Journal of Integer Sequences, Vol. 3 (2000),
Article 00.1.4

On the kernel of tree incidence matrices

M. Bauer and O. Golinelli

Service de Physique Théorique, CEA Saclay,
F-91191, Gif-sur-Yvette, France.

Email addresses: bauer@spht.saclay.cea.fr and golinelli@spht.saclay.cea.fr

Abstract

We give a closed form, a generating function, and an asymptotic estimate for the sequence $(z_n)_{n\geq 1} = 1, 0, 3, 8, 135, 1164, 21035, \ldots$ *that gives the total multiplicity of the eigenvalue* 0 *in the set of* n^{n-2} *tree incidence matrices of size* n.

1. Introduction.

By a classical result in graph theory, the number of labeled trees[1] on $n \geq 1$ vertices is n^{n-2}. We endow the set $\mathcal{T}_n$ of labeled trees on $n \geq 1$ vertices with uniform probability, giving weight n^{2-n} to each tree.

Each tree in $\mathcal{T}_n$ comes with its incidence matrix, the $n \times n$ symmetric matrix with entry ij equal to 1 if there is an edge between vertices i and j and to 0 otherwise. Each such matrix has n (real) eigenvalues, which by definition form the spectrum of the corresponding tree. This leads in turn to $n\, n^{n-2} = n^{n-1}$ eigenvalues counted with multiplicity for $\mathcal{T}_n$ as a whole. In the sequel, we will concentrate on the multiplicity of the eigenvalue 0. Let $Z(T)$ be the multiplicity of the eigenvalue 0 in the spectrum of the incidence matrix of the tree T, i.e. the dimension of the kernel. For each $n \geq 1$, the restriction Z_n of Z to $\mathcal{T}_n$ is a random

[1]Precise definitions for this and the following terms can be found in Section 2.

variable. We set $z_n = \sum_{T\in\mathcal{T}_n} Z_n(T)$. The expectation of $Z_n(T)$ is $\mathbb{E}(Z_n) = z_n/n^{n-2}$.

To illustrate these definitions, we give an explicit enumeration of $z_1, \cdots, z_4$ in Appendix A.

Our aim is to prove :

Theorem 1. *Let z_n be the total multiplicity of the eigenvalue 0 in the spectra of the n^{n-2} labeled trees on n vertices. Then :*

i) Closed form :

$$\begin{aligned} z_n &= n^{n-1} - 2\sum_{2\le m\le n} (-1)^m n^{n-m} m^{m-2} \binom{n-1}{m-1} \\ \frac{z_n}{n^{n-2}} &\equiv \mathbb{E}(Z_n) = n\left(1 - 2\sum_{2\le m\le n} \frac{(-1)^m}{m}\left(\frac{m}{n}\right)^m \binom{n}{m}\right). \end{aligned}$$

ii) Formal power series identity :

$$x^2 + 2x - xe^x = \sum_{n\ge 1} \frac{z_n}{n!}\left(xe^x e^{-xe^x}\right)^n.$$

Corollary 2. *For large n, $\mathbb{E}(Z_n)$ has an asymptotic expansion in powers of $1/n$ whose first two terms are*

$$\mathbb{E}(Z_n) = (2x_* - 1)n + \frac{x_*^2(x_*+2)}{(x_*+1)^3} + O(1/n)\,,$$

where $x_* = 0.567143290409783872999\cdots$ is the unique real root of $x = e^{-x}$. In particular, the average fraction of the spectrum occupied by the eigenvalue 0 in a large random tree is asymptotic to $2x_* - 1 = 0.134286508081956774599\cdots$.

Remark 3. We do not try to show here that the fluctuations in random trees become small when the number of vertices is large. However, it is expected that $\mathbb{E}(Z_n^2) - \mathbb{E}(Z_n)^2$ grows only linearly with the number of vertices, so that in an appropriate sense the fraction of the spectrum occupied by the eigenvalue 0 in an infinite random tree is $2x_* - 1$ with probability 1.

Remark 4. With the explicit formula above, it is easy to list the first terms in the sequence $(z_n)_{n\ge1}$, which are

$$1, 0, 3, 8, 135, 1164, 21035, 322832, 7040943, 153153620, 4048737099, \cdots$$

To prove part i) of Theorem 1 we establish a few preparatory lemmas of independent interest. Then we prove ii) using Lagrange inversion and obtain Corollary 2 by the steepest descent method.

There is an application of Z_n to random graph theory — see Remark 22.

2. Definitions.

Even if ultimately we are interested only in trees, we shall need more general graphs (for instance, forests) in the proofs.

Definition 5. A *simple graph* G is a pair (V, E) where V is a finite set called the set of *vertices* and E is a subset of $V^{(2)} \equiv \{\, \{x, y\},\ x \in V,\ y \in V,\ x \neq y\}$ called the set of *edges*.

Remark 6. The adjective *simple* refers to the fact that there is at most *one* edge between two vertices and that edges are pairs of *distinct* vertices. From now on we use *graph* for *simple graph*.

Definition 7. If V is empty, then we say that the graph G is *empty*. The vertices adjacent to a given vertex x are called the *neighbors* of x. The number of neighbors of a vertex x is called the *degree* of x. A *leaf* of G is a vertex of degree 1. Two edges of G with a common vertex are called *adjacent edges*

Definition 8. A *labeled* graph on $n \geq 1$ vertices is a graph with vertex set $[n] = \{1, \cdots, n\}$.

Remark 9. If the graph G has $|V| = n \geq 1$ vertices[2], any bijection between V and $[n]$ defines a labeled graph. The incidence matrices for different bijections differ only by a permutation of the rows and columns. In particular the eigenvalues are independent of the bijection.

Definition 10. The *spectrum* of a graph is the set of eigenvalues (counted with multiplicity) of any of the associated incidence matrices. By convention, the spectrum of the empty graph is empty.

Definition 11. A *subgraph* of a graph $G = (V, E)$ is a graph (W, F) such that $W \subset V$ and $F \subset E$. An *induced subgraph* of G is a graph (W, F) such that $W \subset V$ and $F = E \cap W^{(2)}$.

[2]For any finite set S , $|S|$ is the number of elements in S.

Definition 12. We say that two vertices x and $x' \in V$ are in the *same component* of G if there is a sequence $x = x_1, \cdots, x_n = x'$ in V such that adjacent terms in the sequence are adjacent in G (taking $n = 1$ shows that luckily x and x are in the same component). This gives a partition of V. Each component defines an induced subgraph of G which is called a *connected component* of G. Then G can be thought of as the disjoint union of its connected components. We say that G is *connected* if it has only one connected component.

Definition 13. A *polygon* in a graph G is a sequence $x_0, x_1, \cdots, x_n$, $n \geq 3$ of vertices such that adjacent terms in the sequence are adjacent in G, $x_0 = x_n$ and $x_1, \cdots, x_n$ are distinct.

Definition 14. A *forest* is a graph without polygons. A *tree* is a non-empty connected forest.

Remark 15. Clearly a subgraph of a forest is a forest. The connected components of a nonempty forest are trees. One shows easily that that a tree with $n \geq 2$ vertices has at least two leaves. Then a simple induction shows that a tree is exactly a connected graph for which the number of vertices is 1 plus the number of edges. A classical theorem of Cayley states that there are n^{n-2} labeled trees on n vertices (see for instance Proposition 5.3.2 in [3]).

3. Two preparatory lemmas.

The first lemma is a characterization of the dimension of the kernel of incidence matrices viewed as a function on forests.

Lemma 16. *The function Z which associates to any forest the multiplicity of the eigenvalue 0 in its spectrum is characterized by the following properties :*

i) The function Z takes the value 0 on $\emptyset$, the empty forest.

ii) The function Z takes the value 1 on $\bullet$, the forest with one vertex.

iii) The function Z is additive on disjoint components, i.e. if the forest F is the union of two disjoint forests F_1 and F_2 then $Z(F) = Z(F_1) + Z(F_2)$

iv) The function Z is invariant under "leaf removal", i.e. if x is a leaf of F, y is its (unique) neighbor, $V' = V \backslash \{x, y\}$, and F' is the subforest of F induced by V' then $Z(F) = Z(F')$.

Remark 17. That the function Z satisfies properties i)–iv) was no doubt known decades ago (see for instance Section 8.1, Hückels theory,

in [1]). We give a proof, because in the sequel we want to emphasize and use the simple fact that these properties characterize the function Z.

Proof of Lemma 16. First, we show that the function Z has properties i)–iv). In fact, this is true for general graphs (not only forests). Properties i) and ii) follow from the definition of Z, property iii) follows from the fact that the incidence matrix can be put into block diagonal form, each block corresponding to a connected component. Property iv) is only slightly more complicated. With an appropriate labeling of the vertices, the incidence matrix $\mathbf{M}$ of F can be decomposed as

$$\mathbf{M} = \begin{pmatrix} 0 & 1 & \mathbf{0} \\ 1 & 0 & \mathbf{N} \\ \mathbf{0} & {}^t\mathbf{N} & \mathbf{M}' \end{pmatrix}$$

where the first row and column are indexed by the leaf x, the second row and column are indexed by its neighbor y, $\mathbf{N}$ describes the edges between this neighbor and V', and $\mathbf{M}'$ is the incidence matrix for V'. Then $\mathbf{v} = {}^t(v_1, v_2, \mathbf{v}')$ is in the kernel of $\mathbf{M}$ if and only if

$$\begin{aligned} v_2 &= 0 \\ v_1 &= -\mathbf{N}\mathbf{v}' \\ \mathbf{M}'\mathbf{v}' &= -{}^t\mathbf{N}v_2. \end{aligned}$$

So $v_2 = 0$ which from the third equation gives $\mathbf{M}'\mathbf{v}' = \mathbf{0}$, implying that $\mathbf{v}'$ is in the kernel of $\mathbf{M}'$, and then the second equation just gives v_1 the appropriate value. So the kernels of $\mathbf{M}$ and $\mathbf{M}'$ have the same dimension. This proves iv).

Now, any tree with more than 1 vertex has leaves, so leaf removal as defined in iv) allows one to reduce the forest F to a (possibly empty) family of isolated vertices (all connected components have only one vertex). Hence there is at most one function, namely Z, that can satisfy properties i)–iv). ∎

Remark 18. Leaf removal and additivity give an efficient algorithm for computing the multiplicity of the eigenvalue 0 for a given forest, especially when this forest is given as a drawing.

The next lemma gives an impractical but theoretically useful formula for the function Z.

Lemma 19. *Let L be the function on forests defined by:*
i') The function L takes the value 0 on $\emptyset$, the empty forest.
ii') The function L takes the value 1 on $\bullet$, the forest with one vertex.
iii') The function L takes the value 0 on disconnected forests.

iv') The function L takes the value $2(-1)^{n-1}$ on trees with $n \geq 2$ vertices.

Then, for any forest F

$$Z(F) = \sum_{F' \subset F} L(F') = \sum_{T' \subset F} L(T')$$

where the first sum is over induced subforests of F, and the second over induced subtrees of F.

Remark 20. For a given forest, there is a much nicer formula, directly connected to the geometry of the forest (again, see for instance Section 8.1, Hückels theory, in [1]). In fact, let $Q(F)$ be the maximum among the cardinalities of sets of pairwise non-adjacent edges in F, and $N(F)$ be the number of vertices in F. Then $Z(F) = N(F) - 2Q(F)$. It is easy to show that $N(F) - 2Q(F)$ satisfies properties i)–iv) of Lemma 16. In particular, a possible way to maximize the number of non-adjacent edges in F in case iv) is to do so on F' and add the edge $\{x, y\}$. This explicit formula allows us to restate our theorems in terms of the random variable Q_n, the restriction of Q to $\mathcal{T}_n$. For instance, in a large random tree on n vertices, one can find about $(1-x_*)n$ pairwise non-adjacent edges. Note that $1 - x_* = 0.4328567095902161270000\cdots$ is not much smaller than 0.5 (the upper bound for $Q(T)/N(T)$ for a given tree because $Z(T) = N(T) - 2Q(T)$ is always nonnegative).

Proof of Lemma 19. Our strategy is to use the characterization of Z in Lemma 16. First, we observe that the second equality is a trivial consequence of i') and iii'). We define a new function Z' on the set of forests by

$$Z'(F) \equiv \sum_{T' \subset F} L(T')$$

(where the sum is over induced subtrees of F) and show that Z' satisfies properties i)–iv) of Lemma 16.

As the empty forest has no non-empty induced subtree i') implies i).

In the same vein, the forest with one vertex has only one non-empty induced subtree, namely itself, so ii') implies ii).

If the forest F is the union of two disjoint forests F_1 and F_2, an induced subtree of F is either an induced subtree of F_1 or an induced subtree of F_2, and the sum defining $Z'(F)$ splits as $Z'(F_1) + Z'(F_2)$, showing that Z' satisfies property iii).

Now, if x is a leaf of F and y its neighbor, we define $V' = V \backslash \{x, y\}$, $V'' = \{x, y\}$ and consider F' and F'', the subforests of F induced by V' and V'' respectively. We split the sum defining $Z'(F)$ into three pieces. The first is over the induced subtrees of F'. This is just the

sum defining $Z'(F')$. The second is over the induced subtrees of F'', which is a tree on two vertices. Its subtrees are itself, with weight $L(F'') = 2(-1)^{2-1} = -2$, and two trees with one vertex, each with weight $L(\cdot) = 1$, so this second sum gives 0. The third sum is over induced subtrees that have vertices in both V' and V''. If this sum is not empty, every tree that appears in it has y as a vertex (by connectivity) and has at least two vertices (because the tree consisting of y alone has already been counted). Then we can group these trees in pairs, a tree containing x being paired with the same tree but with x and the edge $\{x, y\}$ deleted. The function L takes opposite values on the two members of a pair, so the third sum contributes 0. Hence Z' satisfies property iv). So $Z'(F) = Z'(F')$. ∎

Remark 21. These two lemmas have an obvious extension to bicolored forests. If we use black and white as the colors, and count the zero eigenvectors having value zero on white vertices, we only need to replace ii) in Lemma 16 by

ii) The function Z takes the value 1 *on* •*, the forest with one vertex colored in black and* 0 *on* ∘*, the forest with one vertex colored in white,*

and ii') and iii') in Lemma 19 by

ii') The function L takes the value 1 *on* •*, the forest with one vertex colored in black and* 0 *on* ∘*, the forest with one vertex colored in white,*

iii') The function L takes the value $(-1)^{n-1}$ on trees with $n \geq 2$ vertices.

The proofs remain the same.

Remark 22. The formula

$$Z(F) = \sum_{F' \subset F} L(F')$$

can be inverted using inclusion-exclusion to give

$$L(F) = \sum_{F' \subset F} (-1)^{|V(F)|-|V(F')|} Z(F').$$

This identity has an application in random graph theory [2], which led to our interest in Lemma 19.

4. Main proofs.

Proof of Theorem 1. By Lemma 16

$$z_n \equiv \sum_{T \in \mathcal{T}_n} Z_n(T) = \sum_{m=1}^{n} \sum_{T \in \mathcal{T}_n} \sum_{T' \in \mathcal{T}_m}^{T' \subset T} L(T').$$

As the function L depends only on the number of vertices, for fixed m the double sum $\sum_{T\in\mathcal{T}_n}\sum_{T'\in\mathcal{T}_m}^{T'\subset T}$ is simply a multiplicity. We count this multiplicity as follows : we remove from T the edges of T', so we are left with m trees, each with a special vertex, the one belonging to T'. This is what is called a planted forest (or rooted forest) with n vertices and m trees. The number of such objects is $m\binom{n}{m}n^{n-m-1}$ (see for instance Proposition 5.3.2 in [3]). Conversely, starting from such a planted forest with m trees (each with a special vertex) and n vertices, we can build a tree on the special vertices in m^{m-2} ways. So

$$\sum_{T\in\mathcal{T}_n}\sum_{T'\in\mathcal{T}_m}^{T'\subset T} 1 = m^{m-1}\binom{n}{m}n^{n-m-1}.$$

Hence summation over m gives

$$z_n = n^{n-1} - 2\sum_{2\le m\le n}(-1)^m n^{n-m-1}m^{m-1}\binom{n}{m}.$$

Simple rearrangements lead to the two equivalent formulæ in i), the first one making clear that z_n is an integer.

To obtain the generating function in ii), we need a mild extension of the Lagrange inversion formula (see for instance Section 5.4 in [3]), which states that if $f(x)$ is a formal power series in x beginning $f(x) = x + O(x^2)$ and $g(x)$ is an arbitrary formal power series in x, then

$$\left(g\circ f^{-1}\right)(t) = g(0) + \sum_{n\ge1}\frac{1}{n}\left[\frac{x^n g'(x)}{f(x)^n}\right]_{n-1} t^n\,,$$

where $[h(v)]_k$ is by definition the k^{th} coefficient of the formal power series $h(v)$.

As an immediate application, we see that if $t = xe^x$ then

$$x = \sum_{m\ge1}(-m)^{m-1}\frac{t^m}{m!}$$

and

$$-x - x^2/2 = \sum_{m\ge1}(-m)^{m-2}\frac{t^m}{m!}.$$

Now we introduce $y = te^{-t}$ and define a sequence $z'_n, n\ge1$, by

$$x^2 + 2x - xe^x = \sum_{n\ge1} z'_n\frac{y^n}{n!},$$

but instead of directly applying the Lagrange inversion formula to $y = xe^x e^{-xe^x}$, we first substitute the t-expansion (already obtained

by Lagrange inversion) on the left-hand side, which yields

$$-2\sum_{m\geq 1}(-m)^{m-2}\frac{t^m}{m!}-t,$$

and then apply Lagrange inversion with $y=te^{-t}$. The result is

$$\frac{z_n'}{n!}=\frac{1}{n}\left[e^{nt}\left(1-2\sum_{m\geq 2}\frac{(-m)^{m-2}}{(m-1)!}t^{m-1}\right)\right]_{n-1}.$$

Straightforward expansion of this formula shows that $z_n'=z_n$, and this establishes the generating function representation in ii). ∎

Remark 23. The derivation of ii) is quite artificial. It turns out that random graph theory gives a natural proof [2] using the formula mentioned in Remark 22.

Proof of Corollary 2. This time we use Lagrange inversion with $y=xe^xe^{-xe^x}$, in a contour integral representation[3]. So

$$\frac{z_n}{n!}=\frac{1}{n}\oint\frac{dx}{(xe^xe^{-xe^x})^n}(1+x)(2-e^x),$$

where the contour is a small anticlockwise-oriented circle around the origin. For large n we use the steepest descent method to obtain the asymptotic expansion of z_n. As $\frac{d}{dx}xe^xe^{-xe^x}=(1+x)(1-xe^x)e^xe^{-xe^x}$, the saddle points of $xe^xe^{-xe^x}$ are $x=-1$ and the solutions to $x=e^{-x}$. This equation has a unique real root, x_*, which is positive. Numerically, $x_*=0.567143290409783872999\cdots$. On the other hand, $x=e^{-x}$ has an infinite number of complex solutions, in complex conjugate pairs. Asymptotically, the imaginary parts of these zeros are evenly spaced by about 2π, while their real parts are negative and grow logarithmically in absolute value. Consideration of the landscape produced by the modulus of the function $xe^xe^{-xe^x}$ shows that the small circle around the origin can be deformed to give the union of two steepest descent curves, one passing through $x=-1$ and the other through $x=x_*$. These two curves are asymptotic to the two lines $y=\pm\pi$ at $x\to+\infty$. Hence, despite the fact that the value of $xe^xe^{-xe^x}$ is the same, namely $1/e$, at all the complex saddle points and at x_*, the complex saddle points do not contribute to the asymptotic expansion of z_n at large n. Moreover, the point $x=-1$ only gives subdominant contributions because $-e^{-1}e^{e^{-1}}$ is larger than $1/e$ in absolute value. So we concentrate on the asymptotic

[3] We include the factor $\frac{1}{2i\pi}$ in the symbol $\oint$.

expansion around x_*. As

$$\log xe^x e^{-xe^x} = -1 - \frac{(x_*+1)}{2x_*}(x-x_*)^2 + O((x-x_*)^3)$$

we infer that

$$e^{-n}\sqrt{2\pi n}\oint \frac{dx}{(xe^x e^{-xe^x})^n}(1+x)(2-e^x)$$

has an asymptotic expansion in powers of $1/n$. By use of Stirling's formula for $n!$ we conclude that $\mathbb{E}(Z_n) = z_n/n^{n-2}$ has an asymptotic expansion in powers of $1/n$. The first two terms are obtained by brute force. ∎

Appendix A. Examples of direct multiplicity counting.

This appendix enumerates the multiplicities of 0 in the spectrum of trees with $n = 1, 2, 3$ or 4 vertices.

Example 24. For $n = 1$ there is only one tree, , and one way to label it, giving $1 = 1^{1-2}$ tree on one vertex. The incidence matrix is (0), so the eigenvalue 0 occurs with multiplicity $z_1 = 1$.

Example 25. For $n = 2$ there is only one tree, , and one way to label it, again giving $1 = 2^{2-2}$ tree on two vertices. The incidence matrix is

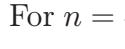
$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

so the eigenvalue 0 occurs with multiplicity $z_2 = 0$.

Example 26. For $n = 3$ there is only one tree, , and three ways to label it, giving a total of $3 = 3^{3-2}$ trees on three vertices. Up to permutation of rows and columns, the incidence matrix for each of these three labeled trees is

$$\begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix},$$

which has zero as an eigenvalue with multiplicity 1 (a corresponding eigenvector is ${}^t(1, 0, -1)$), so there is a total of 3×1 zero eigenvalues, and $z_3 = 3$

Example 27. For $n = 4$ there are two trees,  (12 ways to label it), and (4 ways to label it), giving a total of $12 + 4 = 16 = 4^{4-2}$

trees on three vertices. Up to permutation of rows and columns, the two incidence matrices are

$$\begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}.$$

The first does not have 0 as an eigenvalue, whereas the second has zero as an eigenvalue with multiplicity 2 (corresponding eigenvectors are for instance ${}^t(1,0,-1,0)$ and ${}^t(1,0,0,-1)$), so there is a total of $12 \times 0 + 4 \times 2$ zero eigenvalues, and $z_4 = 8$.

References

[1] D.-M. Cvetković, M. Doob and H. Sachs, *Spectra of Graphs*, Academic Press, New York, 1980.

[2] M. Bauer and O. Golinelli, *On the spectrum of random graphs*, in preparation.

[3] R.-P. Stanley, *Enumerative Combinatorics, Vol II*, Cambridge University Press, Cambridge, 1999.

(Concerned with sequence A053605.)

Received Nov. 10, 1999; published in Journal of Integer Sequences March 2, 2000.

Return to Journal of Integer Sequences home page.