in Fig. 1(b).

Now let $\mathcal{P}$ be a collection of phylogenetic trees. The *label set* of $\mathcal{P}$, denoted $\mathcal{L}(\mathcal{P})$, is the union of the label sets of the trees in $\mathcal{P}$. We say that a phylogenetic X-tree $\mathcal{T}$ *displays* $\mathcal{P}$ if $\mathcal{T}$ displays each of the trees in $\mathcal{P}$, in which case, $\mathcal{P}$ is said to be *compatible*. Furthermore, if $\mathcal{T}$ is the only such tree and $X = \mathcal{L}(\mathcal{P})$, then $\mathcal{P}$ is said to be *definitive*.

Associated with each edge e of a phylogenetic X-tree $\mathcal{T}$ is an *X-split*, that is, a bipartition of X into two non-empty parts. Here the two parts are $X \cap V_1$ and $X \cap V_2$, where V_1 and V_2 are the vertex sets of the two connected components of $\mathcal{T}\backslash e$. We say that a phylogenetic X-tree $\mathcal{T}'$ is a *refinement* of $\mathcal{T}$ if every X-split of $\mathcal{T}$ is an X-split of $\mathcal{T}'$. Note that $\mathcal{T}$ is a refinement of itself. Intuitively (and equivalently), $\mathcal{T}$ can be obtained from $\mathcal{T}'$ by contracting edges (see [10]. We say that a collection $\mathcal{P}$ of phylogenetic trees with $\mathcal{L}(\mathcal{P}) = X$ *identifies* $\mathcal{T}$ if $\mathcal{T}$ displays $\mathcal{P}$ and every phylogenetic X-tree that displays $\mathcal{P}$ is a refinement of $\mathcal{T}$.

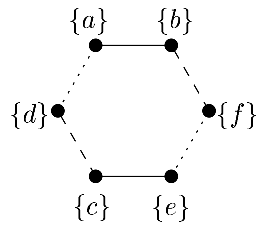

Figure 2: The quartet graph of $\{ab|ce, cd|bf, ef|ad\}$.

2.2 Quartets and the Quartet Graph

Let q be a quartet with label set $\{a, b, c, d\}$. If the path from a to b does not intersect the path from c to d, then we denote q by $ab|cd$ or, equivalently, $cd|ab$. For a collection $\mathcal{Q}$ of quartets with label set X, we define the *quartet graph* of $\mathcal{Q}$, denoted $G_{\mathcal{Q}}$, as follows. The vertex set of $G_{\mathcal{Q}}$ is the set of singletons of X and, for each $q = ab|cd \in \mathcal{Q}$, there is an edge joining $\{a\}$ and $\{b\}$, and an edge joining $\{c\}$ and $\{d\}$ each of which is labelled q. Apart from these edges, $G_{\mathcal{Q}}$ has no other edges. Note that if $q_1 = ab|cd, q_2 = ab|ce \in \mathcal{Q}$, then $G_{\mathcal{Q}}$ has edges $\{a, b\}$ and $\{c, d\}$ labelled q_1, and separate edges $\{a, b\}$ and $\{c, e\}$ labelled q_2. For purposes later in the paper, in reference to q, we sometimes use $\{a, b\}_q$ and $\{c, d\}_q$ to denote the two parts of q.

As an example of a quartet graph, consider the set $\mathcal{Q} = \{ab|ce, cd|bf, ef|ad\}$ of quartets. The quartet graph of $\mathcal{Q}$ is shown in Fig. 2, where, instead of labelling the edges with the appropriate element of $\mathcal{Q}$, we have used solid, dashed, and dotted lines to represent the edges arising from $ab|ce$, $cd|bf$, and $ef|ad$, respectively.

Each edge of $G_{\mathcal{Q}}$ has a partner, namely, the one which is labelled by the same quartet. Another way we could have indicated this is by assigning a distinct colour to each quartet in $\mathcal{Q}$, and then assigning this colour to each of the two edges corresponding to this quartet. In doing this, we observe that the resulting edge colouring of $G_{\mathcal{Q}}$ is a proper edge colouring. From this viewpoint, we say that an edge is *q-coloured* if it is labelled q. Recall that an *edge colouring* of a graph G is an assignment of colours to the edges of G. An edge colouring is *proper* if no two edges incident with the same vertex have the same colour.

2.3 Unification Operation

Central to this paper is a particular graphical operation that 'unifies' vertices. Let X be a non-empty finite set, and let G be an arbitrary graph with no loops and whose vertex set V is a partition of X, where no part is the empty set. In other words, X is the disjoint union of the vertices of G. Furthermore, suppose that G is properly edge-coloured. Let U be a subset of V with the property that if e and f are distinct edges of G with the same colour, then at most one of these edges is incident with a vertex in U. The *unification* of the vertices in U is the graph obtained from G by

(i) replacing the vertices in U together with every edge for which both end-vertices are

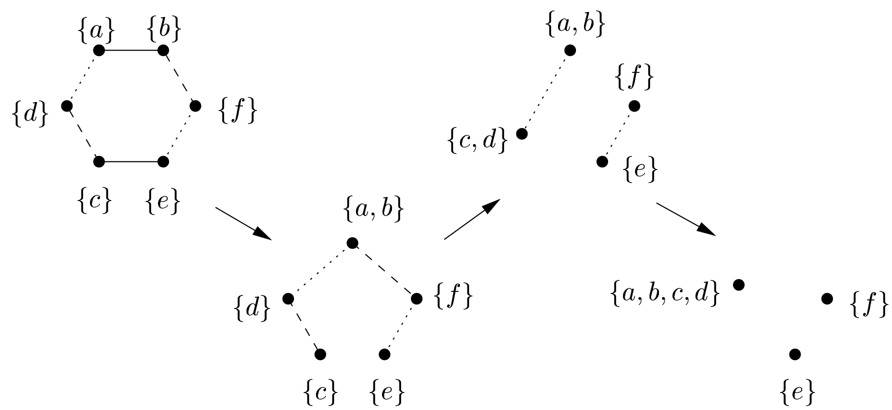

Figure 3: A complete-unification sequence of the quartet graph in Fig. 2.

in U by a single new vertex such that if an edge is incident with exactly one vertex in U, then it is incident with the resulting new vertex;

(ii) labelling the new vertex as the union of the elements in U; and

(iii) for each edge that joins two vertices in U, delete all other edges with the same colour.

Observe that, at the end of (ii), the resulting graph is properly edge-coloured.

Let $\mathcal{Q}$ be a collection of quartets on X. Noting that the quartet graph $G_\mathcal{Q}$ satisfies the above properties, let $G_0 = G_\mathcal{Q}, G_1, \ldots, G_k$ be a sequence $\mathcal{S}$ of graphs, where G_i is obtained from G_{i-1} by a unification for all $i \in \{1, \ldots, k\}$. We will call such a sequence a *unification sequence* of $G_\mathcal{Q}$. If G_k has no edges, then $\mathcal{S}$ is said to be *complete*. As a matter of convenience, for all $i \in \{1, \ldots, k\}$ we denote by $\mathcal{S}_i$ the unification sequence $G_0 = G_\mathcal{Q}, G_1, \ldots, G_i$.

Example 2.1. Consider the quartet graph $G_\mathcal{Q}$ shown in Fig. 2. Figure 3 illustrates a unification sequence of $G_\mathcal{Q}$ beginning with $G_\mathcal{Q}$ on the left and ending with the graph consisting of three isolated vertices on the right. Initially, we unify the vertices $\{a\}$ and $\{b\}$ to get the second graph. The third graph is obtained by unifying $\{c\}$ and $\{d\}$ in the second graph, while the last graph is obtained from the third graph by unifying $\{a, b\}$ and $\{c, d\}$. Since the last graph has no edges, this unification sequence is complete.

The following theorem characterizes the compatibility of a collection of quartets in terms of quartet graphs.

Theorem 2.1. *Let $\mathcal{Q}$ be a set of quartets. Then $\mathcal{Q}$ is compatible if and only if there is a complete-unification sequence of $G_\mathcal{Q}$.*

As an illustration of Theorem 2.1, the set $\mathcal{Q} = \{ab|ce, cd|bf, ef|ad\}$ is compatible since there is a complete-unification sequence of $G_\mathcal{Q}$ (see Fig. 3). Indeed, the phylogenetic tree $\mathcal{T}$ shown in Fig. 1(a) displays $\mathcal{Q}$.

To describe our characterizations of when a set of quartets identifies and defines a phylogenetic tree, we require some further definitions.

2.4 Distinguishing Quartets

Let $\mathcal{T}$ be a phylogenetic tree. We denote by $\mathcal{Q}(\mathcal{T})$ the set of quartets that are displayed by $\mathcal{T}$. Let $q = ab|cd \in \mathcal{Q}(\mathcal{T})$. An interior edge $e = uv$ of $\mathcal{T}$ is *distinguished* by q if, for one end-vertex of e, say u, the labels a and b are in separate components of $\mathcal{T}\backslash u$ and neither of these components contains v, while c and d are in separate components of $\mathcal{T}\backslash v$ and neither of these components contains u. A subset $\mathcal{Q} \subseteq \mathcal{Q}(\mathcal{T})$ distinguishes $\mathcal{T}$ if every interior edge of $\mathcal{T}$ is distinguished by some $q \in \mathcal{Q}$.

Let $\mathcal{T}$ be a phylogenetic X-tree that displays a collection $\mathcal{Q}$ of quartets on X, and let $e = uv$ be an interior edge of $\mathcal{T}$. We define $G_{\mathcal{Q}(u,v)}$ to be the graph that has the neighbours of v except u as its vertex set and where two vertices w_i, w_j are joined by an edge precisely if there is a quartet in $\mathcal{Q}$ that distinguishes e and is of the form $w_iw_j|xy$ for some $x, y \in X$. A set $\mathcal{Q}$ of quartets on X *specially distinguishes* a phylogenetic X-tree $\mathcal{T}$ if $\mathcal{T}$ displays $\mathcal{Q}$ and, for every interior edge $e = uv$ of $\mathcal{T}$, both $G_{\mathcal{Q}(u,v)}$ and $G_{\mathcal{Q}(v,u)}$ are connected.

2.5 Collecting Quartets and Unification Sequences

Let $\mathcal{Q}$ be a collection of quartets on X, and let $G_0 = G_{\mathcal{Q}}, G_1, \ldots, G_k$ be a unification sequence $\mathcal{S}$ of $G_{\mathcal{Q}}$. For all i, let U_i denote the subset of vertices of G_{i-1} that are unified to obtain G_i and let A_i denote the union of the elements of U_i. We will call $U_1, \ldots, U_k$ the sequence of *unifying sets associated with* $\mathcal{S}$. Observe that, for all i and j with $i < j$, we have that either $A_i \subseteq A_j$ or $A_i \cap A_j = \emptyset$. This observation will be used throughout the paper. Furthermore, we call the set

$$\Sigma(\mathcal{S}) = \{A_i|(X - A_i) : i \in \{1, \ldots, k\}\}$$

of X-splits the set of *X-splits induced by* $\mathcal{S}$

Now let $q = ab|cd$ be an element of $\mathcal{Q}$. If, for some j, either $\{a, b\}$ or $\{c, d\}$ is a subset of A_j, but neither $\{a, b\} \subseteq A_i$ nor $\{c, d\} \subseteq A_i$ for all $i < j$, then we say that q has been *collected* by U_j or, more generally, by $\mathcal{S}$. Moreover, if $\{a, b\} \subseteq A_j$ and, for all $i < j$, neither $\{a, b\} \subseteq A_i$ nor $\{c, d\} \subseteq A_i$, we say that A_j or, again more generally, $\mathcal{S}$ *merged* $\{a, b\}_q$. For a subset $\mathcal{Q}'$ of $\mathcal{Q}$, we denote the set

$$\big\{\{a, b\}_q : q = ab|cd \in \mathcal{Q}' \text{ and } \mathcal{S} \text{ merged } \{a, b\}_q\big\}$$

by $M(\mathcal{Q}')_{\mathcal{S}}$.

Lastly, if $\mathcal{S}$ is complete, then $\mathcal{S}$ is said to be *minimal* if there is no other complete-unification sequence $\mathcal{S}'$ with $U'_1, \ldots, U'_l$ as its sequence of unifying sets such that $\{A'_j : j \in \{1, \ldots, l\}\}$ is a proper subset of $\{A_i : i \in \{1, \ldots, k\}\}$, where A'_j is the union of the elements in U'_j for all j.

Theorem 2.2. *Let $\mathcal{Q}$ be a set of quartets on X. Then $\mathcal{Q}$ identifies a phylogenetic X-tree if and only if both of the following conditions hold:*

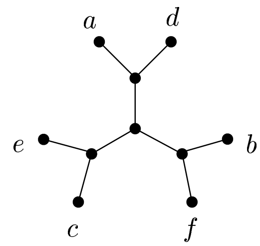

Figure 4: Another phylogenetic tree that displays $\mathcal{Q}$.

(i) There exists a phylogenetic X-tree $\mathcal{T}$ that displays $\mathcal{Q}$ and is specially distinguished by $\mathcal{Q}$.

(ii) Let $\mathcal{Q}'$ be a minimal subset of $\mathcal{Q}$ that specially distinguishes $\mathcal{T}$ and let $q = A|B \in \mathcal{Q}'$. Let $\mathcal{S}$ and $\mathcal{S}'$ be minimal complete-unification sequences of $G_{\mathcal{Q}}$ such that, amongst the quartets in $\mathcal{Q}'$, the quartet q is collected (joint) last and A is merged. Then $M(\mathcal{Q}')_{\mathcal{S}} = M(\mathcal{Q}')_{\mathcal{S}'}$.

Provided (i) holds in Theorem 2.2, we remark here that there is always at least one minimal complete-unification sequence that satisfies the assumption conditions in (ii). (See Lemma 4.5.)

Example 2.2. To illustrate Theorem 2.2, again consider the set of quartets

$$\mathcal{Q} = \{ab|ce, cd|bf, ef|ad\}.$$

As well as the phylogenetic tree $\mathcal{T}$ shown in Fig. 1(a), the phylogenetic tree shown in Fig. 4 also displays $\mathcal{Q}$. Since $\mathcal{Q}$ specially distinguishes $\mathcal{T}$, and the second tree is not a refinement of $\mathcal{T}$, the set $\mathcal{Q}$ does not identify any phylogenetic tree. This fact is realized by Theorem 2.2 as follows.

In addition to the complete-unification sequence $\mathcal{S}_1$ shown in Fig. 3, Fig. 5 shows a second complete-unification sequence $\mathcal{S}_2$ of $G_{\mathcal{Q}}$. Now, $\mathcal{Q}$ specially distinguishes $\mathcal{T}$. In both $\mathcal{S}_1$ and $\mathcal{S}_2$, the quartet $ef|ad$ is the last quartet of $\mathcal{Q}$ that is collected and $\{a, d\}$ is merged. Consider the quartet $ab|ce \in \mathcal{Q}$. In $\mathcal{S}_1$, we have that $\{a, b\}$ is merged, while, in $\mathcal{S}_2$, we have that $\{c, e\}$ is merged. Thus $M(\mathcal{Q})_{\mathcal{S}_1} \neq M(\mathcal{Q})_{\mathcal{S}_2}$. It now follows by Theorem 2.2 that $\mathcal{Q}$ does not identify a phylogenetic tree.

We remark here that the quartet set $\mathcal{Q}$ used in Example 2.2 shows that condition (i) by itself in Theorem 2.2 is not sufficient for a collection of quartets to identify a phylogenetic tree, as $\mathcal{Q}$ specially distinguishes the phylogenetic tree shown in Fig. 1.

It will turn out that a consequence of Theorem 2.2 is the next corollary.

Corollary 2.3. *Let $\mathcal{Q}$ be a set of quartets on X. Then $\mathcal{Q}$ defines a phylogenetic X-tree if and only if both of the following conditions hold:*

(i) There exists a binary phylogenetic X-tree $\mathcal{T}$ that displays $\mathcal{Q}$ and is distinguished by $\mathcal{Q}$.

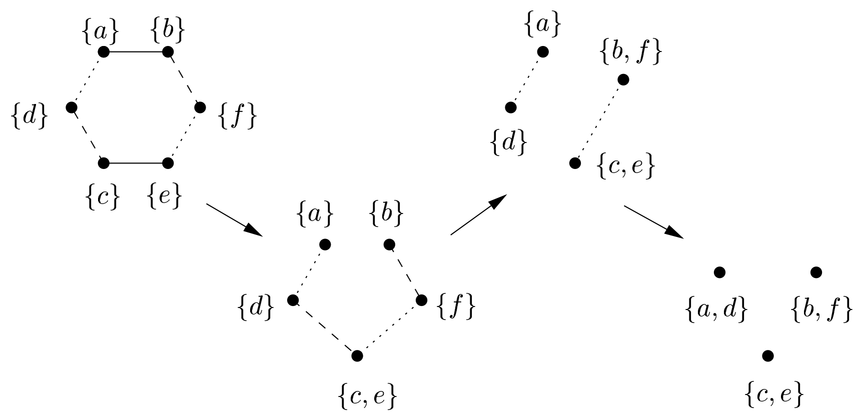

Figure 5: Another complete-unification sequence of the quartet graph in Fig. 2.

(ii) Let $\mathcal{Q}'$ be a minimum-sized subset of $\mathcal{Q}$ that distinguishes $\mathcal{T}$ and let $q \in \mathcal{Q}'$. Let $\mathcal{S}$ and $\mathcal{S}'$ be minimal complete-unification sequences of $G_{\mathcal{Q}}$ such that, amongst the quartets in $\mathcal{Q}'$, the quartet q is collected last. Then $M(\mathcal{Q}' - q)_{\mathcal{S}} = M(\mathcal{Q}' - q)_{\mathcal{S}'}$.

As mentioned in the introduction, in the second half of the paper we consider the problem of determining, for a given phylogenetic tree $\mathcal{T}$, the size of a minimum-sized set of quartets that identifies $\mathcal{T}$. In particular, we establish the following theorem. This corrects [10, Theorem 6.3.9] which incorrectly states that the size of such a set is $|X| - 3$, where X is the label set of $\mathcal{T}$. For binary phylogenetic trees, $|X| - 3$ is the correct size (corresponding to the number of interior edges of $\mathcal{T}$), but, for non-binary trees, the result is somewhat more complicated.

For a phylogenetic tree $\mathcal{T}$, let $\mathring{E}(\mathcal{T})$ denote the set of interior edges of $\mathcal{T}$ and let $d(u)$ denote the degree of a vertex u of $\mathcal{T}$. Let $q(\mathcal{T})$ denote the size of a minimum-sized set of quartets that identifies $\mathcal{T}$.

Theorem 2.4. *Let $\mathcal{T}$ be a phylogenetic X-tree and let $\mathcal{Q}$ be a collection of quartets that identifies $\mathcal{T}$. Then, for each interior edge $e = uv$ of $\mathcal{T}$ with $d(u) \leq d(v)$, the collection $\mathcal{Q}$ contains at least $q(d(u) - 1, d(v) - 1)$ quartets that distinguish e, where*

$$q(r, s) = \left\lceil \frac{r(s-1)}{2} \right\rceil$$

for all $r, s \geq 2$. In particular,

$$|\mathcal{Q}| \geq \sum_{uv \in \mathring{E}(\mathcal{T})} q(d(u) - 1, d(v) - 1).$$

Moreover, there exists a collection of quartets that identifies $\mathcal{T}$ and has size

$$q(\mathcal{T}) = \sum_{uv \in \mathring{E}(\mathcal{T})} q(d(u) - 1, d(v) - 1).$$

Restricting Theorem 2.4 to binary trees, where the notions of identify and define are equivalent, we get the following known result (see [10, Corollary 6.3.10] for example).

Corollary 2.5. *Let $\mathcal{T}$ be a binary phylogenetic X-tree and let $n = |X|$. Let $\mathcal{Q}$ be a collection of quartets that defines $\mathcal{T}$. Then $|Q| \geq n-3$. Moreover, there exists a collection of quartets that defines $\mathcal{T}$ and has size $n-3$.*

We end this section with some additional preliminaries.

2.6 Phylogenetic Trees and Splits

A *partial split* $A|B$ of X is a bipartition of a subset of X into two non-empty parts. If the disjoint union of A and B is X, then $A|B$ is a *split* of X. A partial split is *non-trivial* if $|A|, |B| \geq 2$. Recall that the edges of a phylogenetic X-tree $\mathcal{T}$ give rise to splits of X. The collection of non-trivial X- splits of $\mathcal{T}$ arising in this way is denoted by $\Sigma(\mathcal{T})$. We say that a partial split $A|B$ of X is *displayed* by $\mathcal{T}$ if there is an edge whose deletion results in two components, where A is a subset of the vertex set of one component and B is a subset of the vertex set of the other component. Observe that if $A = \{a_1, a_2\}$ and $B = \{b_1, b_2\}$, then $\mathcal{T}$ displays $A|B$ if and only if it displays the quartet $a_1a_2|b_1b_2$. Consequently, for the purposes of this paper, we will often use the quartet notation for such partial splits.

Buneman [4] showed that every phylogenetic tree is determined by its collection of non-trivial X-splits. A collection Σ of partial splits of X is *compatible* if there is a phylogenetic tree that displays each of the splits in Σ. The following result, which we will refer to as the Splits-Equivalence Theorem, is due to Buneman [4].

Theorem 2.6. *Let Σ be a non-trivial collection of X-splits. Then the following statements are equivalent:*

(i) *there is a phylogenetic X-tree $\mathcal{T}$ such that Σ is the set of non-trivial X-splits of $\mathcal{T}$;*

(ii) *Σ is pairwise compatible;*

(iii) *for each pair $A_1|B_1$ and $A_2|B_2$ of X-splits in Σ, at least one of the sets $A_1 \cap A_2$, $A_1 \cap B_2$, $B_1 \cap A_2$, and $B_1 \cap B_2$ is empty.*

Moreover, if such a phylogenetic X-tree exists, then, up to isomorphism, $\mathcal{T}$ is unique.

A *one-split* phylogenetic X-tree is a phylogenetic tree with exactly one interior edge. For example, a quartet is a one-split phylogenetic tree with four leaves. If the one non-trivial X-split of this tree is $\{a_1, \ldots, a_r\}|\{b_1, \ldots, b_s\}$, then we will denote this tree by $a_1 \cdots a_r|b_1 \cdots b_s$ or $A|B$, where $A = \{a_1, \ldots, a_r\}$ and $B = \{b_1, \ldots, b_s\}$.

3 Chordal Graph Characterizations

In this section, we state the chordal graph analogues of Theorems 2.1 and 2.2, and Corollary 2.3. This section is independent of the rest of the paper and so the reader may wish to initially skip it.

The *partition intersection graph* of a collection $\mathcal{Q}$ of quartets, denoted $\mathrm{int}(\mathcal{Q})$, is the vertex-coloured graph that has vertex set

$$\bigcup_{q=A|B\in\mathcal{Q}} \{(q,A),(q,B)\},$$

and an edge joining (q',B') and (q'',B'') precisely if $B'\cap B''$ is non-empty. Here two vertices are the same colour if they share the same first coordinate.

A graph is *chordal* if none of its vertex-induced subgraphs is isomorphic to a cycle with at least four vertices. A graph G is a *restricted chordal completion* of $\mathrm{int}(\mathcal{Q})$ if G is a chordal graph that can be obtained from $\mathrm{int}(\mathcal{Q})$ by only adding edges between vertices whose first coordinates are distinct. Note that this maintains the property of a proper vertex colouring. Theorem 3.1, the chordal graph analogue of Theorem 2.1, was indicated by Buneman [5] and Meacham [8], and formally proved by Steel [11].

Theorem 3.1. *Let $\mathcal{Q}$ be a set of quartets. Then $\mathcal{Q}$ is compatible if and only if there is a restricted chordal completion of* $\mathrm{int}(\mathcal{Q})$.

A restricted chordal completion G of $\mathrm{int}(\mathcal{Q})$ is *minimal* if, for every non-empty subset F of edges of $E(G)-E(\mathrm{int}(\mathcal{Q}))$, the graph $G\backslash F$ is not chordal. The next theorem is due to Semple and Steel [9].

Theorem 3.2. *Let $\mathcal{Q}$ be a set of quartets on X. Then there is a unique phylogenetic X-tree that displays $\mathcal{Q}$ if and only if the following two conditions hold:*

(i) there is a binary phylogenetic X-tree that displays $\mathcal{Q}$ and is distinguished by $\mathcal{Q}$; and

(ii) there is a unique minimal restricted chordal completion of $\mathrm{int}(\mathcal{Q})$.

To describe the chordal graph analogue of Theorem 2.2 requires some further definitions. Let $\mathcal{T}$ be a phylogenetic X-tree and let $e=u_1u_2$ be an edge of $\mathcal{T}$. Then e is *strongly distinguished* by a one-split phylogenetic tree $A_1|A_2$ if, for each i, the following hold:

(i) A_i is a subset of the vertex set of the component of $\mathcal{T}\backslash e$ containing u_i, and

(ii) the vertex set of each component of $T\backslash u_i$, except for the one containing the other end vertex of e, contains an element of A_i.

For a collection $\mathcal{Q}$ of quartets on X, let $\mathcal{G}(\mathcal{Q})$ denote the collection of graphs

$$\{G : \text{there is a phylogenetic } X\text{-tree } \mathcal{T} \text{ displaying } \mathcal{Q} \text{ with } G=\mathrm{int}(\mathcal{Q},\mathcal{T})\},$$

where $\mathrm{int}(\mathcal{Q},\mathcal{T})$ is the graph that has the same vertex set as $\mathrm{int}(\mathcal{Q})$, and an edge joining two vertices (q,A) and (q',A') if the vertex sets of the minimal subtrees of $\mathcal{T}$ connecting the elements in A and A' have a non-empty intersection. Note that if G is a graph in $\mathcal{G}(\mathcal{Q})$, then G is a restricted chordal completion of $\mathrm{int}(\mathcal{Q})$. There is a partial order $\leq$

on $\mathcal{G}(\mathcal{Q})$ which is obtained by setting $G_1 \leq G_2$ for all $G_1, G_2 \in \mathcal{G}(\mathcal{Q})$ if the edge set of G_1 is a subset of the edge set of G_2. Lastly, a compatible collection $\mathcal{Q}$ of quartets *infers* a one-split phylogenetic tree if every phylogenetic tree that displays $\mathcal{Q}$ also displays this one-split tree. Theorem 3.3 was established by Bordewich *et al.* [6].

Theorem 3.3. *Let $\mathcal{Q}$ be a set of quartets on X. Then $\mathcal{Q}$ identifies a phylogenetic X-tree if and only if the following conditions hold:*

(i) *there is a phylogenetic X-tree that displays $\mathcal{Q}$ and, for every edge e of this tree, there is a one-split phylogenetic tree inferred by $\mathcal{Q}$ that strongly distinguishes e; and*

(ii) *there is a unique maximal element in $\mathcal{G}(\mathcal{Q})$.*

Remark 1. Note that if $\mathcal{Q}$ is a collection of quartets, then $\text{int}(\mathcal{Q})$ is the line graph of the quartet graph $G_{\mathcal{Q}}$ where, for a graph G, the *line graph* of G has vertex set $E(G)$ and two vertices joined by an edge precisely if they are incident with a common vertex in G. The vertex colouring of the partition intersection graph corresponds to the edge colouring of the quartet graph. However, the characterizations of defining and identifying quartet sets described in this section and those ones derived in this paper are quite different and we do not use the duality between the partition intersection graph and the quartet graph to prove the new results.
Remark 2. The results stated in this section were originally proved for general 'characters' (that is, partitions of X) rather than for quartets. The concept of the quartet graph can be extended to this more general setup but then hypergraphs have to be considered. On the other hand, the phylogenetic information of characters can be expressed in terms of quartets thus no generality is lost in restricting our attention to quartets in this paper (see [10, Proposition 6.3.11]).

4 Proofs of Theorems 2.1 and 2.2, and Corollary 2.3

The proof of Theorem 2.1 is an immediate consequence of the next two lemmas.

Lemma 4.1. *Let $\mathcal{Q}$ be a set of quartets on X, and let $\mathcal{S}$ be a unification sequence of $G_{\mathcal{Q}}$. Then the set $\Sigma_{\mathcal{S}}$ of X-splits induced by $\mathcal{S}$ is compatible. Moreover, if $\mathcal{Q}'$ denotes the subset of $\mathcal{Q}$ collected by $\mathcal{S}$, then the phylogenetic X-tree whose set of non-trivial X-splits is $\Sigma_{\mathcal{S}}$ displays each of the quartets in $\mathcal{Q}'$, but no quartet in $\mathcal{Q} - \mathcal{Q}'$.*

Proof. Suppose that $\mathcal{S}$ is the sequence $G_0 = G_{\mathcal{Q}}, G_1, \ldots, G_k$ with unifying sequence $U_1, \ldots, U_k$. For all i, let A_i denote the union of the elements of U_i. The proof of the proposition is by induction on k. If $k = 0$, the result holds trivially. Now suppose that the result holds for all unification sequences of $G_{\mathcal{Q}}$ of smaller length, in particular, the result holds for the unification sequence $G_0 = G_{\mathcal{Q}}, G_1, \ldots, G_{k-1}$. Denote this last sequence by $\mathcal{S}'$.

Consider the X-split $A_k|(X - A_k)$, and note that, by the induction assumption, $\Sigma_{\mathcal{S}'}$ is compatible. Let $A_i|(X - A_i) \in \Sigma_{\mathcal{S}'}$. Since A_i is a subset of a vertex of G_{k-1}, either

$A_i \subseteq A_k$, in which case $A_i \cap (X - A_k) = \emptyset$, or $A_i \cap A_k = \emptyset$. In either case, by the Splits-Equivalence Theorem, $A_i|(X - A_i)$ and $A_k|(X - A_k)$ are compatible. It follows by the induction assumption and the Splits-Equivalence Theorem that $\Sigma_{\mathcal{S}}$ is compatible.

Let $\mathcal{T}$ denote the phylogenetic X-tree whose set of non-trivial X-splits is $\Sigma_{\mathcal{S}}$, and let $\mathcal{T}'$ denote the phylogenetic X-tree whose set of non-trivial X-splits is $\Sigma_{\mathcal{S}'}$. By the induction assumption, $\mathcal{T}'$ displays each of the quartets collected by $\mathcal{S}'$, but no other quartet in $\mathcal{Q}$. Assume that $ab|cd$ is a quartet collected by U_k. Then either $a, b \in A_k$ and $c, d \in X - A_k$, or $c, d \in A_k$ and $a, b \in X - A_k$, and so $\mathcal{T}$ displays $ab|cd$. Since $\mathcal{T}$ is a refinement of $\mathcal{T}'$, it follows that $\mathcal{T}$ displays each of the quartets collected by $\mathcal{S}$. Moreover, if $wx|yz$ is a quartet of $\mathcal{Q}$ not collected by $\mathcal{S}$, then, for all $i \in \{1, \ldots, k\}$,

$$\{w, x, y, z\} \cap A_i \notin \{\{w, x\}, \{y, z\}\},$$

and so $wx|yz$ is not displayed by $\mathcal{T}$. □

Given Lemma 4.1, we call the phylogenetic X-tree $\mathcal{T}$ whose set of non-trivial X-splits is equal to the set of X-splits induced by a unification sequence $\mathcal{S}$ the *phylogenetic X-tree induced by $\mathcal{S}$*.

Lemma 4.1 provides one direction of the proof of Theorem 2.1. The next lemma gives the other direction.

Let $\mathcal{Q}$ be a set of quartets on X and let $\mathcal{T}$ be a phylogenetic X-tree that displays $\mathcal{Q}$. Let v be a vertex of $\mathcal{T}$. Order the elements $A_1|(X - A_1), \ldots, A_k|(X - A_k)$ of $\Sigma(\mathcal{T})$ as follows:

(i) If e_i is the edge of $\mathcal{T}$ that induces $A_i|(X - A_i)$, then A_i is the subset of the vertex set of the component that does not contain v in $\mathcal{T} \backslash e_i$.

(ii) If $i < j$, then either $A_i \subseteq A_j$ or $A_i \cap A_j = \emptyset$.

It is easily checked that such an ordering is possible. Now let $\mathcal{S}_v$ denote the sequence of graphs $G_0 = G_{\mathcal{Q}}, G_1, \ldots, G_k$, where, for all i, the graph G_i is obtained from G_{i-1} by unifying the vertices whose disjoint union is A_i. It is easily seen that $\mathcal{S}_v$ is well-defined. The next lemma shows that $\mathcal{S}_v$ is a complete-unification sequence of $G_{\mathcal{Q}}$.

Lemma 4.2. *Let $\mathcal{Q}$ be a set of quartets on X and let $\mathcal{T}$ be a phylogenetic X-tree that displays $\mathcal{Q}$. Let v be a vertex of $\mathcal{T}$. Then $\mathcal{S}_v$ (as described above) is a complete-unification sequence of $G_{\mathcal{Q}}$.*

Proof. Suppose that $\mathcal{S}_v$ is not such a sequence and let j denote the smallest index for which G_j is not a unification of G_{j-1}. Since G_j is not a unification of G_{j-1}, there is a quartet, $ab|cd$ say, in $\mathcal{Q}$ not yet collected by $\mathcal{S}_v$ such that $|\{a, b, c, d\} \cap A_j| \geq 2$, where, in the case $|\{a, b, c, d\} \cap A_j| = 2$, we have $\{a, b, c, d\} \cap A_j \notin \{\{a, b\}, \{c, d\}\}$. If $|\{a, b, c, d\} \cap A_j| = 2$, then, by the construction of $\mathcal{S}_v$, the tree $\mathcal{T}$ does not display $ab|cd$; a contradiction. So we may assume that $|\{a, b, c, d\} \cap A_j| \geq 3$. But then by our choice of q, U_j contains three distinct vertices each having a non-empty intersection with $\{a, b, c, d\}$. This implies that no X-split of $\mathcal{T}$ displays q; a contradiction. Hence $\mathcal{S}_v$ is a unification sequence of $G_{\mathcal{Q}}$.

To see that $\mathcal{S}_v$ is complete, note that $\mathcal{T}$ displays $\mathcal{Q}$ and so, for each quartet, $ab|cd$ in $\mathcal{Q}$, there exists some i with the property that either $a, b \in A_i$ or $c, d \in A_i$. This establishes the lemma. $\square$

Proof of Theorem 2.1. This is now an immediate consequence of Lemmas 4.1 and 4.2. $\square$

We begin the proof of Theorem 2.2 with three lemmas.

Lemma 4.3. *Let $\mathcal{Q}$ be a collection of quartets on X. If $\mathcal{Q}$ identifies a phylogenetic X-tree $\mathcal{T}$, then $\mathcal{Q}$ specially distinguishes $\mathcal{T}$.*

Proof. Suppose that $\mathcal{Q}$ identifies $\mathcal{T}$, but does not specially distinguish $\mathcal{T}$. Then there exists an interior edge, uv say, of $\mathcal{T}$ such that $G_{\mathcal{Q}(u,v)}$ contains $k > 1$ components $C_1, \ldots, C_k$. We next construct a phylogenetic X-tree $\mathcal{T}'$ from $\mathcal{T}$ that displays $\mathcal{Q}$ but is not a refinement of $\mathcal{T}$.

Recalling the definition of $G_{\mathcal{Q}(u,v)}$, delete v and all its incident edges from $\mathcal{T}$. For each $i \in \{1, \ldots, k\}$, either add a new edge joining u and the vertex of C_i if C_i contains exactly one vertex, or adjoin a new vertex v_i to u via a new edge and, for each vertex w of C_i, add a new edge joining v_i and w. It is now easily seen that the resulting phylogenetic X-tree $\mathcal{T}'$ displays $\mathcal{Q}$. But $\mathcal{T}'$ is not a refinement of $\mathcal{T}$. It now follows that $\mathcal{Q}$ specially distinguishes $\mathcal{T}$. $\square$

A phylogenetic tree is *minimally refined* with respect to displaying a set $\mathcal{Q}$ of quartets if it is not a strict refinement of another phylogenetic tree that displays $\mathcal{Q}$.

Lemma 4.4. *Let $\mathcal{Q}$ be a compatible set of quartets on X. If $\mathcal{S}$ is a minimal complete-unification sequence of $G_{\mathcal{Q}}$, then the phylogenetic X-tree whose set of non-trivial X-splits is $\Sigma_{\mathcal{S}}$ is minimally refined with respect to displaying $\mathcal{Q}$.*

Proof. Suppose that $\mathcal{S}$ is the sequence $G_{\mathcal{Q}} = G_0, G_1, \ldots, G_k$ with unifying sequence $U_1, \ldots, U_k$, and let $\mathcal{T}$ be the phylogenetic X-tree whose set of non-trivial X-splits is $\Sigma_{\mathcal{S}}$. If $\mathcal{T}$ is not minimally refined with respect to displaying $\mathcal{Q}$, then there is an edge e of $\mathcal{T}$ whose contraction results in another phylogenetic X-tree, $\mathcal{T}'$ say, that displays $\mathcal{Q}$. Let $A_e|(X - A_e)$ denote the X-split of $\mathcal{T}$ displayed by e, where, for some i, A_e is the union of the elements of U_i.

Let $\mathcal{S}'$ be the sequence that is obtained from $\mathcal{S}$ by replacing the sequence of unifying sets associated with $\mathcal{S}$ with $U_1, \ldots, U_{i-1}, U'_{i+1}, \ldots, U'_k$, where, for all $j \in \{i+1, \ldots, k\}$,

$$U'_j = \begin{cases} (U_j - A_e) \cup U_i, & \text{if } A_e \text{ is an element of } U_j; \\ U_j, & \text{otherwise.} \end{cases}$$

Note that if, for some j, $U'_j \neq U_j$, then there is exactly one such j. To prove the lemma, it suffices to show that $\mathcal{S}'$ is a complete-unification sequence of $G_{\mathcal{Q}}$.

Clearly, $\mathcal{S}_{i-1}$ is a unification sequence of $G_{\mathcal{Q}}$. Consider G'_{i+1}. If $U'_{i+1} = U_{i+1}$, then it is easily seen that $G_{\mathcal{Q}} = G_0, G_1, \ldots, G_{i-1}, G'_{i+1}$ is a unification sequence of $G_{\mathcal{Q}}$. Therefore assume that $U'_{i+1} \neq U_{i+1}$. If $G_{\mathcal{Q}} = G_0, G_1, \ldots, G_{i-1}, G'_{i+1}$ is not a unification sequence, then there is a quartet, q say, in $\mathcal{Q}$ such that the two q-coloured edges are both incident with vertices in U'_{i+1}. Since $\mathcal{S}_{i+1}$ is a unification sequence of $G_{\mathcal{Q}}$, this implies that one of these q-coloured edges, ab say, is incident with two vertices in U_i, while the other q-coloured edge, cd say, is incident with at least one vertex in $U_{i+1} - A_e$. It now follows that $A_e|(X - A_e)$ is the unique X-split in $\Sigma_{\mathcal{S}}$ that displays q. In turn, this implies that $\mathcal{T}'$ does not display $\mathcal{Q}$; a contradiction. Thus $G_{\mathcal{Q}} = G_0, G_1, \ldots, G_{i-1}, G'_{i+1}$ is a unification sequence of $G_{\mathcal{Q}}$. Moreover, $G'_{i+1} = G_{i+1}$ and, for all $j \in \{i+2, \ldots, k\}$, we have $U'_j = U_j$. It now follows that in this case $\mathcal{S}'$ is a complete-unification sequence of $G_{\mathcal{Q}}$.

Considering, in turn, each of the graphs $G'_{i+2}, \ldots, G'_k$ and repeatedly using the same argument as that in the previous paragraph, we eventually deduce that either $\mathcal{S}'$ is a complete-unification sequence of $G_{\mathcal{Q}}$ or $\mathcal{S}'$ is a unification sequence but not complete. In the latter case, there is a $q' \in \mathcal{Q}$ such that G'_k contains two q'-coloured edges. By Lemma 4.1, the phylogenetic X-tree whose set of non-trivial X-splits is $\Sigma_{\mathcal{S}'}$ does not display q'. But, as $\mathcal{S}'$ is not complete, $U'_j = U_j$ for all j and so $\Sigma_{\mathcal{S}'} = \Sigma_{\mathcal{S}} - A_e|(X - A_e)$. But $\Sigma_{\mathcal{S}'}$ is the set of non-trivial X-splits of $\mathcal{T}'$ and so $\mathcal{T}'$ does not display q'; a contradiction. This completes the proof of the lemma. $\square$

Lemma 4.5. *Let $\mathcal{Q}$ be a set of quartets on X and let $\mathcal{T}$ be a phylogenetic X-tree that displays $\mathcal{Q}$ and is distinguished by $\mathcal{Q}$. Let $q = A|B$ be a quartet in $\mathcal{Q}$ that distinguishes an edge $e = uv$ of $\mathcal{T}$. Then there is a minimal complete-unification sequence of $G_{\mathcal{Q}}$ such that, amongst the quartets in $\mathcal{Q}$, the quartet q is collected (joint) last and A is merged. In particular, by choosing v to be the vertex of $\mathcal{T}$ such that the elements in A are in a different component of $\mathcal{T}\backslash e$ from v, the sequence $\mathcal{S}_v$ described prior to Lemma 4.2 is such a sequence.*

Proof. Suppose that q distinguishes the edge $e = uv$ of $\mathcal{T}$, and let $A_e|(X - A_e)$ denote the X-split induced by e. Without loss of generality, we may assume that the elements in A are in the same component of $\mathcal{T}\backslash e$ as u. Let $\mathcal{S}_v$ be the complete-unification sequence of $G_{\mathcal{Q}}$ as described prior to Lemma 4.2 with the additional proviso that $A_e|(X - A_e)$ is last in the associated ordering of the non-trivial X-splits induced by the edges of $\mathcal{T}$. It is easily seen using Lemma 4.2 that such an ordering and sequence is possible.

To complete the proof of the lemma, we show that $\mathcal{S}_v$ is minimal. If not, then there is a complete-unification sequence $\mathcal{S}$ of $G_{\mathcal{Q}}$ such that $\Sigma_{\mathcal{S}}$ is a proper subset of $\mathcal{S}_v$. But then $\mathcal{T}$ is a proper refinement of the phylogenetic tree whose set of non-trivial X-splits is $\Sigma_{\mathcal{S}}$. Since this last tree also displays $\mathcal{Q}$, we contradict the fact that $\mathcal{Q}$ distinguishes $\mathcal{T}$. Thus $\mathcal{S}_v$ is minimal. $\square$

Proof of Theorem 2.2. First suppose that $\mathcal{Q}$ identifies a phylogenetic tree $\mathcal{T}$. Then, by Lemma 4.3, (i) holds for $\mathcal{T}$. We next show that (ii) holds for $\mathcal{T}$. Let $\mathcal{Q}'$ be a minimal subset of $\mathcal{Q}$ that specially distinguishes $\mathcal{T}$ and let $q = A|B \in \mathcal{Q}'$. Let $\mathcal{S}$ and $\mathcal{S}'$ be two minimal complete-unification sequences of $G_{\mathcal{Q}}$ such that amongst the quartets in $\mathcal{Q}'$, the quartet q is collected (joint) last and A is merged. Let $q' = A'|B' \in \mathcal{Q}'$ and suppose that, in $\mathcal{S}$, the set A' is merged, while, in $\mathcal{S}'$, the set B' is merged. Furthermore, suppose that A_i merged A' and A_j merged A in $\mathcal{S}$, and that $A_{i'}$ merged B' and $A_{j'}$ merged A in $\mathcal{S}'$.

Since $\mathcal{Q}$ identifies $\mathcal{T}$, it follows by Lemma 4.4 that the phylogenetic X-trees whose set of non-trivial X-splits are $\Sigma_{\mathcal{S}}$ and $\Sigma_{\mathcal{S}'}$ are both isomorphic to $\mathcal{T}$, in particular, $\Sigma_{\mathcal{S}} = \Sigma_{\mathcal{S}'}$. Since $\mathcal{Q}'$ is a minimal subset of $\mathcal{Q}$ that specially distinguishes $\mathcal{T}$, both q and q' distinguish edges of $\mathcal{T}$, and so exactly one split of $\Sigma_{\mathcal{S}}$ displays q and exactly one split of $\Sigma_{\mathcal{S}}$ displays q'. This implies that $A_i = (X - A_{i'})$ (so $A_{i'} = (X - A_i)$) and $A_j = A_{j'}$. Up to symmetry, there are two cases to consider:

(I) $A_i \subseteq A_j$ and $A_{i'} \subseteq A_{j'}$; and

(II) $A_i \subseteq A_j$ and $A_{i'} \cap A_{j'} = \emptyset$.

If (I) holds, then A_j contains $X - A_i$. But A_j contains A_i, and so A_j contains X; a contradiction. Consider (II). Since $A_{i'} \cap A_{j'} = \emptyset$, we have $(X - A_i) \cap A_j = \emptyset$. But $A_i \subseteq A_j$, so $A_i = A_j$. Therefore, as q is collected (joint) last amongst the quartets in $\mathcal{Q}'$ in $\mathcal{S}$, $i = j$. Thus, as $A_i = A_j = A_{j'}$, we have $A_{i'} = (X - A_{j'})$. But $i' \leq j'$, and so $\mathcal{S}'$ merges B; a contradiction. Hence (II) does not hold. It now follows that (ii) does indeed hold.

To prove the converse, suppose that, in the size of its label set, $\mathcal{Q}$ is a minimal collection of quartets that satisfies (i) and (ii), but does not identify a phylogenetic tree. Since $\mathcal{T}$ is specially distinguished by $\mathcal{Q}$, it follows that $\mathcal{T}$ is minimally refined with respect to displaying $\mathcal{Q}$. Let $\mathcal{T}'$ be another phylogenetic X-tree that is minimally refined with respect to displaying $\mathcal{Q}$.

We will show that every X-split of $\mathcal{T}$ is also an X-split of $\mathcal{T}'$, contradicting the assumption that $\mathcal{T}'$ is minimally refined and different from $\mathcal{T}$. Assume not. Let $\mathcal{Q}'$ be a minimal subset of $\mathcal{Q}$ that specially distinguishes $\mathcal{T}$, and let $q = ab|cd$ be a quartet in $\mathcal{Q}'$ such that the subset of X-splits in $\Sigma(\mathcal{T}')$ that display q is minimal and does not contain any X-split induced by $\mathcal{T}$. Such a quartet exists, since every quartet in $\mathcal{Q}'$ distinguishes an edge of $\mathcal{T}$ and thus is displayed by exactly one X-split of $\mathcal{T}$. Therefore, a quartet that is displayed by an X-split in $\Sigma(\mathcal{T}) - \Sigma(\mathcal{T}')$ is not displayed by any X-split in $\Sigma(\mathcal{T}) \cap \Sigma(\mathcal{T}')$. Let $A|B$ be the X-split of $\mathcal{T}$ that displays q. Without loss of generality, we may assume that $a, b \in A$. Let H be the graph that has vertex set X and an edge joining to vertices g and h precisely if $\{g, h\} \in M(\mathcal{Q}')_{\mathcal{S}}$, where $\mathcal{S}$ is a minimal complete-unification sequence of $G_{\mathcal{Q}}$ that collects q (joint) last amongst the quartets in $\mathcal{Q}'$ and merges $\{a, b\}$.

We claim that the vertex set of the connected component of H that contains a and b also contains A. Assume the claim is wrong and choose $A'|B' \in \Sigma(\mathcal{T})$ such that A' is minimal with the property that $A' \subseteq A$ and that there is no component of H whose vertex set contains A'. Let $L_1, \ldots, L_k$ be the (pairwise different) maximal proper subsets of A' such that, for all $i \in \{1, \ldots, k\}$, the bipartition $L_i|(X - L_i)$ is an X-split of $\mathcal{T}$. For all i, it follows from the minimality of A' that there is a component of H that contains L_i. Let H' be the graph that has vertex set $L_1, \ldots, L_k$ and an edge joining to vertices L_i and L_j precisely if there is a quartet $gg'|hh' \in \mathcal{Q}'$ with $g \in L_i$, $g' \in L_j$, and $h, h' \in B'$. Since $\mathcal{Q}'$ specially distinguishes $\mathcal{T}$ the graph H' is connected. It now follows by Lemma 4.5 and the fact that (ii) holds for $\mathcal{T}$ that, for all such $gg'|hh'$, we have $\{g, g'\} \in M(\mathcal{Q}')_{\mathcal{S}}$. Hence there is a connected component of H whose vertex set contains A'; a contradiction. This establishes the claim.

By Lemma 4.5, there is a minimal complete-unification sequence $\mathcal{S}'$ of $G_{\mathcal{Q}}$ that collects q (joint) last amongst the quartets in $\mathcal{Q}'$ and merges $\{a, b\}$ such that $\mathcal{T}'$ is the phylogenetic tree induced by $\mathcal{S}'$. Noting that $M(\mathcal{Q}')_{\mathcal{S}'} = M(\mathcal{Q}')_{\mathcal{S}}$, it is easily seen that, as there is a connected component of H whose vertex set contains A, the graph obtained from $\mathcal{T}'$ by deleting all edges corresponding to X-splits that display q has a connected component whose vertex set contains A. By repeating the above argument using $\{c, d\}$ instead of $\{a, b\}$, the same graph also has a connected component whose vertex set contains B. Hence $A|B \in \Sigma(\mathcal{T}')$. This completes the proof of the converse and thus the theorem. $\square$

Proof of Corollary 2.3. Suppose that $\mathcal{Q}$ defines a phylogenetic X-tree $\mathcal{T}$. Then it is clear that (i) holds for $\mathcal{T}$. To show that (ii) holds for $\mathcal{T}$, let $\mathcal{Q}'$ be a minimum-sized subset of $\mathcal{Q}$ that distinguishes $\mathcal{T}$. First note that, for distinct $q, q' \in \mathcal{Q}'$, the quartets q and q' distinguish different edges of $\mathcal{T}$. Let $q = A|B \in \mathcal{Q}'$. Let $\mathcal{S}$ and $\mathcal{S}'$ be two minimal complete-unification sequences of $G_{\mathcal{Q}}$ so that amongst the quartets in $\mathcal{Q}'$, the quartet q is collected last. If both $\mathcal{S}$ and $\mathcal{S}'$ merge A, or both $\mathcal{S}$ and $\mathcal{S}'$ merge B, then, by Theorem 2.2, (ii) holds. Furthermore, making use of the note, the argument for the case that one of the sequences, $\mathcal{S}$ say, merges A and the other sequence, $\mathcal{S}'$ say, merges B is similar to that used in the analogous part in the proof of Theorem 2.2. We omit the straightforward details.

Now suppose that (i) and (ii) hold. Then, by Theorem 2.2, $\mathcal{Q}$ identifies a phylogenetic tree. Since $\mathcal{T}$ is a binary phylogenetic tree that displays $\mathcal{Q}$ and is distinguished by $\mathcal{Q}$, we deduce that $\mathcal{Q}$ defines $\mathcal{T}$. This completes the proof of the corollary. $\square$

5 Minimum Identifying Sets of Quartets

The main result of this section is Theorem 2.4. To establish this result, we begin by describing some partial split (inference) rules. For a set Σ of partial splits, we write $\Sigma \vdash A|B$ if every phylogenetic tree that displays Σ also displays $A|B$. The statement $\Sigma \vdash A|B$ is called a *partial split rule*. The input to the first two rules are quartets (see [7]):

$$\{ab|cd, ab|ce\} \vdash ab|cde; \tag{dc}$$

$$\{ab|de, ac|df, bc|ef\} \vdash abc|def. \tag{tc}$$

These rules are examples of so-called dyadic and triadic rules, respectively. The third rule says that if $A_1|B_1$ and $A_2|B_2$ are partial splits, $A_1 \cap A_2 \neq \emptyset$, and $B_1 \cap B_2 \neq \emptyset$, then

$$\{A_1|B_1, A_2|B_2\} \vdash (A_1 \cap A_2)|(B_1 \cup B_2). \tag{sc}$$

The rule (sc) is "Rule 1" in [8]. Observe that (dc) is a special case of (sc).

The next lemma is obtained by repeated application of (dc). The proof is routine and omitted.

Lemma 5.1. *Let $A|B$ be a non-trivial partial split of a set X, and let*

$$\mathcal{Q}(A|B) = \{aa'|bb' : a, a' \in A \text{ and } b, b' \in B\}.$$

Then $\mathcal{Q}(A|B) \vdash A|B$.

Lemma 5.2 generalizes (tc).

Lemma 5.2. *Let $\Sigma = \{A_1|B_1, A_2|B_2, A_3|B_3\}$ be a set of partial splits of X such that $A_i \cap A_j \neq \emptyset, B_i \cap B_j \neq \emptyset$ for all $i \neq j$. Then*

$$\Sigma \vdash \bigcup_{i \neq j}(A_i \cap A_j) | \bigcup_{i \neq j}(B_i \cap B_j).$$

Proof. By Lemma 5.1, it suffices to show that every $q = xy|wz$, where $x, y \in \bigcup_{i \neq j}(A_i \cap A_j)$ and $w, z \in \bigcup_{i \neq j}(B_i \cap B_j)$, is inferred by Σ. Clearly, this holds if $x, y \in A_i$ and $w, z \in B_i$ for some i. Therefore assume that this does not happen. Then, without loss of generality, we may assume that $x \in A_1 \cap A_2$, $y \in A_1 \cap A_3$, and $z \in B_2 \cap B_3$. By symmetry, there are two cases to consider depending on whether $w \in B_1 \cap B_2$ or $w \in B_2 \cap B_3$.

Let $a \in A_2 \cap A_3$ and $b \in B_1 \cap B_3$. If $w \in B_1 \cap B_2$, then, as $xy|wb \in \mathcal{Q}(A_1|B_1)$, $xa|wz \in \mathcal{Q}(A_2|B_2)$, and $ya|zb \in \mathcal{Q}(A_3|B_3)$, it follows by (tc) that

$$\{xy|wb, xa|wz, ya|zb\} \vdash xya|wzb.$$

Hence, in this case, q is inferred by Σ.

If $w \in B_2 \cap B_3$, then $xa|wz \in \mathcal{Q}(A_2|B_2)$ and $ya|wz \in \mathcal{Q}(A_3|B_3)$. Therefore, by (dc), Σ infers $xya|wz$ which in turn infers q. This completes the proof of the lemma. □

Analogously to a collection of phylogenetic trees, a collection Σ of partial splits *identifies* a phylogenetic tree $\mathcal{T}$ if $\mathcal{T}$ displays Σ and all phylogenetic trees that display Σ are refinements of $\mathcal{T}$.

Lemma 5.3. *Let $\mathcal{T}$ be a one-split phylogenetic tree in which the unique non-trivial split is $A|B$ with $A = \{a_1, \ldots, a_r\}$ and $B = \{b_1, \ldots, b_s\}$. Then, for non-negative integers m and n with $r \leq 2m - 1$ and $s \leq 2n - 1$, the 2-element collection*

$$\Sigma = \left\{a_1 \cdots a_m | b_1 \cdots b_n, a_{r-m+1} \cdots a_r | b_{s-n+1} \cdots b_s\right\}$$

of partial splits together with the collection

$$\mathcal{Q} = \left\{a_i a_{m+i} | b_j b_{n+j} : 1 \leq i \leq r - m, 1 \leq j \leq s - n\right\}$$

of quartets identifies $\mathcal{T}$.

Proof. Let

$$A' = \{a_1, \ldots, a_m\} \cap \{a_{r-m+1}, \ldots, a_r\}$$

and

$$B' = \{b_1, \ldots, b_n\} \cap \{b_{s-n+1}, \ldots, b_s\}.$$

Since $r \leq 2m-1$ and $s \leq 2n-1$, it follows that both A' and B' are non-empty. Therefore, by Lemma 5.2, the two partial splits in Σ together with the quartet $a_i a_{m+i}|b_j b_{n+j}$ infer the partial split

$$(A' \cup \{a_i, a_{m+i}\})|(B' \cup \{b_j, b_{n+j}\}) \tag{1}$$

for all i and j. Furthermore, by repeated applications of (sc), the partial splits of the form (1) infer $(A' \cup \{a_i, a_{m+i}\})|B$ for all i. Repeatedly using (sc) again, these last partial splits infer $A|B$. It now follows that the partial splits in Σ together with the quartets in $\mathcal{Q}$ identify $\mathcal{T}$. $\square$

For a one-split phylogenetic tree $\mathcal{T}$ whose non-trivial split is $A|B$ with $|A| \leq |B|$, we will denote the size of a minimum-sized set of quartets that identifies $\mathcal{T}$ by $q(|A|, |B|)$. Much of the work in proving Theorem 2.4 goes into proving the next lemma, a special case of that theorem.

Lemma 5.4. *Let $\mathcal{T}$ be a one-split phylogenetic X-tree in which the only non-trivial split is $A|B$ with $|A| = r$ and $|B| = s$, where $2 \leq r \leq s$. Then*

$$q(r, s) = \left\lceil \frac{r(s-1)}{2} \right\rceil.$$

Proof. Throughout the proof, we will assume that $A = \{a_1, \ldots, a_r\}$ and $B = \{b_1, \ldots, b_s\}$. We first show that $q(r, s) \geq \lceil \frac{r(s-1)}{2} \rceil$.

Suppose that $\mathcal{Q}$ is a set of quartets that identifies $\mathcal{T}$ with $|\mathcal{Q}| < \frac{r(s-1)}{2}$, and consider the quartet graph $G_{\mathcal{Q}}$. Since $\mathcal{Q}$ identifies $\mathcal{T}$, no edge in $G_{\mathcal{Q}}$ joins a singleton of A to a singleton of B, and, in view of Lemma 4.3, $G_{\mathcal{Q}}$ consists of two components whose vertex sets are the set of singletons of A and the set of singletons of B. Furthermore, if $q \in \mathcal{Q}$, then there is a q-coloured edge joining a pair of singletons of A and a q-coloured edge joining a pair of singletons of B. Since $|\mathcal{Q}| < \frac{r(s-1)}{2}$ and $r \leq s$, there is a vertex $\{a\} \subset A$ that is incident with at most $s-2$ differently coloured edges.

Let G_a be the subgraph of G_Q that is obtained by deleting all of the singletons of A and deleting all edges whose colour is not that of any coloured edge incident with $\{a\}$ in $G_{\mathcal{Q}}$. Hence, G_a has s vertices and at most $s-2$ edges and is therefore disconnected. Let $C_1, \ldots, C_k$ be the connected components of G_a containing at least two vertices. As $\mathcal{Q}$ specially distinguishes $\mathcal{T}$, we have $k \geq 1$. Now consider the unification sequence $\mathcal{S}$ of $G_{\mathcal{Q}} = G_0, G_1, \ldots, G_{k+1}$ in which we make the following unifications:

(i) For $1 \leq i \leq k$, unify the vertices in C_i of G_{i-1} to obtain G_i;

(ii) unify $\{a\}$ together with the set of vertices whose union is B to obtain G_{k+1}.

It is easily checked that $\mathcal{S}$ is a complete-unification sequence of $G_{\mathcal{Q}}$. By Lemma 4.1, the phylogenetic tree $\mathcal{T}'$ whose set of non-trivial X-splits is $\Sigma_{\mathcal{S}}$ displays $\mathcal{Q}$. But $A|B$ is not an X-split of $\mathcal{T}'$, and so $\mathcal{T}'$ is not a refinement of $\mathcal{T}$, contradicting that $\mathcal{Q}$ identifies $\mathcal{T}$. We conclude that $q(r,s) \geq \lceil \frac{r(s-1)}{2} \rceil$.

We next show that $q(r,s) \leq \lceil \frac{r(s-1)}{2} \rceil$ for all r and s. We begin with the case $r=2$.

5.4.1. *For all s, we have $q(2,s) \leq \lceil \frac{2(s-1)}{2} \rceil = s-1$.*

Proof. Here $A|B = \{a_1, a_2\}|\{b_1, \ldots, b_s\}$ and it follows by repeated applications of (sc) that the collection

$$\mathcal{Q} = \big\{a_1a_2|b_1b_i : i \in \{2, \ldots, s\}\big\}$$

of quartets identifies $\mathcal{T}$. As $|\mathcal{Q}| = s-1$, the inequality holds for $r=2$. □

5.4.2. *For all r, we have $q(r,r) \leq \frac{r(r-1)}{2}$.*

Proof. Let $\mathcal{Q}_r$ be the collection $\{a_ia_j|b_ib_j : 1 \leq i < j \leq r\}$ of quartets. Then $|\mathcal{Q}_r| = \binom{r}{2} = \frac{r(r-1)}{2}$. The proof is by induction on r. Clearly, the result holds for $r=2$. Now suppose that $r \geq 3$ and that the result holds for all smaller values of r. Then the partial split $a_1 \cdots a_{r-1}|b_1 \cdots b_{r-1}$ can be identified by $\mathcal{Q}_{r-1}$. By (tc), the quartets in $\mathcal{Q}_{r-1}$ and $\mathcal{Q}_r - \mathcal{Q}_{r-1}$ infer each of the partial splits in

$$\{a_ia_ja_r|b_ib_jb_r : 1 \leq i < j < r\}.$$

Moreover, by repeatedly applying (sc), we deduce that the elements in this set infer $a_1 \cdots a_r|b_1 \cdots b_r$. □

5.4.3. *For all r and all s with $r \leq s \leq 2r-2$, we have $q(r,s) \leq \lceil \frac{r(s-1)}{2} \rceil$.*

Proof. The proof is by induction on r. If $r=2$, then the result holds by (5.4.1). Now suppose that $r \geq 3$, and that the result holds for all smaller values of r. There are five cases to consider.

Case 1. $s = 2l-1$ for some integer $l \geq 2$.

By Lemma 5.3, the 2-element collection

$$\Sigma_1 = \big\{a_1 \cdots a_l|b_1 \cdots b_l, a_{r-l+1} \cdots a_r|b_l \cdots b_s\big\}$$

of partial splits together with the collection

$$\mathcal{Q}_1 = \big\{a_ia_{l+i}|b_jb_{l+j} : 1 \leq i \leq r-l, 1 \leq j \leq l-1\big\}$$

of quartets identify $\mathcal{T}$. By the induction assumption, each partial split in Σ_1 can be identified by a collection of $\frac{l(l-1)}{2}$ quartets. Furthermore, $\mathcal{Q}_1$ contains $(r-l)(l-1)$ quartets. Thus

$$\begin{aligned} q(r,s) &\leq l(l-1) + (r-l)(l-1) \\ &= \frac{r(s-1)}{2}. \end{aligned}$$

Case 2. $r = 2k$ and $s = 2l$ for some integers $k \geq 2$ and $l \geq 3$, where either k is odd or l is even.
By Lemma 5.3, the 2-element collection

$$\Sigma_2 = \left\{a_1 \cdots a_{k+1}|b_1 \cdots b_{l+1}, a_k \cdots a_r|b_l \cdots b_s\right\}$$

of partial splits together with the collection

$$\mathcal{Q}_2 = \left\{a_i a_{k+i+1}|b_j b_{l+j+1} : 1 \leq i \leq k-1, 1 \leq j \leq l-1\right\}$$

of quartets identify $\mathcal{T}$. By the induction assumption, each partial split in Σ_2 can be identified by a collection of $\frac{(k+1)l}{2}$ quartets. Without loss of generality, we may assume that these last collections share the quartet $a_k a_{k+1}|b_l b_{l+1}$. Furthermore, $\mathcal{Q}_2$ contains $(k-1)(l-1)$ quartets. Thus

$$\begin{aligned} q(r,s) &\leq (k+1)l - 1 + (k-1)(l-1) \\ &= \frac{r(s-1)}{2}. \end{aligned}$$

Case 3. $r = 2k-1$ and $s = 2l$ for some integers $k \geq 2$ and $l \geq 3$, where either k is odd or l is even.
By Lemma 5.3, the 2-element collection

$$\Sigma_3 = \left\{a_1 \cdots a_{k+1}|b_1 \cdots b_{l+1}, a_{k-1} \cdots a_r|b_l \cdots b_s\right\}$$

of partial splits together with the collection

$$\mathcal{Q}_3 = \left\{a_i a_{k+i+1}|b_j b_{l+j+1} : 1 \leq i \leq k-2, 1 \leq j \leq l-1\right\}$$

of quartets identify $\mathcal{T}$. By the induction assumption, each partial split in Σ_3 can be identified by a collection of $\frac{(k+1)l}{2}$ quartets. Without loss of generality, we may assume that these last collections share the quartet $a_k a_{k+1}|b_l b_{l+1}$. Furthermore, $\mathcal{Q}_3$ contains $(k-2)(l-1)$ quartets. Thus

$$\begin{aligned} q(r,s) &\leq (k+1)l - 1 + (k-2)(l-1) \\ &= \left\lceil \frac{r(s-1)}{2} \right\rceil. \end{aligned}$$

Case 4. $r = 4k$ and $s = 4l-2$ for integers $k \geq 1$ and $l \geq 2$.
This case includes an anomaly. In particular, when $l = 2$, that is $(r,s) = (4,6)$. We will prove this subcase first before proving Case 4 in general.

Let

$$\begin{aligned} \mathcal{Q}'_1 &= \left\{a_1a_2|b_1b_2, a_1a_3|b_1b_3, a_2a_3|b_2b_3\right\}, \\ \mathcal{Q}'_2 &= \left\{a_2a_3|b_4b_5, a_2a_4|b_4b_6, a_3a_4|b_5b_6\right\}, \end{aligned}$$

and

$$\mathcal{Q}'_3 = \left\{a_1a_2|b_3b_4, a_3a_4|b_3b_4, a_1a_4|b_1b_5, a_1a_4|b_2b_6\right\}.$$

By (tc), $\mathcal{Q}'_1$ and $\mathcal{Q}'_2$ infer the partial splits $a_1a_2a_3|b_1b_2b_3$ and $a_2a_3a_4|b_4b_5b_6$, respectively. Furthermore, together with $\mathcal{Q}'_3$, these partial splits infer $a_1a_2|b_1b_2b_3b_4$ and $a_3a_4|b_3b_4b_5b_6$ by (sc). By (tc), the partial splits $a_1a_2|b_1b_4$, $a_2a_4|b_4b_5$ $a_1a_4|b_1b_5$ infer $a_1a_2a_4|b_1b_4b_5$. Similarly, by (tc), we infer

$$a_1a_2a_4|b_2b_4b_6, a_1a_3a_4|b_1b_3b_5, a_1a_3a_4|b_2b_3b_6.$$

In turn, again using (tc), we infer

$$a_1a_2a_3|b_3b_4b_5, a_1a_2a_3|b_3b_4b_6, a_2a_3a_4|b_1b_3b_4, a_2a_3a_4|b_2b_3b_4.$$

The last eight partial splits now infer $a_1a_2|B$, $a_2a_3|B$, and $a_3a_4|B$ which, by (sc), infers $A|B$. Thus $q(4,6) \leq 10 = \frac{4(6-1)}{2}$.

Now assume that $k \geq 2$ and $l \geq 3$. By Lemma 5.3, the 2-element collection

$$\Sigma_4 = \left\{a_1 \cdots a_{2k+2}|b_1 \cdots b_{2l+1}, a_{2k-1} \cdots a_r|b_{2l-2} \cdots b_s\right\}$$

of partial splits together with the collection

$$\mathcal{Q}_4 = \left\{a_i a_{2k+i+2}|b_j b_{2l+j+1} : 1 \leq i \leq 2k-2, 1 \leq j \leq 2l-3\right\}$$

of quartets identifies $\mathcal{T}$. By the induction assumption, each partial split in Σ_4 can be identified by a collection of $(2k+2)l$ quartets. Consider one of these partial splits, say $a_1 \cdots a_{2k+2}|b_1 \cdots b_{2l+1}$. Since the size of the larger side is $2l+1 \geq 7$ and odd, we may make up the set of $(2k+2)l$ quartets that identify this partial split as in Case 1, where, by (5.4.2), we may assume that this set contains

$$\begin{aligned}\{&a_{2k-1}a_{2k}|b_{2l-2}b_{2l-1}, a_{2k-1}a_{2k+1}|b_{2l-2}b_{2l}, a_{2k}a_{2k+1}|b_{2l-1}b_{2l},\\ &a_{2k-1}a_{2k+2}|b_{2l-2}b_{2l+1}, a_{2k}a_{2k+2}|b_{2l-1}b_{2l+1}, a_{2k+1}a_{2k+2}|b_{2l}b_{2l+1}\}.\end{aligned}$$

Similarly, we may assume the set of $(2k+2)l$ quartets that identifies the other partial split in Σ_4 also contains the six quartets in this set. Since $\mathcal{Q}_4$ contains $(2k-2)(2l-3)$ quartets, it now follows that

$$\begin{aligned} q(r,s) &\leq 2(2k+2)l - 6 + (2k-2)(2l-3) \\ &= \frac{r(s-1)}{2}. \end{aligned}$$

Case 5. $r = 4k-1$ and $s = 4l-2$ for some integers $k \geq 1$ and $l \geq 2$.
By Lemma 5.3, the 2-element collection

$$\Sigma_5 = \left\{a_1 \cdots a_{2k}|b_1 \cdots b_{2l}, a_{2k} \cdots a_r|b_{2l-1} \cdots b_s\right\}$$

of partial splits together with the collection

$$\mathcal{Q}_5 = \left\{a_i a_{2k+i}|b_j b_{2l+j} : 1 \leq i \leq 2k-1, 1 \leq j \leq 2l-2\right\}$$

of quartets identifies $\mathcal{T}$. By the induction assumption, each partial split in Σ_5 can be identified by a collection of $k(2l-1)$ quartets. Furthermore, $\mathcal{Q}_5$ contains $(2k-1)(2l-2)$ quartets. Thus

$$\begin{aligned} q(r,s) &\leq 2k(2l-1) + (2k-1)(2l-2) \\ &= \left\lceil \frac{r(s-1)}{2} \right\rceil. \end{aligned}$$

Combining Cases 1-5, we conclude that $q(r,s) \leq \lceil \frac{r(s-1)}{2} \rceil$ whenever $r \leq s \leq 2r-2$. □

We complete the proof of Lemma 5.4 by showing that, for any fixed r, the result holds for all s with $r \leq s$. By (5.4.3), the result holds whenever $s \leq 2r-2$. Now assume that $s > 2r-2$ and that the result holds for all smaller values of s.

Consider the 2-element collection

$$\Sigma = \left\{ a_1 \cdots a_r | b_1 \cdots b_r, a_1 \cdots a_r | b_r \cdots b_s \right\}$$

of partial splits. Observe that, as $s > 2r-2$, we have $|\{a_1, \ldots, a_r\}| \leq |\{b_r, \ldots, b_s\}|$. By a single application of (sc), Σ infers $A|B$. Furthermore, by (5.4.2), the first partial split in Σ can be identified by a collection of $\frac{r(r-1)}{2}$ quartets and, by the induction assumption, the second partial split in Σ can be identified by a collection of $\lceil \frac{r(s-r)}{2} \rceil$ quartets. Hence

$$\begin{aligned} q(r,s) &\leq \frac{r(r-1)}{2} + \left\lceil \frac{r(s-r)}{2} \right\rceil \\ &= \left\lceil \frac{r(s-1)}{2} \right\rceil. \end{aligned}$$

Running over all values of r, we deduce that

$$q(r,s) \leq \left\lceil \frac{r(s-1)}{2} \right\rceil$$

for all r and all s with $2 \leq r \leq s$. This completes the proof of the lemma. □

The next lemma is an immediate consequence of the definition of identifying quartet sets.

Lemma 5.5. *Let $\mathcal{T}$ be a one-split phylogenetic X-tree in which the only non-trivial split is $A|B$, and suppose that $\mathcal{T}$ displays a collection $\mathcal{Q}$ of quartets. If $\mathcal{Q}$ does not identify $\mathcal{T}$, then there is a phylogenetic X-tree that displays $\mathcal{Q}$, but for which $A|B \notin \Sigma(\mathcal{T})$.*

Before proving Theorem 2.4, we require one further definition. An interior vertex of a tree that is adjacent to exactly k leaves is called a *k-bud*, or more generally a bud.

Proof of Theorem 2.4. First suppose that for some interior edge $e = uv$ of $\mathcal{T}$, the subset $\mathcal{Q}_e$ of $\mathcal{Q}$ containing exactly the quartets that distinguish e has the property that

$$|\mathcal{Q}_e| < q(d(u) - 1, d(v) - 1).$$

Suppose the neighbours of u that are not v are $u_1, \ldots, u_r$ and the neighbours of v that are not u are $v_1, \ldots, v_s$. Let $\mathcal{T}_e$ denote the phylogenetic tree that is the minimal subtree of $\mathcal{T}$ containing the vertices in $\{u_1, \ldots, u_r, v_1, \ldots, v_s\}$. Furthermore, let $\mathcal{P}_e$ be the collection of quartets obtained from $\mathcal{Q}_e$ by replacing each quartet, $aa'|bb'$ say, with $u_i u_j | v_k v_l$, where u_i is on the path from u to a, u_j is on the path from u to a', v_k is on the path from v to b, and v_l is on the path from v to b'. Since $\mathcal{T}$ displays $\mathcal{Q}_e$, it follows that $\mathcal{T}_e$ displays $\mathcal{P}_e$. However, because of the cardinality of $\mathcal{Q}_e$, it follows by Lemma 5.4 that $\mathcal{P}_e$ does not identify $\mathcal{T}_e$.

By Lemma 5.5, there is a phylogenetic tree $\mathcal{T}'_e$ with label set $\{u_1, \ldots, u_r, v_1, \ldots, v_s\}$ that displays $\mathcal{P}_e$ but does not contain the split $\{u_1, \ldots, u_r\}|\{v_1, \ldots, v_s\}$. Let $\mathcal{T}'$ be the phylogenetic X-tree that is obtained by adjoining, for all $w \in \{u_1, \ldots, u_r, v_1, \ldots, v_s\}$, the maximal subtree of $\mathcal{T}$ that contains w and neither u nor v to $\mathcal{T}'_e$ by identifying the common vertices denoted by w. Clearly, $\mathcal{T}'$ displays $\mathcal{Q}_e$. Moreover, it is easily seen by the construction of $\mathcal{T}'$ that every quartet in $\mathcal{Q} - \mathcal{Q}_e$ is also displayed by $\mathcal{T}'$. Since $\mathcal{T}'$ does not contain the split of $\mathcal{T}$ induced by e, we deduce that $\mathcal{Q}$ does not identify $\mathcal{T}$. This contradiction means that, for every interior edge $e = uv$, the collection $\mathcal{Q}$ contains $q(d(u) - 1, d(v) - 1)$ quartets that distinguish e. Thus

$$|\mathcal{Q}| \geq \sum_{e \in \mathring{E}} q(d(u) - 1, d(v) - 1).$$

We prove the second part of the theorem by induction on the number m of interior edges of $\mathcal{T}$. If $m = 1$ and the unique interior edge is uv, then, by Lemma 5.4, there exists a collection of quartets of size $q(d(u) - 1, d(v) - 1)$ that identifies $\mathcal{T}$. Now assume that $m \geq 2$ and that the result holds for every phylogenetic tree with $m - 1$ interior edges.

Let $e = uv$ be an interior edge of $\mathcal{T}$ such that u is a bud of $\mathcal{T}$. First assume that $d(u) \leq d(v)$. Let $r = d(u) - 1$ and $s = d(v) - 1$. Furthermore, let $a_1, \cdots, a_r$ be the leaves of $\mathcal{T}$ adjacent to u, and let $b_1, \cdots, b_s$ be leaves of $\mathcal{T}$ such that, for all distinct i and j, the path from b_i to b_j contains v, but not u. Let $\mathcal{T}' = \mathcal{T}|(X - \{a_2, \cdots, a_r\})$. Now $\mathcal{T}'$ is a phylogenetic tree with precisely $m - 1$ interior edges, and so by our induction assumption $\mathcal{T}'$ can be identified by a collection $\mathcal{Q}'$ of quartets of size $q(\mathcal{T}')$.

Let $\mathcal{Q}_e$ be a minimum-sized set of quartets that identifies the one-split phylogenetic tree whose non-trivial split is $a_1 \cdots a_r | b_1 \cdots b_s$. By Lemma 5.4, $|\mathcal{Q}_e| = q(r, s)$. Consider $\mathcal{Q}_e \cup \mathcal{Q}'$. Clearly, $\mathcal{T}$ displays $\mathcal{Q}_e \cup \mathcal{Q}'$. Let $\mathcal{T}''$ be a phylogenetic tree that displays $\mathcal{Q}_e \cup \mathcal{Q}'$. Since $\mathcal{Q}'$ identifies $\mathcal{T}'$, we have that $\mathcal{T}''|(X - \{a_2, \ldots, a_r\})$ is a refinement of $\mathcal{T}'$. Using this fact and the fact that $\mathcal{T}''$ displays $\mathcal{Q}_e$, it is easily seen that $\mathcal{T}''$ displays the partial split $a_1 \cdots a_r | b_1 \cdots b_s$. It now follows that $\mathcal{Q}_e \cup \mathcal{Q}'$ identifies $\mathcal{T}$. Moreover,

$$|\mathcal{Q}_e \cup \mathcal{Q}'| = q(d(u) - 1, d(v) - 1) + q(\mathcal{T}') = q(\mathcal{T}).$$

The same argument holds if $d(v) < d(u)$. This completes the proof of the theorem. $\square$

Recall that $q(\mathcal{T})$ denotes the size of a minimum-sized set of quartets that identifies a phylogenetic tree $\mathcal{T}$. We end this section with two results that determine, for all n, those phylogenetic trees $\mathcal{T}$ with n leaves for which $q(\mathcal{T})$ is minimized and maximized.

Proposition 5.6. *Let $\mathcal{T}$ be a phylogenetic X-tree with n leaves and at least one interior edge. Then $q(\mathcal{T}) \geq n-3$. Moreover, $q(\mathcal{T}) = n-3$ if and only if*

(i) $\mathcal{T}$ has exactly one interior edge and contains a 2-bud or two 3-buds; or

(ii) $\mathcal{T}$ has at least two interior edges and every vertex with degree at least four is a bud.

Proof. First suppose that $\mathcal{T}$ has exactly one interior edge uv. Let $r = d(u) - 1 \geq 2$ and $s = d(v) - 1 \geq 2$. Without loss of generality, we may assume that $r \leq s$. Then, by Theorem 2.4,

$$q(\mathcal{T}) = q(r, s) = \left\lceil \frac{r(s-1)}{2} \right\rceil.$$

It is easily checked that $q(\mathcal{T}) \geq r + s - 3$. Furthermore, a routine check also shows that $q(\mathcal{T}) = r + s - 3$ if and only if $r = 2$ or $s = 3$. As $r + s - 3 = n - 3$, the proposition holds over all phylogenetic trees with exactly one interior edge.

Next we show that the proposition holds in general. The proof is by induction on n. Clearly, the result holds if $n = 4$. Let $\mathcal{T}$ be a phylogenetic tree with n leaves, where $n \geq 5$, and suppose that $q(\mathcal{T})$ is of minimum size. Suppose that the proposition holds for all phylogenetic trees $\mathcal{T}'$ with fewer leaves for which $q(\mathcal{T}')$ is of minimum size. Since we already know that the result holds if $\mathcal{T}$ has exactly one interior edge, we may assume that $\mathcal{T}$ has at least two interior edges. Since every binary phylogenetic tree with n leaves is defined by $n-3$ quartets (see, for example, [10]), $q(\mathcal{T}) \leq n-3$. Let w be a bud of $\mathcal{T}$ of maximum size. Let j be the size of this bud, let $x_1, \ldots, x_j$ denote the leaves adjacent to w, let v be the non-leaf vertex adjacent to w, and let $\mathcal{T}'$ be the restriction of $\mathcal{T}$ to $X - \{x_j\}$. By the induction assumption, $q(\mathcal{T}') \geq (n-1) - 3 = n - 4$. We consider the two cases $j \geq 3$ and $j = 2$ separately.

Suppose firstly that $j \geq 3$. If $d(w) \leq d(v)$, then, by Theorem 2.4,

$$\begin{aligned} q(\mathcal{T}) - q(\mathcal{T}') &= q(j, d(v)-1) - q(j-1, d(v)-1) \\ &= \left\lceil \frac{j(d(v)-2)}{2} \right\rceil - \left\lceil \frac{(j-1)(d(v)-2)}{2} \right\rceil \\ &\geq 1. \end{aligned}$$

Therefore

$$q(\mathcal{T}) \geq q(\mathcal{T}') + 1 \geq n - 4 + 1 = n - 3. \tag{2}$$

Since $q(\mathcal{T}) \leq n-3$, it follows that equality holds throughout (2) and so $q(\mathcal{T}) = n-3$ and $q(\mathcal{T}') = n-4$. Since $\mathcal{T}$ has at least two interior edges and $k \geq 3$, the phylogenetic tree $\mathcal{T}'$ has at least two interior edges and so, by the induction assumption, (ii) holds for $\mathcal{T}'$. Hence (ii) holds for $\mathcal{T}$. A similar argument also shows that (ii) holds for $\mathcal{T}$ if $d(w) > d(v)$.

Now suppose that $j = 2$. Here every bud of $\mathcal{T}$ has size two. Note that, in this case, $d(w) \leq d(v)$. By Theorem 2.4,

$$q(\mathcal{T}) - q(\mathcal{T}') = q(2, d(v) - 1) = d(v) - 2 \geq 1.$$

Arguing as in (i), we now deduce that $q(\mathcal{T}) = n-3$ and $q(\mathcal{T}') = n-4$. This implies that $d(v) - 2 = 1$ and so $d(v) = 3$. If $\mathcal{T}'$ has at least two interior edges, then (ii) holds for $\mathcal{T}'$ and so (ii) holds for $\mathcal{T}$. Furthermore, if $\mathcal{T}'$ has exactly one interior edge, then $\mathcal{T}'$ is a quartet and again it follows that (ii) holds for $\mathcal{T}$. This completes the proof of the proposition. $\square$

For two non-negative integers k and l with $k + l \geq 3$, we will denote by $\mathcal{T}_k^{2l}$ the phylogenetic tree with $k + 2l$ leaves that has an interior vertex adjacent to k leaves while all other l neighbours are 2-buds.

Theorem 5.7. *Let $\mathcal{T}$ be a phylogenetic X-tree with n leaves. Then $q(\mathcal{T}) \leq \left\lfloor \left(\frac{n}{2} - 1\right)^2 \right\rfloor$. Moreover, $q(\mathcal{T}) = \left\lfloor \left(\frac{n}{2} - 1\right)^2 \right\rfloor$ if and only if $\mathcal{T}$ is isomorphic to*

(i) $\mathcal{T}_2^{n-2}$ if n is even; or

(ii) $\mathcal{T}_1^{n-1}$ or $\mathcal{T}_3^{n-3}$ if n is odd.

Proof. First note that, for $1 \leq k \leq 3$, a routine check using Theorem 2.4 shows that $q(\mathcal{T}_k^{n-k}) = \left\lfloor \left(\frac{n}{2} - 1\right)^2 \right\rfloor$. In other words, $q(\mathcal{T}_2^{n-2}) = \left(\frac{n}{2} - 1\right)^2$ if n is even and $q(\mathcal{T}_1^{n-1}) = q(\mathcal{T}_3^{n-3}) = \frac{(n-1)(n-3)}{4}$ if n is odd. The proof is by induction on n. A simple check shows that the result holds if $n \in \{4, 5\}$. Let $\mathcal{T}$ be a phylogenetic tree with n leaves, where $n \geq 6$, and suppose that $q(\mathcal{T})$ is of maximum size. Note that

$$q(\mathcal{T}) \geq \left\lfloor \left(\frac{n}{2} - 1\right)^2 \right\rfloor. \tag{3}$$

Suppose that the theorem holds for all phylogenetic trees $\mathcal{T}'$ with fewer leaves for which $q(\mathcal{T}')$ is of minimum size. Say $\mathcal{T}$ has exactly one interior edge. Then one of the interior vertices is a j-bud with $j \leq \frac{n}{2}$ and the other interior vertex is an $(n-j)$-bud. Consequently, by Theorem 2.4,

$$q(\mathcal{T}) = \frac{1}{2} j(n-j-1) \leq \frac{1}{2}\left(\frac{n-1}{2}\right)^2 < \left\lfloor \left(\frac{n}{2} - 1\right)^2 \right\rfloor$$

as $n \geq 6$. It now follows that $\mathcal{T}$ has at least two interior edges, which also means that $\mathcal{T}$ has no adjacent buds.

Let w be a bud of $\mathcal{T}$ of maximum size and let k be the size of this bud. Let $x_1, \dots, x_k$ denote the leaves adjacent to w, let v be the non-leaf vertex adjacent to w, and let $\mathcal{T}'$ be the restriction of $\mathcal{T}$ to $X - \{x_k\}$. By the induction assumption, $q(\mathcal{T}') \leq \left\lfloor \left(\frac{n-1}{2} - 1\right)^2 \right\rfloor$. Combining this with (3), we deduce that

$$q(\mathcal{T}) - q(\mathcal{T}') \geq \left\lceil \frac{n-3}{2} \right\rceil. \tag{4}$$

First suppose $k \geq 3$. Then, by Theorem 2.4, $q(\mathcal{T}) - q(\mathcal{T}') = q(k, d(v) - 1) - q(k - 1, d(v) - 1)$ and a routine check shows that $q(\mathcal{T}) - q(\mathcal{T}') \leq \frac{d(v)}{2}$. Together with (4), this implies that $d(v) \geq n - 2$ if n is even and $d(v) \geq n - 3$ if n is odd. Since $\mathcal{T}$ has at least two interior edges and w is adjacent to $k \geq 3$ leaves, this is only possible if n is odd, $k = 3$, and v is adjacent to $n - 5$ leaves and a 2-bud. Assuming n is odd, $n \geq 7$ and so, by Theorem 2.4,

$$q(\mathcal{T}) = q(2, n-4) + q(3, n-4) = \frac{5}{2}(n-5) < \frac{(n-1)(n-3)}{4};$$

a contradiction.

Now suppose that $k = 2$. By Theorem 2.4, $q(\mathcal{T}) - q(\mathcal{T}') = q(2, d(v) - 1) = d(v) - 2$. Therefore, by (4), $d(v) \geq \frac{n+1}{2}$. Assume that $\mathcal{T}$ has an interior vertex $v' \neq v$ such that v' is adjacent to a bud. Then, as v is adjacent to a bud, there are at least $d(v) \geq \frac{n+1}{2}$ leaves ℓ of $\mathcal{T}$ for which v' is not contained in the path from ℓ to v. Interchanging v and v' in this argument, we also deduce that there are at least $d(v) \geq \frac{n+1}{2}$ leaves ℓ of $\mathcal{T}$ for which v is not contained in the path from ℓ to v'. Hence $\mathcal{T}$ has at least $n+1$ leaves; a contradiction.

It follows from the above arguments that $\mathcal{T}$ has exactly one interior vertex that is not a bud and all buds are 2-buds. Thus, for some k, we have that $\mathcal{T}$ is isomorphic to $\mathcal{T}_k^{n-k}$. Now

$$\begin{aligned}
q(\mathcal{T}_k^{n-k}) &= \frac{n-k}{2} q\left(2, \frac{n+k}{2} - 1\right) \\
&= \frac{n-k}{2}\left(\frac{n+k}{2} - 2\right) \\
&= \frac{1}{4}(n - 2 + (k-2))(n - 2 - (k-2))
\end{aligned}$$

and, since k and n must have the same parity, $q(\mathcal{T}_k^{n-k})$ is maximum for $k = 2$ if n is even and for $k \in \{1, 3\}$ if n is odd. This completes the proof of the theorem. □

Acknowledgments

We thank the referees for their constructive and helpful comments.

References

[1] Bininda-Emonds, O. R. P., Gittleman, J. L., and Steel, M. A. (2002), The (super)tree of life: procedures, problems and prospects, *Annual Reviews of Ecology and Systematics*, **33**, 265-289.

[2] Bininda-Emonds, O. R. P. ed. (2004). *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*. Computational Biology Series, Kluwer.

[3] Bodlaender, H. L., Fellows, M. R., and Warnow, T. J. (1993). Two strikes against perfect phylogeny. In: *Proceedings of the International Colloquium on Automata, Languages and Programming*, Lecture Notes in Computer Science, **623**. Springer-Verlag, Berlin, pp.273-283.

[4] Buneman, P. (1971). The recovery of trees from measures of dissimilarity. In *Mathematics in the archaeological and historical sciences* (ed. F. R. Hodson, D. G. Kendall, and P. Tautu). Edinburgh University Press, pp.387-395.

[5] Buneman, P. (1974). A characterization of rigid circuit graphs, *Discrete Math.*, **9**, 205-212.

[6] Bordewich, M., Huber, K. T., and Semple, C. (2005). Identifying phylogenetic trees. *Discrete Math.*, **300**, 30-43.

[7] Dekker, M. C. H. (1986). Reconstruction methods for derivation trees. Unpublished Masters thesis, Vrije Universiteit, Amsterdam, Netherlands.

[8] Meacham, C.A. (1983). Theoretical and computational considerations of the compatibility of qualitative taxonomic characters. In: *Numerical Taxonomy*, NATO ASI Series, Vol. G1, (ed. J. Felsenstein). Springer-Verlag, Berlin, pp.304-314.

[9] Semple, C. and Steel, M. (2002). A characterization for a set of partial partitions to define an X-tree, *Discrete Math.*, **247** 169-186.

[10] Semple, C. and Steel, M. (2003). *Phylogenetics*. Oxford University Press.

[11] Steel, M. (1992). The complexity of reconstructing trees from qualitative characters and subtrees. *J. Classification*, **9**(1) 91-116.