JOURNAL OF COMBINATORIAL THEORY 11, 105–119 (1971)

Comparison of Labeled Trees with Valency Three

D. F. ROBINSON

University of Canterbury, Christchurch, New Zealand

Communicated by J. M. Hammersley

Received August 27, 1969

This paper is concerned with the sets of trees with all interior vertices of valency three and the terminal vertices labeled from a given set. Some of the properties of such trees are investigated and a "crossover" operation defined, leading to a distance between trees. Results towards the calculation of the probability of two trees being at most a certain distance apart are produced. Finally a connection with manipulation in semigroups is pointed out.

1. INTRODUCTION: LABELED TREES

One of the problems of numerical taxonomy (formal methods of classification) is the reconstruction of the way in which creatures evolved from a common ancestor. It would appear that no satisfactory methods of reconstruction yet exist: this paper is intended as a contribution to this theory by indicating how methods can be assessed. As an illustration we will discuss a biological example, though other applications, such as the copying of manuscripts, also come to mind. No knowledge of biology is involved.

Consider the following creatures: cat, dog, seal, horse, ostrich, goose, whale, and platypus, together with their common ancestor, presumably some kind of reptile. The way they evolved may have been as in Fig. 0(i), or less probably as in Fig. 0(ii).

These two trees certainly have something in common, though they are not identical. Indeed if one drew a tree at random and labeled the terminal vertices with these names at random it is most unlikely that it would bear as much resemblance to the tree in Fig. 0(i) as that in Fig. 0(ii) does. This paper begins the process of assigning values to this probability. We confine our attention to trees in which all the interior vertices have valency three, which is probably no disadvantage in practice; and though in the above example it seems natural to work in terms of rooted trees,

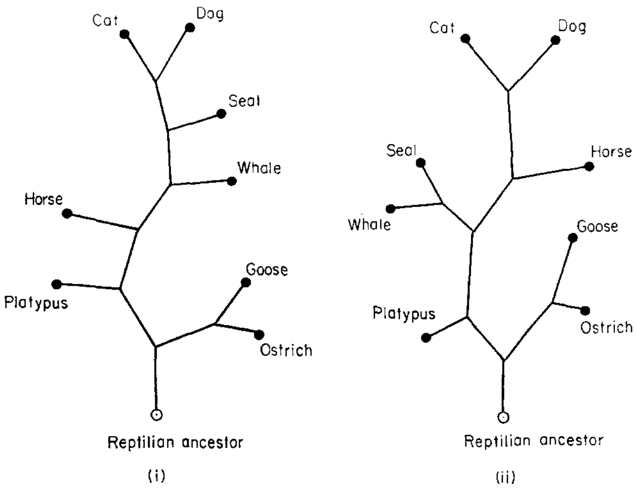

FIGURE 0

the root being the common ancestor, it turns out to be simpler mathematically to ignore any special nature of the root.

Let n be any integer greater than 1, and S any set of n members (labels). The nature of these labels is immaterial. We define $\mathscr{T}_n$ to be the set of all labeled trees T satisfying the conditions:

1. T has n terminal vertices.
2. All interior vertices of T have valency 3.
3. The terminal vertices of T are labeled with the members of S, each label being used just once.

Two trees are considered identical it there is a label-preserving isomorphism between them.

Several of the proofs in this paper employ induction on the number of terminal vertices, employing operations of "deleting" or "inserting" a terminal vertex.

Let T be any tree in $\mathscr{T}_n$, t any terminal vertex, u the vertex adjacent to it and v, w the other vertices adjacent to u. Then the tree T_1 "obtained from T by deleting t" has as vertices the vertices of T except for t and u, and as edges the edges of T except for (t, u), (u, v), and (u, w), and in addition the edge (v, w). Insertion is the inverse of deletion. Given any tree T in $\mathscr{T}_n$ and any edge (v, w) we construct a new tree T_2 which has the vertices of T together with two new vertices t and u, and the edges of T with the exception of (v, w) and the addition of (t, u), (u, v), and (u, w). (See Fig. 1.) These operations are defined basically on unlabeled graphs;

where the labels matter appropriate assignments must be made. The words "insertion" and "deletion" in this paper will always signify the above operations.

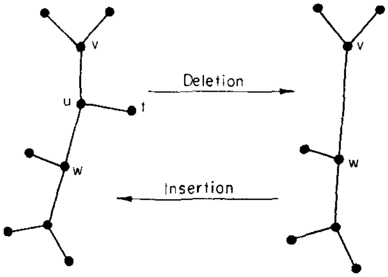

FIGURE 1

THEOREM 1. *Let T belong to $\mathscr{T}_n$ with $n \geqslant 3$. Then T has $(n - 2)$ interior vertices, $(2n - 3)$ edges, and $(n - 3)$ interior edges (edges joining two interior vertices).*

The proof is immediate using deletion and induction. The labels are irrelevant for this theorem, as for the following one.

THEOREM 2. *Let T be a member of $\mathscr{T}_n$. Then the diameter of T is at most $(n - 1)$.*

For each path contains at most two terminal edges and at most all $(n - 3)$ interior edges.*

THEOREM 3. *There are*

$$1.3.5 \cdots (2n - 5)$$

distinct trees in $\mathscr{T}_n$, where $n \geqslant 3$.

Let $T_1 \in \mathscr{T}_{n-1}$, the labels being, for example, $1, 2, \ldots, n - 1$. Then T_1 has $2n - 5$ edges and a new vertex labeled "n" may be inserted into any of these.

Any tree in $\mathscr{T}_n$ can be obtained in this way from precisely one tree in $\mathscr{T}_{n-1}$, insertion into different edges yielding different trees. Thus if $\mathscr{T}_r$ contains t_r members,

$$t_n = (2n - 5)\, t_{n-1},$$

and there is a single tree when n is 2 or 3.

* The author thanks the reviewer for this proof.

2. Crossovers

We now introduce an operation on these trees. Let $\alpha = (a, b)$ be any interior edge in a labeled tree T_0 in $\mathscr{T}_n$. Then α divides the rest of T_0 into four subtrees A, B, C, D, two rooted at each of a and b.

By re-pairing these subtrees we can obtain two new trees T_1 and T_2 (Fig. 2) also in $\mathscr{T}_n$. The operation of deriving T_1 or T_2 from T_0 is termed a *crossover*, and α is known as the *axis*.

We note in Fig. 2 that it is also possible to form T_2 from T_1 by a crossover on the same axis; this will form part of Theorem 9. We also note that the operation is invertible: if T_1 can be obtained from T_0 by a single crossover, T_0 can be obtained from T_1 in the same way.

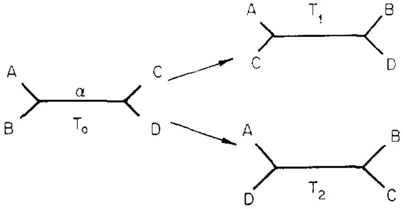

Figure 2

The symbol α used for the axis has the status of a label for we apply it also to the edge after the crossover has been performed. We similarly transfer to T_1 the labels for other edges of T_0.

It takes two crossovers to convert the tree in Fig. 0(i) to that in Fig. 0(ii). If we can find out what proportion of trees in $\mathscr{T}_9$ are within two crossovers of a given tree, we can give a numerical value to the degree of resemblance of these two trees.

Theorem 4. *Let T_1, T_2 be two trees in $\mathscr{T}_n$, each of which can be obtained from the same tree T_0 in some $\mathscr{T}_{n-1}$ by the insertion of a single vertex, having the same label in each case. Then T_1 may be transformed to T_2 by a sequence of at most $n - 3$ crossovers.*

Let T_1, T_2 first be two trees in which the insertions are into edges in T_0 with the vertex w in common. Then the branches from w divide T_0 into three subtrees A, B, C. With the inserted vertex t forming a single-vertex subtree D in T_1 and T_2, they have the same four subtrees A, B, C, D disposed in such a way that each can be obtained from the other by a single crossover.

In the general case there is a path in T_0 containing both the edges into which the insertions are made. A corresponding sequence of crossovers, working along the path, converts T_1 to T_2, and the number of steps required is one less than the length of the path. We have already shown in Theorem 2 that T_0 has diameter at most $n - 2$. So at most $n - 3$ crossovers are required.

This theorem acts as a lemma for the general result:

THEOREM 5. *Let T and U be any two trees in $\mathcal{T}_n$. Then T can be transformed into U by a sequence of crossovers.*

We proceed by induction on n. Suppose the theorem is true for all $n \leqslant k$. Let T and U have $k + 1$ terminal vertices. By Theorem 1, T has only $k - 1$ interior vertices. Hence T has at least one pair of terminal vertices adjacent to the same interior vertex. Let p, q be the labels of such a pair of terminal vertices. We construct new trees T_1, U_1 by deleting from T and U the vertices labeled p. Then T_1 and U_1 have each k terminal vertices and T_1 can be converted to U_1 by a sequence of crossovers, by the induction hypothesis. Now let U_2 be the tree obtained from U_1 by inserting a vertex labeled p into the edge incident with a vertex labeled q. U_2 is in $\mathcal{T}_n$ and each crossover of the sequence which sends T_1 to U_1 corresponds to a crossover in a sequence from T to U_2, preserving throughout the relationship between the vertices labeled p and q. U_2 may be converted to U by a further (possibly trivial) sequence of crossovers, by Theorem 4. Thus T can be converted to U.

The proof is completed by the observation that the assertion is true when $n = 3$, there being only one such tree. It is also true when $n = 2$, but this cannot be used to start the induction process.

We can make further use of this construction to estimate the maximum over all pairs T, U of trees in $\mathcal{T}_n$ of the number of crossovers required to convert T into U. Let m_n be this maximum. Then the above construction requires at most m_{n-1} steps from T to U_2 and by Theorem 4 at most $n - 3$ from U_2 to U. Thus

$$m_n \leqslant m_{n-1} + (n - 3).$$

As $m_3 = 0$,

$$m_n \leqslant \tfrac{1}{2}(n - 2)(n - 3).$$

This value is exact for $n = 2, 3, 4, 5$.

Let T_1, T_2 be two members of $\mathcal{T}_n$ and define $s(T_1, T_2)$ to be the minimum number of terms in any sequence of crossovers converting T_1 to T_2. Then it can easily be established that $(\mathcal{T}_n, s)$ is a metric space which we may depict as a graph whose vertices are the trees and whose edges join trees which differ by a single crossover.

THEOREM 6. *The number of distinct labeled trees obtainable from a given labeled tree with $n \geqslant 4$ terminal vertices by a single crossover is $2n - 6$.*

Given an interior edge α this functions as the axis of two distinct crossovers. Distinct crossovers give rise to distinct trees. The tree has $n - 3$ interior edges. Thus the number of trees obtainable by a single crossover is $2n - 6$.

THEOREM 7. *Let α, β be a pair of distinct interior edges in a tree T in $\mathscr{T}_n$, having no vertex in common. Then T gives rise to four distinct trees at distance $s = 2$ using α and β in either order as the axes of the crossovers.*

We simply observe the results of the eight possible crossovers with α and β as axes on the tree in Fig. 3. The effect of using α then β is the same as using β then α.

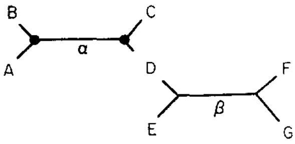

FIGURE 3

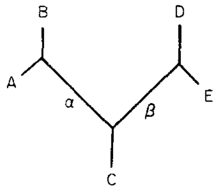

FIGURE 4

THEOREM 8. *Let α, β be a pair of interior edges with common vertex w in a tree T in $\mathscr{T}_n$. Then T gives rise to eight distinct trees at distance 2 as a result of crossovers with axes α and β.*

Again we consider the outcomes of the eight possibilities, this time with reference to the tree in Fig. 4. In this case using α then β is not the same as using β then α.

THEOREM 9. *Let U_1, U_2 be the two trees obtained by single crossovers from the tree T, using the same edge α as axis. Then U_2 can be reached from U_1 by a single crossover. Conversely given any tree U_1 which can be obtained from T by a single crossover there is a unique tree U_2 obtainable from both T and U_1 by single crossovers.*

The first part of the theorem has already been remarked. The second is a corollary of Theorems 7 and 8, for a pair of crossovers with distinct axes never gives the effect of a single crossover.

THEOREM 10. *Let w_2 be the number of trees at distance 2 from a given tree T in $\mathscr{T}_n$. Then*

$$2n^2 - 10n + 8 \leqslant w_2 \leqslant 2n^2 - 8n \textit{ if } n \textit{ is even,}$$

$$2n^2 - 10n + 8 \leqslant w_2 \leqslant 2n^2 - 8n - 2 \textit{ if } n \textit{ is odd.}$$

Each tree at distance 2 arises from a pair of distinct edges in T. Considering the forms obtained from those in Figs. 3 and 4, for example that in Fig. 5, we can see that such a tree could not arise by crossovers with axes other than α and β, for any such axes would separate parts of some of the subtrees $A, B, \ldots, G$. Thus each tree at distance 2 from T arises by crossovers in a unique pair of edges.

If there are x_{II} pairs of edges of the kind mentioned in Theorem 7 and x_V of the kind in Theorem 8,

$$w_2 = 4x_{II} + 8x_V = 4(x_{II} + x_V) + 4x_V. \tag{1}$$

On the other hand, since there are $n - 3$ interior edges,

$$x_{II} + x_V = \tfrac{1}{2}(n - 4)(n - 3). \tag{2}$$

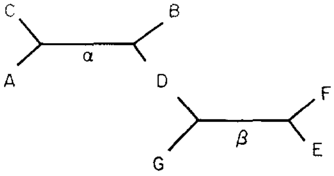

FIGURE 5

Now the interior edges of T form a tree J in which no vertex has valency greater than 3, but in which vertices of valency 2 may occur. J has $n - 2$ vertices: let there be v_i of valency i, where $i = 1, 2, 3$. Then

$$v_1 + v_2 + v_3 = n - 2,$$

$$v_1 + 2v_2 + 3v_3 = 2(n - 3),$$

since the sum of the valencies is twice the number of edges. From these equations,

$$v_3 = v_1 - 2, \qquad v_2 = n - 2v_1.$$

Thus

$$2 \leqslant v_1 \leqslant n/2,$$

the upper bound being exact if n is even and replaceable by $\frac{1}{2}(n-1)$ if n is odd. (Such trees exist.)

We also see that each vertex of valency 2 contributes one pair to x_V, while each vertex of valency 3 contributes three pairs. Thus

$$x_V = v_2 + 3v_3 = v_1 + n - 6,$$

so that, by (1) and (2),

$$\begin{aligned} w_2 &= 2(n-3)(n-4) + 4(v_1 + n - 6) \\ &= 2n^2 - 10n + 4v_1 . \end{aligned}$$

When the bounds for v_1 are inserted the theorem is established.

By streamlining the above approach we can extend this method to a discussion of the number of trees at distance three. Let J again be the tree of interior vertices and edges of T, and let $\sigma = (\alpha, \beta,..., \theta)$ be any finite sequence of edges from J. Then we define $T\sigma$ to be the set of all trees which can be reached by crossovers with axes $\alpha, \beta, \gamma,..., \theta$ in that order. If ω is the empty sequence, $T\omega = \{T\}$, and, if α is any single edge, $T(\alpha)$ has two members. We also know (Theorem 9) that $T(\alpha, \alpha) = T(\alpha)$ and (Theorems 7, 8) that, if β is another edge, $T(\alpha, \beta) = T(\beta, \alpha)$ if and only if α and β have no vertex in common.

Every tree at distance 3 from T is a member of some $T(\alpha, \beta, \gamma)$, but of no $T(\kappa, \lambda)$ for choice of edges $\alpha, \beta, \gamma, \kappa, \lambda$ in J. We must first distinguish the types of triplets $\{\alpha, \beta, \gamma\}$ which occur and then determine the number of distinct sets $T\ (\alpha, \beta, \gamma)$ in each case.

There are four cases when α, β, and γ are all distinct, to which we give the mnemonic symbols *III*, *Y*, *VI*, and *N* (see Fig. 6, where the edges

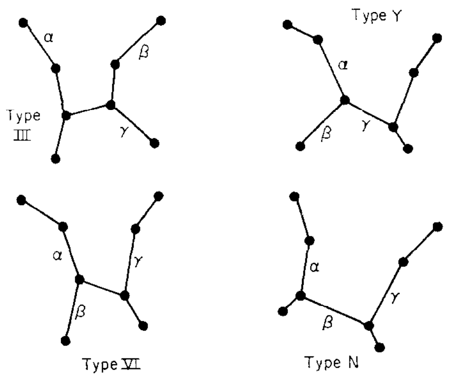

FIGURE 6

shown are those of J). All that is significant is whether or not two edges have a vertex in common. Each set $T(\alpha, \beta, \gamma)$ has eight members.

There are also two types, *II*, *V*, in which an edge is repeated. It follows from Theorem 9 that $T(\alpha, \alpha, \alpha) = T(\alpha) \cup \{T\}$.

In type *III*, crossovers in α, β and γ commute so that $T(\alpha, \beta, \gamma) = T(\alpha, \gamma, \beta) = T(\beta, \alpha, \gamma)$, etc. There are thus precisely eight trees at distance 3 arising from a given set $\{\alpha, \beta, \gamma\}$ of type *III* of edges in J.

In type *VI*, γ commutes with each of α and β, but they do not commute between themselves, so that we have two distinct sets of eight,

$$T(\alpha, \beta, \gamma) = T(\alpha, \gamma, \beta) = T(\gamma, \alpha, \beta)$$

and

$$T(\beta, \alpha, \gamma) = T(\beta, \gamma, \alpha) = T(\gamma, \beta, \alpha).$$

In type *N*, α and γ commute, so that $T(\alpha, \gamma, \beta) = T(\gamma, \alpha, \beta)$. We cannot say that $T(\beta, \alpha, \gamma) = T(\beta, \gamma, \alpha)$, however, for one of $T(\beta)$ has $\{\alpha, \beta, \gamma\}$ in the form of a "*Y*" so α and γ no longer commute. We find that $T(\beta, \alpha, \gamma)$ and $T(\beta, \gamma, \alpha)$ have four trees in common. The other sets of eight are all distinct, so that in all a triplet of type *N* gives rise to 36 trees.

In type *Y*, none of α, β, γ commute at first. But $T(\alpha)$, for instance, has both trees with β, α, γ as a triplet of type *N*, in that order. Hence $T(\alpha, \beta, \gamma) = T(\alpha, \gamma, \beta)$ and the other axes may be treated similarly. Thus a triplet of type *Y* yields 24 trees.

In type *II* and type *V*, there are six possible sequences from $\{\alpha, \beta\}$, namely› (α, β, α), (β, α, β), (α, α, β), (β, α, α), (α, β, β), (β, β, α). In both types all except the first two can be reduced to $T(\alpha, \beta)$ or $T(\beta, \alpha)$, yielding no trees at distance 3. In type *II*, $T(\alpha, \beta) = T(\beta, \alpha)$ also allows the first two sequences to be reduced. On the other hand, in type *V*, these first two are not identical or reducible so that it appears that there are sixteen trees at distance 3, namely, eight each from $T(\alpha, \beta, \alpha)$ and $T(\beta, \alpha, \beta)$. But when we look further we find that each of these sets has only six members, and that four of $T(\alpha, \beta, \alpha)$ are also in (and make up) $T(\beta, \alpha)$, and similarly with α and β interchanged. Finally we discover that the remaining two members of $T(\alpha, \beta, \alpha)$ are the same as the remaining members of $T(\beta, \alpha, \beta)$, so that $\{\alpha, \beta\}$ gives rise to just two trees at distance 3.

If x_{III}, x_Y, etc. are the numbers of triplets of type *III*, *Y* etc. in J, and w_3 is the number of trees at distance 3 from T we have established:

THEOREM 11.

$$w_3 = 8x_{III} + 24x_Y + 16x_{VI} + 36x_N + 2x_V .$$

Our next task is to find x_{III}, x_Y, etc. and appropriate bounds on them so that we can estimate at least a maximum value for w_3.

We know that

$$x_{III} + x_Y + x_{VI} + x_N = (n-3)(n-4)(n-5)/6,$$

for that is the number of triplets obtainable from the $(n-3)$ edges of J.

THEOREM 12. *If v_3 is the number of vertices valency 3 in J,*

$$x_Y = v_3 .$$

Further,

$$x_Y \leqslant \tfrac{1}{2}(n-4) \qquad \text{for} \quad n \geqslant 4,$$
$$x_Y = 0 \qquad \text{for} \quad n = 2, 3.$$

The first part is obvious, since each vertex valency 3 is the "center" of precisely one "Y." The second part follows from the calculations in the proof of Theorem 10.

THEOREM 13. Let y_{rs} $(1 \leqslant r \leqslant s \leqslant 3)$ *be the number of edges in J joining a vertex valency r to a vertex valency s. Then*

$$x_N = y_{22} + 2y_{23} + 4y_{33} .$$

Further,

$$\begin{aligned} x_N &= 0 \qquad \text{for} \quad n < 6 \\ &\leqslant 1 \qquad \text{for} \quad n = 6 \\ &\leqslant 2n - 12 \qquad \text{for} \quad n \geqslant 7. \end{aligned}$$

The equation is proved by relating the valencies r, s to the number of "N" triplets of which the given edge is the central member. This is clearly 0 when either vertex has valency 1, 1 when each has valency 2, 2 when one has valency 2 and the other valency 3, and 4 when each has valency 3.

The second part is dealt with by a construction and inductive argument. The special cases for $n \leqslant 7$ follow from consideration of all the possibilities for J. For larger values of n the tree J can be built up from a tree with four edges $(n = 7)$ by means of operations of the following types:

(i) addition of a new terminal edge and vertex to a terminal vertex;

(ii) addition of a new terminal edge and vertex to a vertex of valency 2 adjacent to a terminal vertex.

Let y'_{rs} be the new value in each case of y_{rs}, and x_N' the new value of x_N.

In this case (i)

$$y'_{23} = y_{23} + 1, \qquad x_N' = x_N + 2;$$

or

$$y'_{22} = y_{22} + 1, \qquad x_N' = x_N + 1,$$

the other values of y_{rs} being unaffected.

In case (ii)

$$y'_{33} = y_{33} + 1, \quad y'_{23} = y_{23} - 1; \qquad x_N' = x_N + 2;$$

or

$$y'_{23} = y_{23} + 1, \quad y'_{22} = y_{22} - 1; \qquad x_N' = x_N + 1.$$

We find that $x_N \leqslant 2$ for $n = 7$ by considering the cases and the result then follows for $n > 7$ by induction on n.

We have already calculated x_V in Theorem 10. Its upper bound is $3(n - 4)/2$.

THEOREM 14. *For $n \geqslant 7$, $x_{VI} \leqslant \frac{3}{2}n^2 - 16n + 42$ if n is even and $x_{VI} \leqslant \frac{3}{2}n^2 - \frac{3}{2}n + 45$ if n is odd.*

The value of x_{VI} is clearly related to that of x_V, the adjacent pairs being weighted by the number of edges not having a vertex in common with either of the edges of the "V." Let for some pair $\{\alpha, \beta\}$ forming a "V," z be the number of edges incident with vertices of the "V," not counting α and β. Then apart from special cases when n is small, $1 \leqslant z \leqslant 5$. The pair $\{\alpha, \beta\}$ contributes $(n - 5 - z)$ to x_{VI}.

But $z = 1$ only for those "V's" containing at least one vertex of valency 1. The sum of values of z is thus at least $v_1 + 2(x_V - v_1)$, that is, $2x_V - v_1$.

So

$$\begin{aligned} x_{VI} &\leqslant (n - 5)\, x_V - (2x_V - v_1) \\ &\leqslant (n - 7)\, x_V + v_1 . \end{aligned}$$

As $x_V = 3v_3 + v_2$,

$$\begin{aligned} x_{VI} &\leqslant (n - 7)(3v_3 + v_2) + v_1 \\ &\leqslant (3n - 21)\, v_3 + (n - 7)\, v_2 + v_1 . \end{aligned}$$

Using the equations

$$v_1 + v_2 + v_3 = n - 2,$$

$$v_1 + 2v_2 + 3v_3 = 2n - 6,$$

$$x_{VI} \leqslant \left(12 - \frac{3n}{2}\right)(n-2) + \left(\frac{3n}{2} - 11\right)(2n-6) + \left(3 - \frac{n}{2}\right)v_2$$

$$\leqslant \frac{3n^2}{2} - 16n + 42 - \left(\frac{n}{2} - 3\right)v_2 .$$

For $n \geqslant 7$ the highest value of this upper bound occurs when v_2 is as small as possible, 0 if n is even, 1 if n is odd. Thus

$$x_{VI} \leqslant \frac{3n^2}{2} - 16n + 42 \qquad \text{if} \quad n \text{ is even}$$

and

$$x_{VI} \leqslant \frac{3n^2}{2} - \frac{33n}{2} + 45 \qquad \text{if} \quad n \text{ is odd.}$$

This approximation is not particularly accurate: for $n = 8$ and 9 it is almost twice the true figure, but proportionately it improves as n increases. For instance the estimate is 322 when $n = 20$, while trees with x_{VI} at least 280 exist. The approximation is not operative until $n = 12$, for the sum of the maxima for x_Y, x_N and x_{VI} exceeds the total number of triplets: the three maxima do not hold for the same trees. An important observation is that, whereas the total number of triplets is a cubic expression in n, x_{VI} is bounded by a quadratic, and x_Y and x_N are bounded by linear expressions. Thus while x_{III} is initially the smallest of the four quantities it must eventually dominate.

In Table 1 we give the value of w_1 and the bounds for w_2 and w_3 for $3 \leqslant n \leqslant 20$. The w_2 bounds are exact, as are those for w_3 for $n \leqslant 10$. The figures in parentheses lower down these columns are values of w_3 for trees which appear to give the extreme values. The final column gives the upper bound for w_3 obtained from the last few theorems. The table also shows in the second column the number t_n of trees in $\mathscr{T}_n$.

THEOREM 15. *Let w_r be the number of trees in $\mathscr{T}_n$ at distance r from a given tree T. Then*

$$w_r \leqslant (2n - 8)\, w_{r-1} .$$

Let U be a tree at distance $r - 1$ from T. There are $2n - 6$ trees one crossover from U. At least one of these, U_1, is only distance $r - 2$

TABLE 1

Total Number of Trees in $\mathscr{T}_n$ and Number at Distances 1, 2, and 3 from a Given Tree.

			w_2		w_3		
			Exact		Exact		Estd.,
n	t_n	w_1	Min.	Max.	Min.	Max.	max.
3	1	0	0	0	0	0	
4	3	2	0	0	0	0	
5	15	4	8	8	2	2	
6	105	6	20	24	30	40	
7	945	8	36	40	110	120	120
8	10395	10	56	64	220	268	268
9	135135	12	80	88	378	454	462
10	2037025	14	108	120	592	730	762
11	34629425	16	140	152			1132
12	6.58×10^8	18	176	192			1624
13	1.37×10^{10}	20	216	232			2130
14	3.16×10^{11}	22	260	280			2774
15	7.89×10^{12}	24	308	328			3480
16	2.13×10^{14}	26	360	384	(3308)	(4100)	4340
17	6.18×10^{15}	28	416	440			5280
18	1.91×10^{17}	30	476	504			6386
19	5.32×10^{18}	32	540	568			7590
20	1.86×10^{20}	34	608	640	(7572)	(8640)	8976

from T, being in the chain of trees from T to U, and another, U_2, is one crossover from both U and U_1, so only $r - 1$ from T. Thus there are at most $2n - 8$ trees reached by one crossover from U and r from T. This establishes the theorem, but the value of w_r is further reduced since each tree r crossovers from T is in general one crossover from several trees like U.

This theorem is a very rough one and not always the best to use. Thus we may use

$$w_4 < (w_2)^2$$

or

$$w_5 < w_2 \cdot w_3 \text{ and so on.}$$

None of these gives a very accurate bound.

The purpose of this paper is the assessment of likeness of two trees. Suppose two trees T_1, T_2 in some $\mathscr{T}_n$ are s crossovers apart: the relevant thing is to estimate the probability that T_1 and T_2 should be at most s crossovers apart if they were a pair randomly chosen from $\mathscr{T}_n$. We have provided data for this for $s \leqslant 3$, and some hints towards an upper bound for higher values of s. The method employed for $s = 3$ can in principle be taken as far as desired; the calculations for $s = 4$ would not be too daunting, but soon the method becomes too cumbrous. Thereafter it would appear that Monte Carlo methods must take over.

We are then faced with a difficulty, but a difficulty which will in any case appear in another part of the calculations. This is the problem of determining the distance between two given trees. There appears to be no way of calculating this other than the construction of a chain of trees starting from one and ending at the other. With a little practice it is not hard to construct such chains which are evidently of minimum length: the proof that they are minimum chains is excessively tedious. Errors arising here are not likely to be serious in practice since we are concerned only with orders of magnitude of probabilities.

In the example of Fig. 0, two crossovers suffice to convert one tree into the other. From Table 1, when $n = 9$, $t_n = 135135$, $w_0 = 1$, $w_1 = 12$ and $w_3 \leqslant 88$. The probability of obtaining so much agreement by chance is thus only 0.00075, showing a very substantial measure of agreement between the trees.

The invention of methods of reconstruction of the evolutionary tree will involve testing with either synthetic models of evolution or the very few examples in which the actual course of evolution is sufficiently known. For a given set S of organisms there is an actual evolutionary tree E. Each method M of reconstruction gives rise to a tree R_M. As a result of this paper, or such future extensions as may be warranted, it will be possible to give at least an estimate of the probability of obtaining a resemblance as close as that between E and R_M as a result of chance. Averaging this probability over a large number of sets S gives an estimate of the accuracy of M.

3. Commutative Semigroups

In conclusion we point out that rooted labeled trees and the crossover operation also arise naturally in the discussion of manipulation in commutative semigroups. To the expression

$$((a + b) + c) + (d + e)$$

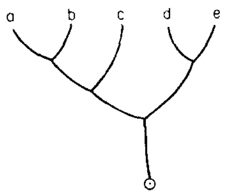

FIGURE 7

belongs the labeled rooted tree shown in Fig. 7. Commutativity between parts of the expression leaves the tree unchanged, but the associative law is connected with the crossover operation:

$$(a + b) + c = a + (b + c).$$

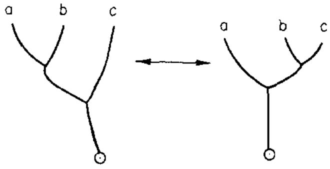

FIGURE 8

The distance between two expressions is then the number of associative manipulations required to convert one into the other.