Scan A7151 Foulds & RWR Ershilmany trees all to 2 segs (56)

At least Pu of Tu

morta

L. R. Foulds Industrial/Systems Engineering University of Florida R. W. Robinson Dept. of Mathematics Southern Illinois University

(Evolutemany
ABSTRACT Trees

Evolutionary trees of biology are represented by a special class of labelled trees. They are characterized by having disjoint subsets of the labelling set assigned to the vertices of a tree in such a way that no pendant vertex or vertex of degree two is assigned an empty set of labels. In this paper exact and asymptotic numbers are determined for two classes of evolutionary trees. The first class is characterized by the properties that: (i) its members have only singleton labels and (ii) there are no vertices of degree two, and (iii) interior vertices may be labelled. The second class has properties: (i) there may be vertices of degree two, and (ii) all the labels are singletons and (iii) only vertices of degree one or two are labelled. This paper continues previous work by the authors on the enumeration of various classes of evolutionary trees.

1. INTRODUCTION

Since the time of Charles Darwin, biologists have postulated that existing biological species have descended from common ancestors. A diagram showing such relationships is termed a phylogenetic or evolutionary tree.

Here we define an <u>evolutionary tree</u> to be a connected acyclic graph together with a function sending a given set {1,2,..., n} of species (called <u>labels</u>) to the vertex set of the tree such that every vertex of degree one or two is in the image. It follows from this definition that in general, some vertices may not be assigned a label and other vertices may have more than one label. The number of vertices in any tree is called its <u>order</u>. The number of labels, denoted by n, is called the <u>magnitude</u> of the tree. A <u>planted</u> evolutionary tree is a tree with a pendant vertex distinguished,

wh s called the <u>root</u>, which corresponds to the common ancestor. As it is already distinguished it is not assigned a label. The construction of such trees and their biological significance have been discussed by many authors, but for the most recent information see the book by Penny, Hendy, and Foulds [10]. Undefined graph theoretic notation and terminology used in this paper is covered in the book by Harary [6].

The previous mentioned book by Penny et al. gives a detailed account of currently-used methods for the identification of evolutionary trees. In the quest for trees satisfyiing certain optimality criteria, it is of interest to know how many possible trees exist which span in given species. The exact and asymptotic numbers of evolutionary trees with given magnitude, along with the average and variance of their orders, was determined by the present authors in [3]. The effect of requiring that each nonempty label be a singelton was also studied. In [4] we performed the same analysis for binary evolutionary trees and also studied the effect of restricting the labelling to pendant vertices. In [5] the same analysis was performed for evolutionary trees which do not have vertices of degree two where the labelling is restricted to the pendant vertices. The effect of restricting the labels to being singletons was also studied. In the present paper we perform this analysis for the first two of the four remaining classes of evolutionary trees which result from taking combinations of the above conditions. These classes are:

Case 1: There are no vertices of degree two, all nonempty labels are singletons, and interior vertices may possibly be labelled.

Case 2: There are no restrictions on vertex degree, all nonempty labels are singletons, and only vertices of degree one or two can be labelled.

<u>Case 3</u>: The same as Case 1 except that labels need not be singletons.

Case 4: The same as Case 2 except that labels need not be singletons.

It is planned to publish the analysis of Cases 3 and 4 elsewhere.

We now introduce some notation, termino y, and objectives which are common to both Cases 1 and 2. Let T_n be the number of different evolutionary trees of magnitude n. Our first objective in each case is to derive recurrence relations by which T_n can be calculated for successive values of n. The exponential generating function defined by

$$T(x) = \sum_{n=1}^{\infty} T_n x^n / n!$$
 (1.1)

will be a useful tool in our analysis. In order to establish the mean and variance of the numbers of vertices in trees of given magnitude we shall also find recurrences involving the number $T_{n,p}$, of different trees with magnitude n and order p. The corresponding generating function is given by

$$T(x,y) = \sum_{n=1}^{\infty} \sum_{p=1}^{2n-2} T_{n,p} x^n y^p / n!.$$
 (1.2)

Thus T(x) is obtained from T(x,y) by setting y to 1. It is not hard to see that $1 \le p \le 2n-2$. A short proof is given in [2].

As is usual in tree counting, the numbers are first determined for trees which have some vertex distinguished as the <u>root</u>. In particular, let a <u>planted evolutionary tree</u> be a tree rooted at a pendant vertex and labelled according to the rules of the case under examination except that the root is not to receive any label. Likewise the root vertex of a planted evolutionary tree is not counted in its order. In the biological context, a planted evolutionary tree corresponds to an evolutionary tree in which a common ancestor is designated. This is represented diagrammatically by orienting all arcs away from the vertex representing the common ancestor.

Let P_n denote the number of different planted evolutionary trees of magnitude n, and let $P_{n,p}$ denote the number of these of order p. The associated generating functions are:

$$P(x) = \sum_{n=1}^{\infty} P_n x^n / n!$$
 (1.3)

$$P(x,y) = \sum_{n=1}^{\infty} \sum_{p=1}^{2n-1} P_{n,p} x^{n} y^{p} / n! . \qquad (1.4)$$

Since an evolutionary tree of magnitude n may have in general, any order from 1 to 2n-2, it is of interest to determine the mean μ_n and the standard deviation ν_n , of the number of vertices among all such trees. This will be done by finding recurrence relations for the first and second moments about the origin, namely

$$T_n(1) = \sum_{p=1}^{2n-2} p T_{n,p}$$
 (1.5)

$$T_n(2) = \sum_{p=1}^{2n-2} p^2 T_{n,p}.$$
 (1.6)

Then as usual we have

$$\mu_{n} = T_{n}^{(1)} / T_{n},$$
(1.7)

$$v_{n} = (T_{n}^{(2)}/T_{n}) - \mu_{n}^{2}. \tag{1.8}$$

The exponential generating functions for the moments are:

$$T^{(i)}(x) = \sum_{n=1}^{\infty} T_n^{(i)} x^n/n!$$
, $i = 1 \text{ or } 2.$ (1.9)

Since

$$T^{(1)}(x) = T_{v}(x,1),$$
 (1.10)

and

$$T^{(2)}(x) = T_{yy}(x,1) + T_{y}(x,1),$$
 (1.11)

we can evaluate the moment generating functions on the basis of previous equations. Define

$$S_n = T_n^{(2)} - T_n^{(1)},$$

 $R_n = T_n^{(1)}, R(x) = T^{(1)}(x), \text{ and } S(x) = T^{(2)}(x) - R(x).$ (1.12)

Using the values of P_n and T_n found by methods outlined earlier and the last few relations we can compute values for $T_n^{(1)}$, $T_n^{(2)}$. S_n , μ_n , and ν_n .

In the next two sections we calculate the above values for Cases 1 and 2.

2. CASE 1: NO VERTICES OF DEGREE TWO, ALL LABELS ARE SINGLETONS.

Our first objective in this section is to derive recurrence relations by which P_n can be calculated for successive values of n. Case 1 differs only slightly from a case published previously by us. In $[5,\ \S\ 2]$ we analyzed the situation which is the same as Case 1 except that the interior vertices are not allowed labels. The obvious approach is to parallel almost exactly the development outlined in that section, with appropriate changes in certain details. However there is a simple relationship between the P(x,y)'s for the two cases which will enable us to obtain the same results much more quickly.

Once again the exponential generating function P(x,y) is a useful tool. Let P(x,y) be the generating function for the similar case in $[5, \S 2]$ which is analogous to P(x,y).

Since labelling is optional for interior vertices and at most one label can be assigned to each, the exponential generating function of the labelling possibilities for an interior vertex is (1+x). For an endpoint the corresponding generating function is just x, since it must be labelled. Hence we can obtain P(x,y) as

$$P(x,y) = \overline{P}(\frac{x}{1+x}, y(1+x)),$$
 (2.1)

where the term $\frac{x}{1+x}$ supplied at each endpoint combines with the factor of (1+x) to give just x for each endpoint. It is shown in [5] that

$$\bar{P}(x,y) = ye^{\bar{P}(x,y)} - y - y \bar{P}(x,y) + xy.$$

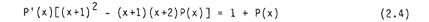
Thu

$$P(x,y) = [y(1+x)e^{P(x,y)} - y - y(1+x)P(x,y)].$$
 (2.2)

Setting y=1 we obtain:

$$x(P(x) - e^{P(x)}) = e^{P(x)} - 2 P(x) - 1.$$
 (2.3)

Then differentiating and simplifying to eliminate the exponential



It is convenient for computational purposes to let $Q(x) = 1/2 P(x)^2$ so that Q'(x) = P'(x)P(x) and the coefficients Q_n of $x^n/n!$ satisfy

$$Q_{n} = \sum_{i \leq n/2} {n \choose i} P_{i} P_{n-i} + 1/2 {n \choose n/2} P_{n/2}^{2}, \qquad n > 0.$$
 (2.5)

Here $P_{n/2}$ is taken to be zero when n is odd. Then (2.4) can be rewritten in terms of Q(x) as

$$P'(x) = 1 + P(x) - (2+x+x^2)P'(x) + (1+x)(2+x) Q'(x).$$

Given that $P_0 = 0$, comparing coefficients of $x^n/n!$ in this equation gives $P_1 = 1$ and

$$P_{n+1} = (1-2n)P_n - n(n-1)P_{n-1} + 2Q_{n+1} + 3nQ_n + n(n-1)Q_{n-1}, n > 1.$$

Then (2.5) and this last equation can be used to compute P_n and Q_n efficiently.

The generating function for the unrooted trees represented by T(x,y) can now be determined in terms of trees in which a vertex is distinguished (vertex-rooted) and those in which an edge is distinguished (edge-rooted). A vertex-rooted tree is associated in a 1-1 fashion with the planted tree obtained by joining a new root vertex to the original. All planted trees are obtained in this way except for those in which the vertex adjacent to the root has degree 3 and has not received a label. The latter are counted by the exponential generating function $1/2 \ y(1+x)P(x,y)^2$, and so vertex-rooted trees rooted at an interior vertex are counted by

$$P(x,y) - 1/2 y(1+x)P(x,y)^2$$

Vertex-rooted trees rooted at a pendant vertex are counted by:

$$xy P(x,y)$$
.

So all vertex-rooted trees are counted by:

$$P(x,y) + xy P(x,y) - 1/2 y(1+x) P(x,y)^2$$
.

An edge-rooted tree can be viewed as the result of identifying the roots of two planted trees, which are then suppressed to form the root edge. These are therefore enumerated by

$$P(x,y)^{2}/2$$

the factor of 2 accounting for the fact that the same edge-rooted tree is obtained by interchanging the two planted trees. The difference between the expression for the vertex-rooted and edge-rooted trees is T(x,y). Thus after simplification,

$$T(x,y) = (1+xy)P(x,y) - 1/2 (1+y + xy)P(x,y)^2$$
 (2.6)

Setting y=1 we obtain

$$T(x) = (1+x)P(x) - (1+x/2)P(x)^{2}$$
 (2.7)

Replacing $1/2 P(x)^2$ by Q(x) we have:

$$T(x) = (1+x)P(x) - (2+x)Q(x)$$
 (2.8)

We can compare coefficients of $x^n/n!$ in this last equation to determine that $T_1=1$, $T_2=1$ and

$$T_n = P_n + nP_{n-1} - 2Q_n - nQ_{n-1}, \quad n > 2$$
 (2.9)

Differentiating (2.6) gives

$$T_y(x,y) = x P(x,y) - \frac{x+1}{2} P(x,y)^2 + [1 + xy - (1+y+xy)P(x,y)]P_y(x,y)$$
(2.10)

Differentiating (2.2) and simplifying gives

$$P_{y}(x,y) = \frac{1}{y} P(x,y)$$
 (2.11)

Communing (2.10) and (2.11) and simplifying we have

$$T_{v}(x,y) = (x + \frac{1}{v}) P(x,y) - \frac{1}{2} (x+1) P(x,y)^{2}$$
 (2.12)

Differentiating again and simplifying with (2.11) yields

$$T_{yy}(x,y) = P(x,y)^2/[y^2(1+xy - P(x,y)(1 + y(1+x))]$$
 (2.13)

Setting y=1 in (2.12) and (2.13) and using (1.11), we have

$$R(x) = (x+1)P(x) - 1/2 (x+1)P(x)^{2}, (2.14)$$

and

$$T^{(2)}(x) = R(x) + P(x)^2/(1+x-P(x)(2+x)).$$
 (2.15)

Recurrence relations are obtained by comparing coefficients of $x^n/n!$. In order to carry out the computation of the R_n 's efficiently we once again introduce the intermediate sequence \mathbb{Q}_n defined in (2.5).

Then

$$R_n = P_n + n P_{n-1} - Q_n - nQ_{n-1}, \quad n > 1.$$
 (2.16)

From (2.15) we have that

$$S(x)[1+x - P(x)(2+x)] = 2Q(x)$$
.

This implies that $S_0 = 0$, and for n > 1 we can determine S_n using the P_n 's and Q_n 's in the recurrence

$$S_n = 2Q_n - n S_{n-1} + \sum_{i=1}^{n-2} {n \choose i} S_{n-i} \{2 P_i + i P_{i-1}\}.$$
 (2.17)

From the values of P_n and T_n mentioned earlier, (2.15), (2.16), and (2.17) we can compute values of R_n and $T_n^{(2)}$, and hence of μ_n and ν_n .

We now determine the asymptotic behavior of $T_n,\ \mu_n,\ and\ \nu_n$ as $n \leftrightarrow \infty$. Since P_n is the basis for the equations defining these quantities, we start with a study of the exponential generating

function P(x). Recall from (2.3) that

$$x (P(x) - e^{P(x)}) = e^{P(x)} - 2P(x) - 1,$$

from which we can solve for x to obtain

$$x = (e^{P(x)} - 2P(x) - 1)/(P(x) - e^{P(x)}).$$

Regarding P as a complex variable, x is clearly analytic for $P \neq e^{p}$, and the derivative

$$\frac{dx}{dP} = (-e^{2P} + (P+2)e^{P} - 2P)/(P - e^{P})^{2}$$

is nonzero when $P \neq e^{-P}$. Also x=0 when P=0, so the inverse function is analytic in some neighborhood of x=0. The power series expansion of this function about x=0 is our generating function P(x), which is now seen to have a positive radius of convergence. say p. Let q be the unique real number such that $q=e^{-q}$, one can calculate that q=0.5671432904... As $P_n > 1$ for n > 1 we see that ρ $< \infty$, since P(x) cannot attain the value of q within the circle of convergence. By Pringsheim's theorem (see Hille [9, p.133]) x=p is a singularity of P(x). Also 0 < P(x) < q for real x, 0 < x < p, and P(x) is strictly increasing in this interval, so P(a) is defined and $P(\rho) \le q$. In fact $P(\rho) = q$, for if not there could be no singularity of P at x=p. Also, for |x|=p and $x\neq p$ we have $|P(x)| \le q$ so that x=p is the sole singularity of P(x) on its circle of convergence. Finally, setting x=p and using the fact that $q \neq -1$ to simplify gives $\rho = (2q-1)/(1-q) = 0.3102333359...$ Note that as expected, this is strictly less than the equivlent value of ρ (which is 2(ln 2)-1 in the similar case in [5, § 2].

So far we have established the results equired of steps 1-11 in the 20 step algorithm of [8], by methods which are more direct than usual for tree counting problems. However the remainder of the development is standard and so we refer to [8] for the explanations of the remaining steps and confine ourselves to performing the necessary calculations.

At P = q the second derivative of x as a function of P can be

cal ated explicitly and is nonzero. so from steps 12 and 13 we have that x=p is a branch point of order 2 for P(x). Thus as in step 14 one has an expansion of the form

$$P(x) = q - a_1(\rho - x)^{1/2} + a_2(\rho - x) + a_3(\rho - x)^{3/2} + \dots$$
 (2.18)

valid in some neighborhood of x= ρ . Substituting into equation (2.4) gives a relation that must be satisfied by this expression. One can then compare coefficients of $(\rho-x)^{-1/2}$, $(\rho-x)^0$, $(\rho-x)^{1/2}$,... to determine as many of $a_1,a_2,...$ as required. In particular we find that

 $a_1^2 = \frac{4}{(\rho+2)} (1+ 1/(2(\rho+1))$

In order for the expansion around 0 to agree with the expansion around p where their circles of convergence overlap. We must take the positive root of this equation giving $a_1=1.017576587\ldots$. Two

more comparisons then establish that $a_2 = -\frac{a_1^2}{6} - -\frac{1}{\rho+2} =$

$$a_3 = \frac{a_2^2}{2a_1^2} - \frac{a_1}{4(\rho+1)(\rho+2)} - \frac{a_2(\rho+1)}{2a_1(\rho+2)^2} + \frac{a_2\rho}{2a_1(\rho+1)(\rho+2)} = 0.3503501726...$$

To evaluate the contribution of a term $(p-x)^{k/2}$ in expansion (2.18), note that the coefficient of x^n in $(1-x)^{-s}$ is just $\Gamma(s+n)/\Gamma(s)\Gamma(n+1)$ provided that s is not a nonnegative integer. From Stirling's formula the latter is

$$\frac{n^{s-1}}{\Gamma(s)} \left(1 + \frac{s(s-1)}{2n} + 0 \left(\frac{1}{n^2}\right)\right)$$

as $n \leftrightarrow \infty$. Thus the term $-a_1(\rho-x)^{1/2}$ contributes

$$1/2 \, a_1(\pi)^{-1/2} \, \rho^{1/2} \, n^{-3/2} \, \rho^{-n} \, \left(1 + \frac{3}{8n} + 0(\frac{1}{n^2})\right).$$

The next term, $a_3(\rho-x)^{3/2}$, contributes

$$3/4 a_3 (\pi)^{-1/2} \rho^{3/2} n^{-5/2} \rho^{-n} (1 + 0(\frac{1}{n}))$$

when taken to the same order. The remaining two scollectively contribute $O(n^{-7/2} \ \rho^{-n})$, as can be seen from Darboux's Theorem (as in Theorem 4 of Bender [1] or from Polya's Lemma (as in [8]). In sum we have

where $A_{p} = \frac{a_{1}}{2} \left(\frac{\rho^{1/2}}{\pi^{1/2}} \right) = 0.1598845156...$ (2.19)

 $B_{p} = \frac{3a_{1}\rho^{1/2}}{16\pi^{1/2}} + \frac{3a_{3}\rho^{3/2}}{4\pi^{1/2}} = \frac{3\rho^{1/2}}{4\pi^{1/2}} - (\frac{a_{1}}{4} + a_{3}\rho)$ = 0.03779276929...

From (2.8) and (2.18) we have directly

$$T(x) = \frac{1}{2}(1+\rho)q + (\frac{1}{2}q^{2} - q - \frac{1+q}{1+\rho})(\rho-x) + \frac{2}{3}(q^{-1}-q)(\rho-x)^{3/2}$$

$$+ \left[\frac{1}{4}a_{1}^{2} - \frac{1+q}{-1+\rho} + (q-q^{-1})a_{2}\right](\rho-x)^{2} + \frac{2}{5}[(q-q^{-1})a_{3}$$

$$+ \left(\frac{1}{(1+\rho)^{2}} - a_{2}\right)a_{1}\right](\rho-x)^{5/2} \pm \dots$$

Following the same reasoning as for the derivation of (2.19), the first two terms of the asymptotic expansion of $T_n/n!$ are determined by the $(\rho-x)^{3/2}$ and $(\rho-x)^{5/2}$ terms, with the result that

$$\frac{T_n}{n!} = n^{-5/2} \rho^{-n} (A_T^+ + \frac{B_T}{n} + 0(\frac{1}{n^2})), \qquad (2.20)$$

where

$$A_{T} = \frac{a_{1}}{2\pi} \frac{1}{1/2} (q^{-1} - q) \rho^{3/2} = 0.05932734740 \dots ,$$

and

$$B_{T} = \frac{15}{8} A_{T} + \frac{3\rho^{5/2}}{4\pi^{1/2}} ((q^{-1}-q)a_{3} + a_{1}(a_{2} - (1+\rho)^{-2}))$$

$$= 0.09321113490...$$

Similarly from (2.18) and (2.14) we get

$$R(x, (1+\rho)(q - (1/2)q^{2}) - qa_{1}(\rho-x)^{1/2} + (1/2 q^{2} + q_{1}(a - 1))$$

$$- (1/2) a_{1}^{2} (1+\rho)(\rho-x) + (a_{1} a_{2}(1+\rho) + q a_{3} + a_{1}(2+\rho)^{-1})(\rho-x)^{3/2} \pm \cdots$$

In the same manner as (2.19) we get

$$\frac{R_n}{n!} = n^{-3/2} \rho^{-n} (A_{\mu} + \frac{B_{\mu}}{n} + 0(\frac{1}{n}2)).$$
 (2.21)

where

$$A_{\mu} = \frac{qa_1}{2} \left(\frac{\rho}{\pi}\right)^{1/2} = 0.09067743025 \dots$$

and

$$B_{\mu} = (3/8) A_{\mu} + \frac{3\rho^{3/2}}{4\pi^{1/2}} - (a_1 a_2 (1+\rho) + qa_3 + \frac{a_1}{(2+\rho)} = 0.06902139685$$

Taking the ratio of (2.21) to (2.20) gives the average number of vertices in a tree of magnitude n as

$$\mu_{n} = n \left(A + \frac{B}{n} + O(\frac{1}{n}2)\right)$$
 (2.22)

where

$$A = q(q^{-1}-q)^{-1}\rho^{-1} = 2q/(a_1^2\rho(\rho+2)) = 1.5284262051$$

and

$$B = -1.2379627565 \dots$$

To analyse the variance similarly we start with exponential generating function $S(x) = T^{(2)}(x) - R(x)$. From (2.14) and (2.15) we have $S(x) = P(x)^2/(1-x-(2+x)P(x))$. Substituting the expansion (2.18) for P(x) yields

$$S(x) = \frac{q^{2}}{a_{1}(2+\rho)} (\rho-x)^{-1/2} + (\frac{(1-q)q^{2}}{a_{1}^{2}(2+\rho)^{2}} + \frac{a_{2}q^{2}}{a_{1}^{2}(2+\rho)})^{-\frac{2q}{2+\rho}}$$

$$+ \frac{1}{a_{1}(2+\rho)} [(\frac{1-q}{a_{1}(2+\rho)} + \frac{a_{2}}{a_{1}})^{2} q^{2} + (\frac{1}{2+\rho} + \frac{a_{3}}{a_{1}})q^{2}$$

$$- 2q a_{1}(\frac{1-q}{a_{1}(2+\rho)} + \frac{a_{2}}{a_{1}}) + a_{1}^{2} + 2qa_{2}](\rho-x)^{1/2} \pm \cdots (2.23)$$

As before, (2.19) can be applied in conjunction the Darboux's Theorem on Polya's Lemma to evaluate the coefficients of S(x) asymptotically. The result is

$$\frac{S_n}{n!} = n^{-1/2} \rho^{-n} \left[A_s + \frac{Bs}{n} + 0 \left(\frac{1}{n^2} \right) \right]$$

$$A_s = q^2 / \left[a_1 (2 + \rho) (\rho \pi)^{1/2} \right] = 0.1385936962 \dots$$

and
$$B_{S} = (1/8) A - \frac{\rho^{\frac{1}{2}} \left[\left(\frac{1-q}{2+\rho} + a_{2} \right)^{2} - \frac{q^{2}}{a_{1}^{2}} + \left(\frac{1}{2+\rho} + \frac{a_{3}}{a_{1}} \right) q^{2} - 2q \left(\frac{1-q}{2+\rho} + a_{2} + a_{1}^{2} + 2q a_{2} \right) \right]}{2a_{1} (2+\rho) \pi^{\frac{1}{2}}}$$

$$= -0.07770161671 \dots$$

Finally $v_n = (S_n/T_n) + \mu_n - \mu_n^2$, so this can be combined with (2.20) and (2.22), giving

$$v_n = An + O(1)$$

where

$$A = \frac{B_S}{A_T} - \frac{A_S}{A_T^2} - 2A_{\mu}B_{\mu} + A_{\mu}$$

$$= 0.33268409010...$$

The $O(n^2)$ terms for v_n cancelled out because of the relation $a_1/2 = (q^{-1} - q)/(\rho + 2).$

3. CASE 2: SINGLETON LABELS, VERTICES OF DEGREE 32 UNLABELLED.

Our first objective in this section is to derive recurrence relations by which P_n can be calculated for successive values of n. Once again the exponential generating function P(x,y), is a useful tool. There is just one planted evolutionary tree of magnitude 1, termed the <u>trivial</u> tree, which has order 1. Thus the exponential generating function for trivial planted trees is xy. Any non-trivial planted tree can be viewed as the result of joining

one or more planted trees at their roots, these being identified as a single ordinary vertex which is then joined to a new root. In the process the original root must receive a label if it becomes a vertex of degree two but cannot be labelled if it becomes of higher degree. Thus the generating function for planted trees in which the vertex adjacent to the root has degree two is

xyP(x,y).

Here xy enumerates the possibilities for the vertex adjacent to the root, while P(x,y) enumerates the possibilities for completing the tree. As is usual in labelled counting problems the product of the exponential generating functions accounts for the number of ways in which a set of labels from the parts can be obtained from the label set of the union. (See for example, chapter 1 of [7].) Similarly if the vertex adjacent to the root has degree k+1 > 3, the number of possibilities is enumerated by

$$yP(x,y)^k/k!$$

In this expression y accounts for the vertex adjacent to the root, since it must not be labelled. There are k planted trees to be joined to this vertex and we divide by k! because the sequence in which they are added is immaterial. Summing over $k \geq 2$, adding in the other two terms and taking advantage of the expontial form of the sum, we have

$$P(x,y) = y[e^{P(x,y)} + (x-1)(1 + P(x,y))].$$
 (3.1)

Setting y=1 gives

$$P(x) = e^{P(x)} + (x-1)(1 + P(x)).$$
 (3.2)

Then differentiating and simplifying to eliminate the exponential yields

$$P'(x)[1 + (x-2)P(x)] = 1 + P(x).$$
 (3.3)

As before we set $Q(x) = 1/2 P(x)^2$, so that eq on (2.5) can be rewritten as

$$P'(x) = 1 + P(x) + (2-x)Q'(x).$$
 (3.4)

Since $P_0 = 0$, comparing coefficients of $x^n/n!$ in this equation gives $P_1 = 1$ and

$$P_{n+1} = P_n - nQ_n + 2Q_{n+1},$$
 $n > 1.$ (3.5)

Then (2.5) and (3.5) together allow P_n and Q_n to be computed recursively in an efficient manner. All planted trees are obtained in the manner explained in section 2 (via vertex-rooted trees) except for those in which the vertex adjacent to the root has degree 3. The latter are counted by the exponential generating function

$$xy P(x,y)^{2}/2$$

and so the vertex rooted trees are counted by

$$P(x,y) + (x-1)y P(x,y)^{2}/2$$

As in section 2, edge-rooted trees are counted by

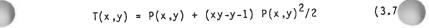
$$P(x,y)^{2}/2$$

and the difference between the expressions for vertex-rooted and edge-rooted trees is T(x,y). On assembling the terms corresponding to a particular unrooted tree of magnitude n and order p we find that

$$px^{p}y^{p}/n! - (p-1)x^{p}y^{p}/n! = x^{n}y^{p}/n!$$

This is because the tree has no automorphisms, having been labelled at all its endpoints, so that all p vertices give distinct vertex-rooted trees while all (p-1) edges give distinct edge-rooted trees. Thus we have

$$T(x,y) = P(x,y) + (x-1)yP^{2}(x,y)/2 - P(x,y)^{2}/2,$$
 (3.6)



Setting y=1 and using $Q(x) = 1/2 P(x)^2$ we obtain

$$T(x) = P(x) + (x-2)Q(x).$$

Differentiating and using (3.4) to simplify, we obtain

$$T'(x) = 1 + P(x) + Q(x)$$
 (3.8)

We can compare coefficients of $x^{n}/n!$ to determine that $T_{1}=1$ and

$$T_{n+1} = P_n + Q_n,$$
 $n > 1.$ (3.9)

Thus (3.9) can be used in conjunction with (3.3) and (3.5) to compute exact values of $T_{\rm n}$. Differentiating (3.6) gives

$$T_{y}(x,y) = 1/2 (x-1)P(x,y)^{2} + [1 + (xy-y-1) P(x,y)] P_{y}(x,y).$$
 (3.10)

Differentiating (3.2) we find, after simplification, that

$$P_{V}(x,y) = \frac{1}{V} P(x,y),$$
 (3.11)

with which (3.10) simplifies to

$$T_{y}(x,y) = [1/2 (x-1)P(x,y) + \frac{1}{y}] P(x,y)$$
 (3.12)

Differentiating again and simplifying with (3.11) yields

$$T_{yy}(x,y) = P(x,y)^2 / [1 + (y(x-1)-1)P(x,y)]y^3$$
 (3.13)

Setting y=1 in (3.12) and (3.13) and using (1.11) and (1.12) we have

$$R(x) = [1/2 (x-1) P(x) + 1] P(x)$$
 (3.14)

and

$$T^{(2)}(x) = [1/2 (x-1)P(x) + 1]P(x) + P(x)^{2}/[1 + (x-2)P(x)].$$
 (3.15)

Recurrence relations are obtained by compare oefficients of $x^n/n!$. In order to carry out the computation of the $R_{n's}$ we can use

$$Q(x) = 1/2 P(x)^2$$
 and (3.8), giving

$$R(x) = T(x) + Q(x),$$

so that

$$R_n = T_n + Q_n,$$
 $n > 1.$ (3.16)

From (3.14) and (3.15) we have that [1 + (x-2) P(x)] S(x) = 2Q(x),

which can be rewritten in the form

$$S(x) = 2Q(x) + (2-x) P(x) S(x)$$
.

This implies that $S_0 = 0$, and

$$S_n = 2Q_n + 2n S_{n-1} + \sum_{k=2}^{n-1} {n \choose k} S_{n-k} \{2P_k - k P_{k-1}\}, n > 1.$$
 (3.17)

From the values of P_n , Q_n , and T_n mentioned earlier, (3.16), and (3.17) we can compute values of R_n and $T_n^{(2)}$, and hence of μ_n and ν_n .

We now determine the asymptotic behavior of T_n , μ_n , and ν_n along the lines of the previous section, i.e. starting with P(x). We can solve equation (3.2) for x to obtain

$$x = [2 P(x) - e^{P(x)} + 1] / (P(x) + 1)$$
 (3.18)

Regarding P as a complex variable, x is clearly analytic for $P \neq -1$, and the derivative

$$\frac{dx}{dP} = [(2-e^{P})(P+1) - 2P-e^{P}+1)] / (P+1)^{2},$$
 (3.19)

is nonzero when $P \neq e^{-P}$. Also x=0 when P=0, so the inverse function is analytic in some neighborhood of x=0. The power series expansion of this function about x=0 is P(x) which then must have positive

s of convergence, denoted by ρ . As in section 2 let q be the unsque real number such that qq=1. We found that $q=0.56714329081\ldots$ By using the same argument as in section 2 we can show that ρ is the sole singularity of P(x) on its circle of convergence. Setting $x=\rho$ in (3.18) yields $\rho=2-q^{-1}=0.2367771669...$ Once again following the algorithm in [8], we can establish that P(x) has form (2.18). We find that

$$a_1 = (2(1+q)q)^{1/2} = 1.3332627671 \dots$$
,
 $a_2 = (1/3)q(2-q) = 0.2708783643 \dots$,
 $a_3 = (\frac{q}{2})^{3/2} (4q^2-q+13) / [9(1+q)^{1/2}] = 0.18387997944 \dots$

The term $-a_1(\rho-x)^{1/2}$ contributes

1/2
$$a_1(\frac{\rho}{\pi})^{1/2} n^{-3/2} \rho^{-n} \left(1 + \frac{3}{8n} + O(\frac{1}{n^2})\right)$$
.

The next term, $a_3(\rho-x)^{3/2}$, contributes

$$3/4 \ a_3(\pi)^{-1/2} \ \rho^{3/2} \ n^{-5/2} \ \rho^{-n} (1 + 0(\frac{1}{n}))$$

when taken to the same order. Once again the remaining terms contribute $O(n^{-7/2} \rho^{-n})$. In sum we have

$$\frac{P_n}{n!} = n^{-3/2} \rho^{-n} A_p \left(1 + \frac{B_p}{n} + O(\frac{1}{n^2})\right), \qquad (3.20)$$

where

$$A_p = (1/2)a_1(\frac{\rho}{\pi})^{1/2} = 0.18301248961 \dots$$

and

$$B_p = [3/8 + \frac{3a_3p}{2a_1}] = 0.4239834558 \dots$$

Since $Q(x) = (1/2)P(x)^2$, equation (2.18) gives an expansion of Q(x) about x=p in powers of $(p-x)^{1/2}$. Then just as for the asymptotic analysis of $P_0/n!$, it follows that

$$\frac{Q_n}{n!} = \frac{a_1 q}{2} \left(\frac{\rho}{\pi} \right)^{1/2} n^{-3/2} \rho^{-n} \left(1 + \frac{3}{n} \right) \left(\frac{1}{8} + \frac{(a_3 - a_1 a_2 q - a_1^{3/2})}{2a_1} \right)$$

$$= 0 \left(\frac{1}{n^2} \right) \right). \tag{3.21}$$

From equation (3.9) we have that

$$\frac{T_n}{n!} = \frac{1}{n} \cdot \left(\frac{P_{n-1}}{(n-1)!} + \frac{Q_{n-1}}{(n-1)!} \right),$$
 $n > 2.$

Now (3.20) and (3.21) imply that

$$\frac{T_n}{n!} = n^{-5/2} \rho^{-n} A_T (1+B_T/n + O(1/n^2)).$$
 (3.22)

where

$$A_T = (1/2) a_1 (1+q) \rho(\frac{\rho}{\pi})^{1/2} = 0.06790930042 \dots$$

and

$$B_T = 15/8 + (1/24) \rho q (4q^2 + 11q-11) (1+q)^{-1} = 1.8625936309 \dots$$

Using (3.16) (3.21), and (3.22) we get

$$\frac{R_n}{n!} = n^{-3/2} \rho^{-n} A_R \left(1 + \frac{1}{n} B_R + 0 \left(\frac{1}{n^2} \right) \right). \tag{3.23}$$

where

$$A_R = (1/2) qa_1(\frac{\rho}{\pi})^{1/2} = .10379430561 \dots$$

and

$$B_{R} = 3/8 + \rho[4q^{4} + 11q^{3} + 25q^{2} + 24q + 24] / (24q(1+q))$$
$$= 0.9086176934 \dots$$

Taking the ratio of (3.23) to (3.22) gives the average number of vertices in a tree of magnitude n as

$$u_n = n A_{\mu} (1 + \frac{1}{n} B_{\mu} + O(\frac{1}{n^2})),$$

(3.24

where

$$A_u = q/(1+q)\rho = 1.5284254876 \dots$$

and

$$B_{\mu} = -3/2 + \frac{\rho}{2q(q+1)} [3q^2 + 2q + 2] = -0.9539759374 \dots$$

We analyze the variance by starting with

$$S(x) = 2Q(x)/[1 + P(x)(x-2)].$$

This yields

$$S(x) = \frac{q^2}{a_1(\rho-2)} (\rho-x)^{-1/2} + (1/3)q^2(q-6) + C_S(\rho-x)^{1/2} \pm \dots (3.25)$$

where

$$c_{S} = \frac{q^{3/2}}{12 \quad 2 \quad (1+q)^{3/2}} \quad (4q^{4} - 27q^{3} - 21q^{2} + 16q + 12).$$

$$= 0.12582363303 \dots$$

Similar to section 2, we can use (3.25) to evalue the coefficients of S(x) asymptotically. The results is

$$\frac{S_n}{n!} = n^{-1/2} \rho^{-n} \left(A_s + \frac{1}{n} B_s + O(\frac{1}{n^2}) \right), \qquad (3.26)$$

where

$$A_s = q^3/a_1(\rho\pi)^{1/2}$$

and

$$B_s = -[1/8 + \rho[4q^4 - 31q^3 - 5q^2 + 48q + 24] / 24q(1+q).$$

Finally, (3.22) and (3.24) can be used in conjunction with (3.26) to produce

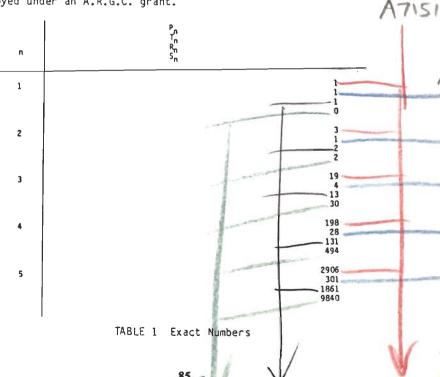
$$\frac{v_n}{n} = -(4q^5 + 12q^4 - 15q^3 - 10q^2 - 12q + 12) v_p^2 (1+q)^3 + 0(\frac{1}{n})$$

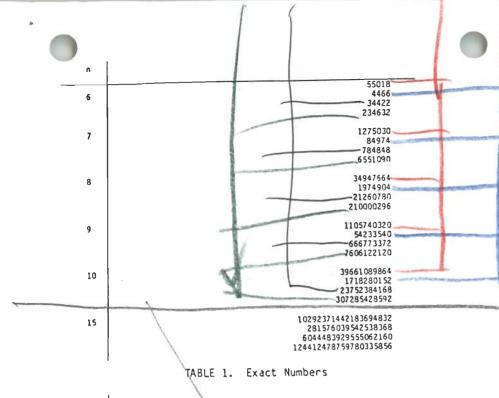
$$= 0.19173964874 \dots + 0(\frac{1}{n})$$
(3.27)

4. NUMERICAL RESULTS

The values of P_n , T_n , R_n , and S_n for $1 \le n \le 0$ and n=15,20, 25,30,35, and 40 for Case 2 are presented in Table 1. The full range of values for $1 \le n \le 40$ is available from the second author as well as similar data for Case 1. Computations for P_n are based on (2.5) and (3.5), for T_n on (3.9), and for R_n on (3.16).

In Table 2 the corresponding values of μ_n and ν_n are given. Asymptotic estimates of the series of exact values in Tables 1 and 2 are provided by (3.20), (3.21), (3.22), (3.24), and (3.27). To indicate the accuracy of these estimates for moderate values of n Table 3 gives relative errors in the asymptotic estimates for n=40. Moreover the expected dependece on n is already exhibited in each case for values of n is the range 10 < n < 40. The computations were performed on a PDP 11/45 by A. Nymeyer while employed under an A.R.G.C. grant.





	\
n	
20	1658651393866913825412347136
	331558482325457407154881024
	9651410476312244702073046400
	272427425074750220273667059456
25	101221761373619922800249422869301911552
	1594048069672352514320983443298130432
	58583850270222052787651170611688084992
	2101394128209830356067285384657612477440
30	1764154837879537542660242990522050740341834303488
	22918826670728264538230908877077055042504704
	1017456530701257631633179282273374276193693382656
	44272238526003063868886828719981453807405390536704
35	7312750677018907562953247815434744392016584820547538714624
	anaxaaqqqqqqqqqqqqqqqqqqqqqqqqqqqqqqqqq
	#20712029870846757189428293982 <u>0</u> 49291364152950823514924087300
	215216354574856734864071208143598086982845362083348942159872
40	63405541022533951267592273434008979603384801773767941774342
	04029.059.072
	61012069737310896276790624496121100929614581843381834791118
	301208688
	364401366961532199066420794688289571691296480722079356668745
	429032960
	214088566327651817150312475146570303448193008959184285060142
	783957041152

TABLE 1 Exact Numbers

n	μ _n	٧n
1 2 3 4 5 6 7 8 9 10 15 20 25	1.00000 2.00000 3.25000 4.67857 6.18272 7.70757 9.23633 10.7655 12.2945 13.8233 21.4666 29.1092 36.7516 44.3939	0.00000 0.00000 187500 432398 64767 838353 1.02175 1.20445 1.38772 1.57154 2.4944 3.41948 4.34524 5.27132
35 40	52.0362 59.6784	6.19756 7.12390

TABLE 2 Mean and Variance of the Number of Vertices

P40	2.4193359	x10 ⁻⁴
T ₄₀	-1.8818933	x10 ⁻³
μ40	0.9	x10-5
¥40	7.6599	x10-2

TABLE 3 Accuracy of Asymptotic Estimates

5. SUMMARY AND CONCLUSIONS

In this paper we have continued our analysis of the enumeration of different classes of phylogenetic (evolutionary) trees. The two cases, which we studied here are: (i) trees with no vertices of degree two in which all labels are singletons and interior points may be labelled, and (ii) trees with no restriction on vertex degree, all nonempty labels are singletons, and only vertices of degree one or two can be labelled. The exact and asymptotic number of trees with given magnitude, along with the average and variance of their order was determined for each case.

In both cases the mean and variance of the order were both O(n) so that the distribution of orders in trees of magnitude n becomes gradually sharper as $n \leftrightarrow \infty$. This is also true in all the earlier

casas studied.

It is planned to publish the analysis of the remaining two cases (Cases 3 and 4) elsewhere.

ACKNOWLEDGEMENT. The second author is grateful for the support of the Australian Research Grants Committee for the project "Numerical Implementation of Unlabelled Graph Counting Algorithms", under which research and computing for the results of this paper were performed.

REFERENCES

- [1] E. A. Bender, Asymptotic methods in enumeration, SIAM Review. 16 (1974) 485-515.
- [2] J. Felsenstein, The numbers of evolutionary trees, Systematic Zoology. 27 (1978) 27-33.
- [3] L. R. Foulds and R. W. Robinson, Determining the asymptotic numbers of phylogenetic trees, Combinatorial Mathematics VII, Lecture Notes in Mathematics 829, Springer-Verlag, Berlin (1980) 110-126.
- [4] L. R. Foulds and R. W. Robinson, Enumeration of binary phylogenetic trees, Combinatorial Mathematics VIII, Lecture Notes in Mathematics 884, Springer-Verlag, Berlin (1981) 187-202.
- [5] L. R. Foulds and R. W. Robinson, Enumeration of phylogenetic trees without points of degree two, Ars Combinatoria (1984) to appear.
- [6] F. Harary, Graph Theory, Addison-Wesley, Reading Mass. (1969).
- [7] F. Harary and E. Palmer, <u>Graphical Enumeration</u>, Academic, New York (1973).
- [8] F. Harary, R. W. Robinson, and A. J. Schwenk, Twenty step algorithm for determining the asymptotic number of trees various species, <u>J. Austral. Math. Soc</u>. Ser. A 20 (1975) 483-503.
- [9] E. Hille, Analytic Function Theory, Vol. 1, Ginn, Boston (1959).
- [10] E. D. Penny, M. D. Hendy, and L. R. Foulds, <u>The Science of Phylogenetic Tree Construction</u>, Springer-Verlag, New York. To appear.