J. Austral. Math. Soc. (Series A) 30 (1980), 229-237

RANDOM LABELLED TREES AND THEIR BRANCHING NETWORKS

G. R. GRIMMETT

(Received 2 January; revised 20 March 1980)

Communicated by W. S. Wallis

Abstract

A random rooted labelled tree on n vertices has asymptotically the same shape as a branching-type process, in which each generation of a branching process with Poisson family sizes, parameter one, is supplemented by a single additional member added at random to one of the families in that generation. In this note we use this probabilistic representation to deduce the asymptotic distribution of the distance from the root to the nearest endvertex other than itself.

1980 Mathematics subject classification (*Amer. Math. Soc.*): primary 05 C 05: secondary 60 J 80.

Keywords: random tree, random graph, branching process.

1. Introduction

There are n^{n-2} different trees on the labelled vertex set $\{v_1, v_2, \ldots, v_n\}$. Let T_n denote a randomly chosen member of this collection, each member having equal probability of being chosen. Meir and Moon have explored many properties of T_n when n is large; their methods are largely enumerative and analytical and make considerable use of the exponential generating function

$$y = \sum_n n^{n-2} x^n/(n-1)!$$

of the numbers of labelled trees. It is the purpose of this note to exploit a probabilistic representation of T_n in terms of branching processes in order to solve a problem concerning the shortest distance between a specified vertex and the endvertices of T_n other than this vertex.

A *random walk* about the vertices of T_n is a sequence $V_0, V_1, V_2, \ldots$ of vertices, possibly with repetition, randomly chosen according to the rule that V_{k+1} is picked uniformly at random from the collection of vertices of T_n which are adjacent to V_k; V_0

https://doi.org/10.1017/S1446788700016517 Published online by Cambridge University Press

is a specified starting point. Conditional upon T_n being given, $\{V_i: i = 0, 1, \ldots\}$ is a Markov chain; it is easy to check that it has stationary distribution $\{\pi_i: i = 1, 2, \ldots, n\}$ given by

$$\pi_i = d_i(2(n-1))^{-1} \quad (i = 1, 2, \ldots, n),$$

where d_i is the degree of the vertex v_i in T_n; a similar result holds for random walks about the vertex set of any finite undirected graph. Meir and Moon have been concerned with certain first passage times associated with the chain. For example, Moon (1973) shows that the time v_n of the first visit to v_2 starting from v_1, given formally by

$$v_n = \min\{m: V_m = v_2\} \quad \text{given } V_0 = v_1,$$

satisfies

$$(1) \qquad E(v_n) \sim (\tfrac{1}{2}\pi n^3)^{\frac{1}{2}}.$$

Also, Meir and Moon (1975) show that the time λ_n of the earliest visit to an endvertex of T_n after the commencement of the walk, given by

$$\lambda_n = \min\{m \geqslant 1: V_m \text{ is an endvertex}\} \quad \text{given } V_0 = v_1,$$

satisfies

$$(2) \qquad P(\lambda_n = k) \to kp^2 q^{k-1}, \quad \text{where } p = 1 - q = e^{-1},$$

and

$$E(\lambda_n) \to 2e - 1.$$

Related work on climbing random trees is described in Moon (1976) and Meir and Moon (1978a). Result (1) contrasts strongly with an earlier result of Meir and Moon (1970, 1978b), which asserts that the number γ_n of edges in the unique path of T_n joining vertices v_1 and v_2 has mean value

$$E(\gamma_n) \sim (\tfrac{1}{2}\pi n)^{\frac{1}{2}}.$$

Here, we answer the corresponding question to result (2): what is the limiting distribution as $n \to \infty$ of the distance between v_1 and the nearest endvertex of T_n other than v_1? In Section 2 we describe an enumerative approach to the problem, in the spirit of Rényi and Szekeres (1967), who used techniques of complex analysis to answer a similar question. In Section 3 we show that, for large n, T_n has asymptotically the same stochastic shape as a random process defined in terms of branching processes. This observation enables us to use easy recursive arguments to solve the problem in question. Other problems may be susceptible to solution by the same method.

Similar combinatorial and probabilistic arguments may be used to find the asymptotic distribution of the distance from v_1 to the nearest endvertex, *including* v_1 if it has degree one.

See Grimmett (1980) for a review of the theory of random trees and other random graphs.

2. A combinatorial approach

Let T be a labelled n-tree with vertices $v_1, v_2, \ldots, v_n$, and let $d(T)$ be the number of edges in the shortest path from v_1 to an endvertex of T, other than v_1 if it has degree one. Let $T(k,n) = |\{T: d(T) \geqslant k\}|$ be the number of n-trees with $d(T) \geqslant k$, and let

$$G_k(x) = \textstyle\sum_{n=1}^{\infty} T(k,n)\, x^n/(n-1)! \tag{3}$$

be the exponential generating function of the $T(k,n)$ for fixed k. It is clear that $T(0,n) = n^{n-2}$, and it follows that $G_0(x)$ satisfies

$$G_0(x) = x \exp(G_0(x)).$$

Furthermore, if $n, k \geqslant 1$,

$$T(k,n) = \textstyle\sum_{r=1}^{n-1} \binom{n-1}{r} \sum_{\mathbf{a}} \binom{n-r-1}{a_1-1, \ldots, a_r-1} \tag{4}$$
$$\times\, T(k-1,a_1)\, T(k-1,a_2) \ldots T(k-1,a_r),$$

where $\sum_{\mathbf{a}}$ sums over all sequences $\{a_1, a_2, \ldots, a_r\}$ with sum $n-1$. This holds because $d(T) \geqslant k$ if and only if each vertex v adjacent to v_1 is at least distance $k-1$ from the nearest endvertex other than v itself in the subtree of T comprising all vertices including v whose unique path to v_1 passes through v. Multiply (4) by $x^n/(n-1)!$ and sum over n to obtain

$$G_k(x) = x(\exp(G_{k-1}(x)) - 1), \tag{5}$$

which is a functional recurrence relation for the G_k subject to the condition

$$G_0(x) = \textstyle\sum_n n^{n-2} x^n/(n-1)! \tag{6}$$

It may be possible to use these relations to determine the asymptotic probabilities

$$\lim_{n\to\infty} T(k,n)\, n^{2-n}$$

that a large random tree T_n has $d(T_n) \geqslant k$. Rényi and Szekeres (1967) have been able to do this in a similar case; however, it seems more natural to proceed by probabilistic techniques.

https://doi.org/10.1017/S1446788700016517 Published online by Cambridge University Press

3. A branching-type process

Let $\{X_i : i \geqslant 0\}$ be the sizes of the generations in a branching process in which $X_0 = 1$ and each family size has the Poisson distribution with parameter λ. The total progeny

$$N = \textstyle\sum_{i=0}^{\infty} X_i$$

is almost surely finite if and only if $\lambda \leqslant 1$. The graphical representation of the process is a tree with N vertices which is similar in appearance to a random labelled N-tree in the following sense. Consider the random tree T_n rooted at v_1. The kth *stratum* of T_n is the collection of vertices of T_n which are exactly distance k from v_1. Let $Z(k, n)$ be the size of the kth stratum.

THEOREM 1 (Kolchin (1977)). *The sequence $\{Z(i,n): 0 \leqslant i < n\}$ has the same joint distribution as the sequence $\{X_i: 0 \leqslant i < \infty\}$ conditional upon the event $\{N = n\}$. That is,*

$$P(Z(i,n) = r_i,\ 0 \leqslant i < n) = P(X_i = r_i,\ 0 \leqslant i < \infty \mid N = n). \tag{7}$$

Kennedy (1956) has studied asymptotic properties of a branching process conditioned upon the value of the total progeny. His results imply that the asymptotic distribution are independent of λ; henceforth we assume that $\lambda = 1$. Furthermore he shows the following

THEOREM 2 (Kennedy (1975)). *If $i_1 < i_2 < \dots < i_k$ then*

$$\lim_{n\to\infty} P(X_{i_j} = r_j,\ j = 1, 2, \dots, k \mid N = n) = r_k\, P(X_{i_j} = r_j,\ j = 1, 2, \dots, k) \tag{8}$$

These two theorems contain information about the likely shape of a random tree as $n \to \infty$. When suitably corrected, Kennedy's Theorem 3 contains as a special case a third solution to the problem, independently solved twice already by Rényi and Szekeres (1967) and Stepanov (1969), of determining the asymptotic distribution of the height of a random tree.

It is not difficult to deduce that the asymptotic form of T_n is the same as that of a branching type process T defined as follows. A population is constructed recursively according to generation number by the stochastic rules:

(a) The zeroth generation contains one member;

(b) For $k \geqslant 0$, the $(k+1)$th generation A_{k+1} is the union of the families of descendants of the members of the kth generation A_k, together with one additional member which is allocated at random to one of these families, each of

https://doi.org/10.1017/S1446788700016517 Published online by Cambridge University Press

the $|A_k|$ families having equal probability of being chosen. Before the allocation of this supplementary member, all family sizes are independent of each other and the past, and are Poisson distributed with parameter one.

Hence, for example, the size Y_k of the kth generation satisfies

$$Y_{k+1} = 1+F_1+F_2+\ldots+F_{Y_k}, \tag{9}$$

where $\{F_i : i \geqslant 1\}$ is a collection of independent Poisson variables; the process can never become extinct. Realizations may be represented graphically as trees rooted at the initial member, with edges joining father–son pairs.

The next theorem concerns the asymptotic shape of T_n for large n. Let k be a positive integer. Draw the subtree of T_n comprising the vertices of T_n which are in one of the first k strata only. Within any family, order the vertices in, say, increasing lexicographic order, and then delete all labels except that of the root; call the resulting random tree T_n^k.

THEOREM 3. *As $n \to \infty$ the characteristics of T_n^k become indistinguishable in distribution from the corresponding characteristics of the tree representation of the first k generations of the branching type process T described above.*

Thus the probabilities of events such as $\{d(T_n) \geqslant k\}$, which are describable in terms of the first l strata of T_n for some l, approach as $n \to \infty$ the probabilities that the corresponding events occur in T. That is, instead of calculating $P(d(T_n) \geqslant k)$ and letting $n \to \infty$, we may first let $n \to \infty$ and then calculate the limit probability directly.

The next theorems are corollaries.

THEOREM 4. *The number $N(k,n)$ of endvertices of T_n in the kth stratum has probability generating function*

$$H_{k,n}(x) = E(x^{N(k,n)}) \to \frac{d}{dy} f_k(y)\bigg|_{y=y_0},$$

where $y_0 = 1-e^{-1}+xe^{-1}$ and f_k is the kth iterate of the Poisson generating function

$$f(y) = \exp(y-1).$$

THEOREM 5. *$P(d(T_n) \geqslant k) \to \pi_k$ as $n \to \infty$ where*

$$\pi_k = \exp(\textstyle\sum_{i=1}^{k-1} \alpha_i)$$

and the α_i are given recursively by

$$\alpha_0 = 0, \quad \alpha_{i+1} = \exp(\alpha_i)-e^{-1}-1.$$

It is not difficult to show that $\alpha_i + 1 \sim Ae^{-i}$ for some $A > 0$.

https://doi.org/10.1017/S1446788700016517 Published online by Cambridge University Press

Proof of theorem 3. We show that the joint distribution of the family sizes of members of the kth generation of the branching-type process, conditional upon the past, is the limit as $n \to \infty$ of the joint distribution of the corresponding quantities associated with the vertices in the kth stratum of T_n, conditional upon the numbers of vertices in earlier strata.

First consider the process $\{Y_i \colon i \geqslant 0\}$. List the family sizes $G_1, \ldots, G_r$ of the r members of the kth generation in some arbitrary order. Their joint probability mass function is

$$(10) \qquad p(a_1, a_2, \ldots, a_r) = \textstyle\sum_{i=1}^{r} a_i (r e^r a_1! a_2! \ldots a_r!)^{-1}$$

$$= \frac{u}{r} \frac{e^{-r}}{a_1! \ldots a_r!}$$

where $u = a_1 + a_2 + \ldots + a_r$.

Now let $\mathbf{z} = (z_1, z_2, \ldots, z_k)$ be a sequence of positive integers, and let $T(\mathbf{z}, n)$ be the number of trees with n labelled vertices such that $Z(i, n) = z_i$ for $i = 1, 2, \ldots, k$. Then putting $r = z_k$ and $s = 1 + z_1 + z_2 + \ldots + z_k$,

$$T(\mathbf{z}, n) = C(\mathbf{z}, n) \textstyle\sum_{\mathbf{b}} \binom{n-s}{b_1 - 1, b_2 - 1, \ldots, b_r - 1} b_1^{b_1 - 2} b_2^{b_2 - 2} \ldots b_r^{b_r - 2},$$

where $\sum_{\mathbf{b}}$ sums over all sequences $(b_1, b_2, \ldots, b_r)$ with sum $n + r - s$ and $C(\mathbf{z}, n)$ is the number of ways of labelling the s vertices in the first k strata.

But consider the number of trees with N labelled vertices $v_1, v_2, \ldots, v_N$ in which v_1 has degree t. The number of such trees is

$$(11) \qquad \binom{N-1}{t} \textstyle\sum_{\mathbf{b}} \binom{N-t-1}{b_1 - 1, b_2 - 1, \ldots, b_t - 1} b_1^{b_1 - 2} \ldots b_t^{b_t - 2} \sim N^{N-2} (e(t-1)!)^{-1}$$

since the probability that v_1 has degree t in T_N approaches $(e(t-1)!)^{-1}$ as $N \to \infty$. Thus

$$(12) \qquad T(\mathbf{z}, n) \sim C(\mathbf{z}, n)(n + r - s + 1)^{n-s-1} r e^{-1}.$$

The number of trees contributing to $T(\mathbf{z}, n)$ which have the property that $(a_1, a_2, \ldots, a_r)$ is the sequence, arbitrarily ordered, of numbers of pendant vertices from the r vertices in the kth stratum is

$$T(\mathbf{z}, \mathbf{a}, n) = C(\mathbf{z}, n) \binom{n-s}{a_1, a_2, \ldots, a_r, a} \textstyle\sum_{\mathbf{b}} \binom{n-s-u}{b_1 - 1, \ldots. b_u - 1} b_1^{b_1 - 2} \ldots b_u^{b_u - 2},$$

where $u = a_1 + a_2 + \ldots + a_r$ and $a = n - s - u$. From (11).

$$(13) \qquad T(\mathbf{z}, \mathbf{a}, n) \sim C(\mathbf{z}, n) \frac{(n-s+1)^{n-s-1}}{a_1! \ldots a_r!} u e^{-1}.$$

https://doi.org/10.1017/S1446788700016517 Published online by Cambridge University Press

Divide (13) by (12) and let $n \to \infty$ to obtain

$$T(\mathbf{z}, \mathbf{a}, n)/T(\mathbf{z}, n) \sim \frac{u}{r}\frac{(1+r(n-s+1)^{-1})^{-n}}{a_1!\dots a_r!}$$

$$\to \frac{u}{r}\frac{e^{-r}}{a_1!\dots a_r!}$$

which coincides with (10) as required. This proves the theorem.

Proof of Theorem 4. From (7), the number $Z(k, n)$ in the kth stratum of T_n has the same distribution as X_k, conditional upon $\{N = n\}$. Hence, from (8),

$$P(Z(k, n) = i) \to iP(X_k = i).$$

Thus, the probability generating function $F_{k,n}(y) = E(y^{Z(k,n)})$ satisfies

$$F_{k,n}(y) = \textstyle\sum_i y^i P(Z(k, n) = i)$$

$$\to \textstyle\sum_i y^i iP(X_k = i)$$

$$= y\frac{d}{dy} f_k(y) = F_k(y).$$

Of course, Y_k also has generating function $F_k(y)$. Furthermore, the number N_k of contributors to Y_k which have family size zero is the sum of $Y_k - 1$ Bernoulli variables, each taking the value 1 with probability e^{-1}; the Y_kth family is that with the compulsory supplementary member. Hence

$$H_{k,n}(x) \to E(x^{N_k})$$

$$= y^{-1} F_k(y)|_{y=y_0}$$

as required.

Proof of theorem 5. Let $\mathbf{i} = (i_1, i_2, \dots, i_k)$ be a sequence of positive integers, and let

$$A_j = \{N_j = 0\} \quad \text{and} \quad B_j = \{Y_j = i_j\} \quad \text{for } j = 1, 2, \dots, k.$$

From Theorem 3, $P(d(T_n) \geqslant k+1) \to P(A_1 A_2 \dots A_k)$. But, by elementary considerations about conditional probabilities,

$$P(A_1 A_2, \dots A_k) = \textstyle\sum_{\mathbf{i}} \prod_{j=1}^{k} P(A_j | A_1 \dots A_{j-1} B_1 \dots B_j) P(B_j | A_1 \dots A_{j-1} B_1 \dots B_{j-1}),$$

with the convention that the intersection of an empty set of events is the whole sample space. Using the Markov property

$$P(A_1 A_2 \dots A_k) = \textstyle\sum_{\mathbf{i}} \prod_{j=1}^{k} P(A_j | B_j) P(B_j | A_{j-1} B_{j-1}) \tag{14}$$

$$= \textstyle\sum_{\mathbf{i}} \prod_{j=1}^{k} (1-e^{-1})^{i_j - 1} C_j(i_j),$$

https://doi.org/10.1017/S1446788700016517 Published online by Cambridge University Press

where $C_j(i)$ is the coefficient of x^i in the generating function $D_j(x)$ of Y_j conditional upon $Y_{j-1} = i_{j-1}$ and $N_{j-1} = 0$. Thus

$$Y_j = 1 + P + F_1 + \ldots + F_{i_{j-1}-1},$$

where P has the Poisson distribution and the F_i are independent variables each with the Poisson distribution conditioned on being nonzero. Hence

$$D_j(x) = xe^{x-1}((e^x - 1)/(e-1))^{i_{j-1}-1}.$$

Summing (14) over i_k gives

$$(15) \qquad P(A_1 A_2 \ldots A_k) = \textstyle\sum_{(i_1,\ldots,i_{k-1})} \prod_{j=1}^{k-1} \beta_1^{i_j-1} C_j(i_j) e^{\beta_1-1} \left(\dfrac{e^{\beta_1}-1}{e-1}\right)^{i_{k-1}-1},$$

where $\beta_1 = 1 - e^{-1}$. Sum (15) over i_{k-1} to obtain

$$(16) \qquad P(A_1 A_2 \ldots A_k) = \textstyle\sum_{(i_1,\ldots,i_{k-2})} \prod_{j=1}^{k-2} \beta_1^{i_j-1} C_j(i_j) e^{\beta_1+\beta_2-2} \left(\dfrac{e^{\beta_2}-1}{e-1}\right)^{i_{k-2}-1}$$

where $\beta_2 = e^{-1}(\exp(\beta_1) - 1)$. Continue, to show that

$$P(A_1 A_2 \ldots A_k) = \exp(\textstyle\sum_{i=1}^{k}(\beta_i - 1)) = \exp(\sum_{i=1}^{k} \alpha_i) \quad \text{for } k \geqslant 1,$$

as required, where $\beta_0, \beta_1, \ldots$ are given by the recursion

$$\beta_0 = 1, \quad \beta_{i+1} = e^{-1}(\exp(\beta_i) - 1),$$

and $\alpha_i = \beta_i - 1$. It is easy to check that the β_i satisfy $\beta_i \sim Ae^{-i}$ for some $A > 0$.

Acknowledgement

I am very grateful to Professor J. W. Moon for indicating and correcting an error in an earlier combinatorial argument which I used in Section 2.

References

G. R. Grimmett (1980), 'Random graphs', *Further selected topics in graph theory*, edited by L. Beineke and R. Wilson (Academic Press), to appear.

D. P. Kennedy (1975), 'The Galton–Watson process conditioned on the total progeny', *J. Appl. Probability* **12**, 800–806.

V. F. Kolchin (1977), 'Branching processes, random trees, and a generalized scheme of arrangements of particles', *Math. Notes* **21**, 386–394.

A. Meir and J. W. Moon (1970), 'The distance between points in random trees', *J. Combinatorial Theory* **8**, 99–103.

A. Meir and J. W. Moon (1975), 'Climbing certain types of rooted trees I', *Proc. 5th British Combinatorial Conf.*, pp. 461–469.

A. Meir and J. W. Moon (1978a), 'Climbing certain types of rooted trees II', *Acta Math. Acad. Sci. Hungar.* **31**, 43 54.

A. Meir and J. W. Moon (1978b), 'On the altitude of nodes in random trees', *Canad. J. Math.* **30**, 997–1015.

J. W. Moon (1970), 'Climbing random trees', *Aequationes Math.* **5**, 68–74.

J. W. Moon (1973), 'Random walks on random trees', *J. Austral. Math. Soc.* **15**, 42–53.

A. Rényi and G. Szekeres (1967), 'On the height of trees', *J. Austral. Math. Soc.* **7**, 497–507.

V. E. Stepanov (1969), 'On the distribution of the number of vertices in strata of a random tree', *Theor. Probability Appl.* **14**, 65–78.

School of Mathematics
University of Bristol
Bristol BS8 1TW
England

https://doi.org/10.1017/S1446788700016517 Published online by Cambridge University Press