QUARTET COMPATIBILITY AND THE QUARTET GRAPH

Stefan Grünewald, Peter J. Humphries and Charles Semple

Department of Mathematics and Statistics
University of Canterbury
Private Bag 4800
Christchurch, New Zealand

Report Number: UCDMS2005/9 OCTOBER 2005

QUARTET COMPATIBILITY AND THE QUARTET GRAPH

STEFAN GRÜNEWALD, PETER J. HUMPHRIES, AND CHARLES SEMPLE

ABSTRACT. A collection $\mathcal{P}$ of leaf-labelled trees is compatible if there exists a single leaf-labelled tree that displays each of the trees in $\mathcal{P}$. Despite its difficulty, determining the compatibility of $\mathcal{P}$ is a fundamental task in evolutionary biology. Attractive characterizations in terms of chordal graphs have been previously given for this problem as well as for the problems of (i) determining if there is a unique tree that displays each of the trees in $\mathcal{P}$, that is $\mathcal{P}$ is definitive and (ii) determining if there is a tree that displays $\mathcal{P}$ and has the property that every other tree that displays $\mathcal{P}$ is a refinement of it, that is $\mathcal{P}$ identifies a leaf-labelled tree. In this paper, we describe new characterizations of each of these problems in terms of edge colourings. Furthermore, for an arbitrary leaf-labelled tree $\mathcal{T}$, we also determine the minimum number of 'quartets' required to identify $\mathcal{T}$, thus correcting a previously published result.

1. INTRODUCTION

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

Let $\mathcal{T}$ and $\mathcal{T}'$ be two phylogenetic trees. We say that $\mathcal{T}'$ *displays* $\mathcal{T}$ if the label set X of $\mathcal{T}$ is a subset of the label set X' of $\mathcal{T}'$ and the minimal subtree of $\mathcal{T}'$ connecting the elements in X is a *refinement* of $\mathcal{T}$, that is $\mathcal{T}$ can be obtained from this subtree by contracting edges. For example, in Fig. 1, the phylogenetic tree on the left displays the phylogenetic tree on the right. If $\mathcal{P}$ is a collection of phylogenetic trees, then $\mathcal{T}'$ *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 the union of the label sets of the trees in $\mathcal{P}$ is X'), then $\mathcal{P}$ is said to be *definitive*.

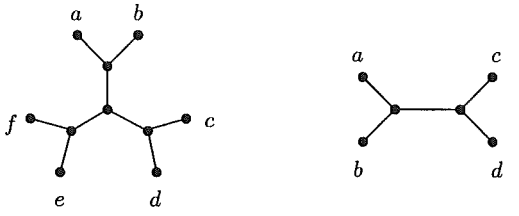
Phylogenetic trees are used in computational biology to represent the evolutionary relationships of a set X of extant species. One fundamental way in which such trees are inferred is by amalgamating a collection $\mathcal{P}$ of smaller phylogenetic trees

Date: 13 October 2005.

1991 *Mathematics Subject Classification.* 05C05; 92D15.

Key words and phrases. Phylogenetic tree, compatibility, restricted chordal completion, identifies.

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.



(a) $\mathcal{T}$ (b)

FIGURE 1. Two phylogenetic trees.

on overlapping subsets of X into a single parent tree. In this context, two natural mathematical problems arise:

(i) Is $\mathcal{P}$ compatible and, if so,
(ii) is $\mathcal{P}$ definitive?

It is well known that every phylogenetic tree is determined by a collection of quartets (see, for example, [8]) and so, in the context of these problems, no generality is loss by viewing $\mathcal{P}$ as a collection of quartets. We will follow this viewpoint throughout the paper.

The first problem is NP-complete [1, 9], while the complexity of the second problem remains open. A variation (and weakening) of (ii) is the following problem:

(iii) If $\mathcal{P}$ is compatible, is there a tree $\mathcal{T}$ that displays $\mathcal{P}$ and has the property that every tree that displays $\mathcal{P}$ is a refinement of it?

If in (iii) there is such a tree $\mathcal{T}$, then we say that $\mathcal{P}$ *identifies* $\mathcal{T}$. Characterizations of each of these problems have been previously given in terms of chordal graphs [3, 4, 6, 7, 9].

In this paper, we introduce the 'quartet graph' and show that these problems can also be characterized in terms of edge colourings via this graph. One of the main motivations for this paper is that it is hoped that the quartet graph may provide new insights not only on the openness of (ii) but also on other quartet problems in phylogenetics. In addition to these characterizations, we also determine, for a given phylogenetic tree $\mathcal{T}$, the size of a minimum-sized set of quartets that identifies $\mathcal{T}$. This corrects a previously published result.

The paper is organized as follows. In the rest of this section, we formally state the main results of this paper. For completeness, Section 2 contains the chordal graph characterizations of problems (i)-(iii). Section 3 contains the proofs of the characterizations of (i)-(iii) in terms of the quartet graph. 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 4 contains the proof of the minimum number of quartets needed to identify a given

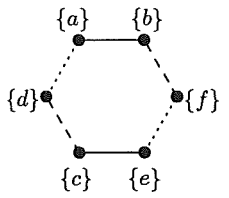


FIGURE 2. The quartet graph of $\{ab|ce, cd|bf, ef|ad\}$.

phylogenetic tree. Throughout the paper, X will always denote a finite set, and notation and terminology follows [8].

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, $cd|ab$. The *label set* of a collection $\mathcal{Q}$ of quartets is the union of the label sets of the quartets in $\mathcal{Q}$. 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. As an example, 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, in particular, the one which is labelled with 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. Under 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.

Central to this paper is a particular graphical operation which preserves proper edge colourings. This operation, called *colour-identification*, is described next. Let X be a finite set, and let G be an arbitrary graph with no loops and whose vertex set V is a partition of X. In other words, X is the disjoint union of the vertices of G. Let U be a subset of V. Then the *identification* of the vertices in U is the graph obtained from G by

(i) deleting every edge in which both end-vertices are in U, and
(ii) replacing the vertices in U with a single vertex which is the union of the elements of U such that if e is incident with exactly one vertex in U, then e is now incident with the vertex that is the union of the elements of U (the other vertex that e is incident with remains unchanged).

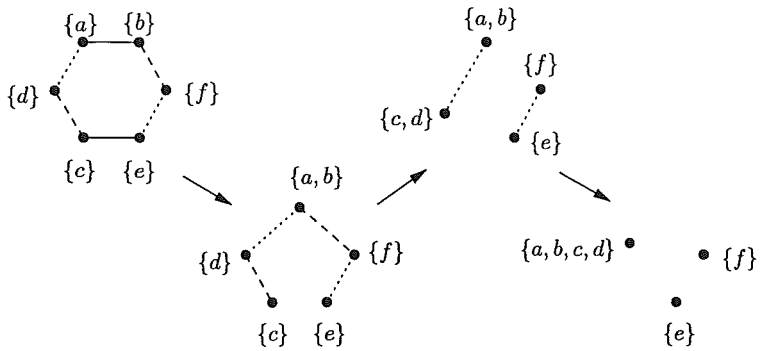


FIGURE 3. A complete colour-identification sequence of the quartet graph in Fig. 2.

If $|U| = p$, then we call this identification a *p-identification.*

Now suppose that the edges of G are coloured and this colouring is proper. Furthermore, suppose that U has 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 *colour-identification* of the vertices in U results in the graph that is obtained from G by identifying the vertices in U and, for each edge that joins two vertices in U, if there is exactly one other edge with the same colour, then this edge is deleted. Observe that, because of the condition imposed on U, the colour-identification of U results in a proper edge-coloured graph. Let $G_0 = G, G_1, G_2, \ldots, G_k$ be a sequence of graphs, where G_i is obtained from G_{i-1} by a colour-identification for all $i \in \{1, 2, \ldots, k\}$. We will call such a sequence a *colour-identification sequence* of G. If G_k has no edges, then this sequence is called a *complete* colour-identification sequence of G.

Example 1.1. Consider the quartet graph $G_{\mathcal{Q}}$ shown in Fig. 2, where $\mathcal{Q} = \{ab|ce, cd|bf, ef|ad\}$. Figure 3 illustrates a colour-identification 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 identify the vertices $\{a\}$ and $\{b\}$ to get the second graph. The third graph is obtained by identifying $\{c\}$ and $\{d\}$ in the second graph, while the last graph is obtained from the third graph by identifying $\{a, b\}$ and $\{c, d\}$. Since the last graph has no edges, this colour-identification sequence is complete.

Theorem 1.1 is the first main result of this paper.

Theorem 1.1. *Let $\mathcal{Q}$ be a collection of quartets. Then $\mathcal{Q}$ is compatible if and only if there is a complete colour 2-identification sequence of $G_{\mathcal{Q}}$.*

As an illustration of Theorem 1.1, the set of quartets $\mathcal{Q} = \{ab|ce, cd|bf, ef|ad\}$ are compatible, as there is a complete colour 2-identification sequence of $G_{\mathcal{Q}}$ as shown in Fig. 3. Indeed, the phylogenetic tree $\mathcal{T}$ shown in Fig. 1(a) displays $\mathcal{Q}$.

To describe the second main result some further definitions are needed. Let $\mathcal{T}$ be a phylogenetic tree. We denote the set of quartets that are displayed by $\mathcal{T}$ by $\mathcal{Q}(\mathcal{T})$. Let $q = ab|cd \in \mathcal{Q}(\mathcal{T})$. An interior edge $e = uv$ of $\mathcal{T}$ is *distinguished* by q if a and b are in separate components of $\mathcal{T}\backslash u$, and c and d are in separate components of $\mathcal{T}\backslash v$. Furthermore, relative to $\mathcal{T}$, a subset of $\mathcal{Q}(\mathcal{T})$ is *distinguishing* if every element in the subset distinguishes some edge of $\mathcal{T}$.

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 with the neighbours of v except u as its vertex set where two vertices w_1, w_2 are adjacent if there is a quartet in $\mathcal{Q}$ that distinguishes e and contains a leaf of the component of $\mathcal{T}\backslash v$ containing w_i for $i \in \{1, 2\}$. 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 $\mathcal{T}$, each of the graphs $G_{\mathcal{Q}(u,v)}$ and $G_{\mathcal{Q}(v,u)}$ is connected.

Let $\mathcal{Q}$ be a collection of quartets on X, and let $G_0 = G_{\mathcal{Q}}, G_1, G_2, \ldots, G_l$ be a colour-identification sequence of $G_{\mathcal{Q}}$. Suppose, for some $j \in \{1, 2, \ldots, l\}$, that G_j is obtained from G_{j-1} by identifying the elements in U_j. If $q = A|B$ is a quartet of $\mathcal{Q}$ and either A or B is a subset of the union of the elements in U_j, we say that q has been *identified* by U_j. Furthermore, this sequence is *minimal* if there is no complete colour-identification sequence $G'_0 = G_{\mathcal{Q}}, G'_1, G'_2, \ldots, G'_k$ where $k < l$ and G'_i is obtained from G'_{i-1} for all $i \in \{1, 2, \ldots, l\}$ by identifying the vertices in U'_i such that, for all i, the union of the elements in U'_i is equal to the union of the elements in a subset of $\{U_1, U_2, \ldots, U_l\}$.

Theorem 1.2. *Let $\mathcal{Q}$ be a set of quartets on X. Then $\mathcal{Q}$ identifies a phylogenetic X-tree if and only if the following hold:*

(i) *there exists a phylogenetic X-tree $\mathcal{T}$ that displays $\mathcal{Q}$ and is specially distinguished by $\mathcal{Q}$; and*
(ii) *if $\mathcal{Q}'$ is a subset of $\mathcal{Q}$ that specially distinguishes $\mathcal{T}$ and is a distinguishing subset of $\mathcal{Q}(\mathcal{T})$, and $q = A|B \in \mathcal{Q}'$, then, whenever the last identification involving a quartet in $\mathcal{Q}'$ in a complete minimal colour-identification sequence of $G_{\mathcal{Q}}$ contains A, the choice of which half of all quartets in $\mathcal{Q}' - \{q\}$ is identified in this sequence is fixed.*

Provided (i) holds in Theorem 1.2, we remark here that there is always at least one complete minimal colour-identification sequence that satisfies the assumption conditions in (ii).

Example 1.2. To illustrate Theorem 1.2, consider Fig. 4 which shows a second complete colour-identification sequence of the quartet graph $G_{\mathcal{Q}}$ shown in Fig. 2, where $\mathcal{Q} = \{ab|ce, cd|bf, ef|ad\}$. As well as the phylogenetic tree $\mathcal{T}$ shown in Fig. 1(a), the phylogenetic tree shown in Fig. 5 also displays $\mathcal{Q}$. Since neither is a refinement of the other, $\mathcal{Q}$ does not identify any phylogenetic tree. This fact is realized by Theorem 1.2 as follows. The set $\mathcal{Q}$ itself specially distinguishes $\mathcal{T}$ and is a distinguishing subset of $\mathcal{Q}(\mathcal{T})$. In both sequences, $ef|ad$ is the last quartet of $\mathcal{Q}$ involved in an identification and this identification contains $\{a, d\}$. Now consider the quartet $ab|ce \in \mathcal{Q}$. In the first sequence $\{a, b\}$ is identified, while in the second sequence $\{c, e\}$ is identified. As the choice of which half of $ab|ce$ that is identified

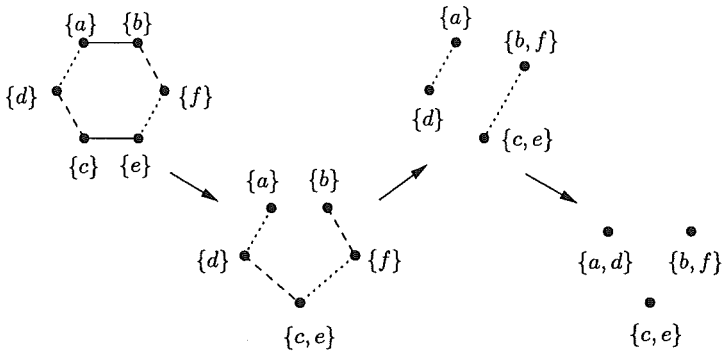


FIGURE 4. Another complete colour-identification sequence of the quartet graph in Fig. 2.

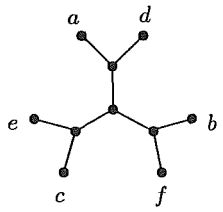


FIGURE 5. Another phylogenetic tree that displays $\mathcal{Q}$.

in such a sequence is not fixed, it follows by Theorem 1.2 that $\mathcal{Q}$ does not identify any phylogenetic tree.

We remark here that the quartet set $\mathcal{Q}$ used in Example 1.2 shows that (i) by itself in Theorem 1.2 is not sufficient to identify a phylogenetic tree; $\mathcal{Q}$ specially distinguishes the phylogenetic tree shown in Fig. 5.

Corollary 1.3. *Let $\mathcal{Q}$ be a set of quartets on X. Then $\mathcal{Q}$ defines a phylogenetic X-tree if and only if the following hold:*

(i) *there exists a binary phylogenetic X-tree $\mathcal{T}$ that displays $\mathcal{Q}$ and is distinguished by $\mathcal{Q}$; and*
(ii) *if $\mathcal{Q}'$ is a minimum-sized subset of $\mathcal{Q}$ that distinguishes $\mathcal{T}$ and $q \in \mathcal{Q}'$, then, whenever the last identification involving a quartet in $\mathcal{Q}'$ in a complete minimal colour-identification sequence contains a half of q, the choice of which half of all quartets in $\mathcal{Q}' - \{q\}$ is identified in this sequence is fixed.*

In the second part of the paper, we consider the problem of determining, for an arbitrary phylogenetic tree $\mathcal{T}$, the minimum number of quartets needed to identify $\mathcal{T}$. In particular, we establish Theorem 1.4. This corrects [8, Theorem 6.3.9] which incorrectly states that the minimum size of such a set is $|X| - 3$, where X is the

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

Theorem 1.4. *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 $e = \{u, v\}$ of $\mathcal{T}$ with $d(u) \leq d(v)$, the collection $\mathcal{Q}$ contains $q(d(u) - 1, d(v) - 1)$ 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,

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

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

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

We remark here that, despite the incorrectness of [8, Theorem 6.3.9], the subsequent corollary [8, Corollary 6.3.10] is still correct (see Theorem 6.8.8 in [8]).

We complete this section with some preliminaries. A *partial split* $A|B$ of X is a bipartition of a subset of X. 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 $|A|, |B| \geq 2$. Phylogenetic X-trees give rise to splits in the following way. Let $\mathcal{T}$ be a phylogenetic X-tree and let $e = u_1u_2$ be an edge of $\mathcal{T}$. Then the split of X *corresponding to* e is the split $X_1|X_2$ where, for each i, the set X_i is the intersection of X and the vertex set of the component of $\mathcal{T}\backslash e$ containing u_i. The collection of non-trivial splits of $\mathcal{T}$ is denoted by $\Sigma(\mathcal{T})$. Buneman [2] showed that every phylogenetic tree is determined by its collection of non-trivial splits. 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.

Let $\mathcal{T}$ be a phylogenetic X-tree and let X' be a subset of X. The *restriction* of $\mathcal{T}$ to X', denoted by $\mathcal{T}|X'$, is the phylogenetic tree that is obtained from the minimal subtree of $\mathcal{T}$ connecting the elements in X' by suppressing all vertices of degree 2.

Lastly, we call a vertex of a tree a *bud* if it is not a leaf and all but one of its neighbors are leaves. An *l-bud* is a bud that is adjacent to l leaves.

2. Chordal Graph Characterizations

In this section we state the chordal graph analogues of Theorems 1.1 and 1.2, and Corollary 1.3. We begin with some definitions.

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

$$\bigcup_{q=A_1|A_2\in\mathcal{Q}} \{(q,A_1),(q,A_2)\},$$

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 it has no vertex induced cycles with at least four vertices. A graph G is a *restricted chordal completion* of $\text{int}(\mathcal{Q})$ if G is a chordal graph that can be obtained from $\text{int}(\mathcal{Q})$ by only adding edges between vertices whose first coordinates are distinct. Note that this maintains the property of a proper vertex colouring. Theorem 2.1, the chordal graph analogue of Theorem 1.1, was indicated by Buneman [3] and Meacham [6], and formally proved by Steel [9].

Theorem 2.1. *Let $\mathcal{Q}$ be a collection of quartets. Then $\mathcal{Q}$ is compatible if and only if there is a restricted chordal completion of* $\text{int}(\mathcal{Q})$.

A restricted chordal completion G of $\text{int}(\mathcal{Q})$ is *minimal* if, for every non-empty subset F of edges of $E(G)-E(\text{int}(\mathcal{Q}))$, the graph $G\backslash F$ is not chordal. The next theorem is due to Semple and Steel [7].

Theorem 2.2. *Let $\mathcal{Q}$ be a collection 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* $\text{int}(\mathcal{Q})$.

To describe the chordal analogue of Theorem 1.2 requires some further definitions. A quartet is a phylogenetic tree with exactly one interior edge an four leaves. More generally, a *one-split* phylogenetic tree is a phylogenetic tree with exactly one interior edge. If the one non-trivial 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, slightly abusing notation, $A|B$ where $A=\{a_1,\ldots,a_r\}$ and $B=\{b_1,\ldots,b_s\}$.

Let $\mathcal{T}$ be a phylogenetic X-tree and let $e=\{u_1,u_2\}$ 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 $\mathcal{T}\backslash u_i$, except for the one containing the other end vertex of e, contains an element of A_i.

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

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

where $\text{int}(\mathcal{Q},\mathcal{T})$ is the graph that has the same vertex set as $\text{int}(\mathcal{Q})$, 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 $\text{int}(\mathcal{Q})$. 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 2.3 was established by Bordewich *et al.* [4].

Theorem 2.3. *Let $\mathcal{Q}$ be a collection 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 loss in restricting our attention to quartets in this paper (see [8, Proposition 6.3.11]).

3. Proofs of Theorems 1.1 and 1.2, and Corollary 1.3

We begin this section with some preliminaries. For $|X| \geq 3$, the *star tree on X*, denoted $\mathcal{S}_X$, is the phylogenetic X-tree with exactly one interior vertex. Let $\mathcal{T}$ and $\mathcal{T}'$ be two phylogenetic X-trees. We say that $\mathcal{T}'$ is a *single-refinement of $\mathcal{T}$* if $\mathcal{T}$ can be obtained from $\mathcal{T}'$ by contracting exactly one edge. Furthermore, if the vertex of $\mathcal{T}$ involved in this refinement is ρ, we also say that $\mathcal{T}'$ is a *single-refinement of ρ*.

Let $\mathcal{Q}$ be a collection of quartets on X, and suppose $G_0 = G_{\mathcal{Q}}, G_1, G_2, \ldots, G_k$ is a colour-identification sequence of $G_{\mathcal{Q}}$. Observe that, for all $i \in \{0, 1, \ldots, k\}$, the union of the vertex sets of G_i is a partition of X. Beginning with the star tree $\mathcal{S}_X$, we next describe the construction of a particular phylogenetic X-tree associated with this sequence.

Label the unique interior vertex of $\mathcal{S}_X$ as ρ. Now G_1 is obtained from G_0 by identifying a subset U_1 of vertices of G_0. Let $\mathcal{T}_1$ be the phylogenetic X-tree obtained from $\mathcal{S}_X$ by a single-refinement of ρ so that the unique non-trivial split

of $\mathcal{T}_1$ is $A_1|(X - A_1)$, where A_1 is the union of the elements of U_1, and ρ is not in the minimal subtree of $\mathcal{T}_1$ that contains the elements in A_1. Observe that if W is the vertex set of a component of $\mathcal{T}_1\backslash\rho$, then $X \cap W$ is a vertex of G_1, and that all vertices of G_1 can be obtained in this way. Now let $q = ab|cd$ be an element of $\mathcal{Q}$ that is identified by U_1. Then either $a, b \in A_1$ or $c, d \in A_1$. In either case, it follows that $\mathcal{T}_1$ displays q. Furthermore, the quartets of $\mathcal{Q}$ that are displayed by $\mathcal{T}_1$ are exactly those which are identified by U_1. We next show by induction that all of these assertions for $\mathcal{T}_1$ can be extended in general.

Suppose that $\mathcal{T}_0 = \mathcal{S}_X, \mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_{i-1}$ is a sequence of phylogenetic X-trees such that the following hold for all $j \in \{1, 2, \ldots, i-1\}$:

(i) $\mathcal{T}_j$ is obtained from $\mathcal{T}_{j-1}$ by a single refinement of ρ so that the unique split of $\Sigma(\mathcal{T}_j) - \Sigma(\mathcal{T}_{j-1})$ is $A_j|(X - A_j)$, where A_j is the union of the elements of the subset U_j of vertices of G_{j-1} that are identified to obtain G_j.
(ii) ρ is not in the minimal subtree of $\mathcal{T}_j$ that contains the elements in A_j.
(iii) If W is the vertex set of a component of $\mathcal{T}_j\backslash\rho$, then $X \cap W$ is a vertex of G_j. Indeed, all vertices of G_j can be obtained in this way.
(iv) $\mathcal{T}_j$ displays all of the quartets identified by U_j as well as all of the quartets identified by $U_1 \cup \cdots \cup U_{j-1}$, but does not display any other quartet of $\mathcal{Q}$.

Suppose that G_i is obtained from G_{i-1} by identifying the elements in A_i. Let $\mathcal{T}_i$ be the phylogenetic X-tree in which $\Sigma(\mathcal{T}_i) = \Sigma(\mathcal{T}_{i-1}) \cup \{A_i|(X - A_i)\}$. Because (iii) holds for $\mathcal{T}_{i-1}$ and G_{i-1}, it follows that $\mathcal{T}_i$ is well-defined and that it can be obtained from $\mathcal{T}_{i-1}$ by a single refinement of ρ with ρ not in the minimal subtree of $\mathcal{T}_i$ that contains the elements in A_i. Thus (iii) holds for $\mathcal{T}_i$ and G_i.

Let A_i be the union of the elements in U_i and let $q = ab|cd$ be a quartet in $\mathcal{Q}$ that is identified by U_i. Then either a and b are elements in distinct members of U_i, or c and d are elements in distinct members of U_i, but not both. Since a, b, c, and d are each in distinct components of $\mathcal{T}_{i-1}\backslash\rho$, it now follows by the construction of $\mathcal{T}_i$ that $\mathcal{T}_i$ displays q. Furthermore, since $\mathcal{T}_i$ is a refinement of $\mathcal{T}_{i-1}$, we have that $\mathcal{T}_i$ displays each of the quartets of $\mathcal{Q}$ identified by $U_1 \cup \cdots \cup U_{i-1}$.

Now let $q' = xy|wz$ be a quartet of $\mathcal{Q}$ that is not identified by $U_1 \cup \cdots \cup U_i$. Then, as there is a q'-coloured edge joining x and y, and a q'-coloured edge joining w and z, none of x, y, w, and z appear in the same vertex of G_i. Therefore, as (iii) holds for $\mathcal{T}_i$ and G_i, the minimal subtree of $\mathcal{T}_i$ containing x, y, w, and z is a star tree, and so $\mathcal{T}_i$ does not display q'. In summary, we have the following proposition.

Proposition 3.1. *Let $\mathcal{Q}$ be a collection of quartets on X, and suppose that*

$$G_0 = G_{\mathcal{Q}}, G_1, G_2, \ldots, G_k$$

is a colour-identification sequence of $G_{\mathcal{Q}}$. Let

$$\mathcal{T}_0 = \mathcal{S}_X, \mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_k$$

be the sequence of phylogenetic X-trees, where, for all $i \in \{1, 2, \ldots, k\}$, $\mathcal{T}_i$ is obtained from $\mathcal{T}_{i-1}$ by a single refinement of ρ so that the unique split of $\Sigma(\mathcal{T}_i) - \Sigma(\mathcal{T}_{i-1})$ is $A_i|(X - A_i)$, where A_i is the union of the elements of the subset U_i of

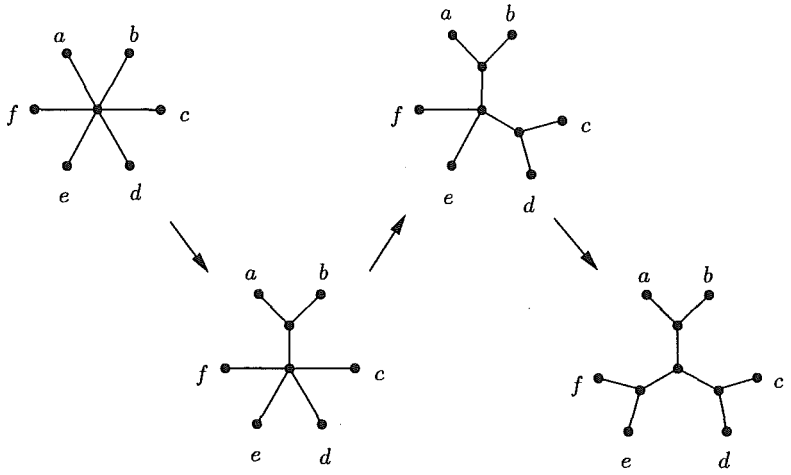


FIGURE 6. The sequence of phylogenetic trees associated with the complete colour-identification sequence shown in Fig. 3.

vertices of G_{i-1} that are identified to obtain G_i. Then, for all $i \in \{1, 2, \ldots, k\}$, the following hold:

(i) *If W is the vertex set of a component of $\mathcal{T}_i \backslash \rho$, then $X \cap W$ is a vertex of G_i. Conversely, if X' is a vertex of G_i, then there is a component of $\mathcal{T}_i \backslash \rho$ whose vertex set W' has the property that $X \cap W' = X'$.*

(ii) *$\mathcal{T}_i$ displays all quartets of $\mathcal{Q}$ that are identified by $U_1 \cup \cdots \cup U_i$, but does not display any other quartet of $\mathcal{Q}$.*

If Φ denotes a colour-identification sequence of a quartet graph $G_{\mathcal{Q}}$, we will denote the sequence of phylogenetic trees described in Proposition 3.1 by Γ_Φ. Furthermore, the last tree in Γ_Φ will be denoted by $\mathcal{T}_\Phi$. To illustrate Proposition 3.1, Fig. 6 shows the sequence of phylogenetic trees corresponding to the complete colour-identification sequence shown in Fig. 3. This sequence begins with the star tree on $\{a, b, c, d, e, f\}$ and ends with the phylogenetic tree $\mathcal{T}$ shown in Fig. 1.

As a partial converse to Proposition 3.1, suppose that $\mathcal{Q}$ is compatible and $\mathcal{T}$ is a phylogenetic X-tree that displays $\mathcal{Q}$. It is easily seen that $\mathcal{T}$ can be constructed from the star tree $\mathcal{S}_X$ with interior vertex ρ by continually applying single refinements of ρ. Let $\mathcal{T}_0 = \mathcal{S}_X, \mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_l = \mathcal{T}$ be the associated sequence of phylogenetic X-trees, where $\mathcal{T}_i$ is obtained from $\mathcal{T}_{i-1}$ by a single refinement of ρ for all $i \in \{1, 2, \ldots, l\}$. For all i, let $A_i | (X - A_i)$ be the unique element in $\Sigma(\mathcal{T}_i) - \Sigma(\mathcal{T}_{i-1})$.

Let $G_0 = G_{\mathcal{Q}}$ be the quartet graph of $\mathcal{Q}$. For all i, let G_i be the graph obtained from G_{i-1} by identifying the vertices whose disjoint union is A_i. We next show that, for all i, this identification is a colour-identification of G_{i-1}. Suppose not,

and that G_j is the first graph that is not a colour-identification of G_{j-1}. Then there is a quartet $q = ab|cd$ in $\mathcal{Q}$ 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 $\{a,b,c,d\} \cap A_j \notin \{\{a,b\},\{c,d\}\}$. But then, by the construction of $\mathcal{T}$ from $\mathcal{S}_X$, it is easily seen that $\mathcal{T}$ does not display $ab|cd$; a contradiction. Hence G_j is a colour-identification of G_{j-1}. Now consider G_l, and suppose that there is a pair of q-coloured edges in G_l, where $q = xy|wz \in \mathcal{Q}$. Since $\mathcal{T}$ displays $\mathcal{Q}$, there is an edge in $\mathcal{T}$ that separates the path from x to y and the path from w to z. It follows that, for some i, either $x, y \in A_i$ or $w, z \in A_i$ and so, in G_i, the elements x and y or the elements w and z are elements of the same vertex. This implies that there is no such pair of q-coloured edges, in particular, G_l has no edges. We have now established the following proposition.

Proposition 3.2. *Let $\mathcal{Q}$ be a compatible collection of quartets. Then there is a complete colour-identification sequence of $G_{\mathcal{Q}}$.*

As before, if $\mathcal{T}$ is a phylogenetic tree that displays a set $\mathcal{Q}$ of quartets and Γ is a sequence of phylogenetic trees starting with the star tree and ending with $\mathcal{T}$ as described above, then we will denote the complete colour-identification sequence corresponding to Γ as Φ_Γ.

Proposition 3.3. *Let $\mathcal{Q}$ be a collection of quartets and let $G_0 = G_{\mathcal{Q}}, G_1, G_2, \ldots, G_k$ be a colour-identification sequence of $G_{\mathcal{Q}}$. Then, for each $i \in \{1,2,\ldots,k\}$, the colour-identification that takes G_{i-1} to G_i can be replaced by a sequence of colour 2-identifications.*

Proof. Suppose that G_i is obtained from G_{i-1} by identifying the vertices in U_i, where $|U_i| = p$. If $p = 2$, then we are done. So assume that $|p| \geq 3$ and, for induction purposes, that if $|U_i|$ is at most $p-1$, then G_i can be obtained from G_{i-1} by a sequence of colour 2-identifications. Choose two vertices, B and B' say, of U_i, and consider the graph G'_{i-1} that is obtained by identifying B and B'. Since G_i is a colour-identification of G_{i-1}, it is easily checked that G'_{i-1} is a colour 2-identification of G_{i-1}. Furthermore, G_i can be obtained from G'_{i-1} by the colour-identification which identifies the vertices in $\big(U_i - \{B,B'\}\big) \cup (B \cup B')$. Since this last identification involves exactly $p-1$ vertices, it now follows by the induction assumption that we can obtain G_i from G_{i-1} by a sequence of colour 2-identifications. This completes the proof of the proposition. □

The next corollary is an immediate consequence of Proposition 3.3.

Corollary 3.4. *Let $\mathcal{Q}$ be a collection of quartets on X. Then there is a complete colour-identification sequence of $G_{\mathcal{Q}}$ if and only if there is a complete colour 2-identification sequence of $G_{\mathcal{Q}}$.*

The proof of Theorem 1.1 now follows by combining Propositions 3.1 and 3.2, and Corollary 3.4.

To prove Theorem 1.2, we begin with two lemmas.

Lemma 3.5. *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. Assume that $\mathcal{Q}$ does not specially distinguish $\mathcal{T}$. Let uv be an interior edge of $\mathcal{T}$ such that $G_{\mathcal{Q}(u,v)}$ contains $k > 1$ components $C_1, C_2, \ldots, C_k$. We next construct a phylogenetic X-tree $\mathcal{T}'$ from $\mathcal{T}$ that displays $\mathcal{Q}$ but is not a refinement of $\mathcal{T}$. Delete v and all its incident edges from $\mathcal{T}$. For each $i \in \{1, 2, \ldots, k\}$, add an edge joining u and the vertex of C_i if C_i contains exactly one vertex; otherwise add a new vertex v_i and edges uv_i and v_iw for every vertex w of C_i. It is now easily checked that the resulting phylogenetic X-tree $\mathcal{T}'$ displays $\mathcal{Q}$. But, clearly, $\mathcal{T}'$ is not a refinement of $\mathcal{T}$. We conclude that $\mathcal{Q}$ specially distinguishes $\mathcal{T}$. □

Lemma 3.6. *Let $\mathcal{Q}$ be a compatible collection of quartets on X.*

(i) *Let Φ be a complete minimal colour-identification sequence of $G_{\mathcal{Q}}$. Then $\mathcal{T}_\Phi$ is a minimally refined phylogenetic X-tree that displays $\mathcal{Q}$.*
(ii) *Let $\mathcal{T}$ be a minimally refined phylogenetic X-tree that displays $\mathcal{Q}$ and let Γ be a sequence $\mathcal{S}_X = \mathcal{T}_0, \mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_l = \mathcal{T}$ of phylogenetic X-trees in which $\mathcal{T}_i$ is obtained from $\mathcal{T}_{i-1}$ by a single refinement of ρ for all $i \in \{1, 2, \ldots, l\}$. Then the complete colour-identification sequence Φ_Γ is minimal.*

Proof. We first prove (i). Let Φ be the sequence $G_0 = G_{\mathcal{Q}}, G_1, G_2, \ldots, G_k$ and, for all i, let U_i be the set of vertices of G_{i-1} that are identified to obtain G_i. Let $\mathcal{T}$ be the canonical phylogenetic X-tree corresponding to Φ, and suppose that $\mathcal{T}$ is not minimally refined with respect to displaying $\mathcal{Q}$. Then there is an edge, e say, of $\mathcal{T}$ such that $\mathcal{T}/e$ displays $\mathcal{Q}$. Let $\mathcal{S}_X = \mathcal{T}_0, \mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_k = \mathcal{T}$ be the sequence of phylogenetic X-trees corresponding to Φ. Because of the way in which $\mathcal{T}$ is constructed from $\mathcal{S}_X$, there is some iteration, i say, where the union of the elements in U_i is equal to the elements of X in the component of $\mathcal{T}\backslash e$ that avoids ρ. Let $\mathcal{S}_X = \mathcal{T}'_0, \mathcal{T}'_1, \mathcal{T}'_2, \ldots, \mathcal{T}'_{k-1}$ be the sequence of phylogenetic trees that is obtained from $\mathcal{S}_X = \mathcal{T}_0, \mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_k = \mathcal{T}$ by applying the same sequence of single refinements of ρ with the one corresponding to e omitted. Clearly, $\mathcal{T}\backslash e$ isomorphic to $\mathcal{T}'_{k-1}$. Now consider the complete colour-identification sequence $G'_0 = G_{\mathcal{Q}}, G'_1, G'_2, \ldots, G'_{k-1}$ corresponding to $\mathcal{S}_X = \mathcal{T}'_0, \mathcal{T}'_1, \mathcal{T}'_2, \ldots, \mathcal{T}'_{k-1}$. For all $j \in \{1, 2, \ldots, k-1\}$, let U'_j denote the set of vertices of G'_{j-1} that are identified to obtain G'_j. It is easily seen that, for all j, the union of the elements in U'_j is equal to the union of the elements in U_i for some i. But this contradicts the minimality of Φ. Thus $\mathcal{T}$ is a minimally refined phylogenetic X-tree with respect to displaying $\mathcal{Q}$.

For the proof of (ii), let Φ be the complete colour-identification sequence $G_0 = G_{\mathcal{Q}}, G_1, G_2, \ldots, G_l$ corresponding to $\mathcal{S}_X = \mathcal{T}_0, \mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_l = \mathcal{T}$. Suppose that Φ is not minimal. Then there is a complete colour-identification sequence $G'_0 = G_{\mathcal{Q}}, G'_1, G'_2, \ldots, G'_k$ of $G_{\mathcal{Q}}$ with $k < l$ and, for all i, the union of the elements in U'_i is equal to the union of the elements in U_j for some j. But this implies that the set of splits of the last phylogenetic X-tree $\mathcal{T}'$ in the sequence of phylogenetic X-trees corresponding to this sequence is a proper subset of the set of splits of $\mathcal{T}$. In other words, $\mathcal{T}$ is a refinement of $\mathcal{T}'$, contradicting the minimality of $\mathcal{T}$. Hence Φ is minimal. □

We now prove Theorem 1.2.

Proof of Theorem 1.2. First suppose that $\mathcal{Q}$ identifies a phylogenetic tree $\mathcal{T}$. Then, by Lemma 3.5, (i) holds. We next show that (ii) holds. Since $\mathcal{Q}$ identifies $\mathcal{T}$, it follows by Lemma 3.6 that if Φ is any complete minimal colour-identification sequence of $G_{\mathcal{Q}}$, then $\mathcal{T}_\Phi$ is isomorphic to $\mathcal{T}$. This fact is implicitly used in the proof of this direction of Theorem 1.2. Let $\mathcal{Q}'$ be a subset of $\mathcal{Q}$ that specially distinguishes $\mathcal{T}$ and is a distinguishing subset of $\mathcal{Q}(\mathcal{T})$, and let $q = A|B \in \mathcal{Q}'$. Now suppose that in some complete minimal colour-identification sequence Φ of $G_{\mathcal{Q}}$ the last identification involving a quartet in $\mathcal{Q}'$ contains A. Let $q' = A'|B' \in \mathcal{Q}' - \{q\}$, and suppose that A' is identified along a q'-coloured edge in Φ.

Let Φ' be any complete minimal colour-identification sequence of $G_{\mathcal{Q}}$ such that the last identification involving a quartet in $\mathcal{Q}'$ contains A. If q and q' distinguish the same edge of $\mathcal{T}$, then the same quartet half of both q and q' must be identified in any colour-identification sequence of $G_{\mathcal{Q}}$ that reconstructs $\mathcal{T}$. This implies that A' is identified along a q'-coloured edge in Φ'. Assume that q and q' distinguish distinct edges of $\mathcal{T}$, and let e be the interior edge of $\mathcal{T}$ distinguished by q. Suppose that B' is identified along a q'-coloured edge in Φ'. Then, as q and q' distinguish different edges of $\mathcal{T}$, the quartet half B' is identified strictly before the quartet half A is identified in Φ'. This means that if e' is the edge of $\mathcal{T}$ distinguished by q', then in $\mathcal{T}\backslash e'$ one component contains the vertices corresponding to the elements in B' and the other component contains the vertices corresponding to the elements in A' and e. However, the same argument applies to the sequence Φ in which A' is identified along a q'-coloured edge and so we also deduce that in $\mathcal{T}\backslash e'$ the component that contains the vertices in B' also contains e'; a contradiction. Thus (ii) holds.

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 $\mathcal{Q}$, the tree $\mathcal{T}$ is a minimally refined phylogenetic tree that displays $\mathcal{Q}$. Let T' be a minimally refined phylogenetic X-tree that displays $\mathcal{Q}$ but is not isomorphic to $\mathcal{T}$.

1.2.1. *There are no two elements $x, y \in X$ such that each of $\mathcal{T}$ and T' have buds which are adjacent to both x and y.*

Proof. Suppose that there are such elements x and y in X. Let z be an element not in X and let $X_z = (X - \{x, y\}) \cup \{z\}$. Let $\mathcal{T}_z$ be the phylogenetic X_z-tree that is obtained from $\mathcal{T}$ by deleting the vertices x and y, adjoining a new leaf z to the bud v neighbouring x and y, and suppressing any resulting vertex of degree 2. Let $\mathcal{Q}_z$ be the collection of quartets obtained from $\mathcal{Q}$ by removing any quartet in which one half contains both x and y, and then replacing both x and y with z in the remaining quartets. We next show that $\mathcal{Q}_z$ satisfies both (i) and (ii).

Since $\mathcal{T}$ displays $\mathcal{Q}$ and is specially distinguished by $\mathcal{Q}$, it is clear that $\mathcal{T}_z$ displays $\mathcal{Q}_z$ and is specially distinguished by $\mathcal{Q}_z$. Thus $\mathcal{Q}_z$ satisfies (i). To show that $\mathcal{Q}_z$ satisfies (ii), let $\mathcal{Q}'_z$ be a subset of $\mathcal{Q}_z$ that specially distinguishes $\mathcal{T}_z$ and is a distinguishing subset of $\mathcal{Q}(\mathcal{T}_z)$, and let $q_z = A_z|B_z \in \mathcal{Q}'_z$. Suppose that Φ_z is a complete minimal colour-identification sequence of $G_{\mathcal{Q}_z}$ in which one half of q, say A_z, is in the last identification involving a quartet in $\mathcal{Q}'_z$. Let Φ be the identification

sequence of $G_{\mathcal{Q}}$ that is obtained from Φ_z by replacing any identification involving z with x and y. In case x and y are the two leaves of a 2-bud, then choose the first identification in Φ to identify x and y. By considering $G_{\mathcal{Q}_z}$ and $G_{\mathcal{Q}}$, it is easily seen that Φ is a complete colour-identification sequence of $G_{\mathcal{Q}}$. Furthermore, as Φ_z is a minimal sequence, it is easily checked that Φ is also minimal.

Because of the way in which $\mathcal{T}_z$ is obtained from $\mathcal{T}$, we can extend $\mathcal{Q}'_z$ to a subset $\mathcal{Q}'$ of $\mathcal{Q}$ that specially distinguishes $\mathcal{T}$ and is a distinguishing subset of $\mathcal{Q}(\mathcal{T})$ by replacing every quartet in $\mathcal{Q}'_z$ that contains z with the quartets of $\mathcal{Q}$ from which it was originally derived and then adding at most one further quartet of $\mathcal{Q}$ so that $G_{\mathcal{Q}(u,v)}$ is connected, where u is the non-leaf vertex of $\mathcal{T}$ adjacent to the bud v. Note that if $q = A|B$ is a quartet obtained from q_z by replacing z with either x or y such that A corresponds to A_z, then A is in the last identification involving a quartet in $\mathcal{Q}'$ in Φ.

Let Φ'_z be an arbitrary complete minimal colour-identification sequence of $G_{\mathcal{Q}_z}$ in which A_z is in the last identification involving a quartet in $\mathcal{Q}'_z$. Let Φ' be the complete minimal colour-identification sequence obtained from Φ'_z in the way described above. If there is a quartet in $\mathcal{Q}'_z$ such that one half is identified in Φ_z but the other half is identified in Φ'_z, then, by construction, there is a quartet in $\mathcal{Q}'$ such that one half is identified in Φ but the other half is identified in Φ'. This contradiction implies that $\mathcal{Q}_z$ satisfies (ii).

Now let $\mathcal{T}'_z$ be the phylogenetic X_z-tree that is obtained from $\mathcal{T}'$ by deleting the vertices x and y, adjoining a new leaf z to the bud neighbouring x and y, and suppressing any vertex of degree 2. Since $\mathcal{T}'_z$ displays $\mathcal{Q}_z$ and it is not a refinement of $\mathcal{T}_z$, we deduce that, in the size of its label set, $\mathcal{Q}_z$ is a smaller counterexample to the converse. This contradiction completes the proof of (1.2.1). □

Let $\mathcal{Q}'$ be a subset of $\mathcal{Q}$ that specially distinguishes $\mathcal{T}$ and is a distinguishing subset of $\mathcal{Q}(\mathcal{T})$. Let $q = ab|cd \in \mathcal{Q}'$ be a quartet such that the common subpath P of the paths from a to c and from b to d in $\mathcal{T}'$ has the property that there is no quartet $xy|wz \in \mathcal{Q}'$ in which the common subpath of the paths from x to w and from y to z in $\mathcal{T}'$ is a proper subpath of P. The phylogenetic X-tree obtained from $\mathcal{T}'$ by identifying all vertices of P displays all quartets in $\mathcal{Q}'$ which do not distinguish P, that is all quartets $ij|kl \in \mathcal{Q}'$ for which the path from i to j intersects P at precisely one terminal vertex of P and the path from k to l intersects P at precisely the other terminal vertex of P. Hence, by Lemma 3.6, there is a complete minimal colour-identification sequence that reconstructs $\mathcal{T}'$, where ab is in the last identification involving a quartet in $\mathcal{Q}'$. By symmetry, this last statement holds if we replace ab with cd. Since $q \in \mathcal{Q}'$, we know that q distinguishes an edge e of $\mathcal{T}$, so we can choose q, or more specifically the half ab, to be in the last identification of some complete minimal colour-identification sequence Φ of $G_{\mathcal{Q}}$ that reconstructs $\mathcal{T}$.

Let r_1 be the vertex of $\mathcal{T}'$ in the intersection of the vertex sets of P and the path from a to b, and let r_2 be the vertex of $\mathcal{T}'$ in the intersection of the vertex sets of P and the path from c to d. Let ρ'_1 and ρ'_2 be the neighbours of r_1 and r_2 in P, respectively. Then a and b are in the same component C'_1 of $\mathcal{T}'\backslash\rho'_1$, and c and d are

in the same component C_2' of $T'\backslash\rho_2'$. Let ρ_1 and ρ_2 be the vertices of $\mathcal{T}$ incident with e such that a and b are in the same component C_1 of $\mathcal{T}\backslash\rho_1$, and c and d are in the same component C_2 of $\mathcal{T}\backslash\rho_2$. Let X_1 be the leaves of C_1, and let H be the subgraph of $G_{\mathcal{Q}'}$ whose vertex set is the set of singleton subsets of X_1 and an edge joins two vertices precisely if it is identified in Φ. Since $\mathcal{Q}'$ specially distinguishes $\mathcal{T}$ and ab is in the last identification of Φ, all vertices of H whose elements are leaves adjacent to the same bud of $\mathcal{T}$ are in the same connected component of H. By viewing each bud and its adjacent leaves in C_1 as a single leaf labelled by these leaves, it is easily seen that all vertices of H whose elements label leaves adjacent to the same bud under this viewpoint are in the same connected component. By iterating this argument, we eventually deduce that H is connected.

Every edge of H is also an edge in $G_{\mathcal{Q}'}$ and, moreover, all of these edges must be contracted either before or simultaneously with ab in Φ. Since H is connected, it follows by (ii) holding that all the leaves in C_1 are also leaves in C_1'. By symmetry, this implication also holds for C_2 and C_2', respectively. Thus each of the elements in $X - X_1$ is a leaf in C_2, and so P contains only the vertices ρ_1' and ρ_2' (with $r_1 = \rho_2'$ and $r_2 = \rho_1'$) and one edge $\rho_1'\rho_2'$. Hence the edge $\rho_1'\rho_2'$ of T' is distinguished by q.

We next construct a new quartet collection $\mathcal{Q}_1$ from $\mathcal{Q}$ as follows. Remove any quartet whose label set contains at least two elements in X_1 and replace any element in X_1 with x_1 in the remaining quartets, where x_1 is an element not in X_1. It is easily checked that the phylogenetic tree $\mathcal{T}_1$ obtained from $\mathcal{T}$ by replacing the minimal subtree containing the elements in X_1 with the leaf x_1 displays $\mathcal{Q}_1$ and is specially distinguished by $\mathcal{Q}_1$. Furthermore, using the approach described in (1.2.1), one can check that $\mathcal{Q}_1$ satisfies (ii). Hence, by the minimality of $\mathcal{Q}$ in the size of X, we have that $\mathcal{Q}_1$ identifies $\mathcal{T}_1$. Now the phylogenetic tree $\mathcal{T}_1'$ obtained from $\mathcal{T}'$ by replacing the minimal subtree containing the elements in X_1 with the leaf x_1 displays $\mathcal{Q}_1$ and is specially distinguished by $\mathcal{Q}_1$. This means that $\mathcal{T}_1'$ is a minimally refined tree that displays $\mathcal{Q}_1$ and so it is isomorphic to $\mathcal{T}_1$. Since $\mathcal{T}_1$ has a bud and $\mathcal{T}_1'$ has the same bud, it follows that there is a bud in $\mathcal{T}$ and in $\mathcal{T}'$ which is adjacent to the same two leaves. This contradiction to (1.2.1) completes the proof of the theorem. □

Proof of Corollary 1.3. Suppose that $\mathcal{Q}$ defines a phylogenetic X-tree $\mathcal{T}$. Then it is clear that (i) holds. The fact that (ii) holds follows from the first part of the proof of Theorem 1.2, where we note, for distinct $q, q' \in \mathcal{Q}'$, the quartets q and q' distinguish different edges of $\mathcal{T}$.

Now suppose that (i) and (ii) hold. Let $\mathcal{Q}''$ be a subset of $\mathcal{Q}$ that distinguishes $\mathcal{T}$ and is a distinguishing subset of $\mathcal{Q}(\mathcal{T})$. Then $\mathcal{Q}''$ contains a subset $\mathcal{Q}'$ of minimum size that distinguishes $\mathcal{T}$. Let $q' = A'|B'$ be an element of $\mathcal{Q}'$ and let $q'' = A''|B''$ be an element of $\mathcal{Q}'' - \mathcal{Q}'$ such that q'' distinguishes the same edge e of $\mathcal{T}$ as q'. Without loss of generality we may assume that the paths in $\mathcal{T}$ connecting the elements in A' and A'' contain the same end vertex of e. Let Φ be a minimal complete colour-identification sequence of $G_{\mathcal{Q}}$ and suppose that A' is the half of q' that is identified in Φ. Then it is easily seen that A'' is the half of q'' that is identified in Φ. Indeed, A' and A'' are identified in the same identification in Φ. Thus (ii) holds for any distinguishing subset of $\mathcal{Q}$. It now follows by Theorem 1.2

that $\mathcal{Q}$ identifies a phylogenetic tree. Since there is a phylogenetic tree $\mathcal{T}$ 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. □

4. Minimum Identifying Sets of Quartets

The main result of this section is Theorem 1.4. 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 [5]):

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

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

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

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

The rule (sc) is "Rule 1" in [6]. 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 4.1. *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 4.2 generalizes (tc).

Lemma 4.2. *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 4.1, 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 upon 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. □

Analogous 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 4.3. *Let $\mathcal{T}$ be a one-split phylogenetic tree in which the unique non-trivial split is $A|B$ with $A = \{a_1, a_2, \ldots, a_r\}$ and $B = \{b_1, b_2, \ldots, b_s\}$. 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 4.2, 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 $\mathcal{T}$. □

For a one-split phylogenetic tree $\mathcal{T}$ whose non-trivial split is $A|B$, we will denote the size of a minimum-sized set of quartets that identifies $\mathcal{T}$ by $q(r, s)$, where $r = |A|$ and $s = |B|$. Much of the work in proving Theorem 1.4 goes into proving the next lemma, a special case of that theorem.

Lemma 4.4. *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 $A = \{a_1, a_2, \ldots, a_r\}$ and $B = \{b_1, b_2, \ldots, b_s\}$. 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 3.5, $G_{\mathcal{Q}}$ 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 $G_{\mathcal{Q}}$ 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 components of G_a. Now consider the colour-identification sequence Φ of $G_{\mathcal{Q}} = G_0$ in which we make the following identifications:

(i) For $1 \leq i \leq k$, identify the vertices in C_i of G_{i-1} to obtain G_i if C_i contains at least two vertices;
(ii) identify $\{a\}$ together with the set of vertices whose union is B to obtain G_{k+1};

It is easily checked that Φ is a complete colour-identification sequence for $\mathcal{Q}$. By Proposition 3.1, $\mathcal{T}_\Phi$ displays $\mathcal{Q}$. But, as G_a is disconnected, the construction of $\mathcal{T}_\Phi$ beginning with $\mathcal{S}_X$ and using Φ implies that $\mathcal{T}_\Phi$ displays a quartet $ab_1|b_2b_3$ or $a_1a_2|ab_1$ where $a_1, a_2 \in A$ and $b_1, b_2, b_3 \in B$. Thus $\mathcal{T}_\Phi$ is not a refinement of $\mathcal{T}$, and so $\mathcal{Q}$ does not identify $\mathcal{T}$. 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 $r = 2$.

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

Proof. Here $A|B = \{a_1, a_2\}|\{b_1, b_2, \ldots, b_s\}$ and it follows by repeated applications of (sc) that the collection

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

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

4.4.2. *Let Q_r be the collection*

$$\{a_ia_j|b_ib_j : 1 \leq i < j \leq r\}$$

of quartets. If $r = s$, then Q_r identifies $\mathcal{T}$. In particular, for all r, we have $q(r,r) \leq \frac{r(r-1)}{2}$.

Proof. First note that $|Q_r| = \binom{r}{2} = \frac{r(r-1)}{2}$. 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

$$\{a_i a_j a_r | b_i b_j b_r : 1 \leq i < j < r\}.$$

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

4.4.3. *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 (4.4.1). Now suppose that $r \geq 3$, and that the result holds for all smaller values of r. There are seven cases to consider.

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

By Lemma 4.3, the 2-element collection

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

of partial splits together with the collection

$$\mathcal{Q}_1 = \{a_i a_{l+i} | b_j b_{l+j} : 1 \leq i \leq r-l, 1 \leq j \leq l-1\}$$

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. $r = 4k-2$ and $s = 2l$ for some integers $k \geq 2$ and $l \geq 3$.

By Lemma 4.3, the 2-element collection

$$\Sigma_2 = \{a_1 \cdots a_{2k} | b_1 \cdots b_{l+1}, a_{2k-1} \cdots a_r | b_l \cdots b_s\}$$

of partial splits together with the collection

$$\mathcal{Q}_2 = \{a_i a_{2k+i} | b_j b_{l+j+1} : 1 \leq i \leq 2k-2, 1 \leq j \leq l-1\}$$

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

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

Case 3. $r = 4k-3$ and $s = 2l$ for some integers $k \geq 2$ and $l \geq 3$.

By Lemma 4.3, the 2-element collection

$$\Sigma_3 = \{a_1 \cdots a_{2k} | b_1 \cdots b_{l+1}, a_{2k-2} \cdots a_r | b_l \cdots b_s\}$$

of partial splits together with the collection

$$\mathcal{Q}_3 = \{a_i a_{2k+i}|b_j b_{l+j+1} : 1 \leq i \leq 2k-3, 1 \leq j \leq l-1\}$$

of quartets identify $\mathcal{T}$. By the induction assumption, each partial split in Σ_3 can be identified by a collection of kl quartets. Without loss of generality, we may assume that these last collections share the quartet $a_{2k-1}a_{2k}|b_l b_{l+1}$. Furthermore, $\mathcal{Q}_3$ contains $(2k-3)(l-1)$ quartets. Thus

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

Case 4. $r = 4k$ and $s = 4l$ for some positive integers $k \geq 1$ and $l \geq 1$.

By Lemma 4.3, the 2-element collection

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

of partial splits together with the collection

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

of quartets identifies $\mathcal{T}$. By the induction assumption, each partial split in Σ_4 can be identified by a collection of $(2k+1)l$ quartets. Without loss of generality, we may assume that these last collections share the quartet $a_{2k}a_{2k+1}|b_{2l}b_{2l+1}$. Furthermore, $\mathcal{Q}_4$ contains $(2k-1)(2l-1)$ quartets. Thus

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

Case 5. $r = 4k-1$ and $s = 4l$ for some integers $k \geq 1$ and $l \geq 1$.

By Lemma 4.3, the 2-element collection

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

of partial splits together with the collection

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

of quartets identifies $\mathcal{T}$. By the induction assumption, each partial split in Σ_5 can be identified by a collection of $(2k+1)l$ quartets. Without loss of generality, we may assume that these last collections share the quartet $a_{2k}a_{2k+1}|b_{2l}b_{2l+1}$. Furthermore, $\mathcal{Q}_5$ contains $(2k-2)(2l-1)$ quartets. Thus

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

Case 6. $r = 4k$ 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 6 in general.

Let

$$\mathcal{Q}_1 = \{a_1a_2|b_1b_2, a_1a_3|b_1b_3, a_2a_3|b_2b_3\},$$

$$\mathcal{Q}_2 = \{a_2a_3|b_4b_5, a_2a_4|b_4b_6, a_3a_4|b_5b_6\},$$

and

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

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 4.3, the 2-element collection

$$\Sigma_6 = \{a_1 \cdots a_{2k+2}|b_1 \cdots b_{2l+1}, a_{2k-1} \cdots a_r|b_{2l-2} \cdots b_s\}$$

of partial splits together with the collection

$$\mathcal{Q}_6 = \{a_ia_{2k+i+2}|b_jb_{2l+j+1} : 1 \leq i \leq 2k-2, 1 \leq j \leq 2l-3\}$$

of quartets identifies $\mathcal{T}$. By the induction assumption, each partial split in Σ_6 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 (4.4.2), we may assume that this set contains

$$\{a_{2k-1}a_{2k}|b_{2l-2}b_{2l-1}, a_{2k-1}a_{2k+1}|b_{2l-2}b_{2l}, a_{2k}a_{2k+1}|b_{2l-1}b_{2l},$$
$$a_{2k-1}a_{2k+2}|b_{2l-2}b_{2l+1}, a_{2k}a_{2k+2}|b_{2l-1}b_{2l+1}, a_{2k+1}a_{2k+2}|b_{2l}b_{2l+1}\}.$$

Similarly, we may assume the set of $(2k+2)l$ quartets that identifies the other partial split in Σ_6 also contains the six quartets in this set. Since $\mathcal{Q}_6$ 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 7. $r = 4k-1$ and $s = 4l-2$ for some integers $k \geq 1$ and $l \geq 2$.

By Lemma 4.3, the 2-element collection

$$\Sigma_7 = \{a_1 \cdots a_{2k}|b_1 \cdots b_{2l}, a_{2k} \cdots a_r|b_{2l-1} \cdots b_s\}$$

of partial splits together with the collection

$$\mathcal{Q}_7 = \{a_ia_{2k+i}|b_jb_{2l+j} : 1 \leq i \leq 2k-1, 1 \leq j \leq 2l-2\}$$

of quartets identifies $\mathcal{T}$. By the induction assumption, each partial split in Σ_7 can be identified by a collection of $k(2l-1)$ quartets. Furthermore, $\mathcal{Q}_7$ 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 1-7, 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 4.4 by showing that, for any fixed r, the result holds for all s with $r \leq s$. By (4.4.3), 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

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

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 (4.4.2), 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 proposition. □

The next lemma is an immediate consequence of the definition of identify.

Lemma 4.5. *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 $A|B$ is not a split.*

At last we prove Theorem 1.4.

Proof of Theorem 1.4. First suppose that for some interior edge $e = \{u, v\}$ 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 $u_1, u_2, \ldots, u_r$ and the neighbours of v that are not u are $v_1, v_2, \ldots, v_s$. 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 $u_iu_j|v_kv_l$, 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 4.4 that $\mathcal{P}_e$ does not identify $\mathcal{T}_e$.

By Lemma 4.5, 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 i, the maximal subtree of $\mathcal{T}$ that contains u_i and not u to $\mathcal{T}'_e$ by identifying the two common vertices, namely u_i and by adjoining, for all j, the maximal subtree of $\mathcal{T}$ that contains v_j and not v to $\mathcal{T}'_e$ by identifying the two common vertices, namely v_j. 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 $e = \{u, v\}$, 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 $\{u, v\}$, then, by Lemma 4.4, 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 $e = \{u, v\}$ 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 4.4, $|\mathcal{Q}_e| = q(r, s)$. 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. □

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 4.6. *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 $\{u, v\}$. 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 1.4,

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

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, [8]), $q(\mathcal{T}) \leq n-3$. Let w be a bud of $\mathcal{T}$ of maximum size. Let j be the size of this bud, let $x_1, x_2, \ldots, x_j$ 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 two cases: a) $j \geq 3$ and b) $j = 2$.

Consider a). If $d(w) \leq d(v)$, then, by Theorem 1.4,

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

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 consider b). Here every bud of $\mathcal{T}$ has size two. Note that, in this case, $d(w) \leq d(v)$. By Theorem 1.4,

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

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. □

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 4.7. *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 1.4 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 an j-bud with $j \leq \frac{n}{2}$ and the other interior vertex is an $(n-j)$-bud. Consequently, by Theorem 1.4,

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

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 $x_1, x_2, \ldots, x_k$ 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 1.4, $q(\mathcal{T}) - q(\mathcal{T}') = q(k, d(v) - 1) - q(k-1, d(v)-1)$ 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 1.4,

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

a contradiction.

Now suppose that $k = 2$. By Theorem 1.4, $q(\mathcal{T}) - q(\mathcal{T}') = q(2, d(v)-1) = d(v) - 2$. 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. □

References

[1] 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.

[2] 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.

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

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

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

[6] 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.

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

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

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

DEPARTMENT OF MATHEMATICS AND STATISTICS, UNIVERSITY OF CANTERBURY, CHRISTCHURCH, NEW ZEALAND

E-mail address: s.grunewald@math.canterbury.ac.nz, pjh96@student.canterbury.ac.nz, c.semple@math.canterbury.ac.nz