PATCHING UP X-TREES

SEBASTIAN BÖCKER, ANDREAS W.M. DRESS, AND MIKE A. STEEL

ABSTRACT. A fundamental problem in many areas of classification, and particularly in biology, is the reconstruction of a leaf-labeled tree from just a subset of its induced subtrees. Without loss of generality, we may assume that these induced subtrees all have precisely four leaves. Of particular interest is the question of determining whether a collection of quartet subtrees uniquely defines a parent tree. Here, we solve this question in case the collection of quartet trees is of minimal size, by studying *encodings* of binary trees by such quartet trees. We obtain a characterization of minimal encodings that exploits an underlying "patchwork" structure. We thereby obtain a polynomial time algorithm for certain instances of the problem of reconstructing trees from subtrees.

1. INTRODUCTION

Trees are widely used to represent evolutionary, historical, or hierarchical relationships in various fields of classification. In biology for example, such trees ("phylogenies") typically represent the evolutionary history of a collection of extant species or the line of descent of some gene [17]. They may also be used to classify individuals (or populations) of the same species. In historical linguistics, trees have been used to represent the evolution of languages [18], while in the branch of philology known as stemmatology, trees may represent the way in which different versions of a manuscript arose through successive copying [13].

In most of these applications, the objects of interest occur at the tips (leaves) of the tree, and all other vertices of the tree correspond to a branching (or speciation) event. From a mathematical perspective, we have a finite set X [1)] of objects of

Date: December 16, 1998.

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

Key words and phrases. Trees, hierarchies, patchworks, quartet encodings, algorithms, consensus methods.

Sebastian Böcker was supported by "DFG – Graduiertenkolleg Strukturbildungsprozesse", Forschungsschwerpunkt Mathematisierung, University of Bielefeld, Germany.

Andreas Dress was supported by "Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie", FKZ 0311642, Germany; presently Distinguished Visiting Professor at City College, CUNY, Dept. of Chem. Engineering.

Mike Steel thanks the New Zealand Marsden Fund for supporting this research, and the Sir Isaac Newton Institute (Cambridge) for its hospitality.

[1)] Throughout this paper, let $\#X$ denote the cardinality of a finite set X, and $\wp(X)$ the set of subsets of X.

interest (species, languages, etc), and we consider triples $T = (V, E; \phi)$ consisting of a tree (V, E) with finite vertex set $V = V_T$ and edge set $E = E_T \subseteq \binom{V}{2}$, and a map $\phi = \phi_T : X \to V$ such that

$$(1) \qquad v \in \phi(X) \quad \text{holds for all} \quad v \in V \quad \text{with} \quad \deg_T(v) \leq 2\,,$$

where $\deg_T(v)$ denotes the degree $\#\{e \in E_T : v \in e\}$ of the vertex $v \in V_T$.

A triple $T = (V, E; \phi)$ which satisfies these conditions is henceforth called an *X-tree*. Denoting the set of vertices of degree i by V_i, we define an X-tree $T = (V, E; \phi)$ to be a *phylogenetic X-tree* if ϕ is a bijection from X onto the set V_1 of *leaves* of (V, E); if, in addition, every vertex in V is of degree 1 or 3 (in which case (V, E) is called a *binary tree*), then we will say that $T = (V, E; \phi)$ is a *binary X-tree*. Two X-trees $T = (V, E; \phi)$ and $T' = (V', E'; \phi')$ are *isomorphic* if there exists a bijection $\alpha : V \xrightarrow{\sim} V'$ which induces a bijection between E and E' and which satisfies $\phi' = \alpha \circ \phi$ (in which case there is exactly one such map α).

Clearly, every X-tree $T = (V, E; \phi)$ gives rise to a unique phylogenetic X-tree $T_{\text{phyl}} = (V_{\text{phyl}}, E_{\text{phyl}}; \phi_{\text{phyl}})$ defined by

$$V_{\text{phyl}} := V \mathbin{\dot\cup} \{x \in X : \deg_T(\phi(x)) \neq 1\}\,,$$
$$E_{\text{phyl}} := E \mathbin{\dot\cup} \{\{\phi(x), x\} : x \in X \text{ and } \deg_T(\phi(x)) \neq 1\}\,,$$
$$\text{and} \quad \phi_{\text{phyl}} : X \to V_{\text{phyl}} \ : \ x \mapsto \begin{cases} \phi(x) & \text{if } \deg_T(\phi(x)) = 1, \\ x & \text{otherwise.} \end{cases}$$

As is well known, there is a canonical and useful one-to-one correspondence between (isomorphism classes of) X-trees and certain set systems, due to Buneman [5], which we shall now recall:

A *split* of X is a subset $\{A, B\} \subseteq \wp(X)$ such that A, B is a bipartition of X into two non-empty, disjoint subsets; a *partial split* of X is a split of some non-empty subset of X. We let $\mathcal{S}(X)$ (resp. $\mathcal{S}_{\text{part}}(X)$) denote the set of all splits of X (resp. all partial splits of X). Two splits S_1, S_2 are called *compatible* if there exist $A_1 \in S_1$ and $A_2 \in S_2$ with $A_1 \cap A_2 = \emptyset$. A partial split $\{A, B\}$ is *trivial* if $\min\{\#A, \#B\} = 1$. A partial split $S_1 = \{A_1, B_1\}$ is said to *extend* a partial split $S_2 = \{A_2, B_2\}$ if $S_2 = \{(A_2 \cup B_2) \cap A_1, (A_2 \cup B_2) \cap B_1\}$ holds (that is, $A_2 \subseteq A_1$ and $B_2 \subseteq B_1$, or $A_2 \subseteq B_1$ and $B_2 \subseteq A_1$) in which case we will also write $S_2 \leq S_1$.

Now, each edge e in an X-tree $T = (V, E; \phi)$ gives rise to a split of X—simply delete e from E and apply ϕ^{-1} to the two connected components of the resulting graph $(V, E - \{e\})$ to obtain a split, which we will call a *T-split* and denote by $\mathcal{S}[e] = \mathcal{S}_T[e]$. Let $\mathcal{S}[T]$ denote the set of all T-splits. It is easily checked that distinct edges induce distinct splits and that the set $\mathcal{S}[T]$ is *compatible*, that is, any two splits from $\mathcal{S}[T]$ are compatible. Furthermore, we have $\#\mathcal{S}[T] \leq 2\#X - 3$ with equality precisely if T is a binary X-tree which follows easily from the following fundamental property of trees: if, for a tree (V, E), we denote the set of *inner edges* not incident with a leaf by $\mathring{E} \subset E$, then

$$\#\mathring{E} \leq \#V_1 + \#V_2 - 3 \tag{2}$$

for every finite tree (V, E) while equality holds for a tree with $\#V_2 = 0$ if and only if that tree is binary.

Buneman established the following fundamental correspondences in [5]:

Lemma 1. *The map $T \rightsquigarrow \mathcal{S}[T]$ induces bijections between:*

(i) *the set of (isomorphism classes of) X-trees and the set of compatible split systems $\mathcal{S} \subseteq \mathcal{S}(X)$;*
(ii) *the set of (isomorphism classes of) phylogenetic X-trees and the set of all those compatible split systems $\mathcal{S} \subseteq \mathcal{S}(X)$ that contain all trivial splits of X;*
(iii) *the set of (isomorphism classes of) binary X-trees and the set of compatible split systems $\mathcal{S} \subseteq \mathcal{S}(X)$ for which $\#\mathcal{S} = 2\#X - 3$ holds, or—equivalently—the set of maximal compatible split systems.*

This correspondence between X-trees and compatible split systems provides a convenient partial order on the set of (isomorphism classes of) X-trees: we write $T' \leq T$ precisely if $\mathcal{S}[T'] \subseteq \mathcal{S}[T]$. Informally, $T' \leq T$ states that T' can be obtained from T by contracting edges, and identifying corresponding vertices—so, $\phi_{T'}$ may be less "refined" than ϕ_T.

Note for instance that $T \leq T_{\text{phyl}}$ holds for every X-tree T and that T_{phyl} is uniquely determined (up to isomorphisms) by the equation

$$\mathcal{S}[T_{\text{phyl}}] = \mathcal{S}[T] \cup \Big\{\{\{x\}, X - \{x\}\} : x \in X\Big\}. \tag{3}$$

Now, given an X-tree $T = (V, E; \phi)$ and a non-empty subset $Y \subseteq X$, we obtain an induced Y-tree $T|_Y$ as follows: first construct the minimal subtree (V', E') of (V, E) that connects all vertices from $\phi(Y)$. Then, make this tree "homeomorphically irreducible" by replacing each maximal path running (except for its two end points) through degree-two vertices from $V' - \phi(Y)$ only, by a single edge (and deleting the superfluous vertices and edges) to obtain a tree (V_Y, E_Y). The restriction $\phi|_Y =: \phi_Y$ maps Y into V_Y and satisfies condition (1) (with X, V, and T replaced by Y, V_Y, and $T|_Y$, respectively). We call the resulting Y-tree $T|_Y = (V_Y, E_Y; \phi_Y)$ the *induced* (Y-)subtree of T. We will say that an X-tree T *displays* a Y-tree T' if $T' \leq T|_Y$ holds.

A more succinct, but less visual description of $T|_Y$ is, in view of Lemma 1, to describe its set of splits:

$$\begin{aligned} \mathcal{S}[T|_Y] &= \{S' \in \mathcal{S}(Y) : S' \leq S \text{ holds for some } S \in \mathcal{S}[T]\} \\ &= \{\{A \cap Y, B \cap Y\} : \{A, B\} \in \mathcal{S}[T] \text{ and } A \cap Y, B \cap Y \neq \emptyset\}. \end{aligned} \tag{4}$$

Our interest lies in the reverse reconstruction problem: Given as input a collection of subtrees, we wish to determine whether there exists some X-tree which induces or—at least—displays these subtrees, and if so, whether there exists exactly one such tree. Formally, let $(Y_1, \dots, Y_k)$ be a family of non-empty subsets of X, and consider a family $\mathcal{F} := (T_1, \dots, T_k)$ where T_j is a Y_j-tree for $j = 1, \dots, k$. We may wish to consider the set $\mathcal{T}(\mathcal{F})$ of all (isomorphism classes of) phylogenetic X-trees T that display every tree in $\mathcal{F}$.

As we will see in the following section, the related and seemingly more special task of reconstructing trees from partial splits is actually equivalent to the task described above; so, the problem of computing $\mathcal{T}(\mathcal{F})$ can always be reduced to this particular version of the reconstruction problem.

There are several reasons why such reconstruction problems arise naturally in applications such as biology. Firstly, we may wish to combine trees that have been reconstructed using distinct, though overlapping collections of species (usually by different researchers, using different data and, as often as not, different reconstruction methods). A second reason is that it is, in general, difficult to accurately reconstruct large trees directly, and we may choose instead to reconstruct trees for small subsets and then combine these in a parent tree (or parent trees), see [2, 9, 10, 16, 19]. A third reason is that, for genetic data, the number of sites that can be accurately aligned across a small number of closely related sequences is generally much larger than the corresponding number of sites for a set that is large and includes rather diverse sequences.

If we were to construct $\mathcal{F}$ by estimating a tree for *every* subset of size 4, then, as $\#\mathcal{T}(\mathcal{F}) \leq 1$ must hold (cf. [2]), we could easily compute $\mathcal{T}(\mathcal{F})$ and would usually find that $\mathcal{T}(\mathcal{F}) = \emptyset$ holds—that is, some of the subtrees must have been incorrectly estimated (this was already known to Colonius and Schulze [6, 7], see also [19]). Thus, we may wish to use only those subtrees which are strongly supported by the data (usually involving closely related objects)—and so we will generally have available trees for only a small number of subsets of X. This makes the reconstruction problem more difficult computationally, but—of course—also more gratifying whenever one is lead this way to a simultaneously non-empty and well-supported set of trees.

Of course, we could examine all X-trees to determine which (if any) of them display (every tree in) $\mathcal{F}$; however, this is computationally infeasible, since even the number of non-isomorphic binary X-trees grows super-exponentially with the number of leaves. Indeed, this number is precisely the product $1 \cdot 3 \cdots (2\#X - 5)$ of the first $(\#X - 2)$ odd numbers, a result which dates back to 1870, see [14]. This motivates the results described below.

2. Partial splits

Given a partial split $S = \{A, B\} \in \mathcal{S}_{\text{part}}(X)$, we define the *support* of S by $\underline{S} := A \cup B$, and for every $x \in \underline{S}$, we define $S(x) := A$ if $x \in A$ and $S(x) := B$ else; for $\mathcal{S} \subseteq \mathcal{S}_{\text{part}}(X)$, we define $\underline{\mathcal{S}} := \bigcup_{S \in \mathcal{S}} \underline{S}$.

For natural numbers i, j, let

$$(5) \qquad \mathcal{S}_{i,j}(X) := \big\{\{A, B\} \in \mathcal{S}_{\text{part}}(X) : \ \{\#A, \#B\} = \{i, j\}\big\}.$$

A partial split $Q \in \mathcal{Q}(X) := \mathcal{S}_{2,2}(X)$ is called a *quartet split*, and we write $Q = xy|wz$ as shorthand for $Q = \{\{x, y\}, \{w, z\}\}$.

Given a subset $\mathcal{S} \subseteq \mathcal{S}_{\text{part}}(X)$ of partial splits of X and an X-tree T, we say that T is *concordant* with $\mathcal{S}$ if, for every $\{A, B\} \in \mathcal{S}$, there exists at least one edge e of T that separates $\phi(A)$ from $\phi(B)$, that is, with $\{A, B\} \leq S[e]$. Clearly, T is concordant with $\mathcal{S}$ if and only if it displays, for every $S \in \mathcal{S}$, the corresponding 2-vertex $\underline{S}$-tree

$$(6) \qquad \big(S, \{S\}; \phi_S : \ \underline{S} \to S : \ x \mapsto S(x)\big).$$

Let $\mathcal{T}_X(\mathcal{S}) = \mathcal{T}(\mathcal{S})$ denote the set of all (isomorphism classes of) phylogenetic X-trees concordant with $\mathcal{S}$. We will say that $\mathcal{S}$ is *strictly arboreal* if $\mathcal{T}(\mathcal{S}) = \{T\}$ holds for some X-tree T in which case $\mathcal{S}$ will also be said to be *strictly T-arboreal*, while any set $\mathcal{S}$ with $\mathcal{T}(\mathcal{S}) \neq \emptyset$ will just be called *arboreal*, or *T-arboreal* if $T \in \mathcal{T}(\mathcal{S})$.

Note that $T \in \mathcal{T}(\mathcal{S})$ and $T \leq T'$ implies $T' \in \mathcal{T}(\mathcal{S})$, hence any X-tree T for which a set $\mathcal{S} \subseteq \mathcal{S}_{\text{part}}(X)$ exists which is strictly T-arboreal must be a binary X-tree.

Note also that a collection $\mathcal{S}$ of partial splits that is strictly T-arboreal for some binary X-tree $T = (V, E)$ must contain at least one partial split for every inner edge which specifically "fits" this edge and, hence, $\mathcal{S}$ must contain at least $\#\mathring{E} = \#X - 3$ distinct non-trivial splits in view of inequality (2). It is easily shown [2, 15] that, for $\mathcal{S} \subseteq \mathcal{S}_{\text{part}}(X)$ and

$$(7) \qquad \mathcal{Q}(\mathcal{S}) := \{Q \in \mathcal{Q}(X) : \ Q \leq S \text{ for some } S \in \mathcal{S}\},$$

the relation

$$\mathcal{T}(\mathcal{S}) = \mathcal{T}(\mathcal{Q}(\mathcal{S}))$$

must hold. Thus, there is no loss of generality in restricting one's attention to quartet splits when reconstructing (phylogenetic) trees from partial splits. Similarly, we have

$$(8) \qquad \mathcal{T}(\mathcal{F}) = \mathcal{T}\Big(\bigcup\nolimits_{i=1,\dots,k} \mathcal{S}[T_i]\Big) = \mathcal{T}\Big(\mathcal{Q}\Big(\bigcup\nolimits_{i=1,\dots,k} \mathcal{S}[T_i]\Big)\Big)$$

for any family $\mathcal{F} = (T_1, \dots, T_k)$ as above, so the problem of reconstructing trees from subtrees also reduces to the problem of reconstructing them from (quartet) splits.

3. Quartet Encodings

In this section, we analyze conditions under which a binary X-tree is uniquely determined by selecting, for each inner edge e, a corresponding representative quartet split Q with $Q \leq S[e]$. In order to describe our main result (Theorem 1), we must introduce some more terminology and preliminary results.

Definition 1. Let $T = (V, E; \phi)$ denote a binary X-tree. A map

$$q : \mathring{E} \to \mathcal{Q}(X)$$

is called a *quartet encoding* of T if $S[e]$ extends $q(e)$ for each $e \in \mathring{E}$. For every $e \in \mathring{E}$ and $F \subseteq \mathring{E}$, we define

$$\underline{q}(e) := \underline{q(e)}\,, \quad q(F) := \{q(e) : \; e \in F\}\,, \quad \text{and} \quad \underline{q}(F) := \bigcup_{e \in F} \underline{q}(e)\,.$$

We will say that q *defines* T if T is the only phylogenetic X-tree concordant with $q(\mathring{E})$, that is, if $q(\mathring{E})$ is strictly T-arboreal. A quartet encoding q is *tight* if, for each edge $e \in \mathring{E}$, there exists no other edge in $\mathring{E}$ separating the two subsets in $q(e)$.

It is easy to see (cf. [15]) that a quartet encoding that defines a tree T is tight; furthermore, given a binary X-tree T and a tight quartet encoding q of T, then $\underline{q}(\mathring{E}) = X$ and $\#q(\mathring{E}) = \#X - 3$ holds. It is also easy to see that a set of quartet splits $\mathcal{Q}$ with $\#\mathcal{Q} = \#X - 3$ is strictly T-arboreal for some (necessarily binary) X-tree T if and only if there exists a quartet encoding q of T with $\mathcal{Q} = q(\mathring{E})$ that defines T.

We now present three instructive examples of tight encodings.

Examples.

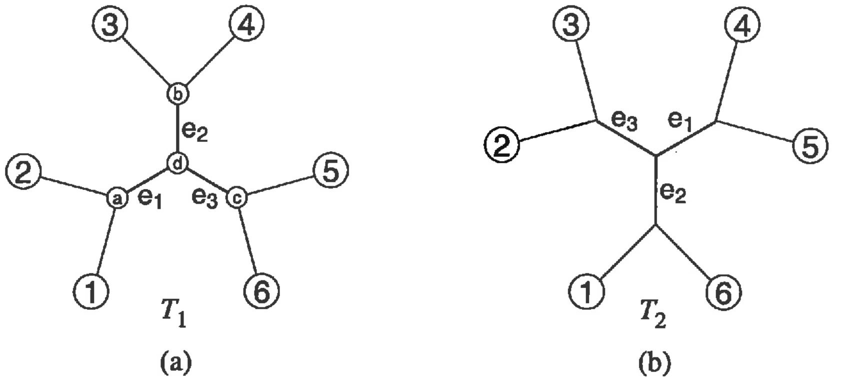
1. For $X := \{1, \dots, 6\}$, consider the binary X-tree $T_1 = (V, E; \phi_1 := Id_X)$ with $E := \{e_1, \dots, e_9\}$ having the nontrivial splits

$$\begin{aligned} S[e_1] &= \{\{1,2\},\{3,4,5,6\}\}\,, \\ S[e_2] &= \{\{3,4\},\{1,2,5,6\}\}\,, \\ \text{and} \quad S[e_3] &= \{\{5,6\},\{1,2,3,4\}\}\,, \end{aligned}$$

plus the six trivial splits as depicted in Fig. 1 (a). Consider the quartet encoding $q : \mathring{E} = \{e_1, e_2, e_3\} \to \mathcal{Q}(X)$ defined by

$$q(e_1) := 12|45, \quad q(e_2) := 34|16 \quad \text{and} \quad q(e_3) := 56|23\,. \tag{9}$$

Then q is a tight encoding of T_1, but does not define T_1, as it also encodes the X-tree T_2 depicted in Fig. 1 (b).



(a) (b)

FIGURE 1. Two possible binary trees for Example 1; also the trees T_1 and T_2 from the proof of Lemma 2.

2. Suppose that $X := \{1, \dots, 7\}$, that $T = (V, E; \phi := Id_X)$ is a caterpillar with seven leaves as depicted in Fig. 2, and that $q : \mathring{E} \to \mathcal{Q}(\{1, \dots, 7\})$ is the following quartet encoding:

$$q(e_1) := 12|36, \;\; q(e_2) := 13|46, \;\; q(e_3) := 24|57 \text{ and } q(e_4) := 25|67$$

Then it is easy to check that q defines T.

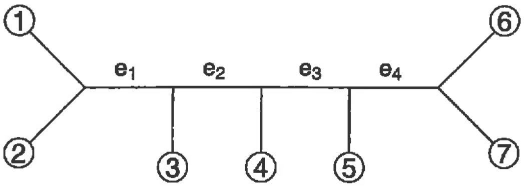


FIGURE 2. Caterpillar with seven leaves.

3. Suppose q is any tight quartet encoding of a binary X-tree with

$$\bigcap_{e \in \mathring{E}} \underline{q}(e) \neq \emptyset .$$

Then it has been observed in [15] that q defines T; see also Corollary 9.

At this point, we want to include a remark on the non-existence of consensus methods that are equivariant and Pareto on subtrees: In phylogenetic analysis, a *consensus method* M is a function that takes a collection of X-trees and returns a single X-tree (which—hopefully—represents some "consensus" or common agreement between the trees). Clearly, a desirable property of such a method is that it should be independent of how the objects in X are ordered. More precisely, given a collection $\mathcal{F}$ of X-trees and a permutation π of X, let $\mathcal{F}^\pi := \{(V, E; \phi \circ \pi) : (V, E; \phi) \in \mathcal{F}\}$. Then M should satisfy the following property of *equivariance*:

$$M[\mathcal{F}] = (V, E; \phi) \implies M[\mathcal{F}^\pi] = (V, E; \phi \circ \pi) \tag{10}$$

A variety of equivariant consensus methods have been proposed, and in case X has a distinguished (out-group) element $o \in X$ (in which case we may regard the trees as rooted), there exist methods which satisfy a further desirable "Pareto" condition: if all the input trees in $\mathcal{F}$ display the same A-tree for a subset A of X containing o, then the consensus tree should display this A-tree, too. Formally, such a method satisfies the following property: whenever $o \in A \subseteq X$, T_A is an A-tree, and $\mathcal{F}$ is a set of X-trees, then:

$$T_A \leq T|_A \text{ for all } T \in \mathcal{F} \implies T_A \leq M[\mathcal{F}]\big|_A \tag{11}$$

Despite some attempts by practitioners, no equivariant method has been found which satisfies (11) if the restriction $o \in A$ is lifted. In fact, as we now show, no such method can exist, even if one restricts oneself to sets $\mathcal{F}$ of phylogenetic or binary trees:

Lemma 2. *When $\#X \geq 6$, there is no equivariant consensus method M which satisfies* (11) *for all subsets A of X of size 4, even if M is restricted to sets $\mathcal{F}$ of binary or phylogenetic X-trees.*

Proof. We consider the case $X = \{1, 2, 3, 4, 5, 6\}$, the general case is similar. Let $T_1 = (V, E; \phi_1 := Id_X)$ denote the tree depicted in Fig. 1 (a), let ϕ_2 denote the permutation $\pi := (2,6) \circ (3,5)$ of X, put $T_2 := (V, E; \phi_2)$ and, finally, put $\mathcal{F} := \{T_1, T_2\}$ (cf. Figure 1). Now, it can be checked that if an X-tree T displays $T_1|_A$ for $A \in \{\underline{q}(e_i) : i = 1, 2, 3\}$ (where q is given as in (9)), then T must be isomorphic to either T_1 or T_2. So, any tree $M[\mathcal{F}]$ satisfying the condition

$$T_1|_A \leq M[\mathcal{F}]\big|_A \quad \text{for all } A \subseteq X \text{ with } \#A = 4 \text{ and } T_1|_A \cong T_2|_A$$

must be isomorphic to either T_1 or T_2, and, hence, it cannot remain invariant under π, while we do have $\mathcal{F}^\pi = \{T_1, T_2\}^\pi = \{T_2, T_1\} = \mathcal{F}$. □

To exploit inequality (2), we now introduce the following definition:

Definition 2. Given a binary X-tree $T = (V, E; \phi)$, a subset $F \subseteq \mathring{E}$, and a quartet encoding $q : \mathring{E} \to \mathcal{Q}(X)$, the ($q$-)*excess* of F is given by

$$\operatorname{exc}(F) = \operatorname{exc}_q(F) := \#\underline{q}(F) - \#F - 3\,. \tag{12}$$

We say F is (q-)*excess-free* if $\operatorname{exc}(F) = 0$ holds.

Note that $\operatorname{exc}(\{e\}) = 0$ holds for every $e \in \mathring{E}$, and that $\operatorname{exc}(\emptyset) = -3$.

Lemma 3. *Suppose that q is a tight quartet encoding of a binary X-tree $T = (V, E; \phi)$. Then,*

(i) $\operatorname{exc}(\mathring{E}) = 0$*;*
(ii) *for every non-empty subset $F \subseteq \mathring{E}$, one has* $\operatorname{exc}(F) \geq 0$*;*

(iii) *if F is excess-free, then F is a connected subset of edges in T, and F equals the set of inner edges of* $T|_{\underline{q}(F)}$.

Sketch of proof. (i) This is merely the statement that a binary tree with n leaves has $n-3$ inner edges.

(ii) Consider $T|_Y = (V_Y, E_Y; \phi_Y)$ with $Y := \underline{q}(F)$ and note that, as q is tight, we have $F \subseteq (E_Y)^\circ$. Hence, (2) implies

$$\#F \leq \#(E_Y)^\circ \leq \#Y - 3 = \#\underline{q}(F) - 3\,,$$

as claimed.

(iii) As above, we consider $T|_Y = (V_Y, E_Y; \phi_Y)$ for $Y := \underline{q}(F)$. If F were not connected, then $F \subsetneq (E_Y)^\circ$ since $(E_Y)^\circ$ is connected. From (2), we would conclude $\#F < \#(E_Y)^\circ \leq \#Y - 3$ and, hence, $\operatorname{exc}(F) > 0$. □

To illustrate the usefulness of these concepts, we show now how they provide further constructions of encodings that define a binary X-tree T. First, we make a further definition: let us say that two distinct elements x, y of X are (a pair of) *twins* of T if the two edges incident with $\phi(x)$ and $\phi(y)$, respectively, share a vertex $v = v(x, y)$—in which case the third edge incident with v is denoted by $e(x, y)$. In Example 1, for instance, ① and ② are a pair of twins for the tree depicted in Fig. 1 (a) with $e(①, ②) = e_1$. If $\#X \geq 4$, every binary X-tree has at least two pairs of twins, and one has $e(x, y) \in \mathring{E}$ for every pair x, y of twins.

Examples.

4. Suppose that $T = (V, E; \phi)$ is a binary X-tree such that the inner edges of T are labeled $\mathring{E} = \{e_1, \dots, e_{n-3}\}$, and that q is a tight quartet encoding of T with

$$\#\Big(\underline{q}(e_i) \setminus \bigcup_{j<i} \underline{q}(e_j)\Big) = 1 \quad \text{for} \quad i = 2, \dots, n-3\,. \tag{13}$$

Then q defines T. To prove this, we apply induction on n. The result holds for $n = 4$, so suppose it holds for $4, \dots, n-1$ and that $\#X = n$. Let T' denote another binary X-tree that is concordant with $q(\mathring{E})$. By assumption, there exists $x \in X$ with $x \in \underline{q}(e_{n-3})$, but $x \notin \underline{q}(e_j)$ for $j = 1, \dots, n-4$. We define $Y := X - \{x\}$ and $F := \{e_1, \dots, e_{n-4}\}$; then $q|_F$ defines $T|_Y$ as well as $T'|_Y$ and, hence, $T|_Y \cong T'|_Y$ by the induction hypothesis. It remains to show that x is attached to the same edge of T and T'. To this end, we infer from Lemma 3 (iii) that F is connected, since q is tight and $\operatorname{exc}(F) = 0$. So, x must have a twin in T, denoted $y \in Y$, and $\{x, y\} \in q(e_{n-3})$ must hold. But the same holds true for T' which indeed implies $T \cong T'$. □

5. For a binary X-tree $T = (V, E; \phi)$, define $\mathrm{clus}(T) := \bigcup_{e \in E} S[e]$, the set of *clusters* of T. Suppose $f : \mathrm{clus}(T) \to X$ is any function which satisfies the condition

$$f(A) \in A \quad \text{for all} \quad A \in \mathrm{clus}(T)\,. \tag{14}$$

Then, f defines a tight quartet encoding of T, denoted q_f, as follows: for each inner edge $e = \{v_1, v_2\}$, deletion of v_1 and v_2 and their incident edges partitions T into four connected components and, thereby, it partitions X into four sets $\{A_1, A_2, B_1, B_2\}$, where we may suppose, without loss of generality, that $S[e] = \{\{A_1 \cup A_2\}, \{B_1 \cup B_2\}\}$. Let

$$q_f(e) := f(A_1)f(A_2)|f(B_1)f(B_2)\,.$$

Clearly, not every tight quartet encoding is of this form; furthermore, q_f does not necessarily define T. However, if we impose a further condition we can achieve this, as follows: suppose f satisfies the condition:

$$f(A) \in B \subseteq A \Longrightarrow f(A) = f(B) \quad \text{for all} \quad A, B \in \mathrm{clus}(T) \tag{15}$$

Then q_f satisfies the condition described in Example 4. In particular, q_f defines T.

Sketch of proof. It suffices to show that the edges of T can be labeled as described in Example 4 (and so as to satisfy (13)) which is again achieved by induction on $n := \#X$. Let $\{x, x'\}$ be a twin in T, and suppose $f(\{x, x'\}) = x'$. We define $Y := X - \{x\}$ and $T' := T|_Y$. Then, by the condition placed on f we obtain a corresponding function $f' : \mathrm{clus}(T') \to Y$ that satisfies (14) with f, T replaced by f', T', and hence the edges of T' can be ordered $\{e_1, \ldots e_{n-4}\}$ so as to satisfy the condition (15). We then label the edge $e(x, x')$ of T incident with the twin $\{x, x'\}$ as e_{n-3}, and verify that this also satisfies (13) for $i = n - 3$. □

This last example generalizes a result from [8] where a specific function f satisfying (15) is considered. In that setting, T has an associated edge weighting, X is given a total ordering, and f selects, for each cluster A, the minimal element (under the imposed ordering on X), taken over all elements of the set A that are nearest to that edge e for which $S[e] = \{A, X - A\}$ holds.

The excess-free subsets of $\mathring{E}$ for a tight quartet encoding q have a useful "patchwork" structure, which we now discuss. Following [3], a collection $\mathcal{C}$ of subsets of a set M is called an *M-patchwork* if it satisfies the following condition:

$$A, B \in \mathcal{C} \quad \text{and} \quad A \cap B \neq \emptyset \quad \Longrightarrow \quad A \cap B, A \cup B \in \mathcal{C} \tag{16}$$

There are quite a number of distinct characterizations of patchworks that are *ample*, that is (cf. [3]), of patchworks $\mathcal{C}$ that satisfy the condition

$$(17) \qquad A, B \in \mathcal{C} \quad \text{and} \quad \#\{C \in \mathcal{C} :\ A \subseteq C \subseteq B\} = 2 \implies B - A \in \mathcal{C}\,;$$

in particular, if M is finite, then

(i) a patchwork $\mathcal{C} \subseteq \wp(M)$ with $\{m\} \in \mathcal{C}$ for all $m \in M$ is ample if and only if, for every cluster $C \in \mathcal{C}$ with $\#C \geq 2$, there exist disjoint non-empty clusters $A, B \in \mathcal{C}$ with $C = A \cup B$;

(ii) given a patchwork $\mathcal{C} \subseteq \wp(M)$ with $\{m\} \in \mathcal{C}$ for all $m \in M$ and $\emptyset, M \in \mathcal{C}$, then $\mathcal{C}$ is ample if and only if $\mathcal{C}$ contains a *maximal hierarchy* $\mathcal{C}'$, that is a maximal subset $\mathcal{C}'$ of $\wp(M)$ for which $A \cap B \in \{\emptyset, A, B\}$ holds for all $A, B \in \mathcal{C}'$.

Note that for $n = \#M$, we can check whether an arbitrary patchwork $\mathcal{C} \subseteq \wp(M)$ is ample in $O(k \cdot n^2)$ steps, provided we can compute, for every $N \subseteq M$, the value $\chi_{\mathcal{C}}(N)$ of the characteristic function $\chi_{\mathcal{C}} : \mathcal{C} \to \{0, 1\}$ in at most k steps (see [3]).

Lemma 4. *Suppose that q is an arbitrary quartet encoding of a binary X-tree $T = (V, E; \phi)$ and assume $F_1, F_2 \subseteq \mathring{E}$. Then,*

$$(18) \qquad \operatorname{exc}(F_1 \cup F_2) + \operatorname{exc}(F_1 \cap F_2) \leq \operatorname{exc}(F_1) + \operatorname{exc}(F_2)\,.$$

Proof. In view of $\underline{q}(F_1 \cap F_2) \subseteq \underline{q}(F_1) \cap \underline{q}(F_2)$, we have

$$\begin{aligned}
&\operatorname{exc}(F_1 \cup F_2) + \operatorname{exc}(F_1 \cap F_2) + 6 \\
&\qquad = \#\underline{q}(F_1 \cup F_2) - \#(F_1 \cup F_2) + \#\underline{q}(F_1 \cap F_2) - \#(F_1 \cap F_2) \\
&\qquad \leq \#\big(\underline{q}(F_1) \cup \underline{q}(F_2)\big) + \#\big(\underline{q}(F_1) \cap \underline{q}(F_2)\big) \\
&\qquad\quad - \big(\#(F_1 \cup F_2) + \#(F_1 \cap F_2)\big) \\
&\qquad = \#\underline{q}(F_1) + \#\underline{q}(F_2) - \#F_1 - \#F_2 \\
&\qquad = \operatorname{exc}(F_1) + \operatorname{exc}(F_2) + 6\,.
\end{aligned}$$

□

Lemma 5. *If q is a tight quartet encoding of $T = (V, E; \phi)$, then the excess-free subsets of $\mathring{E}$ form a patchwork denoted by $\mathcal{C}(q)$.*

Proof. Let $F_1, F_2 \subseteq \mathring{E}$ denote subsets with $\operatorname{exc}(F_1) = \operatorname{exc}(F_2) = 0$ and $F_1 \cap F_2 \neq \emptyset$. By Lemma 3 (ii), we have

$$\operatorname{exc}(F_1 \cap F_2) \geq 0 \quad \text{and} \quad \operatorname{exc}(F_1 \cup F_2) \geq 0\,,$$

yet by Lemma 4,

$$\operatorname{exc}(F_1 \cap F_2) + \operatorname{exc}(F_1 \cup F_2) \leq 0\,.$$

Consequently, $\operatorname{exc}(F_1 \cap F_2) = \operatorname{exc}(F_1 \cup F_2) = 0$, as required. □

It is tempting to conjecture that if q defines a binary tree T, then there always exists an edge $e \in \mathring{E}$ such that $\mathring{E} - \{e\}$ is q-excess-free. In some cases (eg. for encodings constructed as in Examples 4 and 5 above), this is indeed the case, but in general it is not (Example 2 provides a counterexample with seven leaves). Nevertheless, we can still hope that the excess-free subsets of $\mathring{E}$ form at least an ample patchwork, and this turns out to be indeed the case, as we state now as part of our main result:

Theorem 1. *Given a quartet encoding q of a binary X-tree $T = (V, E; \phi)$, then the following two statements are equivalent:*

(i) *T is defined by q;*
(ii) *q is tight, and the patchwork $\mathcal{C}(q)$ of excess-free subsets of $\mathring{E}$ is ample.*

While the proof that (ii) implies (i) is relatively easy, following the lines of thought used already in the previous examples, the converse—except for the fact that q must be tight—is far from trivial: One proceeds by induction relative to $n := \#X$ which allows one to assume that

$$\mathcal{C}(q)_{\subseteq F} := \{F' \in \mathcal{C}(q) : \; F' \subseteq F\}$$

is an ample patchwork for all subsets F of X contained in

$$\mathcal{C}(q)_{\subsetneq X} := \{F' \in \mathcal{C}(q) : \; F' \subsetneq X\}\,.$$

It then follows easily (cf. [3]) that

$$\max\left(\mathcal{C}(q)_{\subsetneq X}\right) := \left\{F \in \mathcal{C}(q)_{\subsetneq X} : \; F \subseteq F' \in \mathcal{C}(q)_{\subsetneq X} \text{ implies } F = F'\right\}$$

must be a partition of X into at least three distinct subsets.

The next step consists of applying the induction hypothesis to trees one derives from T by identifying pairs of twins $x, y \in X$. This way, one is led to study decompositions of $\mathring{E}$ into two disjoint and connected subsets $F_1 = F_1(x, y)$ and $F_2 = F_2(x, y)$ with $e(x, y) \in F_1$,

$$\#\Big(\underline{q}(F_1 - \{e(x,y)\}) \cup \{x, y\}\Big) = \#F_1 + 3\,,$$

and $\operatorname{exc}(F_2) = 0$ or $\#(\underline{q}(F_2) - \{x, y\}) = \#F_2 + 2$. Next, one shows—and this is the most tricky part of the whole proof—that (i) by choosing $F_1 = F_1(x, y)$ maximal subject to these conditions, one can always assume $\operatorname{exc}(F_2) = 0$, and (ii) that $\operatorname{exc}(F_2) = 0$ and $\#F_2 \geq 2$ would in turn imply the existence of a connected subset $F_2' \in \mathcal{C}(q)$ with $F_2 \subseteq F_2'$ and $\mathring{E} - F_2' \in \mathcal{C}(q)$ in contradiction to $\#\max(\mathcal{C}(q)_{\subsetneq X}) \geq 3$. Consequently, we can assume that, for every pair x, y of twins in T, there exists a single edge $f(x, y) \in \mathring{E} - \{e(x, y)\}$ such that $F_1 := \mathring{E} - \{f(x, y)\}$ and $F_2 := \{f(x, y)\}$ is a pair as above, that is, with

$$\begin{aligned}\#\Big(\underline{q}(\mathring{E} - \{e(x,y), f(x,y)\}) \cup \{x,y\}\Big) &= \#(\mathring{E} - \{f(x,y)\}) + 3\\ &= (\#X - 4) + 3 \;=\; \#X - 1\end{aligned}$$

which in turn implies that $f(x,y)$ must be of the form $e(x',y')$ for some pair x',y' of twins (because $F_1 = \mathring{E} - \{f(x,y)\}$ is connected) and with, say, x' the unique element in X missing in $\underline{q}(\mathring{E} - \{e(x,y), f(x,y)\})$. Applying the same argument now to the twins x',y' and repeating this process iteratively will therefore eventually produce a sequence of distinct pairs x_1, y_1; x_2, y_2; $\ldots$; x_k, y_k of twins such that, for $i = 1, \ldots, k$ (mod k) we have

$$\underline{q}\big(\mathring{E} - \{e(x_i,y_i), e(x_{i+1},y_{i+1})\}\big) \cup \{x_i, y_i\} = X - \{x_{i+1}\}$$

which in turn implies easily that we can construct a second X-tree $(V', E'; \phi')$ which is concordant with $q(\mathring{E})$: Note first that our assumptions imply that, for each $i = 1, \ldots, k$, there must exist some $z_i \in X$ with $x_i y_i | x_{i+1} z_i = q(e(x_i, y_i))$, and that $z_i \in \{x_1, \ldots, x_k\}$ cannot hold because—according to our construction—$e(x_i, y_i)$ is the *only* edge in $\mathring{E} - \{e(x_{i+1}, y_{i+1})\}$ with $x_{i+1} \in \underline{q}(e(x_i, y_i))$. Hence, we can cut off, for every $i = 1, \ldots, k$ (mod k), the edge incident with $\phi(x_{i+1})$ and implant it instead into the edge incident with $\phi(z_i)$ as depicted in Fig. 3 below.

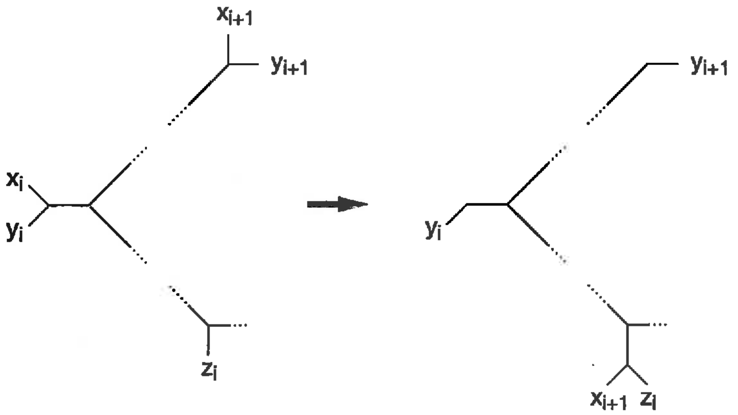


FIGURE 3. Cutting off and re-implanting edges.

If $z_i = z_{i+1}$, we have to make sure that the edge leading to $\phi(x_{i+2})$ gets implanted closer to $\phi(z_i)$ than the edge leading to $\phi(x_{i+1})$—no further special care needs to be taken (and because we cannot have $z_1 = z_2 = \cdots = z_k$, this requirement can always be fulfilled).

Finally, defining V', E' and ϕ' accordingly, we find a second X-tree $(V', E'; \phi')$ which is clearly non-isomorphic with $(V, E; \phi)$—the final contradiction. □

Remark 6. Note that for the tree T_1 considered in Example 1 and in the proof of Lemma 2 (see also Fig. 1), we can choose $f(2i-1, 2i) \in \{e_j, e_k\}$ for all $\{i,j,k\} = \{1,2,3\}$, so we can use the twin sequence 1, 2; 4, 3 as well as the twin sequence 2, 1; 4, 3; 6, 5, leading to $z_1 = 5$ and $z_2 = 6$ or $z_1 = 5$, $z_2 = 1$, and $z_3 = 3$, respectively; and indeed, both rearrangements lead to a tree isomorphic to T_2.

The detailed proof will take up more than 20 pages of reasoning, and so it will be published elsewhere (see [4]). The equivalence of the conditions of the following corollary is proven in [3].

Corollary 7. *Suppose we are given a binary X-tree $T = (V, E; \phi)$ and a set of quartet splits $\mathcal{Q}$ such that $\#\mathcal{Q} = \#\underline{\mathcal{Q}} - 3$. Then, $\mathcal{Q}$ is strictly T-arboreal if and only if there exists a tight quartet encoding q of T with $\mathcal{Q} = q(\mathring{E})$, and one of the following (equivalent) conditions holds true:*

(i) *the patchwork $\mathcal{C}(q)$ of q-excess-free subsets of $\mathring{E}$ is ample;*
(ii) *for every subset $F \in \mathcal{C}(q)$ with $\#F \geq 2$, there exist disjoint subsets $F_1, F_2 \subseteq F$ with $F_1 \cup F_2 = F$ and $F_1, F_2 \in \mathcal{C}(q)$;*
(iii) *given a family $\mathcal{F} = (F_i)_{i \in I}$ (with index set I of cardinality $\#I \geq 2$) of disjoint subsets $F_i \in \mathcal{C}(q)$ for $i \in I$ such that $\bigcup_{i \in I} F_i \in \mathcal{C}(q)$ holds, there exist distinct indices $i, j \in I$ with $F_i \cup F_j \in \mathcal{C}(q)$.*

Corollary 8. *Suppose a set of quartet splits $\mathcal{Q}$ with $\#\mathcal{Q} = \#\underline{\mathcal{Q}} - 3 \geq 2$ is strictly arboreal. Then, $\mathcal{Q}$ is the disjoint union of two proper subsets $\mathcal{Q}_1$ and $\mathcal{Q}_2$ with $\#\mathcal{Q}_i = \#\underline{\mathcal{Q}_i} - 3$ for $i = 1, 2$ that are strictly arboreal.*

4. Implications for tree reconstruction

We return now to the problem that motivated our analysis of minimal quartet encodings of binary trees, namely the problem of reconstructing trees from a collection of quartet splits. As a corollary to Theorem 1, we can derive the following result already mentioned in Example 3:

Corollary 9. *Suppose $\mathcal{Q} \subseteq \mathcal{Q}(X)$ is a subset of cardinality $\#\mathcal{Q} = \#\underline{\mathcal{Q}} - 3$ such that $\bigcap_{Q \in \mathcal{Q}} \underline{Q} \neq \emptyset$ holds. If there exists a binary X-tree T and a tight quartet encoding q of T with $q(\mathring{E}) = \mathcal{Q}$, then $\mathcal{Q}$ is strictly T-arboreal.*

Proof. Assume $x \in \bigcap_{Q \in \mathcal{Q}} \underline{Q}$, define $v \leq u$ for $u, v \in V$ if the path from $\phi(x)$ to u passes through v, and put

$$V(v) := \{u \in V : v \leq u\} \quad \text{and}$$
$$\mathring{E}(v) := \{e \in \mathring{E} : e \subseteq V(v)\}.$$

Now, it is easy to see that

$$\mathcal{C}' := \{\mathring{E}(v) : v \in V\}$$

is a maximal $\mathring{E}$-hierarchy, and that every non-empty element from $\mathcal{C}'$ is in $\mathcal{C}(q)$. In view of the results in [3], this implies that $\mathcal{C}(q)$ is ample. □

Suppose we have a set $\mathcal{Q}$ of quartet splits, and we wish to determine whether or not $\mathcal{Q}$ is strictly arboreal, and if so to construct the unique (binary) tree T that is concordant with $\mathcal{Q}$, where both tasks are to be carried out in polynomial time in $n := \#X$. If the conditions of Corollary 9 are fulfilled, this can be achieved using the algorithm of Aho et al. [1]. However, in general, even the problem of determining whether or not $\mathcal{T}(Q)$ is non-empty is an NP-complete problem [15]. Nevertheless, Theorem 1 allows us to construct a polynomial time algorithm which has the following property: in case $\mathcal{Q}$ contains a strictly arboreal subset $\mathcal{Q}' \subseteq \mathcal{Q}$ with $\underline{\mathcal{Q}'} = X$ and $\#\mathcal{Q}' = \#X - 3$, then the algorithm determines whether $\mathcal{Q}$ is arboreal, and if so reconstructs the unique (necessarily binary) tree in $\mathcal{T}(\mathcal{Q})$.

At this point, two comments are in order. Firstly, even if we suppose that $\#\mathcal{T}(\mathcal{Q}) = 1$ holds, this does not imply that $\mathcal{Q}$ contains a strictly (T-)arboreal subset of size $\#\underline{\mathcal{Q}} - 3$ as the following example demonstrates.

Example.

6. The set $\mathcal{Q} := \{12|35, 24|57, 13|47, 34|56, 15|67\} \subseteq \mathcal{Q}(\{1, \dots, 7\})$ is strictly arboreal for the caterpillar with seven leaves as depicted in Fig. 2, but, for any subset of $\mathcal{Q}$ of size $7 - 3 = 4$, there are at least two trees concordant with the resulting collection. (from [15])

In [4], we will present an algorithm that checks—given a quartet encoding $q : \mathring{E} \to \mathcal{Q}(X)$ of some binary X-tree $T = (V, E; \phi)$—whether $\mathcal{C}(q)$ forms an ample patchwork in $O(n^2)$ time for $n = \#X$. Furthermore, if $\mathcal{Q} \subseteq \mathcal{Q}(X)$ is strictly arboreal and has size $n = \#\underline{\mathcal{Q}} - 3$, it is relatively simple to formulate a recursive, polynomial time (in n) algorithm for reconstructing T from $\mathcal{Q}$ using Theorem 1. However, a much more useful algorithm would reconstruct T when $\mathcal{Q}$ merely contains such a strictly arboreal set of size $n-3$ as a(n unknown) subset. The following approach achieves this objective in polynomial time.

Definition 3. The *dyadic closure* of a collection $\mathcal{Q} \subseteq \mathcal{Q}(X)$ of quartet splits, denoted $cl_2(\mathcal{Q})$, is the minimal subset of $\mathcal{Q}(X)$ that contains $\mathcal{Q}$ and is closed under the following two rules:

(dc1) $\quad ab|cd, ab|ce \in cl_2(\mathcal{Q}) \implies ab|de \in cl_2(\mathcal{Q})$

(dc2) $\quad ab|cd, ac|de \in cl_2(\mathcal{Q}) \implies ab|ce, ab|de, bc|de \in cl_2(\mathcal{Q})$

It is clear that the dyadic closure $cl_2(\mathcal{Q})$ of $\mathcal{Q} \subseteq \mathcal{Q}(X)$ can be computed in polynomial time in $n := \#X$ (for a particular algorithm, see [8]). It is of course possible that $cl_2(\mathcal{Q})$ may contain both $xy|wz$ and $xw|yz$ for $x, y, w, z \in X$—such a pair of quartet splits we say is *contradictory*. In this case, it is clear that $\mathcal{T}(\mathcal{Q}) = \emptyset$ since any tree that induces the input splits to either rule, must also induce the output split, and no tree can induce contradictory splits. Note also that T can be reconstructed from $\mathcal{Q}(\mathcal{S}[T])$ in $O(n \log(n))$ steps, see [11,12]. Thus, the following theorem (which relies on Corollary 8 to Theorem 1, and whose proof we also defer to [4]) provides a polynomial time solution to the tree reconstruction problem under certain conditions:

Theorem 2. *Suppose there exist collections $\mathcal{Q}' \subseteq \mathcal{Q} \subseteq \mathcal{Q}(X)$ with $\underline{\mathcal{Q}'} = X$ and $\#\mathcal{Q}' = \#X - 3$, and a binary X-tree T such that $\mathcal{Q}'$ is strictly T-arboreal. If $\mathcal{T}(\mathcal{Q}) \neq \emptyset$, then $cl_2(\mathcal{Q}) = \mathcal{Q}[T]$ holds; otherwise $cl_2(\mathcal{Q})$ contains a contradictory pair of quartet splits.*

Example 6 shows that even when $\mathcal{Q}$ is strictly T-arboreal, it may not necessarily be the case that $cl_2(\mathcal{Q}) = \mathcal{Q}[T]$. Indeed, the computational complexity of determining whether a general set $\mathcal{Q} \subseteq \mathcal{Q}(X)$ is strictly arboreal (or even strictly T-arboreal for T given) is still not resolved. We also note that $cl_2(\mathcal{Q})$ can equal $\mathcal{Q}[T]$ even if $\mathcal{Q}$ does not contain a strictly T-arboreal subset of size $\#\underline{\mathcal{Q}} - 3$ as the set $\mathcal{Q}^* := \{12|34, 12|45, 26|15, 45|36\}$ testifies. Notice that this example relies on both dyadic rules to reconstruct $\mathcal{Q}[T]$. However, it can be shown that Theorem 2 remains true when $cl_2(\mathcal{Q})$ is replaced by the "semi-dyadic closure" of $\mathcal{Q}$, which is defined in the same way as $cl_2(\mathcal{Q})$, except allowing just rule (dc2). If we denote the semi-dyadic closure of $\mathcal{Q}$ by $scl_2(\mathcal{Q})$ then $scl_2(\mathcal{Q}^*) = \mathcal{Q}^*$ holds for the example mentioned above.

References

1. Alfred V. Aho, Yehoshua Sagiv, Thomas G. Szymanski, and Jeffrey D. Ullman, *Inferring a tree from lowest common ancestors with an application to the optimization of relational expressions*, SIAM J. Comput. **10** (1981), no. 3, 405–421.
2. Hans-Jürgen Bandelt and Andreas Dress, *Reconstructing the shape of a tree from observed dissimilarity data*, Adv. in Appl. Math. **7** (1986), no. 3, 309–343.
3. Sebastian Böcker and Andreas W.M. Dress, *Patchworks*, manuscript, 1998.
4. Sebastian Böcker, Andreas W.M. Dress, and Mike Steel, *Most parsimonious quartet encodings of binary trees*, in preparation.
5. Peter Buneman, *The recovery of trees from measures of dissimilarity*, Mathematics in the Archaeological and Historical Sciences (F.R. Hodson, D.G. Kendall, and P. Tautu, eds.), Edinburgh University Press, Edinburgh, 1971, pp. 387–395.
6. Hans Colonius and Hans-Henning Schulze, *Repräsentation nichtnummerischer Ähnlichkeitsdaten durch Baumstrukturen*, Psych. Beitr. **21** (1979), 98–111.
7. Hans Colonius and Hans-Henning Schulze, *Tree structures for proximity data*, British J. Math. Statist. Psych. **34** (1981), no. 2, 167–180.
8. Péter L. Erdős, László A. Székely, Mike Steel, and Tandy Warnow, *A few logs suffice to build (almost) all trees*, Random Structures and Algorithms, in press.
9. Daniel Huson, Scott Nettles, Laxmi Parida, Tandy Warnow, and Shibu Yooseph, *The disk-covering method for tree reconstruction*, Proceedings of "Algorithms and Experiments" (ALEX98), Trento, Italy, Feb. 9–11, 1998 (R. Battiti and A.A. Bertossi, eds.), 1998, pp. 62–75.
10. Daniel Huson and Tandy J. Warnow, *Obtaining highly accurate topology and evolutionary estimates of evolutionary trees from very short sequences*, accepted for RECOMB 99.
11. Sampath K. Kannan, Eugene L. Lawler, and Tandy J. Warnow, *Determining the evolutionary tree using experiments*, J. Algorithms **21** (1996), no. 1, 26–50.
12. Judea Pearl and Michael Tarsi, *Structuring causal trees*, J. Complexity **2** (1986), no. 1, 60–77, Complexity of approximately solved problems (Morningside Heights, N.Y., 1985).

13. Peter M.W. Robinson, *Computer-assisted stemmatic analysis and 'best-text' historical editing*, Studies in Stemmatology (P.V. Reenen and M. van Mulken, eds.), John Benjamins Publishing, Amsterdam, 1996, pp. 71–103.
14. Ernst Schröder, *Vier Kombinatorische Probleme.*, Z. Math. Phys. **15** (1870), 361–376.
15. Mike Steel, *The complexity of reconstructing trees from qualitative characters and subtrees*, J. Classification **9** (1992), no. 1, 91–116.
16. Korbinian Strimmer and Arndt von Haeseler, *Quartet puzzling: A quartet maximum likelihood method for reconstructing tree topologies*, Mol. Biol. Evol. **13** (1996), no. 7, 964–969.
17. David L. Swofford, Gary J. Olsen, Peter J. Waddell, and David M. Hillis, *Phylogenetic inference*, Molecular Systematics (D.M. Hillis, C. Moritz, and B.K. Marble, eds.), Sinauer Associates, second ed., 1996, pp. 407–514.
18. Tandy Warnow, *Mathematical approaches to comparative linguistics*, Proc. Nat. Acad. Sci. U.S.A. **94** (1997), no. 13, 6585–6590.
19. Stephen J. Willson, *Measuring inconsistency in phylogenetic trees*, J. Theor. Biol. **190** (1998), no. 1, 15–36.

GK Strukturbildungsprozesse, FSP Mathematisierung, Universität Bielefeld,
PF 100 131, 33501 Bielefeld, Germany

E-mail address: `boecker@mathematik.uni-bielefeld.de`

GK Strukturbildungsprozesse, FSP Mathematisierung, Universität Bielefeld,
PF 100 131, 33501 Bielefeld, Germany

Current address: The City College, The City University of New York, Dept. of Chem. Engineering, Convent Avenue at 140$^{\text{th}}$ Street, New York, NY 10031, USA

E-mail address: `dress@mathematik.uni-bielefeld.de`

Biomathematics Research Centre, University of Canterbury, Private Bag 4800,
Christchurch, New Zealand

E-mail address: `m.steel@math.canterbury.ac.nz`

Recent Newton Institute Preprints

NI97038-RAG **M Geck and G Malle**
On special pieces in the unipotent variety

NI97039-NNM **SP Luttrell**
A unified theory of density models and auto-encoders
DERA report DERA/CIS/CIS5/651/FUN/STIT/5-4 31 October 1997

NI97040-NNM **CKI Williams and D Barber**
Bayesian classification with Gaussian processes

NI97041-NNM **TS Richardson**
Chain graphs and symmetric associations
Learning in Graphical Models, MIT Press Jan 98 M Jordan (ed.)

NI97042-NNM **A Roverato and J Whittaker**
An importance sampler for graphical Gaussian model inference

NI97043-DQC **MR Haggerty, JB Delos, N Spellmeyer et al**
Extracting classical trajectories from atomic spectra

NI97044-DQC **S Zelditch**
Level spacings for quantum maps in genus zero

NI97045-DQC **U Smilansky**
Semiclassical quantization of maps and spectral correlations

NI97046-DQC **IY Goldscheid and BA Khoruzhenko**
Distribution of Eigenvalues in non-Hermitian Anderson models
Phys. Rev. Lett. 80 (1998) No.13, 2897-2900

NI97047-DQC **G Casati, G Maspero and DL Shepelyansky**
Quantum fractal Eigenstates

NI98001-STA **N Linden and S Popescu**
Non-local properties of multi-particle density matrices

NI98002-AMG **J-L Colliot-Thélène**
Un principe local-global pour les zéro-cycles sur les surfaces fibrés en coniques au-dessus d'une courbe de genre quelconque

NI98003-AMG **RGE Pinch and HPF Swinnerton-Dyer**
Arithmetic of diagonal quartic surfaces, II

NI98004-AMG **DR Heath-Brown**
The solubility of diagonal cubic diophantine equations

NI98005-AMG **B Poonen and M Stoll**
The Cassels-Tate pairing on polarized Abelian varieties

NI98006-AMG **R Parimala and V Suresh**
Isotropy of quadratic forms over function fields of curves over p-adic fields

NI98007-AMG **E Peyre**
Application of motivic complexes to negligible classes

NI98008-AMG **E Peyre**
Torseurs universels et méthode du cercle

NI98009-RAG **JA Green**
Discrete series characters for $GL(n, q)$

NI98010-DQC **K Zyczkowski, P Horodecki, A Sanpera et al**
On the volume of the set of mixed entangled states

NI98011-DQC **K Zyczkowski and W Slomczyński**
Monge distance between quantum states

NI98012-DAD **JA Sellwood, RW Nelson and S Tremaine**
Resonant thickening of disks by small satellite galaxies

NI98013-DAD **GI Ogilvie and SH Lubow**
The effect of an isothermal atmosphere on the propagation of three-dimensional waves in a thermally stratified accretion disk

NI98014-DAD **JA Sellwood and SA Balbus**
Differential rotation and turbulence in extended H I disks

NI98015-DAD **AM Fridman and OV Khoruzhii**
On nonlinear dynamics of 3D astrophysical disks

NI98016-DQC **RE Prange, R Narevich and O Zaitsev**
Quasiclassical surface of section perturbation theory

NI98017-DQC **A Sedrakyan**
Edge excitations of an incompressible fermionic liquid in a disorder magnetic field

NI98018-DAD **CF Gammie**
Accretion disk turbulence

NI98019-DAD **R Popham and CF Gammie**
Advection dominated accretion flows in the Kerr metric: II. Steady state global solutions

NI98020-DAD **CF Gammie, R Narayan and R Blandford**
What is the accretion rate in NGC 4258?

NI98021-DAD **CF Gammie**
Photon bubbles in accretion disks

NI98022-DAD **EC Ostriker, CF Gammie and JM Stone**
Kinetic and structural evolution of self-gravitating, magnetized clouds: 2.5-dimensional simulations of decaying turbulence

NI98023-DAD **JA Sellwood and EM Moore**
On the formation of disk galaxies and massive central objects

NI98024-STA **VA Vladimirov, HK Moffatt and KI Ilin**
On general transformations & variational principles for the magnetohydrodynamics of ideal fluids. Part IV. Generalized isovorticity principle for three-dimensional flows

NI98025-BFG **M Steel**
Sufficient conditions for two tree reconstruction techniques to succeed on sufficiently long sequences

NI98026-BFG **M Steel**
The emergence of a self-catalysing structure in abstract origin-of-life models

NI98027-RAG **A Marcus**
Derived equivalences and Dade's invariant conjecture

NI98028-DAD **J Goodman and NW Evans**
Stability of power-law disks

NI98029-DAD **C Terquem**
The response of accretion disks to bending waves: angular momentum transport and resonances
Astrophysical Journal, 509:000-000, 1998 December 20

NI98030-NSP **P Flandrin**
Inequalities in Mellin-Fourier signal analysis

NI98031-NSP **MB Kennel and AI Mees**
Testing for general dynamical stationarity with a symbolic data compression technique

NI98032-BFG **G McGuire**
A Bayesian model for detecting past recombination events in multiple alignments

NI98033-NSP **M Paluš and D Novotná**
Sunspot cycle: a driven nonlinear oscillator

NI98034-NSP **RG Baraniuk, P Flandrin, AJEM Janssen et al**
Measuring time-frequency information content using the Rényi entropies

NI98035-BFG **S Böcker, AWM Dress and MA Steel**
Patching up X-trees

Information about Newton Institute Preprints is also available at
`http://www.newton.cam.ac.uk/preprints.html`