Plane recursive trees, Stirling permutations and an urn model

Svante Janson

Fifth Colloquium on Mathematics and Computer Science DMTCS proc. **AI**, 2008, 541–548

Plane recursive trees, Stirling permutations and an urn model

Svante Janson

Department of Mathematics, Uppsala University, PO Box 480, SE-751 06 Uppsala, Sweden svante.janson@math.uu.se

We exploit a bijection between plane recursive trees and Stirling permutations; this yields the equivalence of some results previously proven separately by different methods for the two types of objects as well as some new results. We also prove results on the joint distribution of the numbers of ascents, descents and plateaux in a random Stirling permutation. The proof uses an interesting generalized Pólya urn.

Keywords: Plane recursive trees, Stirling permutations, number of ascents, number of descents, urn model, generalized Pólya urn

1 Introduction

A *plane recursive tree* is a rooted plane (= ordered) tree obtained by starting with the root and recursively adding leaves to the tree. We label the root by 0 and the added vertices by $1, 2, \dots$; a plane recursive tree with $n+1$ vertices is thus a labelled rooted plane tree (with labels $0, \dots, n$) where the labels increase along each branch as we travel from the root. Since a new vertex may be joined to an existing vertex i in $d_i + 1$ positions, where d_i is the outdegree of i, the total number of positions for vertex n is $\sum_{i=0}^{n-1}(d_i + 1) = 2n - 1$ (and in particular not depending on the present shape of the tree), and thus the number of plane recursive trees with $n + 1$ vertices is $(2n - 1)!! := 1 \cdot 3 \cdots (2n - 1)$. A *random plane recursive tree* with $n + 1$ vertices is obtained by randomly (and uniformly) choosing one of these $(2n - 1)!!$ trees; equivalently, it is obtained by recursively adding n vertices to the root, each time choosing the position uniformly at random among the possibilities. Random plane recursive trees were studied by Mahmoud, Smythe and Szymański (12). They obtained, among other results, an asymptotic joint normal distribution for the numbers of vertices of outdegrees 0,1,2; this was later extended to arbitrary outdegrees by Janson (10). More recently, the distribution of the degree of the node with a given label has been studied by Kuba and Panholzer (11). See also (3) and the survey (14), where also other related types of trees are studied.

The well-known *depth first walk* of a rooted plane tree starts at the root, goes first to the leftmost daughter of the root, explores that branch (recursively, using the same rules), returns to the root, and continues with the next daughter of the root, until there are no more daughters left. Note that every edge is passed twice in this walk, once in each direction. We find it convenient to label the edges too in a plane recursive tree, using the labels $1, 2, \dots$ in the order the edges are added to the tree. Thus, edge j is the edge connecting vertex j to some earlier vertex. We code a plane recursive tree by the sequence of the labels of the edges passed by the depth first walk; a plane recursive tree with $n + 1$ vertices is thus

coded by a string of length $2n$, where each of the labels $1,\dots,n$ appears twice. In other words, the code is a permutation of the multiset $\{1,1,2,2,\dots,n,n\}$. Adding a new vertex $n+1$ means inserting the pair $(n+1)(n+1)$ somewhere in the code, at one of $2n+1$ possible places (first, last, or in any of the $2n-1$ gaps between of consecutive labels) corresponding to the $2n+1$ places in the tree where a new leaf might be added. (We will in the sequel call all these $2n+1$ places 'gaps', including the places before the first element and after the last.) This shows that the code determines the plane recursive tree uniquely. Moreover, it follows that the possible codes are exactly the *Stirling permutations* defined by Gessel and Stanley (6): a Stirling permutation is a permutation of $\{1,1,2,2,\dots,n,n\}$ such that for each $i \le n$, the elements occuring between the two occurences of i are larger than i. (The name comes from relations with the Stirling numbers, see (6).)

Letting $\mathcal{T}_n$ be the set of plane recursive trees with $n+1$ vertices (and thus n edges) and $\mathcal{Q}_n$ the set of Stirling permutations of length $2n$, there is thus a simple bijection $\mathcal{T}_n \cong \mathcal{Q}_n$. We let T_n denote a random plane recursive tree with $n+1$ vertices, i.e., a (uniformly) random element of $\mathcal{T}_n$, and let similarly Q_n denote a random Stirling permutation of length $2n$, i.e., a (uniformly) random element of $\mathcal{Q}_n$; the bijection $\mathcal{T}_n \cong \mathcal{Q}_n$ thus yields a correspondence between the random objects T_n and Q_n. One of the purposes of this note is to show how this correspondence connects some previous results on random plane recursive trees and random Stirling permutations. We will also extend some of these results. (The correspondence is very simple, but we have not seen it utilized in the literature before.)

Gessel and Stanley (6) studied the number of descents in Stirling permutations. This was recently continued and extended by Bona (4), using the following definitions: If $a_1a_2\cdots a_{2n}$ is a Stirling permutation, say that an index $i = 0,\dots,2n$ (or the gap $i, i+1$) is an *ascent* if $a_i < a_{i+1}$, a *descent* if $a_i > a_{i+1}$, and a *plateau* if $a_i = a_{i+1}$, where we set $a_0 = a_{2n+1} = 0$. (Thus 0 is always an ascent and $2n$ a descent.) Let X_n, Y_n and Z_n denote the numbers of ascents, descents and plateaux, respectively, in a random Stirling permutation Q_n. Thus

$$X_n + Y_n + Z_n = 2n+1. \tag{1.1}$$

Let N_{nd} denote the number of vertices with outdegree d in the random plane recursive tree T_n and let $L_n := N_{n0}$ be the number of leaves. It is immediately seen that in the correspondence above, leaves in the plane recursive tree correspond to plateaux in the Stirling permutation, and thus the number of leaves in the random plane recursive tree T_n equals the number of plateaux in Q_n, i.e.,

$$L_n = Z_n. \tag{1.2}$$

As said above, L_n was studied by Mahmoud, Smythe and Szymański (12) (note that our L_n is their L_{n+1}). They proved, using simple recursion relations,

$$\mathbb{E}\, L_n = \frac{2n+1}{3}, \tag{1.3}$$

$$\operatorname{Var} L_n = \frac{2(n^2-1)}{9(2n-1)} \sim \frac{n}{9} \tag{1.4}$$

and, using a generalized Pólya urn (see Section 2 and Remark 2.6) the asymptotic normality

$$\frac{L_n - 2n/3}{\sqrt{n}} \xrightarrow{\mathrm{d}} N(0,1/9). \tag{1.5}$$

(All unspecified limits in this paper are as $n \to \infty$.) Of course, in (1.5), we can replace $2n/3$ by the exact mean $(2n+1)/3$ from (1.3), and by (1.4) the result can also be written $(L_n - \mathbb{E}\, L_n)/\sqrt{\operatorname{Var} L_n} \xrightarrow{\mathrm{d}} N(0,1)$.

Bona (4) proved, among other things, that the random variables X_n, Y_n and Z_n have the same distribution (reproved as Corollary 2.2 below), and thus by (1.1)

$$\mathbb{E}\, X_n = \mathbb{E}\, Y_n = \mathbb{E}\, Z_n = \frac{2n+1}{3}, \tag{1.6}$$

and further, by first showing that the probability generating functions have only real roots, that these random variables are asymptotically normally distributed:

$$\frac{X_n - 2n/3}{\sqrt{n}} \overset{\mathrm{d}}{=} \frac{Y_n - 2n/3}{\sqrt{n}} \overset{\mathrm{d}}{=} \frac{Z_n - 2n/3}{\sqrt{n}} \xrightarrow{\mathrm{d}} N(0, 1/9). \tag{1.7}$$

We now see that the equality $L_n = Z_n$ means that the results (1.3) and (1.6) are equivalent, as well as (1.5) and (1.7). Furthermore, (1.4) yields also an exact formula for the variance of X_n, Y_n and Z_n, which was raised as a question in (4).

In Theorem 2.4 we will extend (1.7) to joint convergence to a joint normal distribution.

Remark 1.1 *Let $C_{n,k}$ be the number of plane recursive trees with $n+1$ vertices and k leaves, or, equivalently by the discussion above, the number of Stirling permutations of length $2n$ with k plateaux (or k ascents, or k descents). It is easy to see that $C_{n,k}$ satisfies the recursion*

$$C_{n,k} = kC_{n-1,k} + (2n-k)C_{n-1,k-1} \tag{1.8}$$

for all $n \geq 2$ and $k \geq 1$ (or $k \in \mathbb{Z}$), with $C_{1,k} = \delta_{1k}$ and $C_{n,0} = 0$, see (6), (12), (8). These numbers $C_{n,k}$ are called second-order Eulerian numbers (8, §6.2); in standard notation $C_{n,k} = \left\langle\!\left\langle {n \atop k-1} \right\rangle\!\right\rangle$ (8), and thus (12), for $n \geq 1$,

$$\mathbb{P}(L_n = k) = \mathbb{P}(X_n = k) = \mathbb{P}(Y_n = k) = \mathbb{P}(Z_n = k) = \frac{\left\langle\!\left\langle {n \atop k-1} \right\rangle\!\right\rangle}{(2n-1)!!}. \tag{1.9}$$

Remark 1.2 *Another combinatorial interpretation of the same numbers $C_{n,k}$ is given by Riordan (13) as the number of* trapezoidal words *of length n with k distinct elements, where a trapezoidal word is a word $a_1 \cdots a_n$ with $a_i \in \{1, 2, \ldots, 2i-1\}$, i.e., an element of $[1] \times [3] \times \cdots \times [2n-1]$. (Again, it is easy to verify the recursion* (1.8)*.) Hence L_n etc. also give the distribution of the number of distinct elements in a random trapezoidal word of length n. We do not know any interesting statistics of trapezoidal words that correspond to other statistics of plane recursive trees or Stirling permutations such as N_{nd} ($d \geq 1$) or the triple (X_n, Y_n, Z_n).*

Remark 1.3 *For the connections between the second-order Eulerian numbers $C_{n,k}$ and Stirling numbers, see e.g. (7), (5), (13), (6), (8).*

The correspondence between plane recursive trees and Stirling permutations makes it natural to study also other statistics of one of these objects and see what they correspond to for the other object, thus giving more equalities of the type (1.2). (In some cases the results are, however, disappointing in that the resulting statistics seem uninteresting.)

Example 1.4 *The number of plateaux in a Stirling permutation is, as discussed above, equal to the number of leaves in the corresponding plane recursive tree. It seems natural to consider ascents and descents also. It is straight-forward to see that an ascent corresponds to a non-root vertex in the tree that either has no sister to its left, or else has a higher label than the sister that is nearest to it to the left; similarly, a descent corresponds to a non-root vertex in the tree that either has no sister to its right, or else has a higher label than the sister that is nearest to it to the left. We can thus interpret the results on* (X_n, Y_n, Z_n) *as results on the numbers of such vertices in a random plane recursive tree, but it remains to show that these numbers are interesting.*

Example 1.5 *The distance* D_{nk} *between the two occurences of* k *in a random Stirling permutation of length* $2n$ *is perhaps a more interesting example. It equals* $2S_n^{(k)} - 1$ *where* $S_n^{(k)}$ *is the size of the subtree rooted at vertex* k *of the corresponding random plane recursive tree. The latter variable was studied by Mahmoud, Smythe and Szymański (12) using a Pólya–Eggenberger urn; they obtained both exact formulas for the distribution, mean and variance, as well as an asymptotic beta distribution:* $S_n^{(k)}/n \xrightarrow{\mathrm{d}} \beta(\frac{1}{2}, k)$ *as* $n \to \infty$ *for every fixed* $k \geq 1$. *(Recall that we label the root by 0, so the labels are shifted from (12).) These results immediately tranfer to results about* D_{nk}*, for example, for any fixed* $k \geq 1$*,*

$$\frac{D_{nk}}{2n} \xrightarrow{\mathrm{d}} \beta(\tfrac{1}{2}, k) \qquad \text{as } n \to \infty. \tag{1.10}$$

Example 1.6 *The degree of the root of* T_n *was also studied in (12). This corresponds for a Stirling permutation* Q_n *to the number* d *such that the Stirling permutation is of the form* $a_1 \cdots a_1 a_2 \cdots a_2 \cdots a_d \cdots a_d$*. Again, this is not of any obvious great interest.*

2 An interesting urn model

Mahmoud, Smythe and Szymański (12) used a representation with a generalized Pólya urn to prove the asymptotic normality (1.5). (In fact, they proved joint asymptotic normality of N_{nd} ($d \leq 2$), which was extended in (10) to all d, see also (9, Example 7.6).) We use a similar urn to study the joint distribution of (X_n, Y_n, Z_n).

If we extend a Stirling permutation in $\mathcal{Q}_n$ by inserting the pair $(n+1)(n+1)$ at one of the $2n+1$ possible places (gaps), then the ascent, descent or plateau at that gap is destroyed and replaced by the sequence *ascent, plateaux, descent* at the three resulting new gaps. Consequently, the random vector (X_n, Y_n, Z_n) is described by the following generalized Pólya urn model:

Urn I *Consider an urn with balls of three colours, and let* (X_n, Y_n, Z_n) *be the number of balls of each colour at time* n*. At each time step, draw one ball at random from the urn, discard it, and add one new ball of each colour. Start with* $(X_1, Y_1, Z_1) = (1, 1, 1)$.

This urn model is completely symmetric in the three colours, and we thus immediately see the following.

Theorem 2.1 *For each* $n \geq 1$*, the distribution of* (X_n, Y_n, Z_n) *is exchangeable, i.e., invariant under any permutation of the three variables.*

Corollary 2.2 (Bona (4)) $X_n \stackrel{\mathrm{d}}{=} Y_n \stackrel{\mathrm{d}}{=} Z_n$.

Remark 2.3 *We have stated Theorem 2.1 and Corollary 2.2 for a fixed n, but the results extend to the processes $(X_n)_n$, $(Y_n)_n$, $(Z_n)_n$.*

It is customary and convenient to formulate generalized Pólya urns using drawings with replacement. In the case of Urn I, we thus restate the description above and say instead that we draw a ball and replace it together with one ball each of the two other colours. In other words, Urn I is described by the replacement matrix

$$A = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}. \tag{2.1}$$

We now easily obtain one of our main results.

Theorem 2.4 (X_n, Y_n, Z_n) *are jointly asymptotically normal:*

$$n^{-1/2}(X_n - 2n/3, Y_n - 2n/3, Z_n - 2n/3) \xrightarrow{\mathrm{d}} N(0, \Sigma), \tag{2.2}$$

where the asymptotic covariance matrix is given by

$$\Sigma = \frac{1}{18} \begin{pmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{pmatrix}. \tag{2.3}$$

Proof: It is easily seen that the matrix A in (2.1) has eigenvalues 2, -1, -1. (Note also that A can be regarded as convolution with $(0, 1, 1)$ on the group $\mathbb{Z}_3$, and thus the eigenvalues are given by the Fourier transform of this vector.) In particular, the dominant eigenvalue $\lambda_1 = 2$ (the row sum in A), and all other eigenvalues lie in the half plane $\{\mathrm{Re}\,\lambda < \lambda_1/2\}$. Consequently, (9, Theorem 3.22) applies and shows the asymptotic normality (2.2).

To identify the asymptotic covariance matrix Σ, we may use (9, Lemma 5.4 and (2.15), or Lemma 5.3) and some straight-forward calculations. We may, alternatively, avoid calculations completely by noting that the diagonal entries, which must be equal by Corollary 2.2, are $1/9$ by (1.7); furthermore, the off-diagonal entries also are equal by Theorem 2.1, and the row sums in Σ are 0 as a consequence of (1.1); hence the off-diagonal entries are $-1/18$. □

An exact formula for the covariances is easily obtained too.

Theorem 2.5 *For every $n \geq 1$,*

$$\mathrm{Cov}(X_n, Y_n) = \mathrm{Cov}(X_n, Z_n) = \mathrm{Cov}(Y_n, Z_n) = -\frac{(n^2 - 1)}{9(2n - 1)}. \tag{2.4}$$

Proof: This can be proved using recursion formulas derived from the urn model. It is, however, simpler to use Theorem 2.1 and the already known variance (1.4). Indeed, by symmetry, the three covariances are equal. Further, by (1.1),

$$\mathrm{Var}(X_n) + \mathrm{Cov}(X_n, Y_n) + \mathrm{Cov}(X_n, Z_n) = \mathrm{Cov}(X_n, X_n + Y_n + Z_n) = 0.$$

Since $X_n \overset{\mathrm{d}}{=} Z_n = L_n$, we thus obtain from (1.4)

$$\operatorname{Cov}(X_n, Y_n) = -\frac{1}{2}\operatorname{Var}(X_n) = -\frac{1}{2}\operatorname{Var}(L_n) = -\frac{(n^2-1)}{9(2n-1)}.$$

□

Remark 2.6 *If we only are interested in the univariate distribution of the variables, we may combine two of the colours into one, and thus instead study Urn II: the two-colour urn with replacement vectors* $(0,2)$ *and* $(1,1)$*. This is the urn used by (12) to show the asymptotic normality* (1.5)*.*

Remark 2.7 *Urn I is a rather special type of generalized Pólya urn, since (using the formulation of drawing without replacement), the added balls do not depend on the drawn ball. This is perhaps even more striking in the corresponding continuous-time multitype branching process (see (1), (2,* §*V.9), (9)), which now can be described as follows:* There are individuals of three types (colours). Each individual has an exponential lifetime, independent of all other individuals. When someone dies, one new individual of each of the three colours is born. *Since it does not matter which individuals that die, it might be thought that the result would be like taking a large number* $3n$ *of individuals,* n *of each colour, and randomly removing* $n-1$ *of them. However, this is not correct, not even asymptotically; in fact, a simple calculation of the asymptotic variance in this simplified model (where the number of remaining individuals of a given colour has a hypergeometric distribution) yields* $(4/27)n$ *instead of* $n/9$*. The reason is that although the individuals die independently, they are born together in triplets. Moreover, since the individuals in older triplets have larger probabilities of having died at some given instance, there is a positive correlation between the deaths (up to a given time) for the individuals in the same triplet; since these individuals have different colours, this tends to decrease the variances.*

Consequently, we draw the conclusion (and warning!) that although Urn I is a very simple type of generalized Pólya urn, the fact that the replacements do not depend on the drawn colour does not really simplify the arguments. (At least, we do not see any simplification.) This also emphasizes that it is usually better to formulate generalized Pólya urn models using drawing with replacement; in our case we then have the replacement matrix A *in* (2.1)*, which (although very nice in other respects) describes replacements that do depend on the drawn colour.*

3 Further comments

The proof by Bona (4) of the asymptotic normality (1.7) is completely different and is based on first showing that the probability generating functions have only real roots; thus the random variables can be represented as sums of independent Bernoulli variables. (These Bernoulli variables have in general irrational means and do not have any combinatorial interpretation.) This is a strong property that implies not only asymptotic normality (provided the variance tends to infinity, as in this case); it has other desirable consequences too, for example it leads to explicit error estimates for the convergence to the normal distribution as well as to large deviation estimates (Chernoff bounds).

The success of the generating function in this context suggests that it might be interesting and profitable to study the trivariate probability generating function for (X_n, Y_n, Z_n).

Note that, by Remark 1.1, the univariate probability generating function of L_n (or Z_n) is $(2n-1)!!^{-1}$ times the generating function $\sum_k C_{n,k}x^k$, which has interesting properties studied in (7), (5), (13), (6),

(8). It might be hoped that the trivariate generating function too has some interesting combinatorial properties.

References

[1] K. B. Athreya & S. Karlin, Embedding of urn schemes into continuous time Markov branching processes and related limit theorems. *Ann. Math. Statist.* **39** (1968), 1801–1817.

[2] K. B. Athreya & P. E. Ney, *Branching Processes*. Springer-Verlag, Berlin, 1972.

[3] F. Bergeron, P. Flajolet & B. Salvy, Varieties of increasing trees. *CAAP '92 (Rennes, 1992), Lecture Notes in Comput. Sci.* 581, Springer, Berlin, 1992, 24–48.

[4] M. Bona, Real zeros and normal distribution for statistics on Stirling permutations defined by Gessel and Stanley. Preprint, 2007. `arXiv:0708.3223v1`.

[5] L. Carlitz, The coefficients in an asymptotic expansion. *Proc. Amer. Math. Soc.* **16** (1965) 248–252.

[6] I. Gessel & R. P. Stanley, Stirling polynomials. *J. Combinatorial Theory Ser. A* **24** (1978), no. 1, 24–33.

[7] J. Ginsburg, Note on Stirling's Numbers. *Amer. Math. Monthly* **35** (1928), no. 2, 77–80.

[8] R.L. Graham, D.E. Knuth & O. Patashnik, *Concrete Mathematics*. 2nd ed., Addison–Wesley, Reading, Mass., 1994.

[9] S. Janson, Functional limit theorems for multitype branching processes and generalized Pólya urns. *Stochastic Process. Appl.* **110** (2004), no. 2, 177–245.

[10] S. Janson, Asymptotic degree distribution in random recursive trees. *Random Struct. Alg.* **26** (2005), no. 1–2, 69–83.

[11] M. Kuba & A. Panholzer, On the degree distribution of the nodes in increasing trees. *J. Combin. Theory Ser. A* **114** (2007), no. 4, 597–618.

[12] H. M. Mahmoud, R. T. Smythe & J. Szymański, On the structure of random plane-oriented recursive trees and their branches. *Random Struct. Alg.* **4** (1993), no. 2, 151–176.

[13] J. Riordan, The blossoming of Schröder's fourth problem. *Acta Math.* **137** (1976), no. 1–2, 1–16.

[14] R. T. Smythe and H. Mahmoud, A survey of recursive trees. *Theory Probab. Math. Statist.* **51** (1995), 1–27.