Annals of Combinatorics 10 (2006) 77-96
0218-0006/06/010077-20
DOI 10.1007/s00026-006-0275-9

Annals of Combinatorics

Mapping Edge Sets to Splits in Trees: the Path Index and Parsimony

Andreas Dress[1] and Mike Steel[2]

[1]Department of Combinatorics and Geometry, CAS-MPG Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, 320 Yue Yang Road, Shanghai, China

[1]Max Plank Institute for Mathematics in the Sciences, Inselstrasse 22 -26, D 04103 Leipzig, Germany
dress@mis.mpg.de, dress@sibs.ac.cn

[2]Allan Wilson Centre for Molecular Ecology and Evolution, University of Canterbury, Christchurch, New Zealand
m.steel@math.canterbury.ac.nz

Received August 24, 2004

AMS Subject Classification: 05C05, 05C25, 92D15

We dedicate this paper to Walter Fitch on the occasion of his 75th birthday

Abstract. Associated to any finite tree, there are simple vector spaces over $\mathbb{Z}_2$ and linear transformations between them that relate collections of edge-disjoint paths, sets of leaves of even cardinality, and bipartitions of the leaf set. In this paper, we use this connection to introduce and study an apparently new integer-valued invariant, the 'path index' of a tree. In the case of trivalent (or 'binary') trees, this index has an interesting recursive description that allows its easy calculation implying in particular that the path index of such a tree never exceeds one quarter of the number of its leaves and coincides with that number exclusively for the (unique!) trivalent tree with four leaves. We then show how our algebraic perspective has some other uses — for example, it relates to Hadamard conjugation first described by Mike Hendy, and it provides a way to study a combinatorial optimization problem considered in phylogenetics called the small maximum parsimony problem.

Keywords: X-trees, paths, maximum parsimony, short exact sequence

1. Introduction

Many properties of a tree can be studied by exploiting a vector-space structure (over $\mathbb{Z}_2$) that relates edge-disjoint collections of paths, sets of leaves of even cardinality, and bipartitions (or *splits*) of the leaf set of the tree. We describe how these vector spaces are

related and how they give rise to a short exact sequence (as considered in homological algebra) as well as an integer-valued invariant that appears to have not been studied yet — which we call its *path index*. We show how this index can be computed by a convenient recursion that has a particularly simple form in the case of trivalent trees, implying in particular that the path index of such a tree never exceeds one quarter of the number of its leaves and coincides with that number exclusively for the (unique!) trivalent tree with four leaves. We then apply some of our constructions to derive results concerning the parsimony score of a character, an invariant that is fundamental for the method of *maximum parsimony* used to reconstruct trees from collections of characters ([1,2,8]).

Many of the definitions and arguments in this paper combine algebraic and combinatorial concepts, and although the algebra is elementary (groups, homomorphisms, linear transformations, short exact sequences etc.), readers who are less familiar with this area may wish to consult a suitable text, e.g. [7]. We begin by outlining the general set-up of the paper.

2. The Index of a Map from a Finite Set into a Finite-Dimensional Vectorspace

Suppose we are given

- $\mathbf{k}$: a field,
- U: a finite-dimensional $\mathbf{k}$-vectorspace,
- E: a finite set,
- $f\colon E \to U$, a map.

Based on these entities, let us define

- $\widehat{U} := \mathrm{Hom}_{\mathbf{k}}(U,k)$, the $\mathbf{k}$-dual of U,
- $\widehat{f}\colon \widehat{U} \to \mathbf{k}^E\colon \varphi \mapsto \varphi \circ f$,
- $\mathrm{ind}(f)\colon \widehat{U} \to U\colon \varphi \mapsto \sum_{e\in E} \varphi\big(f(e)\big) f(e)$ $\big(= \sum_{e\in E} \widehat{f}(\varphi)(e) f(e)\big)$,
- $\mathrm{ind}_f := \dim_{\mathbf{k}} \ker\big(\mathrm{ind}(f)\big)$, the $\mathbf{k}$-dimension of the kernel of the linear transformation $\mathrm{ind}(f)$.

Remark 2.1. The map $f\colon E \to U$ induces a linear transformation

$$\begin{aligned} \mathbf{k}[f]\colon \quad \mathbf{k}[E] \quad &\to U \\ \sum_{e\in E} x_e e \;&\mapsto\; \sum_{e\in E} x_e f(e) \end{aligned}$$

from the finite-dimensional $\mathbf{k}$-vectorspace $\mathbf{k}[E]$ freely generated by E into U, and thereby a dual transformation

$$\begin{aligned} \widehat{f} := \widehat{\mathbf{k}[f]}\colon \; \widehat{U} \;&\to\; \mathbf{k}^E \\ \varphi \;&\mapsto\; \big(\varphi \circ f\colon E \to \mathbf{k}\colon e \mapsto \varphi(f(e))\big) \end{aligned}$$

from the dual $\widehat{U}$ of U into the $\mathbf{k}$-vectorspace $\mathbf{k}^E$ consisting of all maps from E into $\mathbf{k}$ $\big($identified with the dual vectorspace $\widehat{\mathbf{k}[E]}$ of $\mathbf{k}[E]$ in the canonical way by associating, to each map $\eta\colon E \to \mathbf{k}$, the linear extension $\mathbf{k}[\eta]\colon \mathbf{k}[E] \to \mathbf{k}$ that maps an element

$\sum_{e\in E} x_e e \in \mathbf{k}[E]$ onto the corresponding linear combination $\sum_{e\in E} x_e \eta(e)$ of the elements $\eta(e)\big)$.

Furthermore, we have the canonical isomorphism

$$\iota_E : \mathbf{k}^E \to \mathbf{k}[E] :$$
$$f \mapsto \sum_{e\in E} f(e)\cdot e$$

giving rise to the following commutative diagram for

$$\mathrm{ind}(f) = \mathbf{k}[f] \circ \iota_E \circ \widehat{\mathbf{k}[f]} :$$

$$\begin{array}{ccc} \widehat{U} & \xrightarrow{\mathrm{ind}(f)} & U \\ \big\downarrow{\scriptstyle \widehat{\mathbf{k}[f]}} & & \big\uparrow{\scriptstyle \mathbf{k}[f]} \\ \mathbf{k}^E & \xrightarrow{\iota_E} & \mathbf{k}[E] \end{array}$$

Remark 2.2. Although $\dim_{\mathbf{k}} U = \dim_{\mathbf{k}} \widehat{U}$ holds, there is no reason to believe that $\mathrm{ind}(f)$ is necessarily an isomorphism. Indeed, as we will see, there are rather naturally occurring instances of such maps whose kernel — much to our own surprise — we found to be non-trivial.

3. The Path Index of a Tree

In this section, we specialise the general set-up described above to the following setting.

(i) We consider a fixed finite set X.

(ii) In this paper, an *X-tree* will be be understood to mean any finite, connected, and cycle-free graph with vertex set V_T and edge set $E_T \subseteq \binom{V_T}{2}$ whose vertex set contains the set X while, conversely, every vertex $v \in V_T$ of degree less than 3 is contained in X. Sometimes, X-trees are defined in a slightly more general way, however we prefer the current simpler definition in the present context.

(iii) For every vertex $v \in V_T$, we define its *co-boundary* $\partial^T v$ relative to T by

$$\partial^T v := \{e \in E_T : v \in e\}$$

so that the degree $\deg_T(v)$ of v relative to T is given by $\deg_T(v) := \#\partial^T v$. Furthermore, we put

$$V_k(T) = \{v \in V_T : \deg_T(v) = k\}$$

for each $k \in \mathbb{N} := \{1, 2, \dots\}$. Of particular interest are *trivalent* X-trees, that is, X-trees T for which X coincides with the set $V_1(T)$ of *leaves* of T while every other vertex of T has degree 3.

Two X-trees T and T' are said to be *canonically isomorphic* if there exists a (necessarily unique) bijection $\psi : V_T \to V_{T'}$ that maps any element $x \in X \subseteq V_T$ onto itself considered as an element of V'_T, and for which the induced map $\binom{\psi}{2} : \binom{V_T}{2} \to \binom{V_{T'}}{2}$ restricts to a bijection $E_T \to E_{T'}$ from E_T onto $E_{T'}$. Any such bijection ψ is said to *induce* an isomorphism from T onto T'.

(iv) We specify $\mathbf{k}$ to be the prime field $\mathbf{k} := \mathbb{F}_2$ of characteristic 2 containing the two distinct elements $0 = 0_{\mathbb{F}_2}$ and $1 = 1_{\mathbb{F}_2}$, only.

(v) We specify U to be the quotient $\mathbb{F}_2^X/\mathrm{const}_{\mathbb{F}_2}(X)$ of the $\mathbb{F}_2$-vectorspace $\mathbb{F}_2^X$ consisting of all maps from X into $\mathbb{F}_2$ modulo the one-dimensional subspace $\mathrm{const}_{\mathbb{F}_2}(X)$ consisting of all (altogether exactly two) constant maps from X into $\mathbb{F}_2$. Recall that a *split* S of X is a subset of the power set $\mathcal{P}(X)$ of X consisting of two disjoint subsets of X, the two 'split halves' of S, whose union is all of X, and let $\mathcal{S}(X)$ denote the set of all splits of X. We may then identify U with $\mathcal{S}(X)$ by associating, to any coset $\overline{\eta} := \eta + \mathrm{const}_{\mathbb{F}_2}(X)$ in U of a map η in $\mathbb{F}_2^X$, the (well-defined!) split $S_{\overline{\eta}} := \{\eta^{-1}(0), \eta^{-1}(1)\}$ that, conversely, determines the coset $\overline{\eta}$ uniquely. Note that, this way, the sum of two splits $S = \{A, B\}$ and $S' = \{A', B'\}$ in $\mathcal{S}(X) = \mathbb{F}_2^X/\mathrm{const}_{\mathbb{F}_2}(X)$ is the split $S\Delta S' = \{A, B\}\Delta\{A', B'\}$ defined by

$$\{A, B\}\Delta\{A', B'\} := \{(A \cap A') \cup (B \cap B'), (A \cap B') \cup (B \cap A')\}$$

or, equivalently, by

$$\{A, B\}\Delta\{A', B'\} := \{A\Delta A', X - (A\Delta A')\},$$

where $A\Delta B$ is the 'symmetric difference' of A and B, defined by

$$A\Delta B := (A \cup B) - (A \cap B) \ \big(= (A - B) \cup (B - A)\big),$$

for all $A, B \in \mathcal{P}(X)$. This allows us to denote the 'Δ-sum' of any family $(S_i)_{i\in I}$ of splits in $\mathcal{S}(X)$ by $\Delta_{i\in I} S_i$. Clearly, two elements $a, a' \in X$ are in the same split half of $\Delta_{i\in I} S_i$ if and only if the number of indices $i \in I$ for which a and a' are in disjoint split halves of S_i is an even number.

(vi) And we specify $f := f_T$ to be the much studied map (cf. [8]) from E_T to $\mathcal{S}(X)$ that associates, to each edge $e = \{u, v\} \in E_T$, the unique split $S_e = \{A_u^e, A_v^e\}$ of X into the two subsets A_u^e and A_v^e consisting of all $x \in X$ that are 'closer' to u than v, or to v than to u, respectively — here, as also later on without further notice, we make use of the simple fact that there exists, for any connected graph $G = (V, E)$ with vertex set $V = V_G$ and edge set $E = E_G \subseteq \binom{V}{2}$, a canonical metric $d_G\colon V \times V \to \mathbb{R}$ on the vertex set V defined, e.g., as the unique (!) largest metric on V with $d_G(u, v) \leq 1$ for all $u, v \in V$ with $\{u, v\} \in E_G$.

With these specifications, the construction above suggests that we consider the induced map

$$\mathrm{ind}(f_T)\colon \widehat{\mathcal{S}(X)} \to \mathcal{S}(X)\colon \varphi \mapsto \Delta_{e\in E_T\colon \varphi(S_e)=1} S_e$$

that maps any $\mathbb{F}_2$-linear map φ from $S(X)$ into $\mathbb{F}_2$ onto the Δ-sum, over all $e \in E_T$ with $\varphi(S_e) = 1$, of the associated splits S_e.

By abuse of notation, we also just write T instead of $\mathrm{ind}(f_T)$ for this map which is justified by the fact to be established below that any two X-trees T and T' are canonically isomorphic if and only if the two associated maps $\mathrm{ind}(f_T)$ and $\mathrm{ind}(f_{T'})$ from $\widehat{\mathcal{S}(X)}$ into $\mathcal{S}(X)$ coincide.

Definition 3.1. *The* path index $\mathrm{i}_{\mathrm{path}}(T)$ *of the* X*-tree* T *is the index* $\mathrm{ind}_T = \mathrm{ind}_{f_T}$ *of the map* $f_T\colon E_T \to \mathcal{S}(X)$ *from the edge set* E_T *of* T *into the set* $\mathcal{S}(X)$ *of all splits of* X (*considered, as described above, as an* $\mathbb{F}_2$*-vectorspace*)*, that is, it is defined by*

$$\mathrm{i}_{\mathrm{path}}(T) := \dim_{\mathbb{F}_2} \ker(T).$$

In the next section, we will show how this index has an alternative, more combinatorial description, namely in terms of the number of even cardinality subsets of X that satisfy various parity conditions based on systems of paths in T.

4. Some Canonical Identifications

In the following, we identify — by further abuse of notation — each subset Z of a set X with

(i) the map

$$Z\colon X \to \mathbb{F}_2\colon x \mapsto \begin{cases} 1, & \text{in case } x \in Z, \\ 0, & \text{in case } x \notin Z, \end{cases}$$

(ii) the map

$$Z\colon \mathcal{P}(X) \to \mathbb{F}_2\colon Y \mapsto \langle Z|Y\rangle_X := \sum_{y\in Y} Z(y)\ (= \#(Z\cap Y)\cdot 1_{\mathbb{F}_2}),$$

(iii) the element

$$Z = \sum_{z\in Z} z \in \mathbb{F}_2[X]$$

in the $\mathbb{F}_2$-vectorspace $\mathbb{F}_2[X]$ freely generated by X,

this way identifying the set $\mathcal{P}(X)$ of all subsets of X with

(i) the $\mathbb{F}_2$-vectorspace $\mathbb{F}_2^X$ of all maps from X into $\mathbb{F}_2$ — with vector addition in $\mathcal{P}(X)$ given by the symmetric difference operation Δ,

(ii) the $\mathbb{F}_2$-dual $\widehat{\mathcal{P}(X)} = \mathrm{Hom}(\mathcal{P}(X), \mathbb{F}_2)$ of $\mathcal{P}(X)$ (considered as a vectorspace over $\mathbb{F}_2$ using the identification described above) consisting of all maps

$$\eta\colon \mathcal{P}(X) \to \mathbb{F}_2 \text{ with } \eta(A\Delta B) = \eta(A) + \eta(B) \text{ for all } A, B \in \mathcal{P}(X),$$

(iii) and the $\mathbb{F}_2$-vectorspace $\mathbb{F}_2[X]$ freely generated by X.

In particular, the one-dimensional subspace $\mathrm{const}_{\mathbb{F}_2}(X)$ of $\mathbb{F}_2^X$ consisting of all constant maps from X into $\mathbb{F}_2$ gets, in this way, identified with the one-dimensional subspace $\{\emptyset, X\}$ of $\mathcal{P}(X)$, thus inducing a canonical identification of the quotient space $\mathbb{F}_2^X/\mathrm{const}_{\mathbb{F}_2}(X)$ of $\mathbb{F}_2^X$ with the set of cosets $Y\Delta\{\emptyset, X\} = \{Y\Delta\emptyset, Y\Delta X\} = \{Y, X - Y\}$ of the subsets Y of X relative to $\{\emptyset, X\}$ and thus — once more and in an even more direct way — with the set $\mathcal{S}(X)$ of splits of X.

Thus, as $Z(X) = \langle Z|X\rangle_X$ vanishes for some subset Z of X if and only if the cardinality of Z is even, this implies further that

- the set $\mathcal{P}_{\text{even}}(X)$ consisting of all subsets of $\mathcal{P}(X)$ of even cardinality forms a subvectorspace of $\mathcal{P}(X)$ of co-dimension 1,
- every set $Z \in \mathcal{P}_{\text{even}}(X)$ induces a canonical linear form $\mathcal{S}(X) \to \mathbb{F}_2$ also denoted, by abuse of notation, by Z from $\mathcal{S}(X) = \mathbb{F}_2^X/\text{const}_{\mathbb{F}_2}(X)$ into $\mathbb{F}_2$ defined by

$$Z(\{A,B\}) := \langle Z \,|\, A\rangle_X \ \big(= \#(A \cap Z) \cdot 1_{\mathbb{F}_2} = \#(B \cap Z) \cdot 1_{\mathbb{F}_2} = \langle Z \,|\, B\rangle_X\big)$$

  for every split $\{A,B\}$ in $\mathcal{S}(X)$,
- the induced pairing

$$\mathcal{P}_{\text{even}}(X) \times \mathcal{S}(X) \to \mathbb{F}_2 \colon (Z,S) \mapsto \langle Z \,|\, S\rangle^X := Z(S)$$

  is a non-degenerate pairing of $\mathbb{F}_2$-vectorspaces,
- and the induced map

$$\mathcal{P}_{\text{even}}(X) \to \widehat{\mathcal{S}(X)} \colon Z \mapsto (Z \colon \mathcal{S}(X) \to \mathbb{F}_2 \colon S \mapsto \langle Z \,|\, S\rangle^X)$$

  is a canonical isomorphism from $\mathcal{P}_{\text{even}}(X)$ onto the $\mathbb{F}_2$-dual $\widehat{\mathcal{S}(X)}$ of the quotient space $\mathbb{F}_2^X/\text{const}_{\mathbb{F}_2}(X) = \mathcal{S}(X)$ referred to in the definition of the map $T = \text{ind}(f_T)$ in the previous section.

Let us now assume that, as in Section 3, T is an X-tree with vertex set V_T and edge set $E_T \subseteq \binom{V_T}{2}$. Then, the induced map $\widehat{f_T}$ from $\mathcal{P}_{\text{even}}(X) = \widehat{\mathcal{S}(X)}$ into $\mathbb{F}_2^{E_T}$ maps every $Z \in \mathcal{P}_{\text{even}}(X)$ onto the map

$$Z \circ f_T \colon E_T \to \mathbb{F}_2 \colon e \mapsto Z\big(f_T(e)\big) = Z(S_e) = \langle Z \,|\, S_e\rangle^X$$

and thus, identifying $\mathbb{F}_2^{E_T}$ with $\mathcal{P}(E_T)$ as explained above, onto the subset

$$[Z] = [Z]_T := \{e \in E_T \colon \langle Z \,|\, S_e\rangle^X = 1\}$$

of E_T. It follows that $\widehat{f_T}$ maps any 2-subset $\{x,y\}$ of X onto the set $[x,y]$ of edges $e \in E_T$ on the *path* from x to y in T and, more generally, any $Z \in \mathcal{P}_{\text{even}}(X)$ onto the (unique!) *smallest* subset F of E_T for which, for every $z \in Z$, there exists some $z' \in Z - \{z\}$ such that the edges on the path from z to z' are all in F:

Indeed, it is fairly obvious that, indexing the elements of a subset Z of X of cardinality $2k$ for some $k \in \mathbb{N}$ as $z_1, z_2, \ldots, z_{2k}$, one has

$$[Z]_T = [\{z_1,z_2\} \Delta \cdots \Delta \{z_{2k-1}, z_{2k}\})]_T = [z_1,z_2]_T \Delta \cdots \Delta [z_{2k-1}, z_{2k}]_T,$$

and there exists some such labelling (unique up to permutations π of the index set $\{1,2,\ldots,2k-1,2k\}$ with $|\pi(2i) - \pi(2i-1)| = 1$ for all $i = 1,2,\ldots,k$ — or, equivalently (!), with $\pi(2i) = \pi(2i-1)+1$ in case $\pi(2i-1)$ is odd, and $\pi(2i) = \pi(2i-1)-1$ in case $\pi(2i-1)$ is even, $i = 1,2,\ldots,k$ — in case T is trivalent) such that $[Z]_T$ coincides with the disjoint union of the sets $[\{z_1,z_2\}]_T, \ldots, [\{z_{2k-1},z_{2k}\}]_T$.

This explains the term 'path index' that we suggest to call the index of the map $f_T \colon E_T \to \mathcal{S}(X)$ and the notation $[Z] = [Z]_T$ for the set of edges e in E_T with $\langle Z \,|\, S_e\rangle^X = 1$.

Furthermore, it follows that the $\mathbb{F}_2$-linear extension

$$\Delta_T := \mathbb{F}_2[f_T]\colon \mathcal{P}(E_T) \to \mathcal{S}(X)$$

of the map $f_T\colon E_T \to \mathcal{S}(X)$ maps any subset F of E_T onto the Δ-sum

$$\Delta_T F := \Delta_{e\in F}\, S_e$$

while the map $T\colon \widehat{\mathcal{S}(X)} = \mathcal{P}_{\text{even}}(X) \to \mathcal{S}(X)$ maps any even-cardinality subset Z of X onto the Δ-sum $\Delta_T[Z] = \Delta_{e\in[Z]}\, S_e$ of the splits in the family $(S_e)_{e\in[Z]}$, i.e., we have

$$T(Z) = \Delta_T[Z] = \Delta_{e\in[Z]}\, S_e$$

for every even-cardinality subset Z of X.

In particular, as Δ_T maps a given subset F of E_T onto the zero element $\{X, \emptyset\}$ in $\mathcal{S}(X)$ if and only if the cardinality of $[x, y]_T \cap F$ is even for all $x, y \in X$ (or, equivalently, if and only if the cardinality of $[x, y]_T \cap F$ is even for some fixed element $x \in X$ and all elements $y \in X - \{x\}$), we see also that the path index of T coincides with the $\mathbb{F}_2$-dimension of the collection of subsets $Z \in \mathcal{P}_{\text{even}}(X)$ with $\#\big([x, y]_T \cap [Z]_T\big) \equiv 0 \mod 2$ for all $x, y \in X$ (or, equivalently, for some fixed elements $x \in X$ and all elements $y \in X - \{x\}$) — where this collection is, of course, considered as an $\mathbb{F}_2$-subvectorspace of $\mathcal{P}(X)$.

Using these notations, it is now easy to show that an X-tree T is indeed uniquely determined by the associated map $T\colon \mathcal{P}_{\text{even}}(X) \to \mathcal{S}(X)$. For the following lemma, consider, for any $a \in X$, the following subset of $\mathcal{P}_{\text{even}}$

$$\mathcal{P}_2(X|a) := \big\{\{a, b\}\colon b \in X - \{a\}\big\}.$$

Lemma 4.1. *Two X-trees T and T' are canonically isomorphic if and only if the induced maps T and T' from $\mathcal{P}_{\text{even}}(X)$ into $\mathcal{S}(X)$ coincide. Moreover this is the case if and only if the restriction of these two maps to $\mathcal{P}_2(X\,|\,a)$ coincide for one* (*or, as well, for all*) *$a \in X$* (*since $\mathcal{P}_2(X\,|\,a)$ forms an $\mathbb{F}_2$-basis of $\mathcal{P}_{\text{even}}(X)$*).

Proof. First note that

$$T\big(\{x, y\}\big) = \big\{X - \{x, y\}, \{x, y\}\big\}$$

holds for a 2-subset $\{x, y\}$ of X if and only if $\{x, y\} \subseteq V_1(T)$ and $d_T(x, y) = 2$ holds. Thus, it suffices to observe that, given any vertex $v \in V_T$ and any two distinct leaves x, y of T with $\{x, v\}, \{y, v\} \in E_T$, the map $T^{x,y}$ associated with the $X^{x,y}$-tree $T^{x,y} := (V^{x,y}, E^{x,y})$ defined by

$$V^{x,y} := V_T - \{x, y\}, \quad E^{x,y} := E_T - \big\{\{x, v\}, \{y, v\}\big\},$$

for the set

$$X^{x,y} := \{v\} \cup (X - \{x, y\})$$

is completely determined by the map T because, for every element a in $X^{x,y} - \{v\} = X - \{x, y\}$, one has

$$T^{x,y}(\{v, a\}) = \big\{A - \{x\}, B \cup \{v\} - \{y\}\big\}$$

where A and B denote the two subsets of X with $T\big(\{x, a\}\big) = \{A, B\}$ and $x \in A$ (and, hence, $y \in B$).

Using these facts, Lemma 4.1 can easily be established by induction on $\#X$ in a well-known recursive fashion. ∎

Remark 4.2. It might be of some interest to find a structural characterization of those $\mathbb{F}_2$-linear maps $\tau\colon \mathcal{P}_{\text{even}}(X)\to\mathcal{S}(X)$ for which some X-tree T with $\tau = T$ exists.

5. Examples

(i) Suppose that T is a trivalent X-tree with at least four leaves such that every leaf has distance 2 to some other leaf. Then, $\#X$ must be even, and X is the only non-empty subset Z in $\mathcal{P}_{\text{even}}(X)$ with $\#\big([x,y]_T \cap [Z]_T\big) \equiv 0 \mod 2$ for all $x, y \in X$, so $\mathrm{i}_{\text{path}}(T) = 1$ must hold.

More generally, if $x, y \in X$ are any two leaves of distance 2, we have either $x, y \in Z$ or $x, y \notin Z$ for any subset $Z \in \ker(T)$. Indeed, it is easily seen that, for any two such leaves $x, y \in X$, $\ker(T)$ is the disjoint union of its two subsets $\{Z \subseteq X - \{x,y\}\colon Z \in \ker(T)\}$ and $\{Z \cup \{x,y\}\colon Z \subseteq X - \{x,y\},\ T(Z) = \{X - \{x,y\}, \{x,y\}\}\}$.

(ii) Let T_n be the trivalent X-tree that has $n \geq 2$ leaves and for which at most two interior vertices are incident with more than one leaf (the class of 'caterpillar trees', as in [8]). Then, a simple inductive argument (see also Section 6 below) shows that

$$\mathrm{i}_{\text{path}}(T_n) = \begin{cases} 1, & \text{if } n \equiv 1 (\mathrm{mod} 3), \\ 0, & \text{if } n \not\equiv 1 (\mathrm{mod} 3). \end{cases}$$

holds.

(iii) The smallest trivalent X-tree with $\mathrm{i}_{\text{path}}(T) > 1$ has 9 leaves and is shown in Figure 5.1. In the next section, we describe how to construct trivalent X-trees with arbitrarily large path index.

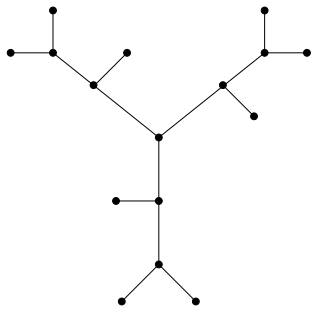

Figure 5.1: The smallest trivalent X-tree having $\mathrm{i}_{\text{path}}(T) > 1$.

6. Computing the Path Index of a Trivalent X-Tree

To compute the path index of a trivalent X-tree T, we first note that any trivalent X-tree that has five or more leaves has a subtree of one of the four types shown in Figure 6.1 where x, y, x', y', z, z' are leaves (this result may be established as follows: Choose two vertices (necessarily leaves) x and v in V_T at maximal distance (≥ 4), consider the unique vertex $u = u_3 \in V_T$ with $d_T(u_3, x) = 3$ and $d_T(u_3, v) = d_T(x, v) - 3$ as well all vertices $w \in V_T$ with $d_T(w, v) = d_T(w, u_3) + d_T(u_3, v)$, and list the resulting cases).

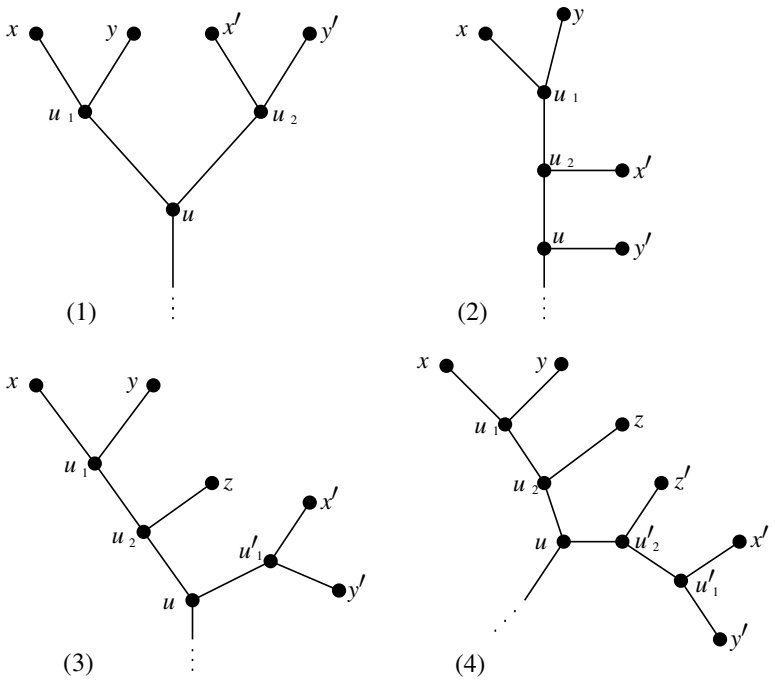

Figure 6.1: The four cases for a 'pendant' subtree in a trivalent X-tree.

Given a subset V' of V_T, let $T\,|\,V' := (V', E_T \cap \binom{V'}{2})$ denote the subgraph of the tree (V_T, E_T) induced by V' that we consider as a (trivalent) X'-tree for $X' := V_1(T\,|\,V')$ in case that subgraph is connected and $V' = V_1(T\,|\,V') \cup V_3(T\,|\,V')$ holds in which case its path index $\mathrm{i}_{\mathrm{path}}(T\,|\,V')$ is well defined. Using the notation introduced in Figure 6.1, this is in particular the case for each of the four types $1, 2, 3, 4$ of subtrees shown in

Figure 6.1 for the respective subsets

$$V(1) := V_T - \{x, y, x', y'\},$$
$$V(2) := V_T - \{x, y, x', y', u_1, u_2\},$$
$$V(3) := V_T - \{x, y, z, x', y', u_1, u_2, u_1'\},$$
$$V(4) := V_T - \{x, y, z, x', y', z', u_1, u_2, u_1', u_2'\}$$

giving rise, for each $j \in \{1, 2, 3, 4\}$, to a trivalent X_j-tree $T|V(j)$, with $X_j := \{u\} \cup (X \cap V(j))$ in case $j = 2, 3, 4$ and $X_j := \{u_1, u_2\} \cup (X \cap V(j))$ in case $j = 1$. Note that $\#X_j = \#X - j - 1$ holds for all $j = 1, 2, 3, 4$.

Theorem 6.1. *For all $j = \{1, 2, 3, 4\}$, we have*

$$\mathrm{i}_{\mathrm{path}}(T) = \mathrm{i}_{\mathrm{path}}(T|V(j)) + \delta_{j,4}.$$

Proof. The proof relies on considering the four cases $j = 1, 2, 3, 4$ described above separately. In each of these cases, we let T' denote the trivalent X_j-tree $T|V(j)$, and we consider a subset $Z \in \ker(T)$. In case $j = 1$, we have either (a) $\{x, y, x', y'\} \subseteq Z$ or (b) $\{x, y, x', y'\} \cap Z = \emptyset$. In subcase (a), put $Z' := Z - \{x, y, x', y'\} \cup \{u_1, u_2\}$ while, in subcase (b), put $Z' := Z$; either way, we have $Z' \in \ker(T')$, since $\#\big([x'', y''']_{T'} \cap [Z']_{T'}\big) \equiv 0 \mod 2$ for all $x'', y'' \in X - \{x, y, x', y'\} \cup \{u_1, u_2\}$. Conversely, if $Y \in \ker(T')$ holds, then there is a unique subset $Z \in \ker(T)$ with $Z' = Y$. Thus, $\mathrm{i}_{\mathrm{path}}(T) = \mathrm{i}_{\mathrm{path}}(T')$.

In case $j = 2$, either (a) $Z \cap \{x, y, x', y'\} = \{x, y, x'\}$ or (b) $Z \cap \{x, y, x', y'\} = \emptyset$ holds. In subcase (a), put $Z' := (Z - \{x, y, y'\}) \cup \{u\}$ while, in subcase (b), put $Z' := Z$; either way, we have $Z' \in \ker(T')$. Conversely, if $Y \in \ker(T')$ holds, there is again a unique subset $Z \in \ker(T)$ with $Z' = Y$. Thus, $\mathrm{i}_{\mathrm{path}}(T) = \mathrm{i}_{\mathrm{path}}(T')$ holds also in this case.

In case $j = 3$, either (a) $Z \cap \{x, y, x', y', z\} = \{x, y, z\}$ or (b) $Z \cap \{x, y, z, x', y'\} = \emptyset$ holds. In subcase (a), put $Z' := Z - \{x, y, z\} \cup \{u\}$ while, in subcase (b), put $Z' := Z$; either way, we have $Z' \in \ker(T')$. Conversely, if $Y \in \ker(T')$ holds, there is once again a unique subset $Z \in \ker(T)$ with $Z' = Y$. Thus, $\mathrm{i}_{\mathrm{path}}(T) = \mathrm{i}_{\mathrm{path}}(T')$ holds also in this case.

Finally, in case $j = 4$, either (a) $Z \cap \{x, y, x', y', z, z'\} = \{x, y, x', y', z, z'\}$ or (b) $Z \cap \{x, y, x', y', z, z'\} = \emptyset$ or (c) $Z \cap \{x, y, x', y', z, z'\} = \{x, y, z\}$ or (d) $Z \cap \{x, y, x', y', z, z'\} = \{x', y', z'\}$ holds. In subcase (a), put $Z' := Z - \{x, y, x', y', z, z\}$, in subcase (b), put $Z' := Z$, in subcase (c), put $Z' := Z - \{x, y, z\} \cup \{u\}$, and in subcase (d), put $Z' := Z - \{x', y', z'\} \cup \{u\}$. In each of these four subcases, we have $Z' \in \ker(T')$. Conversely, if $Y \in \ker(T')$ holds, there are precisely two subsets $Z \in \ker(T)$ with $Z' = Y$. Thus, $\mathrm{i}_{\mathrm{path}}(T) = \mathrm{i}_{\mathrm{path}}(T') + 1$ holds in case $j = 4$. ∎

Clearly, Theorem 6.1 allows us to compute the path index of any trivalent X-tree in a simple recursive fashion as well as to construct all trivalent X-trees with a given number of leaves and a given path index starting from the trivalent X-trees T_2, T_3, and T_4 with exactly 2, 3, or 4 leaves, respectively. Thus, it implies in particular the following noteworthy consequences:

Corollary 6.2. *Suppose T is a trivalent X-tree with $n = n_T$ leaves. Then*

$$\mathrm{i}_{\mathrm{path}}(T) \leq \frac{n+1}{5}.$$

Furthermore, this bound is achieved exactly for $n = 4, 9, 14, 19, 24, \ldots$, i.e., whenever this number $\frac{n+1}{5}$ is an integer, and the trees for which this bound is achieved are exactly those that can be constructed by starting with the trivalent tree T_4 with 4 leaves, and then making repeated application (in any order) of the reverse of operations of type $T \mapsto T\,|\,V(4)$ described above. In particular, the quotient $\frac{\mathrm{i}_{\mathrm{path}}(T)}{n_T}$ is maximized exclusively for T_4 where it attains the value 0.25.

Corollary 6.3. *More generally, given any two non-negative integers n, i with $n \geq 6$ and $i \leq \frac{n+1}{5}$, there exists a trivalent X-tree with n leaves and path index i, and any such tree can be constructed by starting with one of the three trees T_k, $k = 2, 3, 4$ and making repeated application (in any order) of the reverse of operations of type $T \mapsto T\,|\,V(j)$ described above for $j \in \{1, 2, 3, 4\}$ applying the reverse of an operation of type $T \mapsto T\,|\,V(j)$ exactly m_j times, for some non-negative integers m_1, m_2, m_3, m_4 that satisfy the conditions*

(i) $k + 2m_1 + 3m_2 + 4m_3 + 5m_4 = n$,
(ii) $\delta_{4,k} + m_4 = i$.

Corollary 6.4. *If T is a trivalent X-tree with at least 4 leaves, its path index either coincides with 1, or there exists at least one leaf x for which no leaf y of distance 2 exists in which case the path index is always smaller than the number $s(T)$ of such leaves x; in particular, we have $\mathrm{i}_{\mathrm{path}}(T) = 0$ for every trivalent X-tree T with $s(T) = 1$, and the inequality $\mathrm{i}_{\mathrm{path}}(T) \leq \max\big(1, s(T) - 1\big)$ holds for every trivalent X-tree T.*

Corollary 6.5. *The class of trivalent X-trees with $\mathrm{i}_{\mathrm{path}}(T) = 0$ is precisely the class of trees that can be constructed by starting with T_2 or T_3, and then making repeated application (in any order) of the reverse of operations of type $T \mapsto T|V(j)$, where $j \in \{1, 2, 3\}$, described above.*

The enumeration of the isomorphism classes of trivalent X-trees T with $\mathrm{i}_{\mathrm{path}}(T) = 0$ might be an interesting exercise. Similarly, a simple 'structural' characterization of such trees would be desirable, as well as further insight into uniqueness and non-uniqueness of the numbers m_1, m_2, m_3, m_4 discussed in Corollary 6.3.

In particular, it might be of some interest to investigate the structure of the infinite edge-labelled graph whose vertices are the (isomorphism classes of) trivalent trees, with two such vertices T, T' connected by a directed edge labelled by $j \in \{1, 2, 3\}$ from T to T' if (up to isomorphism) T' results from T by an operation of type j, and to characterize those (we expect, finitely many) trivalent trees T for which edges $T \overset{j'}{\to} T'$ and $T \overset{j''}{\to} T''$ exist such that no trivalent tree T''' with edges $T' \overset{j''}{\to} T'''$ and $T'' \overset{j'}{\to} T'''$ exist. For example, define two directed paths

$$T_{(0)} \overset{j_1}{\to} T_{(1)} \overset{j_2}{\to} T_{(2)} \cdots \overset{j_n}{\to} T_{(n)}$$

and

$$T'_{(0)} \overset{j'_1}{\to} T'_{(1)} \overset{j'_2}{\to} T'_{(2)} \cdots \overset{j'_n}{\to} T'_{(n)}$$

to be equivalent if one can be transformed into the other by a repeated exchange of a subpath of the form

$$T_{(\nu-1)} \overset{j_\nu}{\to} T_{(\nu)} \overset{j_{\nu+1}}{\to} T_{(\nu+1)}$$

by the form

$$T_{(\nu-1)} \overset{j_{\nu+1}}{\to} T'_{(\nu)} \overset{j_\nu}{\to} T_{(\nu+1)}.$$

Then we expect that there exists a small constant C such that the number $C(T,e)$ of inequivalent paths leading from any given tree T containing a given fixed edge e to one of the (at most seven!) smallest trivalent subtrees T' of T that still contain the edge e and are not further reducible by any one of the operations of type 1, 2, 3, or 4 without eliminating e is, independently of the choice of T and e, bounded from above by C.

For the sake of completeness, we note that it is, of course, possible to compute $\mathrm{i}_{\mathrm{path}}(T)$ for any X-tree T by using a slightly more general recursive procedure.

7. Further Applications

We continue to consider a fixed finite X-tree T with vertex set V_T and edge set E_T, we put $V_T^* := V_T - X$, and we continue writing $[Z]$ instead of $[Z]_T$ for any subset $Z \in \mathcal{P}_{\mathrm{even}}(X)$ and T for the map $\mathrm{ind}(f_T) := \Delta_T \circ [\ldots] \colon Z \mapsto \Delta_T[Z]$ that is the topic of this note.

7.1. A Short Exact Sequence

Given any vertex in $v \in V_T^*$, the co-boundary $\partial^T v \subseteq E_T$ of v is easily seen to be contained in the kernel of Δ_T. In fact, we can say quite a bit more here by describing what is known in homological algebra as a 'short exact sequence'. To do so, we consider the $\mathbb{F}_2$-linear extension of the map $\partial^T : V_T \to \mathcal{P}(E_T) \colon v \mapsto \partial^T v$ that we will also denote by ∂^T:

$$\begin{aligned} \partial^T \colon \ \mathcal{P}(V_T) &\to \mathcal{P}(E_T), \\ U &\mapsto \partial^T U := \Delta_{u \in U}\, \partial^T u, \end{aligned}$$

and its restriction to $\mathcal{P}(V_T^*)$ that we denote by Δ^T. Obviously, ∂^T maps any subset U of V_T onto the set of all edges $e \in E_T$ for which $\#(e \cap U) = 1$ holds and, as mentioned just above, any vertex $v \in V_T^*$ and, thus, any subset of V_T^* onto an element in the kernel of Δ_T. More generally, we have $\Delta_T(\partial^T U) = \{U \cap X, X - U\}$ for any subset U of V_T and, hence, $\Delta_T(\partial^T A) = \{A, X - A\}$ for every subset A of X.

For the next proposition, recall that a sequence of linear transformations

$$\cdots C_i \overset{\Psi_i}{\to} C_{i+1} \overset{\Psi_{i+1}}{\to} C_{i+2} \cdots$$

is *exact* if $\Psi_i(C_i) = \ker(\Psi_{i+1})$ holds for all applicable i. Furthermore, it is called *a short exact sequence* if, in addition, it is of the form

$$0 \to C_0 \overset{\Psi_0}{\to} C_1 \overset{\Psi_1}{\to} C_2 \to 0.$$

For instance, given any embedding α of one finite set Φ into another finite set Ψ and putting $\Psi^* := \Psi - \alpha(\Phi)$, one has a canonical short exact sequence of $\mathbb{F}_2$-vectorspaces

$$0 \to \mathcal{P}(\Psi^*) \xrightarrow{\alpha^*} \mathcal{S}(\Psi) \xrightarrow{\alpha_*} \mathcal{S}(\Phi) \to 0, \tag{7.1}$$

that is given by the $\mathbb{F}_2$-linear transformations

$$\alpha^* \colon \mathcal{P}(\Psi^*) \to \mathcal{S}(\Psi) \colon U \mapsto \{U, \Psi - \alpha(U)\}$$

and

$$\alpha_* \colon \mathcal{S}(\Psi) \to \mathcal{S}(\Phi) \colon \{A, B\} \mapsto \{\alpha^{-1}(A), \alpha^{-1}(B)\}.$$

In particular, any X-tree T gives rise to a canonical short exact sequence

$$0 \to \mathcal{P}(V_T^*) \xrightarrow{\alpha^T} \mathcal{S}(V_T) \xrightarrow{\alpha_T} \mathcal{S}(X) \to 0, \tag{7.2}$$

given by the linear transformations

$$\alpha^T \colon \mathcal{P}(V_T^*) \to \mathcal{S}(V_T) \colon U \mapsto \{U, V_T - U\}$$

and

$$\alpha_T \colon \mathcal{S}(V_T) \to \mathcal{S}(X) \colon \{A, B\} \mapsto \{A \cap X, B \cap X\}.$$

Here, we want to establish the following, closely related result:

Proposition 7.1. *The sequence*

$$0 \to \mathcal{P}(V_T^*) \xrightarrow{\Delta^T} \mathcal{P}(E_T) \xrightarrow{\Delta_T} \mathcal{S}(X) \to 0 \tag{7.3}$$

is exact.

Proof. Recall first from standard graph theory that, given any finite connected simple graph $G = (V, E)$ with vertex set V and edge set $E \subseteq \binom{V}{2}$, one has a canonical $\mathbb{F}_2$-linear map

$$\begin{aligned} \partial^G \colon \mathcal{P}(V) \to \mathcal{P}(E) \colon U \mapsto \partial^G U &:= \Delta_{u \in U} \, \partial^G u \\ &= \{e \in E \colon \#(e \cap U) = 1\} \\ &= E \cap \{\{u, v\} \in \binom{V}{2} \colon u \in U, v \in V - U\} \end{aligned}$$

that induces an isomorphism

$$\iota_G \colon \mathcal{S}(V) \xrightarrow{\sim} \{F \subseteq E \colon \#(F \cap C) \equiv 0 \mod 2 \text{ for all 'cycles' } C \subseteq E\}$$

from the set $\mathcal{S}(V)$ of splits of the vertex set V of G onto the set of subsets F of E that intersect any 'cycle' C of G in an even number of edges.* We remark in passing that this

* A 'cycle' of G is any subset of E of the form $\{\{v_0, v_1\}, \{v_1, v_2\} \ldots \{v_{k-1}, v_k\}\{v_k, v_0\}\}$ where $v_0, v_1, \ldots, v_k$ are vertices in V and the subsets $\{v_0, v_1\}, \{v_1, v_2\} \ldots \{v_{k-1}, v_k\}, \{v_k, v_0\}$ are edges in E.

can be viewed as a generalization of the rather familiar simple fact that G is bipartite — that is, there exists a split $\{U, W\}$ of V with $E \subseteq \{\{u, w\} : u \in U, w \in W\}$ — if and only if every cycle of G has even cardinality.

In particular, if $G = (V, E)$ is a tree, ι_G maps $\mathcal{S}(V)$ isomorphically onto $\mathcal{P}(E)$. Using this in the special case where $G = (V_T, E_T)$ for some X-tree T and denoting the resulting isomorphism $\mathcal{S}(V_T) \to \mathcal{P}(E_T)$ by ι_T, we see that short exact sequence (7.2) gives rise to another short exact sequence

$$0 \to \mathcal{P}(V_T^*) \to \mathcal{P}(E_T) \to \mathcal{S}(X) \to 0 \tag{7.4}$$

where the map $\mathcal{P}(V_T^*) \to \mathcal{P}(E_T)$ is the map $\iota_T \circ \alpha^T : \mathcal{P}(V_T^*) \to \mathcal{S}(V_T) \to \mathcal{P}(E_T)$ and the map $\mathcal{P}(E_T) \to \mathcal{S}(X)$ is the map $\alpha_T \circ \iota_T^{-1} : \mathcal{P}(E_T) \to \mathcal{S}(V_T) \to \mathcal{S}(X)$.

Thus, all that remains to be observed is that — using the additivity of all the maps involved and representing one-element sets by the single element they contain — $\iota_T(\alpha^T(v)) = \Delta^T(v) = \partial^T(v)$ holds for every $v \subseteq V_T^*$ which is obvious, and that $\alpha_T \circ \iota_T^{-1}(e) = \Delta_T(e) = S_e$ holds for every edge $e = \{u, v\}$ of E which follows from the fact that $\iota_T(\{V_T(u, v), V_T(v, u)\} = \{e\}$ holds for the split of V_T into the two connected components $V_T(u, v) := \{w \in V_T : d_T(w, u) < d_T(w, v)\}$ and $V_T(v, u) := \{w \in V_T : d_T(w, u) < d_T(w, v)\}$ of the graph $(V_T, E_T - \{e\})$. ∎

The relationship between the maps described in this proof is summarized by the commutative diagram shown in Figure 7.1. Note also the short exact sequence (7.3) is related to the maps $\widehat{f_T}$ and ind_{f_T} by the commutative diagram presented in Figure 7.2.

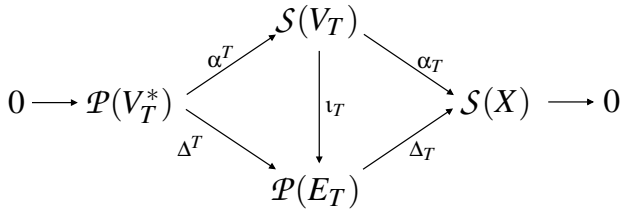

Figure 7.1.

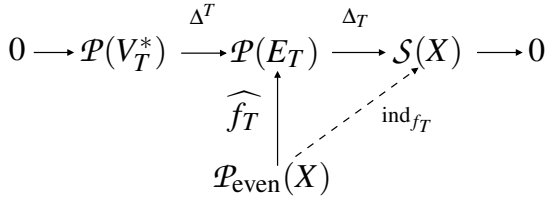

Figure 7.2.

7.2. Hadamard Conjugation

The linear structure we have described so far provides a convenient way to state the main 'Hadamard transform' identity due to Mike Hendy [5]. This is an identity between two apparently quite different polynomials (the original proof of this identity in [5] involves identifying two probabilities; a purely combinatorial proof which fits into the framework developed above can be found in [9]).

Theorem 7.2. *For each edge e in the edge set E_T of the trivalent X-tree T let x_e denote an indeterminant. Then, the identity*

$$\sum_{F\subseteq E_T\,:\,\Delta_T(F)=S}\prod_{e\in F}x_e\prod_{e\in E_T-F}(1-x_e)=\frac{1}{2^{n-1}}\sum_{Z\in\mathcal{P}_{\mathrm{even}}(X)}(-1)^{<Z|S>}\prod_{e\in[Z]}(1-2x_e)$$

holds for any given $S\in\mathcal{S}(X)$.

8. Parsimonious Edge Sets

We continue with the assumptions and notations introduced above. A subset F of E_T is called *parsimonious* if $\#F\leq\#F'$ holds for every subset F' of E_T with $\Delta_T(F')=\Delta_T(F)$, and *strictly parsimonious* if $F'=F$ holds for any such subset F' of E_T with $\#F=\#F'$. In other words, F is parsimonious if $\#F\leq\#(\Delta^T U\,\Delta\,F)$ holds for all subsets U of V_T^*, and strictly parsimonious if $\#F<\#(\Delta^T U\,\Delta\,F)$ holds for every non-empty subset U of V_T^*.

Note that $2\#(\partial^T v\cap F)\leq\#\partial^T v=\deg_T(v)$ holds for any parsimonious subset F of E_T, in particular, no two edges in a parsimonious set $F\subset E_T$ are incident in a trivalent X-tree T.

Our first result in this section describes basic properties of parsimonious sets, using simple arguments that exploit the linearity of the operator Δ_T.

Proposition 8.1.

(i) *Any subset F' of a (strictly) parsimonious subset F of E_T is (strictly) parsimonious.*
(ii) *Any two subsets F',F'' of a parsimonious subset F of E_T with $\Delta_T F'=\Delta_T F''$ coincide.*

Proof. Suppose that $\Delta_T F'=\Delta_T F''$ holds for some subset F'' of E_T, put

$$F'':=F\,\Delta\,F'\,\Delta\,F'',$$

and note that

$$\Delta_T F'''=(\Delta_T F)\,\Delta\,(\Delta_T F')\,\Delta\,(\Delta_T F'')=\Delta_T F$$

as well as $F'''=(F-F')\Delta F''$ (by definition) and, hence,

$$F'''\subseteq(F-F')\cup F''\text{ as well as }\#F'''\leq\#F-\#F'+\#F''$$

holds. Hence, if F is parsimonious, we must have $\#F\leq\#F'''$ and, therefore, $\#F'\leq\#F''$, so F' must be parsimonious, too. Further, if in addition F is either strictly parsimonious and $\#F''\leq\#F'$ holds, or if $F''\subseteq F$ and, hence, also $F'''\subseteq F$ holds, we must have $F'''=F$ and, therefore, $F'=F''$. ∎

It follows that the collection of parsimonious subsets as well as the collection of strictly parsimonious subsets of E_T form a simplicial complex whose combinatorial-topological properties might be a good topic for further investigations.

It is possible to determine in linear time (in $\#X$) by applying the well-used 'Fitch-Hartigan' algorithm ([2, 4]) whether or not F is parsimonious and, if so, whether it is strictly parsimonious. Because this algorithm is recursive, it is of interest to ask for a structural characterization of the collection of subsets F of E_T that are (strictly) parsimonious. We now provide such a characterization (Proposition 8.3), and derive some simple consequences from it.

Definition 8.2. *Given a subset U of V_T, and a subset F of E_T, put*

$$\|U\|_F := \#\big((E_T - F) \cap \partial^T U\big) - \#\big(F \cap \partial^T U\big).$$

Proposition 8.3. *Let F be any subset of E_T. Then the following holds*:

(i) *F is parsimonious if and only if $\|U\|_F \geq 0$ holds for every subset U of V_T^*.*

(i′) *F is parsimonious if and only if $\|U\|_F \geq 0$ holds for every subset U of V_T^* for which $T|U$ is connected. In particular, any connected component U of the graph $(V_T, E_T - F)$ contains at least one element from X (as $U \subseteq V_T^*$ and $\emptyset \neq \partial^T U \subseteq F$ would otherwise hold implying that $\|U\|_F$ is negative).*

(ii) *F is strictly parsimonious if and only if $\|U\|_F > 0$ holds for every non-empty subset U of V_T^*.*

(ii′) *F is strictly parsimonious if and only if $\|U\|_F > 0$ holds for every non-empty subset U of V_T^* for which $T|U$ is connected.*

(iii) *There exists a canonical bijection between the set*

$$\{F' \subseteq E_T : \#F' = \#F \text{ and } \Delta_T F' = \Delta_T F\}$$

and the set

$$\{U \subseteq V_T^* : \|U\|_F = 0\}$$

that is given by associating to each such subset U of V_T^* the symmetric difference $\Delta^T U \,\Delta\, F$ of $\Delta^T U$ and F.

In particular, F is strictly parsimonious if and only if F is parsimonious and $\|U\|_F \neq 0$ holds for every non-empty subset U of V_T^*.

Proof. Note first that

$$\#\Delta^T U = \#\partial^T U = \#\big((E_T - F) \cap \partial^T U\big) + \#\big(F \cap \partial^T U\big) = \|U\|_F + 2\#\big(F \cap \partial^T U\big)$$

holds for every subset U of V_T^* and every subset F of E_T. Thus,

$$\#(\Delta^T U \,\Delta\, F) = \#F + \#\Delta^T U - 2\#(\Delta^T U \cap F) = \#F + \|U\|_F,$$

and, therefore,

$$\#(\Delta^T U \,\Delta\, F) - \#F = \|U\|_F \tag{8.1}$$

holds for all F and U as above. In consequence, the assertions (i) and (ii) follow from the fact that, as observed above, F is parsimonious if and only if $\#(\Delta^T U \Delta F) \geq \#F$ holds

for all $U \subseteq V_T^*$, and F is strictly parsimonious if and only if $\#(\Delta^T U \Delta F) > \#F$ holds for every non-empty subset U subset of V_T^*.

Next, note that

$$\|U \cup U'\|_F = \|U\|_F + \|U'\|_F$$

holds for every subset F of E_T and any two subsets U, U' of V_T^* with

$$\partial^T U \cap \partial^T U' = \emptyset,$$

which 'additivity' readily implies that (i) holds if and only if (i′) holds, and that (ii) holds if and only if (ii′) holds.

This establishes the first four claims while the last one follows also immediately from (8.1). ∎

Proposition 8.3 leads to the following simple sufficient condition for a set of edges to be strictly parsimonious where we denote, for any pair of edges e, e' of T, the 'T-distance' of their 'midpoints' by $d_T(e, e')$, noting that

$$d_T(e, e') = \frac{1}{4} \sum_{u \in e, u' \in e'} d_T(u, u') = 1 + \min(d_T(u, u') \colon u \in e, u' \in e')$$

holds for any distinct edges $e, e' \in E_T$ (and that this defines a (proper) metric on E_T).

Corollary 8.4. *Let F be any subset of E_T. Suppose that, for any two distinct edges $e, e' \in F$, we have $d_T(e, e') \geq 4$. Then F is strictly parsimonious.*

Proof. By Proposition 8.3 (ii′), it suffices to show that, for any subset U of V_T^* for which $T \,|\, U$ is connected, we have $\partial_{E_T - F} U > \partial_F U$. However, this is easily seen to be a simple consequence of the fact that, for any finite tree without vertices of degree 2 and, thus, in particular, for the tree $T' = (\bigcup_{e \in E, e \cap U \neq \emptyset} e, \{e \in E \colon e \cap U \neq \emptyset\})$, one has

$$\#L_1 < \#(V_1(T') - L_1)$$

for any subset L_1 of $V_1(T')$ with $d_{T'}(x, y) \geq 5$ for all $x, y \in L_1$ (and, thus, in particular, for the set $\{v \in V - U \colon \{u, v\} \in F \text{ for some } u \in U\}$) which in turn follows from the fact that

$$\#F'' < \#V_1(T'')$$

holds for every subset F'' of the edge set E'' of a finite tree T'' without vertices of degree 2 for which $e \cap e' = \emptyset$ holds for any two distinct edges e, e' in F'' (applied to a tree T'' that is constructed in an appropriate way from T'). For more details, see [9]. ∎

The condition on F described in Corollary 8.4 is similar to a condition studied by Huber in [6]. Note that if we weaken this condition on F in Corollary 8.4 to $d_T(e, e') \geq 3$, one cannot even guarantee that F is parsimonious as the example provided by Figure 8.1 illustrates. However, if T is a 'caterpillar tree' as described above in Section 5 (ii), it can be shown quite easily using Proposition 8.3 that a subset F of its edge set E_T is strictly parsimonious if and only if $d_T(e, e') \geq 3$ holds for any two distinct edges e, e' in E_T.

Another characterization of parsimonious edge sets (which follows from an appropriate version of Menger's Theorem) is described in the following result.

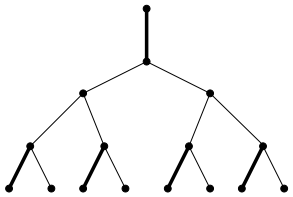

Figure 8.1: Example to illustrate a limitation to extending Corollary 8.4.

Proposition 8.5. *Given a set F of edges of cardinality k and a split $\{A,B\}$ of X with $\Delta_T(F)=\{A,B\}$, the following assertions are equivalent*:

(i) *F is parsimonious, i.e., $\#F \leq \#F'$ holds for any subset F' of F with $\Delta_T(F')=\{A,B\}$,*

(ii) *$k \leq \#F'$ holds for any subset F' of E_T that* separates *A from B, i.e., for which no connected component of the graph $(V_T, E_T - F')$ has a non-empty intersection with both, A and B,*

(iii) *there exist k edge-disjoint paths in T that each have one endpoint in A and one endpoint in B any one of which contains exactly one edge from F,*

(iii′) *there exist k edge-disjoint paths in T that each have one endpoint in A and one endpoint in B any one of which contains at least one edge from F,*

(iv) *there exist k edge-disjoint paths in T all of whose endpoints are elements from X and any one of which contains exactly one edge from F,*

(iv′) *there exist k edge-disjoint paths in T all of whose endpoints are elements from X and any one of which contains at least one edge from F,*

(v) *there exist $2k$ distinct elements $z_1, z_2, \ldots, z_{2k-1}, z_{2k}$ in X such that the edge set $[\{z_1, z_2, \ldots, z_{2k-1}, z_{2k}\}]$ is the disjoint union of the k paths $[\{z_{2i-1}, z_{2i}\}]$, $i=1,\ldots,k$, and each path $[\{z_{2i-1}, z_{2i}\}]$ has a non-empty intersection with F.*

In other words, one can construct all parsimonious edge sets $F \subseteq E_T$ by listing first all even-order subsets Z of X, decomposing the edge set $[Z]$ into a union of edge-disjoint paths X-to-X paths, and then choosing, in all possible ways, one edge in each of these disjoint paths in $[Z]$.

Proof. (i) $\Rightarrow$ (ii): As this is a well-known, yet important fact, we just sketch a proof: Assume that F' is any subset of E_T that separates A from B and consider the connected components of the graph $(V_T, E_T - F')$. There are three types among these connected components, those of A-type that contain some vertex from A, but none from B, those of B-type that contain some vertex from B, but none from A, and those that contain neither a vertex from A nor a vertex from B — as F' was supposed to separate A from B, there can't be any connected components of $(V_T, E_T - F')$ that contains simultaneously vertices from A and from B. Moreover, if no proper subset F'' of F' separates A from B, there cannot be any connected component U of $(V_T, E_T - F')$ that contains neither a vertex from A nor a vertex from B as dropping any one of the edges in $\partial^T U$ would produce a proper subset F'' of F' that would also separate A from B. Similarly, there

can be no edge $e = \{u, v\}$ in F' such that both, u and v, are contained in a connected component of A-type, nor one such that both, u and v, are contained in a connected component of B-type because $F' - \{e\}$ would also separate A from B in this case. Thus, any path from a vertex in A to one in A must contain an even number of edges from F', and any path from a vertex in A to one in B must contain an odd number of edges from F' which readily implies that $\Delta_T F' = \{A, B\}$ must hold in this case. Thus, if F is parsimonious, $k = \#F \leq \#F'$ must hold for any edge set F' of E_T that separates A from B.

(ii) $\Rightarrow$ (iii): One form of Menger's celebrated theorem (cf. [3]) implies that, for any two disjoint subsets of vertices A and B in a graph G (not necessarily a tree), the minimum number of edges that must be deleted from G in order to separate all vertices in A from all vertices in B is precisely the same as the maximum number of edge-disjoint paths that have the property that each path has one endpoint in A and the other endpoint in B. Consequently, if $k = \#F \leq \#F'$ holds for any subset F' of E_T that separates A from B, there must exist a set P of k edge-disjoint paths, each of which connects a vertex from A with a vertex from B and each of which must therefore contain exactly one edge from F.

The implications (iii) $\Leftrightarrow$ (iii$'$), (iii) $\Leftrightarrow$ (iv), and (iv) $\Leftrightarrow$ (iv$'$), and (iii) $\Leftrightarrow$ (v) are trivial.

(iii) $\Rightarrow$ (i): Any subset F' of E_T with $\Delta_T(F') = \{A, B\}$ must have a non-empty intersection with any of those k paths whose existence is asserted in (iii). So, $\#F' \geq k = \#F$ must indeed hold for any subset F' of E_T with $\Delta_T(F') = \{A, B\}$. ∎

We end this section by mentioning how parsimonious edge sets are connected with the 'small parsimony problem' in phylogenetics. In this setting, we have a function $\chi\colon X \to \{0, 1, \ldots, r-1\}$ that we wish to extend to a function $\overline{\chi}$ from V_T into $\{0, 1, \ldots, r-1\}$ so as to minimize the size of the set

$$\mathrm{Ch}(\overline{\chi}) := \{e = \{u, v\} \in E_T : \overline{\chi}(u) \neq \overline{\chi}(v)\}.$$

Any such extension is called a *most parsimonious extension* of χ.

Proposition 8.6. *With T and X as above, consider a map $\overline{\chi}\colon V_T \to \{0, 1, \ldots, r-1\}$ that extends the map $\chi\colon X \to \{0, 1, \ldots, r-1\}$.*

(i) *If $r = 2$, $\mathrm{Ch}(\overline{\chi})$ is a parsimonious set of edges of T if and only if $\overline{\chi}$ is a most parsimonious extension of χ.*
(ii) *If $r > 2$ and $\mathrm{Ch}(\overline{\chi})$ is a parsimonious set of edges of T, then $\overline{\chi}$ is a most parsimonious extension of χ.*

Proof. The first assertion follows immediately from the fact that associating, to each extension $\overline{\chi}$ of χ, the edge set $\mathrm{Ch}(\overline{\chi})$ defines a canonical one-to-one correspondence between extensions $\overline{\chi}$ of χ and subsets F of E_T with $\Delta_T F = \{\chi^{-1}(0), \chi^{-1}(1)\}$.

The second assertion follows from the fact established in Proposition 8.5 that, if $F := \mathrm{Ch}(\overline{\chi})$ is parsimonious, it is an edge set of minimal cardinality separating the two subsets A and B of X with $\Delta_T F = \{A, B\}$ and that $k := \#F$ edge-disjoint paths must exist in T that each have one endpoint in A and one endpoint in B such that any one of them contains exactly one edge from F. This implies that $\chi(a) \neq \chi(b)$ must hold for

any pair a, b of endpoints of those k edge-disjoint paths and that, therefore, $\mathrm{Ch}(\overline{\chi}')$ must contain at least one edge from each of these k paths for each extension $\overline{\chi}'$ of χ implying that $\#\mathrm{Ch}(\overline{\chi}') \geq k = \#\mathrm{Ch}(\overline{\chi})$ must hold for each extension $\overline{\chi}'$ of χ. So, $\overline{\chi}$ must indeed be most parsimonious if $F = \mathrm{Ch}(\overline{\chi})$ is parsimonious. ∎

Acknowledgments. We wish to thank Mike Hendy for a number of valuable discussions over the years that have made clear the relevance that even-ordered sets have in phylogenetic analysis. We also thank the referees for some helpful suggestions.

References

1. J. Felsenstein, Inferring Phylogenies, Sinauer Press, 2004.
2. W.M. Fitch, Towards defining the course of evolution: minimum change for a specific tree topology, Syst. Zool. **20** (1971) 406–416.
3. F. Harary, Graph Theory, Addison-Wesley Publishing Company, 1969.
4. J.A. Hartigan, Minimum mutation fits to a given tree, Biometrics **29** (1973) 53–65.
5. M.D. Hendy, The relationship between simple evolutionary tree models and observable sequence data, Syst. Zool. **38** (1989) 310–321.
6. K.T. Huber, Recovering trees from well-separated multi-state characters, Discrete Math. **278** (2004) 151–164.
7. S. Maclane and G. Birkoff, Algebra, 2nd Ed., Macmillan Publishing Co. Inc., New York, 1979.
8. C. Semple and M. Steel, Phylogenetics, Oxford University Press, 2003.
9. M.A. Steel, Distributions on bicoloured evolutionary trees, Ph.D. Thesis, Massey University, Palmerston North, New Zealand, 1989.
10. C. Tuffley and M.A. Steel, Links between maximum likelihood and maximum parsimony under a simple model of site substitution, Bull. Math. Biol. **59** (3) (1997) 581–607.