ADVANCES IN APPLIED MATHEMATICS 13, 375–389 (1992)

# Evolutionary Trees: An Integer Multicommodity Max-Flow – Min-Cut Theorem

PÉTER L. ERDÖS

*Department of Applied Mathematics, University of Twente, Enschede, The Netherlands and Hungarian Academy of Sciences, H-1364 Budapest, P.O.B 127, Hungary*

AND

LÁSZLÓ A. SZÉKELY

*Department of Computer Science, Eötvös Loránd University, H-1088 Budapest, Hungary*

In biomathematics, the extensions of a leaf-colouration of a binary tree to the whole vertex set with minimum number of colour-changing edges are extensively studied. Our paper generalizes the problem for trees; algorithms and a Menger-type theorem are presented. The LP dual of the problem is a multicommodity flow problem, for which a max-flow–min-cut theorem holds. The problem that we solve is an instance of the NP-hard multiway cut problem. © 1992 Academic Press, Inc.

## 1. INTRODUCTION

Let $T$ be a tree. A *leaf* is a vertex of degree 1. Let $L$ denote the set of leaves, $|L| = l$. We term the nonleaf vertices *branching vertices*. The set of branching vertices is denoted by $B = V(T) - L$. Set $|B| = b$, $n = b + l$. Assume $C$ be a set of $r$ colours. A map $\chi: L \to 2^C$ is a *leaf-colouration* of $T$. A leaf-colouration is *simple*, if it has only singleton colours in its range. A map $\bar{\chi}: V(T) \to C$ is termed *colouration* of $T$ if for $\forall l \in L: \bar{\chi}(l) \in \chi(l)$. The *changing number* of the colouration $\bar{\chi}$ is the number of edges whose end-vertices have different colours according to $\bar{\chi}$. The colouration is an *optimal colouration* according to the leaf-colouration $\chi$, if it has the minimum changing number among all colourations. We term the minimum changing number the *length* of the tree $T$, $l(T)$ (according to $\chi$).



0196-8858/92 $9.00

The notions of changing number and length came from biomathematics. In genetics, an evolutionary tree is a leaf-coloured binary tree, where leaves represent species and colours represent some genetical properties. It is assumed by the parsimony principle, that the most likely evolutionary trees built on the given, coloured leaf set, are those with minimum possible length. (See, e.g., Felsenstein [Fe], Carter *et al.* [CHPSzW], Steel [St].) Fitch [Fi] and Hartigan [Ha] gave an algorithm to determine the length of the evolutionary tree according to the given simple leaf-colouration. Section 2 provides an $O(nr)$ algorithm to find an optimal colouration for a given $\chi$ leaf-colouration and describes the structure of all of them. We included Section 2 in this paper, for the following three reasons: (1) our main result requires some corollaries of the algorithm which are not explicit in [Fi, Ha], (2) Fitch did not prove the correctness of his algorithm, and (3) the journals where they published may not be easily accessible for those who work in combinatorial optimization. We make use of the (virtual) generality of leaf-colouration in comparison with simple leaf-colouration in a more convenient inductive hypothesis in Section 2.

In the whole paper but Section 2 we restrict our interest to simple leaf-colourations. In this special case ($|\chi(l)| = 1$ for all $l \in L$), which seems to be of highest importance for applications, the length of $T$ 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. For $r = 2$, this property is known as the $k$-edge-connectivity of the tree between two colour-classes of leaves, and Menger's theorem [Me, LP] gives that the length of $T$ is the maximum number of edge-disjoint paths connecting the colour-classes. The main objective of the present paper is the generalization of this special case of Menger's theorem.

We say that an oriented path is *colour-changing*, if its endpoints are vertices differently coloured by $\chi$ (leaves). Two colour-changing paths are in *conflict*, if either they use an edge in opposite directions or they use an edge in the same direction and the endpoints to which they are directed have the same colour.

THEOREM 1. *The length of $T$ is equal to the maximum number of colour-changing paths, with no two in conflict.*

Section 3 provides the proof of Theorem 1. The proof is algorithmic and yields a polynomial algorithm to construct $l(T)$ colour-changing paths with no two in conflict. We also show that in a straightforward generalization of the problem for graphs, the cardinality of any colour-changing path system without conflict provides a lower bound for the changing number (see Section 3 for the definitions). Section 4 describes our problem in terms of linear programming. The dual of the problem of the optimal colouration is

a multicommodity flow problem, for which Theorem 1 turns into a max-flow–min-cut theorem. We would like to know about a biological interpretation of the max-flow–min-cut theorem, if there is any.

Section 5 sets our problems in terms of multiway cuts. The Fitch–Hartigan algorithm solves the NP-hard multiway cut problem for a large class of graphs in polynomial time and Theorem 1 provides a min–max theorem for the minimum multiway cut in these instances.

What the biologists really would like to have, is the minimum length evolutionary tree over a set of species, in a much more complicated situation, where the colour of a species is a *word* made of colours and extensions of this partial colouration to internal vertices are considered. Every bit on every edge may contribute by one to the changing number and the *length* of the tree is the minimum possible changing number over all extensions. To decide, if a tree with given length (in this more general setting) exists, is NP-complete (see Graham and Foulds [GF, FG]). Hence, we would like to see statistical information on the length as a tool to decide if an evolutionary tree is "acceptable." Statistical information on evolutionary trees with one-letter colour-words (i.e., evolutionary trees in our sense) in leaves does help, see Steel [St]. The statistical analysis of the most likely evolutionary trees requires the enumeration of evolutionary trees of length $k$ built on a given, coloured leaf set. For $r = 2$ and $k = r - 1$, the enumeration was done by Carter *et al.* [CHPSzW], using multivariate Lagrange inversion and computer algebra. Steel [St] has found a combinatorial decomposition based on Menger's theorem to solve the enumeration problem for $r = 2$ and he also solved the enumeration problem for $k = r$. Erdös and Székely [ESz] gave a simple presentation of Steel's decomposition. We understood that for the solution of the enumeration problem for arbitrary $r$ and $k$ it is necessary (but not sufficient) to come up with the proper generalization of Menger's theorem. It was the starting point of the present paper. We are indebted to Professors A. Frank, E. Györi, and L. Lovász for fruitful discussions on the topics of the present paper.

## 2. An Algorithm to Find Optimal Colourations

Let $T$ be a tree, $\chi$: $L \to 2^C$ be a leaf-colouration.

Definition. The triplet $\chi^e = (\chi_1, \chi_2, \rho)$ is an *extension* of the left-colouration $\chi$ if $\chi_1$ and $\chi_2$ are maps from $V(T)$ to $2^C$, the map $\rho$: $B \to \{1, \ldots, b\}$ is a bijection, and, furthermore, for $l \in L$ we have $\chi_1(l) = \chi(l)$ and $\chi_2(l) = \varnothing$, and finally $\chi_1(v) \cap \chi_2(v) = \varnothing$ for all $v \in V(T)$.

Definition. The map $\bar{\chi}$: $V(T) \to C$ is an *implementation* of the extension $(\chi_1, \chi_2; \rho)$ if it can be derived by the following algorithm:

IMPLEMENTATION ALGORITHM.

*Step* 1. Set $\bar{\chi}(v) \leftarrow \emptyset$ for $v \in V(T)$. Set $A \leftarrow \emptyset$.

*Step* 2. If $u$ is a neighbour of a $v \in A$, $u \notin A$ (and hence $\bar{\chi}(u) = \emptyset$), then

—if $\bar{\chi}(v) \in \chi_1(u)$ then set $\bar{\chi}(u) \leftarrow \bar{\chi}(v)$, and $A \leftarrow A \cup \{u\}$.

—if $\bar{\chi}(v) \in \chi_2(u)$ then (it is up to you), set $\bar{\chi}(u) \leftarrow \bar{\chi}(v)$, and $A \leftarrow A \cup \{u\}$, or do not do anything.

*Step* 3. Repeat Step 2 with the current set $A$. Every vertex of $T$ will be examined at most once. If every neighbour of the current $A$ has been examined, then go to Step 4.

*Step* 4. If $B \setminus A \neq \emptyset$ then take $v \in B \setminus A$ for which $\rho$ is maximal.

*Step* 5. Let $\bar{\chi}(v)$ be an arbitrary element of $\chi_1(v)$. Let $A \leftarrow A \cup \{v\}$. Go to Step 2.

*Step* 6. If $B \subseteq A$ then repeat Step 5 with $v \leftarrow l$ for every leaf $l$ with $\bar{\chi}(l) = \emptyset$. If $\bar{\chi}$ nowhere equals to $\emptyset$ then finish the algorithm.

DEFINITION. An extension $\chi^e$ is *proper extension* of leaf-colouration $\chi$ if

(i) every implementation of $\chi^e$ is an optimal colouration,

(ii) every optimal colouration is an implementation of $\chi^e$.

THEOREM 2. *Let $T$ be an arbitrary tree with leaf-colouration $\chi$: $L \to 2^C$. Than there exists a proper extension $\chi^e$ of $\chi$.*

*Proof.* We construct a proper extension by mathematical induction on $b$. For a star with midpoint $m$, set $\rho(m) = 1$, $\chi_1(m) =$ the set of colours occurring with maximum multiplicity in the sets $\chi(l)$: $l \in L$, and finally set $\chi_2(m) = \emptyset$. This is obviously a good choice for $\chi^e$. Assume $T$ be a tree with diameter at least 3, with a given leaf-colouration $\chi$.

Let $v$ be a vertex of $T$ whose every neighbour (except a unique vertex $w$) is leaf. We denote the set of these leaves by $L_v$. (The existence of such a $v$ is obvious.) Let $M_1$ denote the set of colours of $C$ which occur with maximum multiplicity in the sets $\chi(l)$: $l \in L_v$, let $M_2$ denote the set of colours of $C$ which occur with maximum minus one multiplicity in the sets $\chi(l)$: $l \in L_v$.

Let us define the tree $T'$ that we obtain by deleting the set $L_v$ from $T$. Define the leaf-colouration $\chi'$ on $T'$ by

$$\chi'(u) = \begin{cases} \chi(u) & \text{if } u \in V(T') \cap (L \setminus L_v), \\ M_1 & \text{if } u = v. \end{cases}$$

Now the tree $T'$ has fewer branching points than the tree $T$ had; therefore, due to the hypothesis, the leaf-colouration $\chi'$ has an $\chi'^e = (\chi'_1, \chi'_2, \rho')$ proper extension. Define the extension $\chi^e = (\chi_1, \chi_2, \rho)$ of $\chi$ as

follows:

$$\chi_1(u) = \begin{cases} \chi_1'(u) & \text{if } u \in T' \\ \chi(u) & \text{if } u \in L_v, \end{cases}$$

$$\chi_2(u) = \begin{cases} \chi_2'(u) & \text{if } u \in T' \setminus \{v\} \\ M_2 & \text{if } u = v \\ \varnothing & \text{if } u \in L_v, \end{cases}$$

$$\rho(u) = \begin{cases} 1 & \text{if } u = v \\ \rho'(u) + 1 & \text{if } u \in B \setminus \{v\}. \end{cases}$$

For completeness, define $\chi_2(u) = \varnothing$ if $\rho(u) = |B|$.

We have to prove that $\chi^e$ is a proper extension.

(A) Let $\bar{\chi}$ be an implementation of $\chi^e$. We show that $\bar{\chi}$ is an optimal colouration. We distinguish two cases.

(a1) $\bar{\chi}(w) \in \chi_1(v)$. In that case $\bar{\chi}|T'$ is an implementation of $\chi'^e$ and the changing number of $\bar{\chi}$ on the tree $T'$ equals to $l(T')$. Let $T_v$ denote the subtree spanned by $\{v, L_v\}$. The changing number of $\bar{\chi}$ on the tree $T_v$ is equal to $l(T_v)$ (with respect to the leaf-colouration $\chi$ on $L_v$). If we apply the inequality $l(T) \geq l(T') + l(T_v)$ (which can be proved easily), then we obtain that the changing number of $\bar{\chi}$ on the tree $T$ is less than or equal to $l(T)$. It proves the claim.

(a2) $\bar{\chi}(w) \notin \chi_1(v)$. Now two further subcases are distinguished: (i) If $\bar{\chi}(w) = \bar{\chi}(v)$ then the changing number of $\bar{\chi}|T'$ is equal to $l(T') - 1$. The changing number of $\bar{\chi}$ on the tree $T_v$ is equal to $l(T_v) + 1$. The repetition of the previous argument proves the claim. (ii) Assume $\bar{\chi}(w) \neq \bar{\chi}(v)$. Then, due to the implementation algorithm, $\bar{\chi}(v) \in \chi_1(v)$ and $\bar{\chi}|T_v$ is an optimal colouration. The same holds for $\bar{\chi}|T'$; therefore $\bar{\chi}$ itself is optimal again.

(B) Assume $\bar{\chi}$ is an optimal colouration of $T$. We have to show, that it is an implementation of the extension $\chi^e$. It is easy to see that $\bar{\chi}(v)$ must belong to $\chi_1(v) \cup \chi_2(v)$. Furthermore, we know that $l(T') + l(T_v) \leq l(T)$. Therefore, if $\bar{\chi}(v) \in \chi_1(v)$, then $\bar{\chi}|T'$ is an optimal colouration. Then, due to the hypothesis, $\bar{\chi}|T'$ is an implementation of $\chi'^e$, and therefore $\bar{\chi}$ is an implementation of $\chi^e$. If $\bar{\chi}(v) \in \chi_2(v)$, then the map $\bar{\chi}'$: $V(T') \to C$, where

$$\bar{\chi}'(u) = \begin{cases} \bar{\chi}(u) & \text{if } u \neq v, \\ \in \chi_1(v) & \text{if } u = v, \end{cases}$$

is an optimal colouration of $T'$. So $\bar{\chi}'$ is an implementation of $\chi'^e$. But the colouration $\bar{\chi}$ can be derived from $\bar{\chi}'$ in a way that conforms to the implementation algorithm. Therefore, $\bar{\chi}$ is again an implementation of $\chi^e$. □

The combination of the implementation algorithm and the construction of the proof of the theorem gives us an algorithm to determine all the optimal colourations. However, if we need only one optimal colouration, e.g., we need the length of the tree, then we may do it more simply.

If this is the case, we do not construct the colour sets of $\chi_2$ and it suffices to determine only one implementation of $\chi^e$. Clearly, we can do it in polynomial time. On the other hand, to determine all implementations (i.e., all optimal colourations) has a higher complexity by the length of the output.

DEFINITION. We say that a $v \in V(T)$ *is of order* $k$ $(k = 1, 2)$, if there exists an edge $e$ adjacent to $v$, such that in the connected component of $V(T) - e$, containing $v$, $B_k(v, e)$, the longest $v$ − leaf distance is of length $k$. Let $T_k(v, e)$ denote the union of the other connected component and $e$ (see Fig. 1).

COROLLARY 1. *If $v$ is of order one, then there exists an optimal colouration $\bar{\chi}$ with $\bar{\chi}(v) \in \chi_1(v)$.*

COROLLARY 2. *If $v$ is of order two and $B_2(v, e)$ is totally multicoloured, then $\chi_1(v)$ is the set of colours of $B_2(v, e)$, and there is an optimal colouration which assigns the same element of that set to $v$ and its non-leaf neighbours in $B_2(v, e)$.*

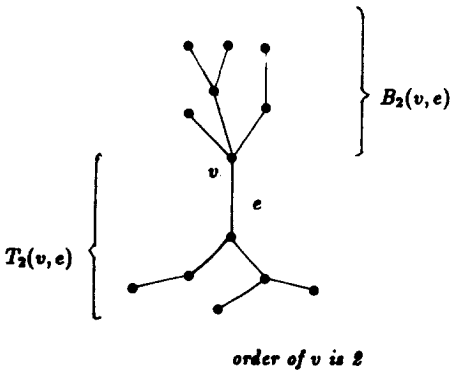



FIGURE 1

Assume $w_i$, $i = 1, \ldots, k$, is the list of those neighbours of $v$ in $B_2(v, e)$, for which $w_i$ is of order one by the edge $vw_i$.

COROLLARY 3. *If $v$ is of order two and $B_2(v, e)$ is not totally multicoloured, but $B_1(w_i, w_i v)$ is $(i = 1, \ldots, k)$, then $\chi_1(v)$ is the set of colours occurring with maximum multiplicity in $B_2(v, e)$, and there is an optimal colouration which assigns the same element of that set to $v$ and its nonleaf neighbours in $B_2(v, e)$.*

*Remark.* After slight modification, the algorithm of the present section yields optimal colouration for the generalized problem, where some branching points also have an admissible colour set assigned and the colouration has to use an admissible colour in every vertex. Either we have to transform the tree before using the same algorithm by hanging sufficiently many leaves of every colour of the admissible colour set from the vertex, which the admissible colour set was assigned to, or, alternatively, we may restrict the set $\chi_1(v)$ to its intersection with the set of admissible colours in $v$ everywhere in the algorithm. We leave the details for the reader.

## 3. THE PROOF OF THEOREM 1

Note that from now on every leaf-colouration is simple. We are going to prove an equivalent form of Theorem 1. First, we give a general lower bound for the length. Suppose $G$ is a graph, $L \subset V(G), \chi\colon L \to C$ is a partial colouration with the elements of a set $C$. Define the length of $G, l(G)$, as the minimum number of colour-changing edges over all extensions of the colouration to $V(G)$. The notions of optimal colouration, colour-changing path, and conflict are naturally generalized for $G$.

We say, that $\mathcal{F}$ is a *jungle*,[1] if it consists of rooted subtrees of a leaf-coloured tree $T$, such that

—the leaves of the trees are in $L$, and no colour is repeated in any of the trees,

—the roots are leaves in the trees,

—if an edge belongs to more than one tree, then the edge does not separate the roots of the trees,

—if an edge belongs to more than one tree, then the root-leaf paths of the trees containing that edge end in leaves of different colours.

[1]Note that a jungle is thicker than a forest.

For $F \in \mathscr{F}$, let $c(F)$ denote the number of different colours present in the leaves of $F$. We claim

THEOREM 3. (a) *For arbitrary* $G$, $L \subset V(G)$ *and* $\chi: L \to C$, *we have* $l(G) \geq$ *maximum number colour-changing paths, with no two in conflict*; (b) *for a tree* $T$,

$$l(T) \geq \max_{\mathscr{F}} \sum_{F \in \mathscr{F}} (c(F) - 1),$$

*where the maximum is taken for jungles.*

*Proof.* Assume we are given a colouration $\bar{\chi}: V(G) \to C$, such that $\bar{\chi}|L = \chi$, and a set of colour-changing paths with no two in conflict. Assign its last colour-changing edge to every colour-changing path, with respect to the orientation of the path and the given colouration. Clearly we have an injection.

For a tree $T$, note, that the root-leaf paths of the trees of $\mathscr{F}$ make a colour-changing path system with no two in conflict. □

THEOREM 4. *For a tree* $T$ *with a simple leaf-colouration* $\chi$, $l(T) = \max_{\mathscr{F}} \Sigma_{F \in \mathscr{F}}(c(F) - 1)$, *where the maximum is taken for jungles.*

It is easy to see the equivalence of Theorems 1 and 4 by building the trees of a jungle from the union of paths with the same starting point, taking the starting point for the root.

*Proof.* We have $l(T) \geq \max_{\mathscr{F}} \Sigma_{F \in \mathscr{F}}(c(F) - 1)$ from Theorem 3. We are left with the nontrivial inequality, $l(T) \leq \max_{\mathscr{F}} \Sigma_{F \in \mathscr{F}}(c(F) - 1)$. We assume that a proper extension of $\chi$ has already been determined by Section 2.

We apply mathematical induction on $b$ and $n$. The base case of the induction is $b = 1$, i.e., $T$ is a star. Now $l(T) = |L|$ − maximum colour multiplicity in $L$. We build a jungle $\mathscr{F}$ with $l(T) = \Sigma_{F \in \mathscr{F}}(c(F) - 1)$. It is possible to cover $L$ with edge-disjoint trees that are totally multicoloured (no colour is repeated in any of them), with no more trees, than the maximum colour multiplicity in $L$. Take any leaf for root in any of them.

Our hypothesis is, that we already know Theorem 4 for every leaf-coloured tree $T'$ with a fewer number of branching points or with the same number of branching points and fewer number leaves. We distinguish cases.

(A) There exists a vertex $v$ of order one and two leaves of the same colour in $V(T) \setminus T_1(v, e)$. Assume $m$ is the maximum multiplicity of a

colour in $V(T) \setminus T_1(v, e)$. Chop off $m - 1$ leaves of every colour in $V(T) \setminus T_1(v, e)$, (or less, but all of them, if the multiplicity of the colour is less than $m - 1$), and call the leftover tree $T'$. Note that the surviving colours are exactly the elements of $\chi_1(v)$. Recall Corollary 1, we have an optimal colouration $\bar{\chi}$ with $\bar{\chi}(v) \in \chi_1(v)$. Hence we have $l(T) = l(T') + d_T(v) - |\chi_1(v)| - m$. By the hypothesis of the induction on $n$, there is a jungle $\mathcal{F}'$ in $T'$ with a sum at least $l(T')$. In order to obtain a suitable jungle $\mathcal{F}$ for $T$, add $m - 1$ totally multicoloured edge-disjoint stars that contain all the deleted leaves to $\mathcal{F}'$, and

$$\begin{aligned} l(T) &= l(T') + d_T(v) - |\chi_1(v)| - m \\ &\leq \sum_{F \in \mathcal{F}'} (c(F) - 1) + d_T(v) - |\chi_1(v)| - m \\ &= \sum_{F \in \mathcal{F}} (c(F) - 1). \end{aligned}$$

The choice for the roots of the stars added is arbitrary, since they are edge-disjoint to $\mathcal{F}'$ and each to the other.

(B) We may assume that for all vertices $v$ of order one, $V(T) \setminus T_1(v, e)$ is totally multicoloured. It is easy to see that if $T$ is not a star, then it has a vertex of order two, say $v$, and edge $e$ shows it. Assume $w_i$, $i = 1, \ldots, k$, is the list of those neighbours of $v$ in $B_2(v, e)$, for which $w_i$ is of order one by the edge $vw_i$.

(B1) Assume $B_2 = B_2(v, e)$ is totally multicoloured. Let $T'$ denote the tree that we obtain from $T$ by deleting $w_i$, $i = 1, \ldots, k$, and joining the resulting isolated vertices to $v$. Let $B'$ denote what we obtain from $B_2$ in the same way. We also think of $B'$ as a subtree of $T'$. By hypothesis, we have a jungle $\mathcal{F}'$ in $T'$, which produces a sum that is at least $l(T')$. By Corollary 2, we have $l(T) = l(T')$.

DEFINITION. We say an $F \in \mathcal{F}'$ is *exterior*, if it has at least one leaf in both of $T_2(v, e)$ and $B'$. If it has all its leaves from $B'$, we say it is *interior*.

We have to build a jungle $\mathcal{F}$ in $T$ with the same sum as $\mathcal{F}'$ had. There is a natural way to extend a tree $F \in \mathcal{F}'$ to $T$ by subdivision of edges. We call it the *lift-up* of $F$ and denote it by $F^*$. Lifting up every element of $\mathcal{F}'$, we obtain $\mathcal{F}'^*$, which is, unfortunately, not necessarily a jungle. However, we want to take advantage of the idea. Note that either all the exterior trees are rooted in $B'$, or none of them, since the edge $e$

cannot be used in opposite directions by the definition of jungle. We distinguish further subcases.

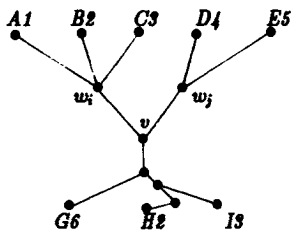
(B1$\alpha$) The total number of exterior and interior trees equals one. Take $\mathcal{F} = \mathcal{F}'^*$.

(B1$\beta$) Every exterior tree is rooted in $B'$. Let $F_1, \ldots, F_k$ be the exterior tress in $\mathcal{F}'$. $\mathcal{F}'$ being a jungle, all the leaves of these exterior trees in $V(T) \setminus B'$ have different colours. If $\mathcal{F}'$ has an interior tree, then it is easy to see, by the optimality of $\mathcal{F}'$, that it has exactly one. In order to build $\mathcal{F}$, we start with $\mathcal{F}'^*$ and do some surgery on it. Note that at most one exterior tree is rooted in $B_1(w_j, w_j v)$; otherwise we contradict the optimality of $\mathcal{F}'$. Assume that the exterior tree $F_i^*$ is rooted in $B_1(w_j, w_j v)$, then we want to add the whole $B_1(w_j, w_j v)$ to $F_i^*$ (and delete it from every tree of the jungle that contained it before). Only a leaf of colour $c$ causes a problem, if $F_i^*$ already has a leaf of colour $c$ in $V(T) \setminus B_2(v, e)$. Delete the latter leaf from $F_i^*$, and join it to another exterior tree. We can do it, since no other exterior tree may contain the colour $c$ after the alterations. If we do not find another exterior tree, then we have an interior tree; otherwise we are in (B1$\alpha$). Join the latter leaf to the interior tree (it does not contain colour $c$ as $B_2(v, e)$ is totally multicoloured and its colour $c$ has already been used up).

We may have had an interior tree. It did not survive the last paragraph unless it contained some $B_2(w_j, w_j v)$. If it did survive, lift it up and take for its root any leaf of that $B_1(w_j, w_j v)$ (see Fig. 2).

(B1$\gamma$) No exterior tree is rooted in $B'$. If $\mathcal{F}'$ has no interior trees, take $\mathcal{F} = \mathcal{F}'^*$. If $\mathcal{F}'$ has an interior tree, then it is easy to see by the optimality of $\mathcal{F}'$, that it has exactly one. Construct $\mathcal{F}$ in the following way: restrict the exterior trees to $V(T) \setminus B'$ and add the star $B'$ with arbitrary root to them and finally, lift them up.

(B2) $B_2 = B_2(v, e)$ is not totally multicoloured. Since we are not in case (A), every $B_1(w_j, w_j v)$ is totally multicoloured. Assume $m$ is the maximum multiplicity of a colour in $B_2$, $m \geq 2$. Delete every neighbour of $v$ but edge $e$, and join to $v$ a representative leaf of every colour occurring with maximum multiplicity. Call the new tree $T'$. By Corollary 3, we have $l(T') = l(T) - (|L(T)| - |L(T')|) + (m - 1)$. By hypothesis, there is a jungle $\mathcal{F}'$ in $T'$, which produces a sum equal to $l(T')$. Call the star of $v$ and the representative leaves $B'$; we speak about interior and exterior trees of $\mathcal{F}'$ again.




$\mathcal{F}'^* = \{A \to G, H, I; D \to B; E \to C\}$

$\mathcal{F} = \{A \to B, C, G; D \to H; E \to I\}$

*letters mean leaves, numbers mean colours*

FIGURE 2

Define a colour-preserving injection $\tau$ from the representative leaf set of $T'$ into the leaf set of $B_2$, whose range does not intersect at least $m - 1$ components of $B_2 \setminus \{v\}$. Such a $\tau$ can be defined by the greedy algorithm; always try to define $\tau$ on a leaf by choosing a leaf of the same colour from the already spoiled components. Whenever we have to pick a new component for a new colour, it means that there are $m$ intact components containing that colour. Define the lift-up $F^*$ of $F \in \mathcal{F}'$ by joining $F \cap T_2(v, e)$ to $\tau(p)$ in $T$ for any representative leaf $p$ of $\mathcal{F}$. We may assume, that $\mathcal{F}'$ covers all the representative leaves; if not, we may add trees of singleton leaves to $\mathcal{F}'$. (It is easy to see that there is at most one tree of a singleton leaf.)

*First extension.* Define $\mathcal{F}$ in the following way. Take $\mathcal{F}'^*$ and add $m - 1$ internal trees to it that cover all leaves of $B_2$ not covered by $\mathcal{F}'^*$ in the following way: pick $m - 1$ roots from $m - 1$ arbitrary but fixed components of $B_2 \setminus \{v\}$ not spoiled by $\tau$, and build $m - 1$ totally multi-coloured rooted trees (stars), such that the selected $m - 1$ unspoiled components considered above wholly belong to one tree each, and then cover with the trees all the not covered leaves in $B_2$.

The first extension may fail to give the required jungle $\mathcal{F}$ only if $\mathcal{F}'$ has an interior tree or exterior tree rooted in a representative leaf. (Again, either all exterior trees are rooted in a representative leaf or none of them.) The problem may be the use of some $w_i v$ edges in both directions, by members of $\mathcal{F}$. In this case we further modify the result of the first extension.

*Second extension.* If a $B_1(w_i, w_i v)$ contains the root of a lift-up tree, then we would like to remove the leaves of $B_1(w_i, w_i v)$ from the other trees rooted elsewhere and add them to the lift-up tree. If we have two

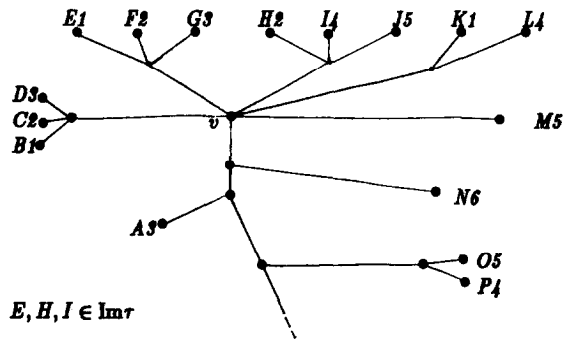



FIGURE 3

lift-up trees rooted in $B_1(w_i, w_iv)$, then we can add any leaf to one or the other, since their colours are all distinct in $T_1(w_i, w_iv)$. If we have only one lift-up tree $F^*$ rooted in $B_1(w_i, w_iv)$, then we may fail in adding a leaf of colour $c$ (belonging to a tree $S$) from $B_1(w_i, w_iv)$ to the tree; i.e., the tree $F^*$ already has a leaf of colour $c$ in $T_1(w_i, w_iv)$. Then switch the leaves of colour $c$ between $F^*$ and $S$ (see Fig. 3).

It is easy to see that the second extension eliminated the conflicts that may have survived the first extension and did not create any new ones. □

## 4. THE LP CONNECTION

For every edge $pq$ of the tree $T$ and every pair of distinct colours $ij$ define a variable $z_{pq,ij}$. If $q \in L$, set $z_{pq,ij} = 0$ for every $j \neq \chi(q)$. Identify $z_{pq,ij}$ and $z_{qp,ji}$. Consider the linear program

$$z_{pq,ij} \geq 0,$$

for every $P_{ab}$ colour-changing path

$$\sum_{pq \in P_{ab}} \sum_{i:\, i \neq \chi(b)} z_{pq,i\chi(b)} \geq 1,$$

where the sum is for $pq$ edges of the path with $p$ reached first (remember, that colour-changing paths are oriented paths):

$$\min \sum z_{pq,ij}.$$

To describe the dual linear program, for every colour-changing path $P_{ab}$, introduce a variable $\lambda_{ab}$,

$$\lambda_{ab} \geq 0.$$

If $p, q$ are not leaves, then for every $i, j \in C$, $i \neq j$, have

$$\sum_{\chi(b)=j} \lambda_{ab} + \sum_{\chi(v)=i} \lambda_{uv} \leq 1;$$

if $q$ is a leaf, then for $j = \chi(q)$, $i \neq j$, have

$$\sum_{\chi(b)=j} \lambda_{ab} + \sum_{\chi(v)=i} \lambda_{uv} \leq 1$$

(the first sums are taken for colour-changing paths $P_{ab}$ that contain $pq$ and reach $p$ first, while the second sums are taken for colour-changing paths $P_{uv}$ that contain $pq$ and reach $q$ first):

$$\max \sum \lambda_{ab}.$$

It is easy to see that the maximum number of colour-changing paths, no two in conflict $\leq \max \Sigma \lambda_{ab}$: $\lambda_{ab}$ integer $\leq \max \Sigma \lambda_{ab} = \min \Sigma z_{pq,ij} \leq \min \Sigma z_{pq,ij}$: $z_{pq,ij}$ integer $\leq l(T)$. Only the first and last inequalities require proof from the chain of inequalities above. The first one holds, since the root-leaf paths of a jungle always provide a feasible integer solution for the second linear program; the last one holds, since we have an optimal colouration $\bar{\chi}$ with $l(T)$ colour-changing edges, define $z_{pq,ij} = 1$, if $pq$ is one of the $l(T)$ colour-changing edges, $\bar{\chi}(p) = i$, $\bar{\chi}(q) = j$. By Theorem 1, each of the inequalities holds by equality.

Finally, our second linear program is a multicommodity flow problem, for which a max-flow–min-cut theorem holds.

## 5. The Multiway Cut Problem

The *multiway cut problem* [CR] is the following one: Given a connected simple graph $G$ with positive edge weights and a set of vertices $N \subset V(G)$, find a minimum-weight edge set that separates all pairs of $N$. Having a constant weight, the multiway cut problem is the problem of finding an

optimal colouration which extends a partial colouration that assigns different colours to the members of $N$. Having an instance of the problem of finding an optimal colouration, identify the vertices to which the partial colouration assigned the same colour and represent the arising multiple edges by multiple weights in order to obtain an instance of the weighted multiway cut. People working on the multiway cut problem seem to be unaware of the related work in biomathematics and vice versa.

Dahlhaus *et al.* [DJPSY] proved that the multiway cut problem is NP-hard even for $|N| = 3$ and equal edge weights. It easily implies that the problem "is $l(G) \leq m$?" is NP-complete. Dahlhaus *et al.* [DJPSY] also proved that the multiway cut problem is polynomially solvable for planar graphs with fixed $|N|$. Note that identifying some leaf sets of a tree, we may obtain nonplanar graphs. Conversely, by splitting vertices of $N$ into as many leaves as their degree, we obtain

THEOREM 5. *If $N$ intersects every cycle of $G$, then the Fitch–Hartigan algorithm described in Section 2 provides a solution in polynomial time for the multiway cut problem with equal edge weights. It also provides a way to obtain all minimum multiway cuts. We have $l(G)$ = the maximum number of colour-changing paths with no two in conflict.* □

We already have a version of the Fitch–Hartigan algorithm for edge-weighted trees with positive integer weights and we also have the weighted version of Theorem 5.

## REFERENCES

[CR] S. CHOPRA AND M. R. RAO, On the multiway cut polyhedron, *Networks* **21** (1991), 51–89.

[CHPSW] 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.

[DJPSY] E. DAHLHAUS, D. S. JOHNSON, C. H. PAPADIMITRIOU, P. SEYMOUR, AND M. YANNAKAKIS, The complexity of multiway cuts, extended abstract 1983.

[ESz] P. L. ERDÖS AND L. A. SZÉKELY, Counting bichromatic evolutionary trees, *Discrete Appl. Math.*, to appear.

[Fe] J. FELSENSTEIN, Phylogenies from molecular sequences: Inference and reliability, *Annu. Rev. Genetics* **22** (1988), 521–565.

[Fi] W. M. FITCH, Towards defining the course of evolution. Minimum change for specific tree topology, *Syst. Zool.* **20** (1971), 406–416.

[FG] L. R. FOULDS AND R. L. GRAHAM, The Steiner problem in phylogeny is NP-complete, *Adv. Appl. Math.* **3** (1982), 43–49.

[GF] 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.

[Ha] J. A. HARTIGAN, Minimum mutation fits to a given tree, *Biometrics* **29** (1973), 53–65.

[LP] L. LOVÁSZ AND M. D. PLUMMER, "Matching Theory," Akadémiai Kiadó, Budapest; North-Holland, Amsterdam, 1986.

[Me] K. MENGER, Zur allgemeinen Kurventheorie, *Fund. Math.* **10** (1926), 96–115.

[St] M. A. STEEL, Distributions on bicoloured binary trees arising from the principle of parsinomy, *Discrete Appl. Math.*, to appear.