

Simple Proofs for Catalanian Forests and Paths

Len Smiley
University of Alaska Anchorage

April 8, 2005

1 Introduction

”Catalania” is a recent coinage [8, p. 256] related to the sequence of Catalan numbers $C_n = \frac{1}{2n+1} \binom{2n+1}{n}$. A visitor to Catalania may also encounter a three-parameter family of positive integers $F(n, t, k) = \frac{k}{tn+k} \binom{tn+k}{n}$ (so that $C_n = F(n, 2, 1)$). The fun in this discovery is in describing F as a *combinatorial*, as well as arithmetical, generalization of C .

Given integers $t \geq 2$, $k \geq 1$, consider the set $\mathcal{P}(n, t, k)$ of lattice paths with steps $(1, 0)$ and $(0, 1)$ in the (x, y) plane from $(0, 0)$ to $((t-1)n + k - 1, n)$ that never pass above the line $L_t : (t-1)y = x$. It is well known that C_n enumerates these for $t = 2$, $k = 1$.

Similarly, we have the set $\mathcal{A}(n, t, k)$ of plane (unlabelled) forests of exactly k components, each of which is a t -ary tree (all vertices have either 0 or t descendants), such that the total number of internal (non-leaf) vertices is n .

Again C_n famously counts the case $t = 2$, $k = 1$.

In this note, we produce a bijection between paths $\mathcal{P}(n, t, k)$ and forests $\mathcal{A}(n, t, k)$, using an n -sequence of integers as intermediary, then give a quick verification that both families are counted by $F(n, t, k)$.

2 Forests and b -sequences

A plane t -ary forest consists of an ordered set of rooted (ordered) trees whose vertices are either “internal” (having exactly t descendants), or are leaves (having 0 descendants). The vertices are unlabelled, so a forest may be specified by the string (d_1, \dots, d_N) , of 0’s and t ’s, one character per vertex, which results when the forest is traversed in preorder (root first, then descendent-trees; components left-to-right), and the number of descendants of each visited vertex is appended to the string. For example, ’03000’ has two components: an isolated vertex followed by a single root with its three children. The string ’303000000’ has three: a tree, the second child of whose root is internal, followed by two isolated vertices. Because a rooted tree has one more vertex than children (edges), it’s clear that $N = tn + k$, where n is the number of internal vertices, and k is

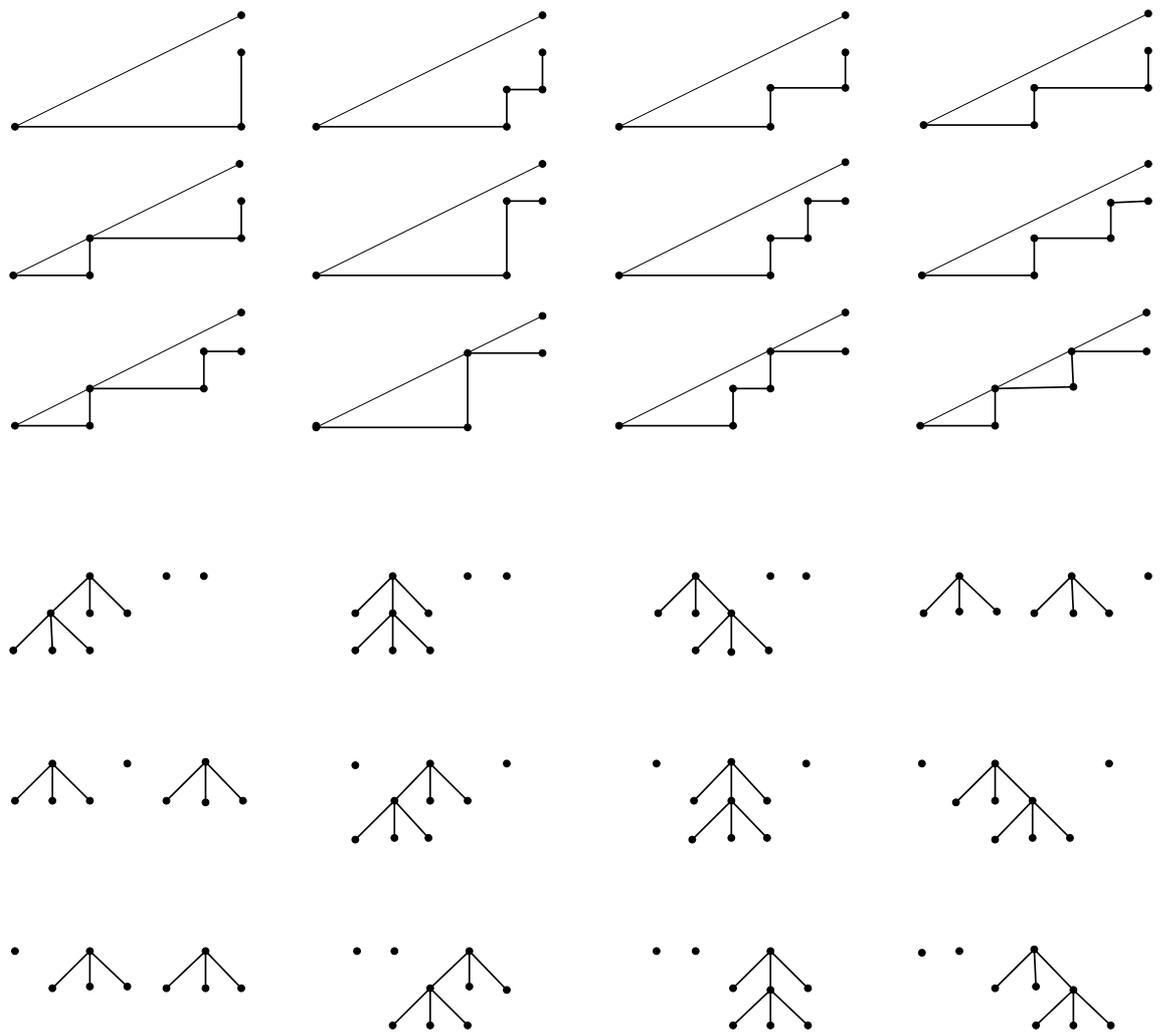


Figure 1: $\mathcal{P}(n, 3, 3)$ and $\mathcal{A}(n, 3, 3)$

the number of components, so that the leaf-count $(t-1)n+k$. Not every string comes from a forest, however.

There is a simple necessary and sufficient condition:

Proposition 1 *An N -string of nonnegative integers comes from an ordered rooted forest with N vertices by the above procedure if and only if the sum of the last m integers in the string is less than m , for $1 \leq m \leq N$.*

Proof. The proof is algorithmic and inductive. A forest is reconstructed beginning with the N th vertex, which must be a leaf, so $d_N = 0$. As we step backward in the string, we maintain an "orphan count". If $d_i = 0$, we place an isolated vertex to the left of the current forest and declare it an orphan, increasing the orphan count by 1. If $d_i > 0$ we place a new vertex as the parent of the d_i most recently declared orphans, and declare this parent an orphan. This lowers the orphan count by $d_i - 1$. In either case, the orphan count is incremented by $1 - d_i$. Clearly the only error this algorithm can encounter is "not enough orphans": the placement of a vertex resulting in a nonpositive orphan count at the m th placement, or, equivalently, $\sum_{j=1}^m (1 - d_{N-j+1}) \leq 0$. QED

For our specialized forests, we know that $d_1 + \dots + d_N = tn$ and so the forest criterion is equivalent to $d_1 + \dots + d_\ell \geq \ell - k + 1$ for $1 \leq \ell \leq N - 1$ and $d_N = 0$. A string $d_1 d_2 \dots d_N$ is determined by the positions of the n t 's among the $(t-1)n+k$ 0 's. Following Zaks [9], we denote by z_i the position of the i th t , but immediately derive from it the reduced, still nonnegative, code $b_i = z_i - i$. The key step in relating forests to lattice paths is the interpretation of the forest criterion on d 's in terms of the b 's.

Proposition 2 *A string d_1, \dots, d_{nt+k} of n t 's and $(t-1)n+k$ 0 's represents a t -ary forest with n internal vertices and k components if and only if, for the sequence (b_1, \dots, b_n) such that the i th t of the string is in position $i + b_i$, we have $0 \leq b_i \leq (i-1)t - i + k$.*

Proof. Assume that ℓ is the smallest index for which the forest condition is violated, that is, $d_1 + \dots + d_\ell < \ell - k + 1$. Then d_ℓ must be a 0 (because ℓ is smallest), and there must be at least one t to the right of d_ℓ in the string (since otherwise $d_1 + \dots + d_\ell = nt$). If the first t to the right of d_ℓ is the j th t in the string, then the forest condition must also be violated at the position immediately before this t . This is position $b_j + (j-1)$. The violation reads

$$(j-1)t = d_1 + \dots + d_{b_j+j-1} < (b_j + j - 1) - k + 1 = b_j - k + j.$$

So $b_j > (j-1)t - j + k$ and the stated condition is equivalent to the impossibility of a violation. QED

3 b -sequences and Paths

Using the b -sequence encodings, we quickly obtain a bijection between $\mathcal{A}(n, t, k)$ and $\mathcal{P}(n, t, k)$. This bijection is reflected geographically in Figure 1. If, for any

path in $\mathcal{P}(n, t, k)$, we view its vertical steps from the highest to the lowest, and record their distances to the line $x = (t - 1)n + k - 1$, we obtain a sequence of n nonnegative integers satisfying exactly the conditions on the sequence of b 's derived above. Also, all the b -sequences encode a path not crossing the slanted barrier line. We see that the parameter k is equal to $(1 + \text{the horizontal distance from the target point of the path to the barrier } (t - 1)y = x)$.

The classical counts of rooted plane forests [8, Th. 5.3.10] and/or lattice paths specialize to give $F(n, t, k) = \frac{k}{nt+k} \binom{nt+k}{n}$. For completeness, we give a quick verification by an ‘‘old school’’ induction.

Fix $t \geq 2$ and denote $|P(n, t, k)|$ by $g(n, k)$. Then

$$g(n, k) = \sum_{i=t}^{i=t+k-1} g(n-1, i)$$

because every path in $P(n, t, k)$ has its last vertical step from $((t-1)(n-1) - 1 + i, n-1) \rightarrow ((t-1)(n-1) - 1 + i, n)$ for some $i = t, \dots, t+k-1$, and the summands count the legal paths from $(0, 0)$ to $((t-1)(n-1) - 1 + i, n-1)$. It is easy to check that $g(1, k) = k$, so by induction we only need to calculate $\sum_{i=t}^{i=t+k-1} g(n-1, i)$:

$$\sum_{i=t}^{i=t+k-1} \frac{i}{(n-1)t+i} \binom{(n-1)t+i}{n-1} = \sum_{j=0}^{j=k-1} \frac{j+t}{nt+j} \binom{nt+j}{n-1} \quad (1)$$

$$= \sum_{j=0}^{j=k-1} \frac{(nt+j) - t(n-1)}{nt+j} \binom{nt+j}{n-1} \quad (2)$$

$$= \sum_{j=0}^{j=k-1} \binom{nt+j}{n-1} - t \binom{nt+j-1}{n-2} \quad (3)$$

$$= \binom{nt+k}{n} - \binom{nt}{n} - t \binom{nt+k-1}{n-1} + t \binom{nt-1}{n-1} \quad (4)$$

$$= \frac{k}{nt+k} \binom{nt+k}{n} \quad (5)$$

The last equality is verified by algebraic manipulation. The summation step comes from the polynomial identity [4, (1.48)]

$$\sum_{p=0}^m \binom{p+x}{r} = \binom{m+x+1}{r+1} - \binom{x}{r+1}$$

which is usually established by counting binary bitstrings of length $m+x+1$ with exactly $r+1$ 1's, not all occurring in the last x positions (the index p represents $(m+1) - \{\text{the position of the first 1 in the bitstring}\}$). We emphasize that our simple proof did not *calculate* $F(n, t, k)$, but only verified the formula for paths, hence for forests. Lagrange inversion and/or the Cycle lemmas have proven effective for deriving formulas from models such as ours.

4 Remarks

I hope this economy tour of Catalonia will entice the newcomer to further exploration. There could be no better guide than the online brochure of Stanley [7], introducing $\approx 2^7$ points of interest, of which we have visited but three [namely his d)(trees), h)(paths), and s)(b -sequences)].

It is easy to check that all three of our models extend to the case $t = 1$, so there is a balanced pyramid of Catalan Forest numbers $F(n, t, k)$, $1 \leq n, t, k$. Two projects are immediately available: a) extend other models in Stanley's list to objects counted by $F(n, 2, k)$, $F(n, t, 1)$, or, optimally, $F(n, t, k)$; b) extend the pyramid to more than three indicial dimensions. The second proposal is trivial for each model separately. The interest is in new bijections between models.

It is amusing to note that an ancient identity [4, (3.142)], attributed to Rothe and Hagen [6, 3]

$$\sum_{m=0}^n \frac{k}{mt+k} \binom{mt+k}{m} \frac{\ell}{(n-m)t+\ell} \binom{(n-m)t+\ell}{n-m} = \frac{k+\ell}{nt+k+\ell} \binom{nt+k+\ell}{n}$$

becomes almost obvious when interpreted using Catalanian forests. For other proofs, see [5, (1.2.6) Example 4],[1],[10].

Acknowledgement. I owe my introduction to this topic to my editing of MONTHLY Problem 11071 [2] - essentially the case $F(n, 3, 2)$ - and thank all the contributors to the solution file.

References

- [1] Erik S. Andersen, Mogens E. Larsen, *Rothe-Abel-Jensen-identiter*, Normat, vol. 42, no. 3, 1994, pp. 116-128.
- [2] Emeric Deutsch, *Problem 11071*, American Mathematical Monthly, March, 2004.
- [3] John G. Hagen, *Synopsis der Hoheren Mathematik, Band I*, Berlin, 1891.
- [4] Henry W. Gould, *Combinatorial Identities*, Morgantown, 1972.
- [5] Donald E. Knuth, *The Art of Computer Programming, vol. 1, 3rd Ed.*, Addison-Wesley, 1997, pp.126-7.
- [6] Heinrich A. Rothe, *Formulae de serierum ...*, Leipzig, 1793.
- [7] Richard P. Stanley, <http://www-math.mit.edu/~rstan/ec/catalan.pdf>, <http://www-math.mit.edu/~rstan/ec/catadd.pdf>
- [8] Richard P. Stanley, *Enumerative Combinatorics, vol. 2*, Cambridge University Press, 1999.

- [9] Shmuel Zaks, *Lexicographic generation of ordered trees*, vol, 10, no. 1, 1980, pp. 63-82.
- [10] Jiang Zeng, *Multinomial Convolution Polynomials*, Discrete Mathematics, 160, no. 1-3, 1996, pp. 219-228.