THE CODIFYING OF TREE-STRUCTURE

By E. H. NEVILLE

Received 18 November 1952

1. The mathematician's tree is a collection of pairs of elements, conveniently called points. A pair of points is a link, and a collection of links is a tree if any two of the points from which the links are formed can be connected by a chain, and if there are no closed chains. A tree is determined if there is provided a complete schedule of the links composing the tree. Conversely, no description of a tree can be adequate unless a complete schedule is recoverable from the description. An arbitrary schedule of links does not usually describe a tree, and the object of this note is to give canonical schedules for the most general finite tree, and to specify these schedules as economically as possible. The fundamental result established is that the structure of a tree with n points can be encoded in a row of $n-2$ symbols each of which may stand for any one of the points. Cayley's theorem, that the number of different trees with n given points is n^{n-2}, is an immediate corollary.

2. The points of the tree are lettered or numbered without reference to the linkage, and the symbols have a serial order of their own. One point is required to play a special part in the codification; any point of the tree will serve, and in exposition it is convenient to suppose that this is the point to which the last of the symbols is assigned. The special point is called the root of the tree.

A point other than the root is called a free end if there is only one link to which it belongs. An infinite tree need have no free ends, but a finite tree has at least one such point. Let $A_1, A_2, \ldots, A_\alpha$ be free ends of a tree T—not necessarily all of the free ends —and let $A_1a_1, A_2a_2, \ldots, A_\alpha a_\alpha$ be the links to which these ends contribute; the α points $A_1, A_2, \ldots, A_\alpha$ are all distinct, but any number of their companions $a_1, a_2, \ldots, a_\alpha$ may coincide. If A denotes the set of links $A_1a_1, A_2a_2, \ldots, A_\alpha a_\alpha$, then $T-A$ is the set obtained by removing these links from T, and $T-A$ is itself a tree. The point A_r belongs to no link except $A_r\alpha_r$ and therefore does not occur in $T-A$. It follows, since A_r is connected with the points which do occur in $T-A$, that a_r occurs in $T-A$. Thus the points $A_1, A_2, \ldots, A_\alpha$ are the points which occur in T and not in $T-A$. Now let $B_1, B_2, \ldots, B_\beta$ be free ends of the tree $T-A$ and subtract a set of links $B_1b_1, B_2b_2, \ldots, B_\beta b_\beta$ to produce a tree $T-A-B$, take a set $C_1, C_2, \ldots, C_\gamma$ of free ends of this tree and form a tree $T-A-B-C$, and so on. Ultimately a tree is reached which consists of μ points $M_1, M_2, \ldots, M_\mu$, each linked directly to the root, and we can write a structural formula

$$T = A+B+C+\ldots+M,$$

in which, reading from left to right, each link except the last is formed of one point which occurs for the last time and one point which is to occur again later.

In the process of dissection, each point of T, except the root which we may denote now by N_0, plays once and only once the part of a free end. That is, the points

$$A_1, A_2, \ldots, A_\alpha; \quad B_1, B_2, \ldots, B_\beta; \quad \ldots; \quad M_1, M_2, \ldots, M_\mu; \quad N_0$$

https://doi.org/10.1017/S030500410002853X Published online by Cambridge University Press

are the points of the tree, and it follows that a tree with n points has $n-1$ links. The dissection does not impose any order within the groups $A, B, \ldots, M$, and we standardize the notation by requiring the free points in each group to have the order which they have in the complete series of points.

3. Consider now the relation of the set

$$(\mathrm{U}) \quad A_1, A_2, \ldots, A_\alpha; \quad B_1, B_2, \ldots, B_\beta; \quad \ldots; \quad M_1, M_2, \ldots, M_\mu$$

to the set

$$(\mathrm{L}) \quad a_1, a_2, \ldots, a_\alpha; \quad b_1, b_2, \ldots, b_\beta; \quad \ldots; \quad m_1, m_2, \ldots, m_\mu.$$

A point other than N_0 which is *not* a free end of T occurs in at least two of the links $A_1a_1, \ldots, M_\mu m_\mu$, and therefore occurs at least twice in the two sets (U), (L) taken together; since every point except N_0 occurs once and only once in (U), a point other than N_0 which is *not* a free end necessarily occurs in (L), and a free end occurs in (U) and not in (L). *Whatever the principle of dissection, the free ends of T are the points of T which do not occur among the points*

$$a_1, a_2, \ldots, a_\alpha; \quad b_1, b_2, \ldots, b_\beta; \quad \ldots; \quad m_1, m_2, \ldots, m_\mu.$$

It follows that the free ends of $T-A$ are the points of $T-A$ which do not occur among

$$b_1, b_2, \ldots, b_\beta; \quad \ldots; \quad m_1, m_2, \ldots, m_\mu,$$

or in other words are the points of T which do not occur among

$$A_1, A_2, \ldots, A_\alpha; \quad b_1, b_2, \ldots, b_\beta; \quad \ldots; \quad m_1, m_2, \ldots, m_\mu,$$

that the free ends of $T-A-B$ are the points of T which do not occur among

$$A_1, \ldots, A_\alpha; \quad B_1, \ldots, B_\beta; \quad c_1, \ldots, c_\gamma; \quad \ldots; \quad m_1, \ldots, m_\mu,$$

and so on.

If the *composition* of the subsets in the row (U) is known, then since the order within each subset is known also, the correlation between the rows (U), (L), which is the structure of the tree, is implicit in the order of the elements in the row (L). But we have just seen that the composition of the subsets in (U) can be inferred from the composition of the subsets in (L). The problem of encoding the structure is therefore to define a sequence of operations in which the punctuation of (L) can be inferred and need not be recorded. In this form the problem can be solved along two different lines: the subsets may be required to be as small as possible, or to be as large as possible. In the latter case, no question of choice arises, since we are taking at each stage all the free ends available; in the former case we are taking away the links one at a time, and we have to lay down a rule determining the free end to be chosen.

4. Perhaps the simplest rule for dissecting a tree with a view to encoding its structure is to remove the links one at a time, choosing always the link whose free end has the smallest number for its symbol. The free ends, in order of removal, form a row

$$(\mathrm{U}_1) \quad ABC \ldots M,$$

in which each of the numbers $1, 2, \ldots, n-1$ occurs once and only once, and the points associated with them as they are removed form a row of $n-1$ numbers,

$$(\mathrm{L}_1) \quad abc \ldots m,$$

of which all we can say is that the last element is n and each of the elements is one of the numbers $1, 2, \ldots, n$.

https://doi.org/10.1017/S030500410002853X Published online by Cambridge University Press

Given the row (L_1), we can identify first A as the smallest of the numbers $1, 2, \ldots, n-1$ which does not occur among $abc \ldots m$, then B as the smallest which does not occur among $Abc \ldots m$, then C as the smallest which does not occur among $ABc \ldots m$, and so on. The set of links $Aa, Bb, Cc, \ldots, Mm$ is a tree determined uniquely from (L_1).

For example, with seven points, 7 3 7 3 3 7 specifies the tree whose links are

$$17, \quad 23, \quad 47, \quad 53, \quad 63, \quad 37 \quad (\textit{Example } 1),$$

and 7 3 6 3 7 7 specifies the tree whose links are

$$17, \quad 23, \quad 46, \quad 53, \quad 37, \quad 67 \quad (\textit{Example } 2).$$

5. The dissection link by link may be varied to follow more closely the structure of the tree. The first free end to be removed is again the free end with the smallest symbol; this point A^1 is linked to a definite point a^1. Instead of taking next the first free end of the tree $T - A^1a^1$, let us take a^1 if a^1 is in fact a free end of $T - A^1a^1$, that is, if a^1 occurs in only two links of T; if a^1 is not available, we revert to the principle used earlier. Writing A^2 for a^1 if a^1 is a free end of $T - A^1a^1$, and so on, we are removing from T a chain A_* consisting of the links $A^1A^2, A^2A^3, \ldots, A^\alpha a^\alpha$, where A^r is a free end of

$$T - A^1A^2 - A^2A^3 - \ldots - A^{r-1}A^r,$$

but a^α is not a free end of $T - A_*$. We make a fresh start with the tree $T - A_*$, removing first the free end B^1 with the smallest symbol, and ultimately we have

$$T = A_* + B_* + C_* + \ldots + M_*,$$

where each of the components $A_*, B_*, \ldots, L_*$ is a chain whose last point comes again in a later member of the set $A_*, B_*, \ldots, L_*, M_*$, and M_* is a chain whose last point is the root n.

Given the order of precedence of the free ends, the rows

$$(U_2) \quad A^1A^2 \ldots A^{\alpha-1}A^\alpha B^1B^2 \ldots B^{\beta-1}B^\beta C^1 \ldots M^{\mu-1}M^\mu,$$

$$(L_2) \quad A^2A^3 \ldots A^\alpha a^\alpha B^2B^3 \ldots B^\beta b^\beta C^2 \ldots M^\mu m^\mu$$

are implicit in the structure of the tree. Conversely, in the row (L_2), the term a^α is recognizable as the first term whose symbol is repeated later in the row, the term b^β is recognizable by the same property in the row $B^2B^3 \ldots m^\mu$, and so on. Since $A^1, B^1, \ldots, M^1$ are known as the points which do not occur in (L_2), the row (U_2) is recoverable from (L_2), the only condition imposed on (L_2) being that the last term m^μ is the root. By its formation from (L_2), the row

$$(V_2) \quad A^2A^3 \ldots A^\alpha B^2B^3 \ldots B^\beta C^2 \ldots M^\mu m^\mu$$

includes each point of (L_2) once and only once, and therefore, since the number of terms in (L_2) is $n-1$, the number of the points $1, 2, \ldots, n$ which do not occur in (L_2) exceeds by unity the number of the terms which are dropped from (L_2) to make (V_2). That is, since the terms dropped are $a^\alpha, b^\beta, \ldots, l^\lambda$, but not m^μ, the number of the terms $a^\alpha, b^\beta, \ldots, l^\lambda, m^\mu$ is the same as the number of the points $1, 2. \ldots, n$ which are absent from (L_2), and this is a direct verification that any row of $n-1$ symbols can be read as the (L_2) of a tree with n points if the last symbol in the row is the symbol for the root.

In the first example in § 4, the free ends are 1, 2, 4, 5, 6, the chains are 17; 23; 47; 54; 63, 37, and (L_2) is 7 3 7 3 3 7, identical, as it happens, with (L_1). In the second example, the free ends are 1, 2, 4, 5, the chains are 17; 23; 46, 67; 53, 37, and the specification (L_2) is 7 3 6 7 3 7.

https://doi.org/10.1017/S030500410002853X Published online by Cambridge University Press

6. The dissection in which all the free ends of T are in the set of links A, all the free ends of $T-A$ in the set of links B, and so on, gives rows

$$(U_3) \quad A_1 A_2 \dots A_\alpha B_1 B_2 \dots B_\beta C_1 \dots M_\mu,$$
$$(L_3) \quad a_1 a_2 \dots a_\alpha b_1 b_2 \dots b_\beta c_1 \dots m_\mu,$$

in which $A_1, A_2, \dots, A_\alpha$ are the points of T which do not occur in (L_3), $B_1, B_2, \dots, B_\beta$ are the points which occur among $a_1, a_2, \dots, a_\alpha$ but not later in (L_3), $C_1, C_2, \dots, C_\gamma$ are the points which occur among $b_1, b_2, \dots, b_\beta$ but do not occur again among $c_1, c_2, \dots, m_\mu$, and so on. Since *all* the available points are taken at each stage, the number α is implicit in the composition of (L_3), the number β is implicit, when α is known, in the relation of the first α terms of (L_3) to the remainder of the row, the number γ is implicit when α and β have been found, and so on. In other words, a row $p_1 p_2 \dots p_{n-1}$ subject only to the condition that p_{n-1} is n can be punctuated as the row (L_3) of a tree, and the tree can then be reconstructed.

In our first example, the first set of free ends is 1, 2, 4, 5, 6, and the first *fringe*, as it may be called, consists of the links 17, 23, 47, 53, 63; the first section of L_3 is 7 3 7 3 3. When the fringe is removed, there remains only the link 37, of which 3 is the free end. Thus the rows (U_3), (L_3) are 1 2 4 5 6 3 and 7 3 7 3 3 7. In the second example, the first fringe is 17, 23, 46, 53; when this is removed, there remain the two links 37, 67, and since 7 is the root, 3 and 6 are both free ends, and the two links form the second fringe. Thus (U_3), (L_3) are 1 2 4 5 3 6 and 7 3 6 3 7 7. In reconstruction of Example 2 from (L_3), the points missing from (L_3) are 1, 2, 4, 5, and these are linked, in this order, with the first four points of (L_3), namely, 7, 3, 6, 3; since the remainder of (L_3) consists only of the root 7, the other points in the set 7 3 6 3 are linked with 7.

7. Very simple examples show inadequately the processes of analysis and reconstruction. For the tree shown in the diagram, the three rows (L_1), (L_2), (L_3) are

(L_1)

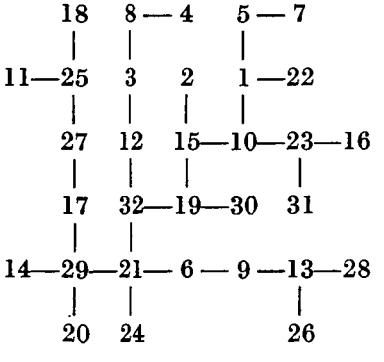
15	8	5	1	3	12	25	32	29	23	25	29	1	10	21			
		27	13	17	29	13	9	6	21	21	32	19	23	10	15	19	32

(L_2)

15	8	3	12	32	5	1	25	29	23	25	27	17	29	29			
		21	1	10	21	13	13	9	6	21	32	19	23	10	15	19	32

(L_3)

15	8	5	25	29	23	25	29	1	21	13	13	19	23	1			
		3	9	10	27	10	12	6	17	21	15	32	29	19	21	32	32



https://doi.org/10.1017/S030500410002853X Published online by Cambridge University Press

From each of these rows the corresponding row (U) can be deduced, and therefore any one of the rows is implicitly a specification of the structure of the tree.

8. The reconstruction of a tree from the row (L) given by one or other of our processes of dissection makes use of the root at the end of the row, but if the row is wanted only as a code for the structure, this invariable part of the symbol is superfluous. Hence *the structure of a tree with n given points is expressible by a row of* $n-2$ *symbols, each of which may be the symbol of any one of the points.*

If C is a row of $n-2$ points which encodes the structure of a tree T, the row (L) is completed by the addition to C of the symbol for the root, and the row (U) consists of the symbols for the other points of the tree, each symbol occurring once and only once in the row (U). Hence the symbol of every point of the tree occurs once more in the two rows (U), (L) together than in the symbol C. But in the two rows together a point occurs once for every link to which it belongs. Hence *a point whose symbol occurs* p *times in a row of symbols by which the structure of a tree is encoded is linked to* $p+1$ *other points of the tree.*

THE UNIVERSITY
READING

https://doi.org/10.1017/S030500410002853X Published online by Cambridge University Press