

# Counting bichromatic evolutionary trees

Péter L. Erdős\*

*Hungarian Academy of Sciences, Budapest, Hungary; and Institute für Ökonometrie und Operations Research, Rheinische Friedrich-Wilhelms Universität, Bonn, Germany*

L.A. Székely\*

*Department of Computer Science, Eötvös L. University, Budapest, Hungary; and Institute für Ökonometrie und Operations Research, Rheinische Friedrich-Wilhelms Universität, Bonn, Germany*

Received 13 December 1990

Revised 17 September 1993

## *Abstract*

We give a short and transparent bijective proof of the bichromatic binary tree theorem of Carter, Hendy, Penny, Székely and Wormald on the number of bichromatic evolutionary trees. The proof simplifies M.A. Steel's proof.

Evolutionary trees are extensively studied structures in biostatistics. (These are leaf-coloured binary trees. For details see, e.g., Felsenstein [4], Steel [10] or Carter et al. [1].)

In general, the mathematical problems arising here are hard (see [6]). One of the very beginning steps is to count evolutionary trees. For two colours it was done by Carter et al. [1]. Their work is based on the generating function method and on a lengthy, computer-assisted application of the multivariate Lagrange inversion. Recently Steel [10] gave a bijective proof for the bichromatic binary tree theorem pioneering the application of Menger's theorem in enumerative theory. Unfortunately, his solution is rather involved. The goal of the present paper is to give a simple and transparent bijective proof for the bichromatic binary tree theorem. Our work was inspired by Steel's work, actually we simplify some crucial steps in his proof and the rest of the proof is identical to his one. The proof uses more graph theory than proofs in enumerative theory usually do.

*Correspondence to:* Professor P.L. Erdős, Hortensiastraat 3, 1338 ZP Almere, Netherlands.

\* Research supported in part by Alexander v. Humboldt-Stiftung.

### Preliminaries and the bichromatic binary tree theorem

In this section we introduce some definitions and notations which may not be common, and state the theorem of Carter et al.

In a tree, a vertex of degree 1 is a *leaf*. A tree is *binary* if every nonleaf vertex of the tree has degree 3. A tree is *rooted binary* if it has exactly one vertex of degree 2 and the other nonleaf vertices have degree 3. The vertex of degree 2 is the *root* of the tree. By definition, a singleton vertex is a binary tree and also a rooted binary tree. In this degenerate tree above, the singleton vertex is a leaf, and in the rooted case it is a root as well.

A (rooted) binary tree with labelled leaves is termed a (rooted) *semilabelled* tree. Hereafter we identify the set of leaves and the set of labels and denote both by  $L$ . A *semilabelled rooted binary forest* is a forest containing rooted semilabelled binary trees, where the label sets of distinct trees are pairwise disjoint. The following facts are well known. (The details can be found in several books and papers, e.g., see [1, 2, 3].)

**Lemma 0.** (a) Any binary tree  $T$  with  $n$  leaves has  $2n - 2$  vertices and  $2n - 3$  edges.

(b) Any rooted binary tree  $T$  with  $n$  leaves has  $N(T) = 2n - 1$  vertices and  $2n - 2$  edges.

(c) The total number of semilabelled binary trees with  $n$  leaves is

$$b(n) = (2n - 5)!!.$$

(d) The total number of semilabelled rooted binary forests with  $n$  leaves and  $k$  trees is

$$N(n, k) = \binom{2n - k - 1}{k - 1} (2n - 2k - 1)!!.$$

Let  $T$  be a semilabelled binary tree. We term a map  $\chi: L \rightarrow \{A, B\}$  a *leaf-colouration*. A colouration  $\bar{\chi}: V(T) \rightarrow \{A, B\}$  is an *extension* of the leaf-colouration  $\chi$  if the two maps are identical on the set  $L$ . The *changing number* of the colouration  $\bar{\chi}$  is the number of edges whose endvertices have different colours according to  $\bar{\chi}$ . An extension is a *minimal colouration* according to the leaf-colouration  $\chi$  if its changing number is minimal among the changing numbers of all extensions of  $\chi$ . We refer to the minimal changing number as the *length* of the tree  $T$  (according to  $\chi$ ). An efficient algorithm for calculating the length of a tree and finding a minimal colouration, due to [5], is established in [7].

Let us fix now a 2-colouration  $\chi$  of the set  $L$  and denote by  $L_A$  and  $L_B$  the nonempty colour classes ( $L_A \cup L_B = L$ ). Set  $a = |L_A| > 0$  and  $b = |L_B| > 0$ . The question is: What is the number of (unrooted) semilabelled binary trees whose leaf set is  $L$  and length is exactly  $k$  (according to  $\chi$ )? Let  $f_k(a, b)$  denote the number in question. Carter, Hendy, Penny, Székely and Wormald proved [1], that

**Theorem.**

$$f_k(a, b) = (k - 1)!(2n - 3k)N(a, k)N(b, k) \frac{b(n)}{b(n - k + 2)}$$

where  $a + b = n$ ,  $a > 0$ ,  $b > 0$ .

In the rest of our paper we prove this theorem. The proof is based on a method developed by Steel [10].

**Steel's decomposition**

In this section we describe the structure of the bichromatic semilabelled trees of length  $k$ .

Let  $\chi$  be a 2-colouration of the set  $L$ . The length of the tree  $T$  is equal to  $k$  iff the deletion of  $k$  well-chosen edges decomposes  $T$  into subtrees with one colour being present in each, but the deletion of less than  $k$  edges cannot do it. Due to Menger's theorem [8], this means that the maximum number of edge-disjoint paths from  $L_A$  to  $L_B$  is  $k$ . Since  $T$  is binary, two edge-disjoint paths between leaves are also vertex-disjoint. Therefore there exist  $k$  (but no more than  $k$ ) vertex-disjoint paths from  $L_A$  to  $L_B$ . A second application of Menger's theorem guarantees the existence of a  $k$ -element vertex set which covers every  $L_A \rightarrow L_B$  path. Any such set is called a *minimal covering system*. It is easy to see that incidence defines a one-to-one correspondence between any minimal covering system and any  $k$  vertex-disjoint paths from  $L_A$  to  $L_B$ .

The following lemma helps to understand the minimal covering systems.

**Lemma 1.** *Suppose  $M$  is a minimal covering system. Set*

$$\mu(T) = \left\{ \bigcap_{\pi \in \Pi} \{P : m \in P \in \pi\} : m \in M \right\},$$

where  $\Pi$  is the family of sets of  $k$  edge-disjoint paths connecting  $L_A$  and  $L_B$ . Then

(a)  $\mu(T)$  is independent of the choice of  $M$ , the members of  $\mu(T)$  are vertex-disjoint paths in  $T$ .

(b) Assume  $v_0 \in \bigcup \mu(T)$ . Define the set  $M_0$  by picking the vertex closest to  $v_0$  from every path of  $\mu(T)$ . Then  $M_0$  is a minimal covering system, hence, any point of any member of  $\mu(T)$  belongs to some minimal covering system.

(c)  $v_0 \in M_0$  and  $M_0$  is unique as long as  $v_0$  is given.

**Proof.** Notice the following consequence of Menger's theorem: for minimal covering systems  $M'$ ,  $M''$ , a set of  $k$  edge-disjoint paths from  $L_A$  to  $L_B$  defines a matching between  $M'$  and  $M''$  by the relation "being on the same path".

To prove (a), we have to see that any set of  $k$  edge-disjoint paths from  $L_A$  to  $L_B$  define the *same* matching.

On the contrary, assume that two path systems define two different matchings of  $M', M''$ . The two matchings define a graph  $G$  on the vertex set  $M' \triangle M''$  with edges taken from the matchings.  $G$  contains a cycle of length longer than 2. Recall that the edges of this cycle can be represented by subpaths of the two path systems. Since  $T$  is cycle-free, these subpaths altogether cover twice a path  $P$  of  $T$ . This contradicts to the disjointness of the path systems.

We have proved that  $\mu(T)$  is independent of the choice of  $M$ . Finally, note that a nonempty intersection of paths in a tree is a path itself.

(We do not need this explicitly, but you may observe that any system of representatives of  $\mu(T)$  covers every path of every  $\pi$  and clearly every minimal covering system  $M$  occurs as such a system of representatives—just define  $\mu(T)$  by this  $M$ ! Unfortunately, not every system of representatives is a minimal covering system. This makes life more difficult.)

To prove (b) notice that every  $L_A \rightarrow L_B$  path intersects at least one member of  $\mu(T)$ . If a path  $P'$  from  $L_A$  to  $L_B$  intersects two members of  $\mu(T)$ , then one member separates the other member from  $v_0$ . Now by definition, the first intersection of  $P'$  with the other member belongs to  $M_0$  and covers the path  $P'$ . Hence we may assume that  $P'$  intersects a unique  $P \in \mu(T)$ . We claim that  $P'$  contains the whole  $P$ . Hence  $P \cap M_0 \in P'$ .

In order to prove the latter claim, we consider two cases. Either  $P' \in \pi$  for some  $\pi \in \Pi$ , or not. In the first case,  $P'$  occurs in the intersection that defines  $P$ , hence  $P \subset P'$ . In the second case,  $P'$  intersects two paths from every  $\pi \in \Pi$ , otherwise we may exchange  $P'$  with the only path  $\pi$  intersected by  $P'$  to get a  $P' \in \pi' \in \Pi$ . It is easy to conclude that there exist  $P_1, P_2 \in \mu(T)$ , such that  $P'$  intersects two paths from every  $\pi$ , which contain  $P_1, P_2$ , respectively. Finally,  $P'$  intersects both  $P_1, P_2$ , a contradiction.  $\square$

Take  $M_0$  from Lemma 1. Define the semilabelled forest  $\mathcal{F}' = \{T'_v : v \in M_0\}$  of pairwise disjoint subtrees of  $T$  as follows: For every vertex  $u$  of the tree  $T$  the unique path  $u \rightarrow v_0$  contains at least one element of  $M_0$ . Let  $u$  belong to  $T'_v$  iff  $v$  is the nearest vertex to  $u$  among these vertices. Finally, let the tree  $T_v$  ( $v \in M_0$ ) be the subtree of  $T'_v$  which is spanned by those leaves of  $T'_v$  which also belong to  $L$ .

**Lemma 2.** *The semilabelled forest  $\mathcal{F} = \{T_v : v \in M_0\}$  satisfies the following conditions:*

- (a) *The leaf set of  $\mathcal{F}$  coincides with  $L$ .*
- (b) *If  $v \in M_0$  then  $v \in T_v$  and the path  $v_0 \rightarrow T_v$  reaches the tree  $T_v$  at the vertex  $v$ .*
- (c) *The degree of the vertex  $v \in (M_0 \setminus \{v_0\})$  in the tree  $T_v$  is equal to 2.*
- (d) *Every tree  $T_v$  is bichromatic (that is it has two colours) according to the leaf-colouration  $\chi$ . Removing the vertex  $v$  from the tree  $T_v$ , the remaining two (or if  $v = v_0$ , then two or three) subtrees are monochromatic according to  $\chi$ .*

**Proof.** Parts (a) and (b) directly follow from the definition of  $\mathcal{F}$ . Part (c) follows from (b). Part (d) contains the essence of this lemma. The set  $M_0$  is a covering system, therefore the subtrees derived by removing the vertex  $v$  must be monochromatic (i.e., they cannot contain leaves of different colours). On the other hand, these subtrees must show two different colours, otherwise any path  $P: L_A \rightarrow L_B$  covered solely by vertex  $v$  out of the elements of  $M_0$  must be closer to the vertex  $v_0$  than the subtree  $T_v$  itself. Therefore the neighbour  $v'$  of vertex  $v$  in the direction of  $v_0$  also covers  $P$ . So the choice of  $v$  from  $M_0$  was wrong,  $v'$  must have been chosen.  $\square$

In the next step we derive a new semilabelled forest from  $\mathcal{F}$ : for every vertex  $v \in M_0$  we contract the vertices of degree 2 in the tree  $T_v$ , except the vertex  $v$  itself. Finally if the degree of  $v_0$  in the tree  $T_{v_0}$  is equal to 3 then we add a root into this tree which covers every  $L_A \rightarrow L_B$  path in  $T_{v_0}$ . Denote  $\mathcal{F}^S$  the derived semilabelled forest consisting of  $k$  rooted binary trees. This forest is the *Steel decomposition* of the tree  $T$  (with respect to the leaf-colouration  $\chi$  and the vertex  $v_0$ ). We call the tree derived from  $T_{v_0}$  the *kernel* of that decomposition.

**Lemma 3.** *For any given  $v_0$ , the Steel decomposition of the tree  $T$  is unique. Moreover, if  $v_0, v'_0 \in P \in \mu(T)$ , then they define the same Steel decomposition.*

**Proof.** By definition, the forest  $\mathcal{F}^S$  is determined by the minimal covering system  $M_0$ . We have already proved the uniqueness of  $M_0$ . Changing  $v_0$  for  $v'_0$ , we end up with  $M'_0 = M_0 - \{v_0\} \cup \{v'_0\}$ .  $\square$

Let  $\mathcal{F} = \{T_0; T_1, \dots, T_{k-1}\}$  be an arbitrary semilabelled rooted binary forest with leaf set  $L = L_A \cup L_B$ . Let  $e_i$  ( $i = 1, \dots, k-1$ ) denote the number of edges in the tree  $T_i$ , and let  $e_0$  be (edge number of  $T_0$ )  $- 1$ . An *extension* of the forest  $\mathcal{F}$  is a semilabelled binary tree whose Steel decomposition is the forest  $\mathcal{F}$  with kernel  $T_0$ .

The first question is: How can we find extensions of the forest  $\mathcal{F}$ ? Let  $B$  be a binary tree and let  $B_1$  be a rooted binary tree. The *insertion* of  $B_1$  into  $B$  is the following operation: subdivide by a new vertex one of the edges of  $B$  and connect the new vertex to the root of  $B_1$  by a new edge.

**Lemma 4.** *Let  $\mathcal{F} = \{T_0; T_1, \dots, T_{k-1}\}$  be a semilabelled rooted binary forest. Let  $\hat{T}_0$  be the binary tree derived from  $T_0$  by deleting the root and joining its neighbours. Insert recursively the trees  $T_1, T_2, \dots, T_{k-1}$  into the actual tree, where the initial actual tree is  $\hat{T}_0$ , and later on the actual tree is the result of the last insertion. Let  $T$  be the semilabelled binary tree which is the last actual tree. Then there is a vertex  $v_0$  in  $T$ , such that the Steel decomposition of the tree  $T$  according to  $v_0$  coincides with the forest  $\mathcal{F}$ .*

**Proof.** Let  $v_0$  be any neighbour of the root of  $T_0$  in  $\hat{T}_0$ . This vertex covers every path  $L_A \rightarrow L_B$  in the tree  $\hat{T}_0$ . The vertex  $v_0$  together with the original roots of  $T_1, \dots, T_{k-1}$  form a minimal covering system in the tree  $T$ . It is easy to see that this system also

satisfies the minimum distance condition with respect to the vertex  $v_0$ . Therefore the Steel decomposition of  $T$  with respect to  $v_0$  is  $\mathcal{F}$ .  $\square$

**Lemma 5.** *Let  $\text{Ext}(T_0; T_1, \dots, T_{k-1})$  denote the set of extensions of the forest  $\mathcal{F}$ . We have*

$$|\text{Ext}(T_0; T_1, \dots, T_{k-1})| = e_0 \frac{b(n)}{b(n-k+2)}.$$

**Proof.** We apply mathematical induction on  $k$ . If we use the abbreviation  $T(e_0, k-1) = |\text{Ext}(T_0; T_1, \dots, T_{k-1})|$ , then we have to prove, that:

- (a)  $T(e_0, 1) = 1$ ;
- (b)  $T(e_0, k-1) = (2n - 2k + 1) T(e_0, k-2)$ .

Case (a) is trivial, because the unique extension of the forest  $\{T_0\}$  is the tree  $\hat{T}_0$  itself.

(b) Suppose  $T$  is an extension of  $\mathcal{F}$ . Define a directed tree  $T^c$  as follows: The vertices of  $T^c$  are  $\hat{T}_0, T_1, \dots, T_{k-1}$ . An arbitrary ordered pair  $(T_i, T_j)$  (or  $(\hat{T}_0, T_j)$ ) is an arc if the last root of the trees  $\hat{T}_0, T_1, \dots, T_{k-1}$  before  $v_j$  on the path  $v_0 \rightarrow v_j$  in the tree  $T$  is the vertex  $v_i$ . Every vertex of  $T^c$  (except the vertex  $\hat{T}_0$ ) has in degree exactly one, and the corresponding arc tells us where the tree  $T_j$  is inserted in this extension. Examine the insertion of the tree  $T_1$ . We distinguish two disjoint subcases:

(b1) There is an  $i \in \{2, \dots, k-1\}$  for which  $(T_i, T_1)$  is an arc in  $T^c$ . Then there are  $e_i$  different insertions of  $T_1$  into  $T_i$ . After any of these insertions we have a forest of  $k-1$  trees (one of them is the kernel  $T_0$ ). By the inductive hypothesis any forest built has  $T(e_0, k-2)$  different extensions. So the total number of extensions of these types is

$$(e_2 + e_3 + \dots + e_{k-1}) T(e_0, k-2).$$

(b2) The ordered pair  $(T_0, T_1)$  is an arc in  $T^c$ . In this case the tree  $T_1$  is inserted into the tree  $\hat{T}_0$ . We have  $e_0$  different ways to realize this insertion. After the insertion we have a forest of  $k-1$  trees, where the kernel has  $e_0 + e_1 + 2$  edges. Therefore any of the forests built can be extended in

$$(e_0 + e_1 + 2) \frac{b(n)}{b(n - [k-1] + 2)}$$

ways. Therefore the total number of extensions of this type is

$$(e_0 + e_1 + 2) T(e_0, k-2).$$

Adding up the numbers from the subcases, the total number of the extensions is

$$\begin{aligned} T(e_0, k-1) &= (e_0 + e_1 + \dots + e_{k-1} + 2) T(e_0, k-2) \\ &= (2n - 2k + 1) T(e_0, k-2). \end{aligned}$$

(In the last step we used Lemma 0(a) and (b).)  $\square$

### The proof of the Theorem

Let  $\chi$  be an arbitrary but fixed 2-colouration of the set  $L$  with colour classes  $L_A$  and  $L_B$ , where  $|L_A| = a$  and  $|L_B| = b$ . Denote  $\mathcal{F}_k(a, b)$  the set of semilabelled binary trees of length  $k$  (according to  $\chi$ ) with leaf set  $L$ . Let

$$\mathcal{F}_k^*(a, b) = \{(T, P): T \in \mathcal{F}_k(a, b), P \in \mu(T)\}.$$

Let  $\mathcal{H}(a, b, k)$  denote the collection of semilabelled rooted binary forests of  $k$  trees with leaf set  $L$ , such that every tree has two oppositely coloured, monochromatic subtrees if its root is removed. Finally let

$$\mathcal{G}_k(a, b) = \{(\mathcal{F}, T_0, T): \mathcal{F} \in \mathcal{H}(a, b, k), T_0 \in \mathcal{F}, T \in \text{Ext}(T_0; \mathcal{F} \setminus \{T_0\})\}.$$

**Lemma 6.** *There exists a bijection  $\psi$  from  $\mathcal{F}_k^*(a, b)$  onto  $\mathcal{G}_k(a, b)$ .*

**Proof.** For  $(T, P) \in \mathcal{F}_k^*(a, b)$  let  $\psi(T, P) = (\mathcal{F}, T_0, T)$  where  $\mathcal{F}$  is the Steel decomposition of  $T$  according to vertex  $v_0 \in P$  and  $T_0$  is the kernel of the decomposition. Since the Steel decomposition is unique and  $P$  is connected, the map  $\psi$  is well defined. If  $\psi(T, P) = \psi(T', P')$  then  $T = T'$  by the definition of  $\psi$ . The kernels of the decompositions are identical. Therefore  $P = P'$ , since both of them are an element of  $\mu(T)$  which is in the kernel. So  $\psi$  is injective. Finally, Lemma 4 proves that  $\psi$  is onto.  $\square$

**Lemma 7.**

$$f_k(a, b) = (k-1)!(2n-3k)N(a, k)N(b, k) \frac{b(n)}{b(n-k+2)}.$$

**Proof.** We know that  $|\mathcal{F}_k(a, b)| = f_k(a, b)$ . Therefore  $|\mathcal{F}_k^*(a, b)| = kf_k(a, b)$ . Now we have

$$\begin{aligned} |\mathcal{G}_k(a, b)| &= \sum_{\mathcal{F} \in \mathcal{H}(a, b, k)} \sum_{T_0 \in \mathcal{F}} |\text{Ext}(T_0; \mathcal{F} \setminus \{T_0\})| \\ &= \sum_{\mathcal{F} \in \mathcal{H}(a, b, k)} \sum_{T_0 \in \mathcal{F}} e_0 \frac{b(n)}{b(n-k+2)} \\ &= (2n-3k)|\mathcal{H}(a, b, k)| \frac{b(n)}{b(n-k+2)}. \end{aligned}$$

Furthermore, we know that  $|\mathcal{H}(a, b, k)| = k!N(a, k)N(b, k)$ . (The forests of  $\mathcal{H}(a, b, k)$  can be built as follows: take a semilabelled forest of  $k$  rooted binary trees with leaf set  $L_A$  and a semilabelled forest of  $k$  rooted binary trees with leaf set  $L_B$ , match them up and make bichromatic rooted binary trees from the pairs.) Now Lemma 6 finishes the proof.  $\square$

**References**

- [1] M. Carter, M. Hendy, D. Penny, L.A. Székely and N.C. Wormald, On the distribution of lengths of evolutionary trees, *SIAM J. Discrete Math.* 3 (1990) 38–47.
- [2] P.L. Erdős, A new bijection on rooted forests, *Discrete Math.* 111 (1993) 179–188.
- [3] P.L. Erdős and L.A. Székely, Application of antilexicographic order I, An enumerative theory of trees, *Adv. Appl. Math.* 10 (1989) 488–496.
- [4] J. Felsenstein, Phylogenies from molecular sequences: Inference and reliability, *Ann. Rev. Genetics* 22 (1988) 521–565.
- [5] W.M. Fitch, Towards defining the course of evolution: Minimum change for specific tree topology, *Systems Zool.* 20 (1971) 406–416.
- [6] R.L. Graham and L.R. Foulds, Unlikelihood that minimal phylogenies for a realistic biological study can be constructed in reasonable computational time, *Math. Biosci.* 60 (1982) 133–142.
- [7] J.A. Hartigan, Minimum mutation fits to a given tree, *Biometrics* 29 (1973) 53–65.
- [8] K. Menger, Zur allgemeinen Kurventheorie, *Fund. Math.* 10 (1926) 96–115.
- [9] J.W. Moon, Counting Labelled Trees, *Canadian Mathematical Congress*, Montreal, Que. (1970).
- [10] M.A. Steel, Distributions on bicoloured binary trees arising from the principle of parsimony, *Discrete Appl. Math.* 41 (1993) 245–261.