DIMACS Technical Report 96-43
September 1996

Local quartet splits of a binary tree infer all quartet splits via one dyadic inference rule[1]

by

Péter L. Erdős
Mathematical Institute of the Hungarian Academy of Sciences
Budapest P.O.Box 127, Hungary-1364
E-mail: `elp@math-inst.hu`

Michael A. Steel
Biomathematics Research Centre
University of Canterbury
Christchurch, New Zealand
E-mail: `m.steel@math.canterbury.ac.nz`

László A. Székely
Department of Mathematics
University of South Carolina
Columbia SC 29208, USA
E-mail: `laszlo@math.sc.edu`
Department of Computer Science
Eötvös University
Budapest, Hungary-1088
E-mail: `szekely@cs.elte.hu`

Tandy J. Warnow
Department of Computer and Information Science
University of Pennsylvania
Philadelphia PA 19104-6389, USA
E-mail: `tandy@central.cis.upenn.edu`

DIMACS is a partnership of Rutgers University, Princeton University, AT&T Research, Bellcore, and Bell Laboratories.
DIMACS is an NSF Science and Technology Center, funded under contract STC–91–19999; and also receives support from the New Jersey Commission on Science and Technology.

Abstract

A significant problem in phylogeny is to reconstruct a semilabelled binary tree from few valid quartet splits of it. It is well-known that every semilabelled binary tree is determined by its set of all valid quartet splits. Here we strengthen this result by showing that its local (i.e. small diameter) quartet splits infer by a dyadic inference rule all valid quartet splits, and hence determine the tree. The results of the paper also present a polynomial time algorithm to recover the tree.

Keywords.semilabelled binary trees, subtrees, phylogeny, quartets.

[1]**Acknowledgment**. This research started when the authors enjoyed the hospitality of DIMACS during the Special Year for Mathematical Support to Molecular Biology. The second author gratefully acknowledges the New Zealand Ministry of Research, Science and Technology (MORST) for support to visit Budapest under ISAC Programme grant 94/22. Research of the first and third authors was supported in part by the Hungarian National Science Fund contract T 016 358 and by the European Communities (Cooperation in Science and Technology with Central and Eastern European Countries) contract ERBCIPACT 930 113.The fourth author was supported in part by a Young Investigator Award from the National Science Foundation, CCR-9457800, and by grant SBR-9512092 from the Linguistics program at NSF, and by generous financial support from Paul Angello.

1 Introduction

We first provide a summary of notations used throughout this paper. The set $[n]$ denotes $\{1, 2, \ldots, n\}$ and for any set S, $\binom{S}{k}$ denotes the collection of subsets of S of size k.

A *semilabelled binary tree* T is a tree whose *leaves* (vertices of degree 1) are labelled by the number $1, 2, \ldots, n$, and whose remaining internal vertices are unlabelled and of degree three. Let $B(S)$ denote the set of semilabelled binary trees on leaf set S, and let $B(n) = B([n])$. For $T \in B(n)$ and $S \subseteq [n]$, there is a unique minimal subtree of T which contains all the elements of S. We call this tree the *subtree of* T *induced by* S, and denote it by $T_{|S}$. We obtain the *binary subtree of* T *induced by* S, denoted by $T^*_{|S}$, if we substitute edges for all maximal paths of $T_{|S}$ in which every internal vertex has degree two. Thus, $T^*_{|S} \in B(S)$. If $|S| = k$, then we refer to $T^*_{|S}$ as a *binary* k*-subtree.*

Given a semilabelled binary tree T with leaf set S, deleting an edge e of T disconnects T into two components, and thereby induces a bipartition of S consisting of the leaves of the two components. This bipartition is called a *split* of T induced by the edge e; the split is called *non-trivial* if both components contain at least 2 leaves. Buneman [3] showed that each semilabelled binary tree T is uniquely defined by its non-trivial splits.

For a semilabelled binary tree $T \in B(n)$, and for a quartet of leaves, $q = \{a, b, c, d\} \in \binom{[n]}{4}$, we say that $t_q = ab|cd$ is a *valid quartet split* of T, if $ab|cd$ is a split of $T^*_{|q}$. It is easy to see that

$$\text{if } ab|cd \text{ is a valid quartet split of } T, \text{ then so are } ba|cd \text{ and } cd|ab, \tag{1}$$

and we understand these three splits as identical. If (1) holds, then $ac|bd$ and $ad|bc$ are not valid quartet splits of T, and we say that any of them *contradicts* (1).

2 Tree reconstruction from an incomplete set of valid quartet splits

Let $Q(T) = \left\{t_q : q \in \binom{[n]}{4}\right\}$ denote the set of valid quartet splits of T. It is a classical result that $Q(T)$ determines T (Colonius and Schulze [4], also Bandelt and Dress [1]); indeed for each $i \in [n]$, $\{t_q : i \in q\}$ determines T, and T can be computed in polynomial time. For example, a simple algorithm for reconstructing T from $Q(T)$ is simply to build up T recursively from the tree with leaf set 1,2,3 by attaching (in any order) the remaining elements from $[n]$ as new leaves to the tree so far constructed. In this way, one uses $Q(T)$ to determine the unique edge of each partial tree to which the new leaf must be attached by bisecting the edge and making the recently created vertex adjacent to the new leaf.

An extension of Colonius and Schultze's result [4] is that for any $T \in B(n)$, a carefully chosen subset of $Q(T)$ of cardinality $n-3$ determines T (Steel [9]). Another extension is that an unknown semilabelled binary tree T with n leaves can be constructed by asking at most $O(n \log n)$ queries of the form: "what is t_q?" for a choice of q that depends on the answers to the queries so far asked (Pearl and Tarsi [7], Kannan, Lawler, and Warnow [6]).

It would be useful to tell from a set of quartet splits if they are valid quartet splits of any semilabelled binary tree. Unfortunately, this problem is NP-complete (Steel [9]). It also would be useful to know which subsets of $Q(T)$ determine T and which subsets would allow for a polynomial time procedure to reconstruct T. A natural step in this direction is to define *inference*: a set of quartet splits A infers a quartet split t_q if whenever $A \subseteq Q(T)$ for a semilabelled binary tree T, then $t_q \in Q(T)$ as well.

Setting a complete list of inference rules seems hopeless (Bryant and Steel [2]). However, having just some valid quartet splits of T, it is often possible to infer additional valid quartet splits of T, for example (see [1], [2] or [5]):

$$\begin{gathered}\text{if } ab|cd \text{ and } ac|de \text{ are valid quartet splits of } T,\\ \text{then so are } ab|ce,\ ab|de, \text{ and } bc|de;\end{gathered} \tag{2}$$

$$\text{if } ab|cd \text{ and } ab|ce \text{ are valid quartet splits of } T, \text{ then so is } ab|de; \tag{3}$$

$$\text{if } ab|cd,\ ab|ef \text{ and } ce|df \text{ are valid quartet splits of } T, \text{ then so is } ab|df. \tag{4}$$

In (2) and (3) we infer a valid quartet split from *two* other quartet splits. These rules are called *second order* or *dyadic* rules. In (4) we see a *third order* rule. These rules are due to Dekker [5]. A set of quartet splits A *dyadically infers* a quartet split t, if t can be derived from A by repeated applications of rules (1), (2) and (3).

It is worth mentioning that for every integer r there are inference rules of order r that cannot be inferred by repeated application of lower-order inference rules. (See Dekker [5] and Bryant and Steel [2].)

We say that a set of quartet splits A *semidyadically infers* a quartet split t, if t can be derived from A by repeated applications of rules (1) and (2). Quartet splits (semi)dyadically inferred by a set of quartet splits can be computed in polynomial time, and quartet splits (semi)dyadically inferred by a set of valid quartet splits of a tree are valid. We denote by $cl_2(A)$ the set of all quartet splits semidyadically inferred by the set A of quartet splits. We say that a set of quartet splits A *(semi)dyadically determine* T if they (semi)dyadically infer *all* valid quartet splits of T, i.e. $Q(T)$; in other words, Q fully determines the tree T.

3 Tree reconstruction from local quartets

For a semilabelled binary tree $T \in B(n)$, and a quartet of leaves, $q \in \binom{[n]}{4}$, let $L_T(q,e)$ denote the *length* (the number of edges) of the path P_e of $T_{|q}$ which turned into the edge e of $T^*_{|q}$. We will abuse the notation somewhat and let $L_T(q)$ denote the length of the longest path of $T_{|q}$ which is turned into an edge of $T^*_{|q}$, i.e. $L_T(q) = \max_e L_T(q,e)$. In [10] Steel *et al.* proved the following extension of the classical result of Colonius and Shultze:

Theorem 1 *For a semilabelled binary tree T on $[n]$ $(n \geq 4)$, let*

$$D(T) = \{q \in \binom{[n]}{4} : L_T(q) \leq 18 \log n\}.$$

Then $S(T) = \{t_q \text{ valid quartet split of T} : q \in D(T)\}$ *semidyadically determines T. In particular, T can be reconstructed from $S(T)$ in polynomial time.*

The interesting point in this proposition is that the local quartets fully determine the underlying binary tree. Based on this fact we built a reconstruction method for Cavender–Farris trees (see [10]). Our main goal is to strengthen Theorem 1. We need some more definitions.

The *depth* of an edge e in a semilabelled binary tree T is the number of edges on the path from e to the nearest leaf. The *depth* of T, $d(T)$, is the maximum depth of any edge e in T. For example, the depth of a complete semilabelled binary tree on n leaves is $\lceil log_2 n \rceil$. By contrast, a *caterpillar* on n leaves (the tree defined by a path $p = v_1, v_2, \ldots, v_{n-2}$ in which v_1 and v_{n-2} each has two

adjacent leaves and the neighbor of each remaining nodes on p is a leaf) has depth 1.
A *cherry* in a binary tree is a pair of leaves sharing a common neighbor, i.e. a pair of leaves at distance two in the tree.
The following theorem is the main result of this paper:

Theorem 2 *For a semilabelled binary tree T on $[n]$, let*

$$D(T) = \left\{ q \in \binom{[n]}{4} : L_T(q) \leq 2d(T) + 1 \text{ and } L_T(q, e) = 1, \right.$$
$$\left. \text{where } e \text{ is the internal edge of } T^*_{|q} \right\}.$$

Then $p(T) := \{T^*_{|q} : q \in D(T)\}$ semidyadically determines T. In particular, T can be reconstructed from $p(T)$ in polynomial time.

Proof. We use induction on n. The result holds for $n = 4$, so we suppose $n > 4$. We distinguish two cases:

(a) Every leaf of T is in a cherry, i.e. the leaves of T can be matched $(l_1, l_2), \ldots, (l_{n-1}, l_n)$, such that every pair (l_{2i-1}, l_{2i}) forms a cherry.

(b) There is a leaf l not covered by any cherry, i.e. l is separated from any other leaf by at least three edges.

In Case (a), let λ_i be the common neighbour of the leaves l_{2i-1} and l_{2i}. The deletion of all leaves of T results in a subtree T', whose leaves are just the λ_i's. Note that if E denotes the set of the T-leaves, $E = \{l_{2i} : i = 1, \ldots, n/2\}$, then T' is isomorphic to $T^*_{|E}$. It is clear that $d(T') = d(T) - 1$.

During the proof we assign to quartet splits of T' certain quartet splits of T, and call this operation *extension.* (The point of definition is to extend a valid quartet split into valid quartet splits.)

For a quartet of T'-leaves, $q' \in Q(T')$, where $t_{q'} = \lambda_a\lambda_b|\lambda_c\lambda_d$, we define the *standard general extension* of $t_{q'}$ by the quartet split $t_q = l_{2a}l_{2b}|l_{2c}l_{2d} \in Q(T)$. Now for any quartet of T'-leaves $q' \in D(T')$, we have

$$L_T(q) \leq L_{T'}(q') + 1 \leq [2(d-1) + 1] + 1 < 2d + 1,$$

and, if e is the internal edge of $T^*_{|q}$, then $L_T(q,e) = L_{T'}(q',e) = 1$. Thus the standard general extension t_q of the valid quartet split $t_{q'} \in p(T')$ belongs to $p(T)$. We define the *non-standard general extensions* of the valid quartet split $t_{q'}$ similarly, but we allow the substitution of one or more l_{2j} with l_{2j-1}. It is clear that every non-standard general extension belongs to $p(T)$ as well. Therefore if $p'(T)$ is the set of all general extensions of $t_{q'} \in p(T')$, then $p'(T)$ is a subset of $p(T)$.

For each leaf λ_j of T', let X_j, Y_j denote the leaf sets of the two other rooted subtrees of T' incident with the unique neighbour of l_j in T', v_j. Since $p(T')$ determines T', there is a quartet q'_j in $D(T')$ containing λ_j, λ_{x_j} and λ_{y_j} where $\lambda_{x_j} \in X_j$ and $\lambda_{y_j} \in Y_j$. We define the *standard special extension* of $t_{q'_j}$ by $t_{q_j} = l_{2j-1}l_{2j}|l_{2x_j}l_{2y_j}$. It is easy to see that

$$L_T(q_j) \leq L_{T'}(q'_j) + 1 \leq [2(d-1)+1] + 1 < 2d+1.$$

In addition, if e denotes the internal edge of $T^*_{|q_j}$ then $L_T(q_j, e) = 1$. Thus $t_{q_j} \in p(T)$ holds. We define the *non-standard special extensions* of the previous valid quartet split $t_{q'_j}$ similarly, but l_{2x_j} may be substituted by l_{2x_j-1} and/or l_{2y_j} may be substituted by l_{2y_j-1}. All the non-standard special extensions belong to $p(T)$ as well. Let $p^*(T)$ denote the set of all special extensions of $t_{q'_j}$ for every $j = 1, 2, \ldots, n/2$. Then $p^*(T) \subseteq p(T)$. Therefore

$$cl_2\left((p'(T) \cup p^*(T)\right) \subseteq cl_2(p(T)). \tag{5}$$

To finish the proof in Case (a), we now show that the left hand side of (5) equals $Q(T)$, so that $cl_2(p(T)) = Q(T)$, as claimed. For this purpose, let $t_q = l_a l_b | l_c l_d$ denote an arbitrary valid quartet split in T. Let $\lambda_a, \lambda_b, \lambda_c$ and λ_d be the neighbours of these T-leaves, respectively. If these four T'-leaves are pairwise distinct, then $t_{q'} = \lambda_a\lambda_b|\lambda_c\lambda_d$ is a valid split, and since (by hypothesis) $cl_2(p(T')) = Q(T')$, there is a sequence of inferences in T' yielding $t_{q'} \in Q(T')$ from $p(T')$, using rules (1) and (2). Repeating the same sequence of inferences with the general extensions of these quartet splits (and working in $Q(T)$), we infer t_q as well.

If $\lambda_a = \lambda_b = \lambda_j$ (where j is an integer between 1 and $n/2$), then for every $\lambda_c \in X_j$ the valid quartet split $l_{2j}l_{2j-1}|l_c l_{2y_j}$ belongs to the left hand side of (5). If the neighbour of l_c happens to be λ_{x_j}, then this is true by the definition of the special extension. So we may assume that $\lambda_c \neq \lambda_{x_j}$. By the preceding part of this case analysis, the valid quartet split $l_{2x_j}l_c|l_{2y_j}l_{2j}$ belongs to the left hand side of (5) (the neighbours of the four leaves are pairwise distinct). Using rule (1) for the special extension t_{q_j}, we infer $l_{2x_j}l_{2y_j}|l_{2j}l_{2j-1}$. The application of the third consequence in rule (2) infers $l_c l_{2y_j}|l_{2j}\lambda_{2j-1}$. Finally, a second application of rule (1) gives the required valid quartet split.

Similarly, the valid quartet split $l_{2j}\lambda_{2j-1}|l_c l_{2x_j}$ (again, $\lambda_c \neq \lambda_{x_j}$) belongs to the left hand side of (5), as it is shown by the application of the same second order inference rule for $l_{2x_j}l_c|l_{2y_j}l_{2j}$ and for the "opposite" of the valid quartet split t_{q_j}.

If we change the role of λ_{x_j} and λ_{y_j}, we obtain analogous inferences. (Namely, we can infer the quartet splits $l_{2j}l_{2j-1}|l_c l_{2y_j}$, where $\lambda_c \neq \lambda_{y_j}$.) Furthermore, since in the use of inference rules λ_{x_j} and λ_{y_j} do not play any special role, changing the role of l_{2x_j} (or l_{2y_j}) with an arbitrary leaf $l_d \in X_j$ (or $\in Y_j$, respectively), then we obtain analogous inferences. Therefore the only remaining case is $\lambda_c = \lambda_d$. Without loss of generality we may assume that $\lambda_c \in X_j$. Due to the previous argument, we have already inferred $l_{2j}l_{2j-1}|l_{2y_j}l_c$ and $l_{2j}l_{2y_j}|l_c l_d$. The application of the second consequence of the inference rule (2) infers $l_{2j}l_{2j-1}|l_c l_d$, which finishes the proof of Case (a).

In Case (b), we use the following notations: let l denote a leaf not covered by any cherry, let Λ be the neighbour of l, and let Δ and Γ be the other two neighbours of Λ (by the choice of l, these vertices exist and are of degree three). Let the two subtrees attached to Δ and disjoint from Λ be denoted A and B, and assume that the number of leaves in A is at most the number of leaves in B. Similarly, let the two subtrees attached to Γ be C and D, and assume that the leaves in C is at most the number of leaves in D.

Let the semilabelled binary trees T_Γ and T_Δ be defined in the following way: let T_Γ be the semilabelled binary tree generated by A, B, and the leaves l and Γ. The semilabelled binary tree T_Δ is generated by C, D, and the leaves l and Δ.

By induction, $cl_2(p(T_i)) = Q(T_i)$ hold for $i = \Delta, \Gamma$. Let the leaf $c \in C$ be the closest leaf to Γ from C, and let the leaf $a \in A$ be the closest leaf to Δ from A. Let R_c be obtained from $p(T_\Gamma)$ by omitting the quartet splits of the form $\Gamma x|yz$, where $x \neq l$, and substituting valid quartet splits $\Gamma l|yz \in p(T_\Gamma)$ with quartet splits $cl|yz \in Q(T)$. Similarly, let R_a be obtained from $p(T_\Delta)$ by omitting quartet splits $\Delta x|yz$ for $x \neq l$ and substituting quartet splits $\Delta l|yz$ with $al|yz \in Q(T)$. We define the *lift-up* of valid quartet splits from $p(T_\Delta) \cup p(T_\Gamma)$ considering them being quartets from $Q(T)$, substituting Γ

by c and Δ by a whenever necessary, i.e. whenever Γ or Δ belong to the quartets. Therefore, the quartets in R_a and R_c are lift-ups from some quartet splits of the subtrees T_Δ and T_Γ, respectively. We now show that $R_a \cup R_c$ is a subset of $p(T)$. For this purpose, let t_{q_a} be the lift-up of the valid quartet split $t_{q'_a} \in p(T_\Delta)$ and similarly, let t_{q_c} be the lift-up of the valid quartet split $t_{q'_c} \in p(T_\Gamma)$.

It is easy to see that $L_T(q_a) = L_{T_\Delta}(q'_a)$ except for some quartet splits of the form $t_{q_a} = la|\gamma\delta$. Similarly, $L_T(q_c) = L_{T_\Gamma}(q'_c)$ except for some quartet splits of the form $t_{q_c} = cl|\alpha\beta$. What remains is to show that $L_T(la|\gamma\delta) \leq 2d(T)+1$ and $L_T(cl|\alpha\beta) \leq 2d(T)+1$. Due to symmetry it is sufficient to prove the first claim only. We will use the notation $t_q = cl|\alpha\beta$.

For the pendant edge e of $T^*_{|q}$ incident with either α and β, we have $L_T(q,e) \leq 2d(T)+1$ since

$$L_T(q,e) = L_{T_\Gamma}(q,e) \leq L_{T_\Gamma}(q) \leq 2d(T_\Gamma)+1 \leq 2d(T)+1.$$

For the edges $e = (\Delta, \lambda)$ and $e = (\lambda, l)$ we have $L_T(q,e) = 1$. Thus, it remains to establish that

$$L_T(q,e) \leq 2d(T)+1 \tag{6}$$

for the edge e of T^*_q incident with c. If $|C| = 1$, then we have nothing to prove since $L_T(q,e) = 2 \leq 2d(T)+1$.

Now for $|C| > 1$ suppose on contrary to (6) that $L_T(q,e) > 2d(T)+1$. Then $d_T(\lambda, c) > 2d(T)+1$ (where $d_T(x,y)$ denotes the *distance* of x and y in the tree T, that is the length of the path from x to y). Since c is the closest leaf in C to λ, all leaves in C are at distance $> 2d(T)+1$ from λ. Let $e^* = (x,y)$ be an edge of C for which $d_T(\Gamma, x) = d-1$ and $d_T(\Gamma, y) = d$. By the definition of $d(T)$, the depth of e^* is at most $d(T)$, therefore there must be a leaf l^* of T at distance at most $d(T)$ from e^*. On the one hand l^* cannot belong to C since all leaves of C must be at distance $> d(T)+1$ from e^* (by the assumption $L_T(q,e) > 2d+1$). On the other hand $l^* \neq l$ since the distance $d_T(l,x) = d+1$. Finally, for every leaf $l^* \in D$, the distance $d_T(l^*, x) > d$ because the path from x to l^* uses at least two edges of D since D has at least two leaves. This contradiction proves that $L_T(q,e) \leq 2d(T)+1$ and therefore $R_c \subseteq p(T)$, and a similar argument shows that $R_a \subseteq p(T)$. Therefore we have

$$cl_2(R_a \cup R_c) \subseteq cl_2(p(T)). \tag{7}$$

To finish the proof in Case (b), we are going to show that the left hand side of (7) equals $Q(T)$, so that $cl_2(p(T)) = Q(T)$, as claimed. For this purpose, let b denote the B-leaf in T, which is closest to Δ. Similarly, let d denote the closest leaf to Γ from D in T. We note that the distance $d_T(b, \lambda) \leq 2d(T)+2$ because we can repeat the proof of formula (6) except that A can be of cardinality one. Therefore $d(b, \Delta) \leq 2d(T)+1$. A similar condition holds for the leaf $d \in D$ which is the closest one to Γ. From know on, the letters a, b, c and d always refer to these fixed leaves.

At first we show that the valid quartet split $cx|yz \in Q(T)$, which is the lift-up of the quartet split $\Gamma x|yz \in p(T_\Gamma)$, belongs to the LHS of (7). Because $\Gamma x|yz \in p(T_\Gamma)$, therefore if x belongs to A, then $y, z \in B$, and they are on different subtrees of the neighbour Φ of Δ. Furthermore, without loss of generality we may assume that b is on the same subtree of Φ as z.)

We know, that $d(x, \Delta) \leq 2d(T)+1$, since $\Gamma x|yz \in p(T_\Gamma)$. We just show that $d(b, \Delta) \leq 2d+1$. Therefore the valid quartet split $lc|xb$ belongs to R_c. Similarly, $lx|yb \in R_c$ also holds. Applying the first consequence of inference rule (2), we have $lc|xy \in cl_2(R_c)$. Putting this together with $lx|yz \in R_c$ (which follows from the fact that $d(\Gamma, \Delta) = d(l, \Delta)$) and applying again rule (2), consequence 3, we proved that $cx|yz \in cl_2(R_c)$. The symmetric claim $au|vw \in cl_2(R_a)$ holds for the lift-up of $\Delta u|vw \in p(T_\Delta)$. Thus, we have proved:

$$\text{the lift-up version of any element of } p(T_\Delta) \text{ or } p(T_\Gamma) \text{ belongs to } cl_2(R_a) \text{ or } cl_2(R_c). \tag{8}$$

Let α and β denote two leaves in $A \cup B$. Since $\alpha\beta|l\Gamma \in Q(T_\Gamma)$ therefore, due to (8), there is a "lifted up" inference sequence for $\alpha\beta|lc$ by semidyadic inference rules. For $\gamma, \delta \in C \cup D$ we have a similar result. Thus, we have proved:

$$\text{for leaves } \gamma, \delta \in C \cup D \;\; al|\gamma\delta \in cl_2(R_a), \tag{9}$$

$$\text{for leaves } \alpha, \beta \in A \cup B \;\; \alpha\beta|lc \in cl_2(R_c). \tag{10}$$

From now on, α, β, γ and δ always refer to leaves like above, but they are not fixed leaves.

Assume, that $a \neq \alpha$ (if this is not true, then exchange the names α and β). Similarly, we may assume, that $c \neq \delta$. Applying the choice $\beta = a$, for property (10), we have $a\alpha|lc \in cl_2(R_c)$. Similarly, for $\gamma = c$ and $\delta = d$ in (9) we have $al|cd \in cl_2(R_a)$. The application of the first consequence of rule (2) gives:

$$a\alpha|ld \in cl_2(R_a \cup R_c). \tag{11}$$

The substitution $\delta = d$ in (9) gives $al|\gamma d \in cl_2(R_a)$. This, together with (11), through the application of the third consequence of (2) gives:

$$\alpha l|\gamma d \in cl_2(R_a \cup R_c). \tag{12}$$

(10) together with (12) (where $\gamma = c$) gives (through rule (2), first consequence)

$$\alpha\beta|ld \in cl_2(R_a \cup R_c). \tag{13}$$

Applying the symmetry rule (1) for (12) and (13) and using again the semidyadical rule (2) with its third consequence and taking again its symmetric form, we have

$$\alpha\beta|l\gamma \in cl_2(R_a \cup R_c). \tag{14}$$

Since γ was not involved in the proof of (14) (except that $\gamma \in R_a \cup R_c$), the following symmetric claim can also be inferred through similar reasoning:

$$\alpha l|\gamma\delta \in cl_2(R_a \cup R_c). \tag{15}$$

Properties (14) and (15) together with our inductive hypothesis give:

$$\text{for any } q_t \in Q(T) \text{ such that } l \in q, q_t \in cl_2(R_a \cup R_c). \tag{16}$$

Furthermore (14) and (15) and the application of (2), first consequence, proves:

$$\alpha\beta|\gamma\delta \in cl_2(R_a \cup R_c). \tag{17}$$

Finally, let x, y and z be leaves of $A \cup B$. Let x be on a subtree of Δ, where y and z are not. By (14) (and with symmetry) we have $l\gamma|xy \in cl_2(R_a \cup R_c)$. Moreover, $lx|yz \in R_c$ due to our inductive hypothesis. These two together, through rule (2), third consequence, give $x\gamma|yz \in cl_2(R_a \cup R_c)$. By symmetry, we also know the analogous result with subtrees T_Δ and T_Γ exchanged; therefore we have proved:

$$\begin{aligned}&\text{if a quartet } q \in \binom{[n]}{4} \text{ contains three leaves from } T_\Delta, \text{ one leaf}\\ &\text{from } T_\Gamma, \text{ or vice versa, but } l \notin q, \text{ then } t_q \in cl_2(R_a \cup R_c).\end{aligned} \tag{18}$$

Now, for the quartet $q = \{s, u, v, w\}$ such that every leaf $\in A \cup B \cup \{l\}$, it is easy to see that $q \in Q(T_\Gamma)$; therefore $t_q \in cl_2(p(T_\Gamma))$, and the "lifted up" version of this proof ensures that $t_q \in cl_2(R_c)$. This fact (and its analogues for the other half-graph) together with properties (16), (17) and (18) finish the proof of Case (b), and we are done. $\square$

It is worth noting that Theorem 2 strengthens Theorem 1, as it is shown by the following result:

Lemma 3 *For any semilabelled binary tree T on $[n], d(T) \leq \log_2 n - 1$.*

Proof. Suppose edge e of T has maximal depth $d = d(T)$. Then there is a set V_e of at least 2^d vertices at distance $d-1$ from e, and none of these can be a leaf of T. For $v \in V_e$ let $S(v)$ be set of leaves of T that become separated from e upon deletion of v. Since $S(v) \cap S(v') = \emptyset$ for $v \neq v'$, and $|S(v)| \geq 2$, we have $n = |\cup_{v \in V_e} S(v)| \geq 2|V_e| \geq 2 \times 2^d = 2^{d+1}$, as claimed. $\square$

References

[1] Bandelt, H.-J. - Dress, A.: Reconstructing the shape of a tree from observed dissimilarity data, Advances in Applied Mathematics, Vol. 7, 1986. pp. 309–343.

[2] Bryant, D. J. - Steel, M. A.: Extension operations on sets of leaf-labelled trees, Advances in Applied Mathematics, Vol. 16, 1995. pp. 425–453.

[3] Buneman, P.: The recovery of trees from measures of dissimilarity. In: Hodson, F. R., Kendall, D. G., Tautu, P. (Eds): Mathematics in the Archaeological and Historical Sciences, Edinburgh University Press, Edinburgh, 1971, pp. 387–395.

[4] Colonius, H. - Schultze, H. H.: Tree structure for proximity data, Brit. J. Math. Stat. Psychol., Vol. 34, 1981. pp. 167–180 .

[5] Dekker, M. C. H.: Reconstruction methods for derivation trees, Master's Thesis, Vrije Universiteit, Amsterdam, 1986.

[6] Kannan, S., Lawler, E. and T. Warnow: Determining the Evolutionary Tree Using Experiments. Journal of Algorithms **21**, 1996, pp. 26-50.

[7] Pearl, J. - Tarsi, M.: Structuring causal trees, J. Complexity, Vol. 2, 1986. pp. 60–77.

[8] Philippe, H. - Douzery, E.: The pitfalls of molecular phylogeny based on four species, as illustrated by the cetacea/artiodactyla relationships, J. Mammal. Evol., Vol 2, 1994. No. 2, pp. 133–152.

[9] Steel, M. A.: The complexity of reconstructing trees from qualitative characters and subtrees, J. Classification, Vol. 9, 1992. pp. 91–116.

[10] Steel, M. A. - Székely, L. A. - Erdős, P. L.: The number of nucleotide sites needed to accurately reconstruct large evolutionary trees, submitted to Random Structures and Algorithms 1996.