Note

A New Proof of Cayley’s Formula for Counting Labeled Trees

PETER W. SHOR

AT & T Bell Laboratories, Room 2D-149, 600 Mountain Avenue,
Murray Hill, New Jersey 07974

Communicated by the Managing Editors

Received July 29, 1993

We give a new proof of Cayley’s formula, which states that the number of labeled trees on \( n \) nodes is \( n^{n-2} \). This proof uses a difficult combinatorial identity, and it could equally well be regarded as a proof of this identity that uses Cayley’s formula. The proof proceeds by counting labeled rooted trees with \( n \) vertices and \( j \) improper edges, where an improper edge is one whose endpoint closer to the root has a larger label than some vertex in the subtree rooted on the edge. © 1995 Academic Press, Inc.

A well-known theorem of Cayley gives the number of labeled trees on \( n \) nodes. A number of proofs of Cayley’s formula are known [5], and we add a new one to the collection. This proof uses a combinatorial identity which is not easy to derive; the proof could equally well be regarded as a combinatorial proof of this identity. Although the proof as given below does not prove the identity in its full generality, it can be altered to do so; we discuss this in the second section of this paper.

1. The Proof

Cayley’s formula states that the number of labeled trees on \( n \) nodes is \( n^{n-2} \). Since we can root a tree at any of its nodes, this is trivially equivalent to the fact that the number of labeled rooted trees on \( n \) nodes is \( n^{n-1} \). Our starting point is a combinatorial identity. Define a function \( Q(i, j) = Q(i, j, 0) \) as

\[
Q(1, 0) = 1 \\
Q(i, -1) = 0, \quad i \geq 1, \\
Q(1, j) = 0, \quad j \geq 1,
\]

154
and
\[ Q(i, j) = (i - 1) Q(i - 1, j) + (i + j - 2) Q(i - 1, j - 1), \]
otherwise.

Then,
\[ \sum_{j=0}^{i-1} Q(i, j) = i^{i-1}. \] (1)

This is a specialization of an identity due to Meir [3, p. 259], with the notational change \( S_z(m, n) = Q(m + 1, n, z - 1) \); the full identity is given in the next section. Two proofs of this identity can be found in [3]. Given this identity, all we need for a proof of Cayley's formula is a classification of labeled rooted trees that partitions the trees on \( i \) nodes into classes, where the \( j \)th class contains \( Q(i, j) \) elements. Such a classification is given in Theorem 1.

We assume that the nodes of a tree are labeled 1, 2, ..., \( n \). We call the node labeled \( i \) the node \( i \). For any edge of the tree, we call the node closer to the root the parent node of the edge, and the node farther from the root the child node of the edge. For a node \( x \) we let \( \beta(x) \) be the smallest label on any node in the subtree rooted at \( x \) (it is possible that \( \beta(x) = x \)). We call a tree edge \( e \) a proper edge if the label on the parent node of \( e \) is smaller than the label on all descendant nodes of \( e \), i.e., if \( \beta(\text{child}(e)) > \text{parent}(e) \). Otherwise, we will call the edge improper.

We can now state the main theorem of this section.

**Theorem 1.** The number of labeled rooted trees on \( n \) nodes with \( j \) improper edges is \( Q(n, j) \).

Mallows and Riordan have defined a similar quantity [4, 6]: they consider the number of inversions in labeled rooted trees, where an inversion is a pair of vertices one of which is an ancestor of the other and in which the ancestor has the larger label. By weighting each improper edge by the number of descendant vertices which "make" it improper, one obtains the number of inversions.

**Proof of Theorem 1.** We give a mapping \( \phi \) from trees on \( n \) nodes to trees on \( n - 1 \) nodes which gives rise to the recurrence for \( Q(n, j) \). To find \( \phi(T) \), where \( T \) is a labeled tree, find the node \( n \) in \( T \) which has the largest label. If this node is a leaf, remove it. If it is not a leaf, then from the children of \( n \), say \( x_1, x_2, ..., x_d \), choose the child \( x_a \) with the largest \( \beta(x_a) \). Contract the edge \( nx_a \) (so that the former children of both \( n \) and \( x_a \) are now siblings) and let the resulting node have label \( x_a \). This is now a tree on \( n - 1 \) labeled nodes. Note that in the first case (when \( n \) is a leaf) the number of improper edges is conserved and in the second case this number is reduced by one.
We now consider the inverse of the map $\phi$. First, if we have a tree $R$ on $n-1$ nodes, we can add node $n$ as the child of any of the nodes of $R$ and obtain a tree $T$ in which $n$ is a leaf and $\phi(T) = R$. We obtain $n-1$ preimages $T = \phi^{-1}(R)$ of every tree $R$ on $n-1$ nodes in this way. For these preimages, $T$ and $R$ have the same number of improper edges.

Next, suppose we have a tree $R$ on $n-1$ nodes with $j$ improper edges. To find a $T = \phi^{-1}(R)$ with $j+1$ improper edges, we can take any node $x$, put node $n$ in place of node $x$, and make node $x$ a child of node $n$. We now have to decide what to do with the children of $x$ in $R$. In $T$, they must become children of either $n$ or $x$. Let $b$ be the number of improper edges out of $x$ in $R$, and let these edges have child nodes $x_1, x_2, \ldots, x_b$, with $\beta(x_1) < \beta(x_2) < \cdots < \beta(x_b) < x$. In order to construct a $T$ with $\phi(T) = R$ we must make all the proper children of $x$ in $R$ be children of $x$ in $T$, and we can choose an $a$, $0 \leq a \leq b$, and make the nodes $x_1, x_2, \ldots, x_a$ children of $n$ in $T$ and the nodes $x_{a+1}, \ldots, x_b$ children of $x$ in $T$. It is easy to verify these are the only ways of partitioning the children of $x$ in $R$ between $x$ and $n$ so as to make $\beta(x)$ the largest $\beta$ value among the children of $n$ in $T$. There are $b+1$ such ways; thus we have one preimage $T$ for each improper edge out of $x$ and one additional preimage. Summing over all nodes $x$ gives $j + n - 1$ preimages $T = \phi^{-1}(R)$, where $T$ has one more improper edge than $R$.

We have shown that any tree $R$ with $n-1$ nodes and $j$ improper edges has $n-1$ preimages $\phi^{-1}(R)$ with the same number, $j$, of improper edges, and $n + j - 1$ preimages $\phi^{-1}(R)$ with $j+1$ improper edges. Thus if $Q'(n, j)$ is the number of trees on $n$ nodes with $j$ improper edges, $Q'(n, j)$ satisfies

$$Q'(n, j) = (n-1) Q'(n-1, j) + (n+j-2) Q'(n-1, j-1),$$

which is the recurrence for $Q$. To show that $Q = Q'$ and thus complete the proof, we must show that the boundary conditions of $Q$ are satisfied for trees with $n$ nodes and $j$ improper edges; i.e., we verify that there is exactly one labeled rooted tree on one node, which has no improper (or proper) edges.

The number of improper edges in a random tree on $n$ vertices can be calculated using the generating function given in Ruehr's proof [3] of the identity (1). The expected number of improper edges in a random tree is

$$n \left[ \left(1 + \frac{1}{n}\right)^n - 2 \right] \approx (e-2)n$$

and the variance of this quantity is

$$n \left[ n \left(1 + \frac{2}{n}\right)^{n+1} - n \left(1 + \frac{1}{n}\right)^{2n} - 3 \left(1 + \frac{1}{n}\right)^n + 1 \right] \approx (e^2 - 3e + 1)n.$$
2. THE GENERAL FORM OF THE IDENTITY

The identity (1) given above is actually a special case \((k=0)\) of the following identity due to Meir \([3, p. 259]\). Define a function \(Q(i, j, k)\) as

\[
Q(i, 0, k) = 1 \\
Q(i, -1, k) = 0, \quad i \geq 1, \\
Q(1, j, k) = 0, \quad j \geq 1,
\]

and

\[
Q(i, j, k) = (i - k - 1) Q(i - 1, j, k) \\
+ (i + j - 2) Q(i - 1, j - 1, k), \quad \text{otherwise.}
\]

Then,

\[
\sum_{j=0}^{i-1} Q(i, j, k) = (i + k)^{i-1}. \tag{2}
\]

The previous section can be viewed as a proof of this identity for the case \(k = 0\). We now give a combinatorial proof of this identity for integers \(k \geq 1\). Since \(Q(i, j, k)\) is a polynomial in \(k\), by interpolation this proves (2) for all \(k\).

We need the following generalization of Cayley's formula \([1]\).

**Theorem 2 (Cayley).** Given \(n\) labeled nodes of which \(k\) are designated as roots, the number of forests of \(k\) rooted trees that can be formed on these nodes is \(k^n n^{-k-1}\).

To be more specific, we have a set of \(n\) labeled nodes, of which \(k\) have been designated as "roots." We are counting acyclic graphs with \(k\) components, where each component contains exactly one root. We will assume that the nodes 1, ..., \(k\) are those designated as roots.

Theorem 2 is easily derived from Clarke's proof of Cayley's formula \([2, 5]\), which proves an equivalent statement by reverse induction on \(k\) (keeping \(n\) fixed). Cayley's formula follows from the case \(k = 1\).

We can label the edges in a rooted forest proper and improper, using the same definition as in the proof of Theorem 1; note that all edges out of a root are proper. By essentially the same proof as for Theorem 1, we get the following.

**Theorem 3.** The number of forests of \(k\) rooted trees on \(n\) labeled nodes with \(j\) improper edges and with roots 1, ..., \(k\) is \(kQ(n-k, j, k)\).
The proof is by induction on $n$, leaving $k$ fixed. The factor of $k$ arises in the base case of the induction, when there are $k$ such forests on $k+1$ nodes. Together, Theorem 2 and Theorem 3 give a proof of the identity (2).

A remarkable fact about the function $Q(i, j, k)$ is that in addition to the recurrence given above, it also satisfies the recurrence

$$Q(i, j, k) = (k-j+1) Q(i-1, j, k+1) + (i+j-2) Q(i-1, j-1, k+1).$$  \hspace{1cm} (3)

Identity (2) follows directly from this recurrence, which can be proved by induction [3]. Although Clarke's proof of Cayley's formula provides a combinatorial proof that

$$i-1 \sum_{j=0}^{i-1} Q(i, j, k) = (i+k) \sum_{j=0}^{i-2} Q(i-1, j, k+1),$$

we do not know any combinatorial interpretation of the recurrence (3); it would be interesting to find one.

\textbf{ACKNOWLEDGMENT}

I thank the referee for the suggestion that Ruehr's generating function could be used to compute the expectation and variance of the number of improper edges and for the Refs. [4, 6] on inversions in labeled trees.

\textbf{REFERENCES}