THE INVERSE PROBLEM FOR CERTAIN TREE PARAMETERS

ÉVA CZABARKA, LÁSZLÓ SZÉKELY, AND STEPHAN WAGNER

ABSTRACT. Let p be a graph parameter that assigns a positive integer value to every graph. The inverse problem for p asks for a graph within a prescribed class (here, we will only be concerned with trees), given the value of p. In this context, it is of interest to know whether such a graph can be found for all or at least almost all integer values of p. We will provide a very general setting for this type of problem over the set of all trees, describe some simple examples and finally consider the interesting parameter "number of subtrees", where the problem can be reduced to some number-theoretic considerations. Specifically, we will prove that every positive integer, with only 34 exceptions, is the number of subtrees of some tree.

1. Introduction and a general construction principle

Numerous graph parameters are known and have been thoroughly studied. Many of them are defined as the number of vertex or edge subsets of some sort. For instance, Klazar [5] describes twelve different tree parameters and analyses them by means of a generating functions approach. Recently, the *inverse problem* has gained a certain importance: given a graph parameter p that assigns an integer value to every graph within a given family $\mathcal{G}$ of graphs and an integer n, determine some $G \in \mathcal{G}$ such that $p(G) = n$. This is of interest in combinatorial chemistry, where graphical parameters are used as molecular descriptors. For uses in the design of combinatorial libraries and molecular recognitions, we refer to [7] and the references therein. In this context, acyclic graphs play a major role, and we will also concentrate on trees in this paper.

We will generally consider rooted trees, which will turn out to be useful for the treatment of the various parameters. Some of them are even defined for rooted trees only. All rooted trees can be constructed recursively by means of the following two operations:

- Attach a new vertex to the root, which also serves as the root of the new tree (see Fig. 1)
- Merge two rooted trees by identifying their roots (see Fig. 2)

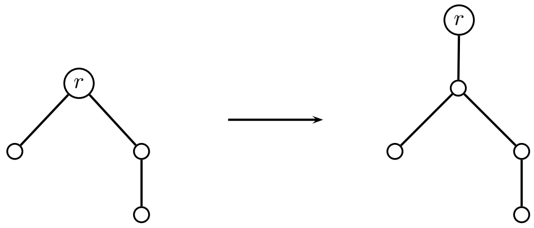

FIGURE 1. Attaching a new vertex to the root

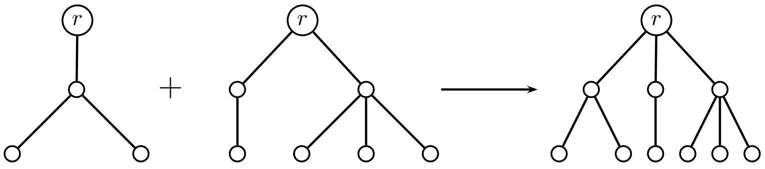

FIGURE 2. Merging two rooted trees

Many graph parameters can be described recursively in a natural way by means of these two operations. Let us present a few examples.

(1) The *number of* (nonempty) *subtrees* has been studied in various papers, see for instance [11]; in order to obtain a recursion, we distinguish between subtrees containing the root and those not containing the root. The respective numbers form a pair (a, b) associated with a tree. If we perform Operation 1 now, the number of subtrees of the former type is 1 + the number of subtrees containing the root in the old tree (1 for the subtree that consists only of the root). On the other hand, a subtree that doesn't contain the root can be any subtree of the old tree. This means that Operation 1 maps (a, b) to $(a + 1, a + b)$. Similar reasoning shows that if Operation 2 is applied to two trees with associated pairs (a_1, b_1) and (a_2, b_2), one gets the new pair $(a_1a_2, b_1 + b_2)$.

(2) The *number of matchings* (independent edge subsets, including the empty set) is a popular parameter in combinatorial chemistry (where it is known as "Hosoya index", see for instance [12]) and statistical physics (so-called dimer-monomer model [4]), and of course it also has interesting mathematical properties. Again, a recursive characterization is possible. For this purpose, we distinguish between matchings which cover the root and those which don't, and we associate a pair (a, b) to each tree again. Then, Operation 1 corresponds to the map
$$(a, b) \mapsto (b, a + b),$$
whereas Operation 2 maps the pair (a_1, b_1), (a_2, b_2) to $(a_1b_2 + b_1a_2, b_1b_2)$.

(3) The *number of independent* (vertex) *subsets* (including the empty set again) has similar properties to the number of matchings and is also a popular chemical parameter (the "Merrifield-Simmons index" [9, 12]). Yet again it is useful to distinguish two cases: we associate a pair (a, b) to each tree, where a and b are the number of independent subsets containing resp. not containing the root. Then, the effects of the two operations are
$$(a, b) \mapsto (b, a + b)$$
(note that this is the same as for the number of matchings!) and
$$((a_1, b_1), (a_2, b_2)) \mapsto (a_1a_2, b_1b_2).$$

(4) The *number of maximal independent subsets*, see for instance [16], needs three different auxiliary parameters: A triple (a, b, c) is associated to each tree, where a is the number of maximal independent sets containing the root, b is the number of maximal independent sets not containing the root, and c is the number of independent sets which are not maximal but become maximal if the root (that is necessarily not contained in the set) is removed from the tree. Similar reasoning to the previous examples yields the two maps
$$(a, b, c) \mapsto (b + c, a, b)$$
and
$$((a_1, b_1, c_1), (a_2, b_2, c_2)) \mapsto (a_1a_2, b_1b_2 + b_1c_2 + b_2c_1, c_1c_2)$$
corresponding to the two operations.

(5) The *number of dominating subsets* satisfies a very similar recursion: again, three auxiliary parameters are necessary, representing the number of dominating sets which contain the root, those which do not contain the root, and those which dominate all vertices except for the root. Then, our operations induce the two maps
$$(a, b, c) \mapsto (a + b + c, a, b)$$
and
$$((a_1, b_1, c_1), (a_2, b_2, c_2)) \mapsto (a_1a_2, b_1b_2 + b_1c_2 + b_2c_1, c_1c_2)$$
on triples associated to trees.

(6) The *Wiener index* is one of the most popular graph-theoretical parameters, see for instance the survey paper [1]. It is defined as the sum of all the distances between pairs of vertices. As shown in [2], it also satisfies certain recursive relations: this time, we associate a triple (a, b, c) consisting of the number of vertices, the sum of all distances from the root and the Wiener index itself to a tree. It is not difficult to show that this yields the two maps
$$(a, b, c) \mapsto (a + 1, a + b, a + b + c)$$

and

$$((a_1,b_1,c_1),(a_2,b_2,c_2)) \mapsto (a_1+a_2-1, b_1+b_2, c_1+c_2+b_1(a_2-1)+b_2(a_1-1))$$

corresponding to Operation 1 and Operation 2 respectively.

(7) The *number of chains* [5] is a parameter that only makes sense for rooted trees: a rooted tree can be regarded as a partially ordered set, where $u \leq v$ holds for two vertices u and v iff u lies on the unique path between v and the root. A chain is a nonempty set of pairwise comparable vertices. The number of chains in a rooted tree satisfies very simple recursions with respect to our two operations—no auxiliary parameters are necessary. Operation 1 induces the map

$$a \mapsto 2a+1$$

(the new root can be added to any chain and also to the empty set), Operation 2 yields the map

$$(a_1,a_2) \mapsto a_1+a_2-1$$

(a chain in the new tree is a chain in one of the old trees, and the chain that only includes the root is counted twice).

(8) Finally, we consider the analogous parameter "*number of antichains*" [5]: an antichain is a nonempty set of mutually incomparable vertices. Again, no auxiliary parameters are necessary, and we obtain the two maps

$$a \mapsto a+1$$

(under Operation 1) and

$$(a_1,a_2) \mapsto a_1a_2$$

(under Operation 2).

Note that all these recursive characterizations share a common scheme: in each case, a k-tuple $(c_1,c_2,\dots,c_k)$ is associated with each tree, the parameter we are actually interested in is some polynomial of the c_i (typically linear), and our two operations induce polynomial maps $f:\mathbb{R}^k \to R^k$ and $g:\mathbb{R}^k \times \mathbb{R}^k \to \mathbb{R}^k$.

2. Examples for the inverse problem over the set of trees

2.1. **Number of matchings.** For this parameter, the problem is essentially trivial, as has also been noticed in [7]: every integer is the number of matchings of some star, since a star with n vertices has exactly n matchings (the empty set and $n-1$ single edges). Nevertheless, the problem can become much harder if one imposes further restrictions. For instance, it is not quite clear whether almost every positive integer is the number of matchings of some binary tree.

2.2. **Number of antichains.** This example is also quite trivial, since it is easily observed that the number of antichains of a path, rooted at one of its ends, it just the number of its vertices (the only antichains are single vertices in this case).

2.3. **Number of chains.** This example is also very simple, but it shows a new phenomenon: it is easy to check from the recursive definition that the number of chains in any rooted tree is always odd. On the other hand, the inverse problem can be solved for all odd integers: indeed, one finds that a star with n vertices, rooted at its center, has exactly $2n-1$ chains (note that all chains are either single vertices or pairs consisting of the root and any leaf), proving the claim.

2.4. **Wiener index.** This problem is somewhat harder than the previous ones. In their paper [6], Lepović and Gutman conjecture that every positive integer, with only 49 explicit exceptions, is the Wiener index of some tree. This conjecture was verified independently in [13] and [15] and further extended to the class of trees with maximum degree 3 in [14]. It should be noted that the Wiener index $W(T)$ of a tree T with n vertices can be estimated above and below as follows:

$$(n-1)^2 \leq W(T) \leq \binom{n+1}{3}.$$

This implies that the minimum does not increase linearly any more, which explains why no simple construction as in the previous examples can be given. However, since the growth is only polynomial, the problem can be reduced to number-theoretic considerations similar to those presented in the following section.

2.5. **Number of independent subsets.** Here, the situation is much harder. It is conjectured, yet unproven that almost every positive integer is the number of independent subsets of some tree. This is a question that is also of relevance in combinatorial chemistry: [7] provides an algorithm for constructing such a tree provided that one exists. The biggest difference to the previously considered parameters lies in the fact that the minimum over all trees of size n grows exponentially with n, making it less amenable to the Diophantine methods applied in [14] and also in this paper. However, Linek [8] gives a partial result in this direction: he shows that every positive integer is the number of independent sets of some bipartite graph.

2.6. **Number of subtrees.** This problem will form the main part of the paper. "On average", the behavior of this parameter is similar to the number of independent subtrees, but the minimum only grows polynomially, which allows us the construction of a reasonably rich set of trees for which the inverse problem can be reduced to a question of elementary number theory. Specifically, we will prove the following theorem in the next section:

Theorem 1. *All positive integers, except for the* 34 *numbers* 2*,* 4*,* 5*,* 7*,* 8*,* 9*,* 12*,* 13*,* 14*,* 16*,* 18*,* 19*,* 22*,* 23*,* 26*,* 27*,* 29*,* 31*,* 33*,* 35*,* 38*,* 39*,* 42*,* 43*,* 46*,* 50*,* 52*,* 54*,* 60*,* 65*,* 68*,* 72*,* 77 *and* 89*, can be expressed as the number of subtrees of some tree.*

3. Proof of Theorem 1

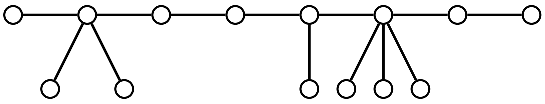

Figure 3. The caterpillar tree $C(1,0,3,1,0,0,2)$

We will show that it is sufficient to consider *caterpillar trees*, see Fig. 3. These can be constructed recursively by means of our two operations, where we only use a particular special case of Operation 2. In each step, we attach exactly one new vertex to the root; there are two possibilities for doing so, namely

- The attached vertex becomes the new root—this is exactly our Operation 1.
- The old root is also the root of the new tree—this is a special case of Operation 2 (merge with a two-vertex tree).

These operations have the following effects on pairs (a, b) as described in the previous chapter:

$$(a, b) \mapsto (a + 1, a + b) \text{ or } (a, b) \mapsto (2a, b + 1).$$

To construct caterpillars, we proceed as follows: start with a single vertex. Now apply Operation 1 a_0 times, then our special case of Operation 2, then a_1 times Operation 1 again, etc. The final step involves a_k applications of Operation 1. The variables a_i can be interpreted as the number of edges between subsequent "legs" of the caterpillar. Let the resulting tree be denoted by $C(a_0, \ldots, a_k)$.

Straightforward induction yields the following lemma:

Lemma 2. *The number of subtrees of $C(a_0, \ldots, a_k)$ containing resp. not containing the root is*

$$2^k \left(1 + \sum_{j=0}^{k} 2^{-j} a_j\right)$$

resp.

$$k + \sum_{j=0}^{k} \left(2^j a_j + \frac{a_j(a_j-1)}{2} \right) + \sum_{j=0}^{k} \sum_{i=0}^{j-1} 2^{j-i} a_i a_j .$$

Thus the total number of subtrees of $C(a_0, \dots, a_k)$ *is given by*

$$2^k + k + \sum_{j=0}^{k} \left(2^{k-j} a_j + 2^j a_j + \frac{a_j(a_j-1)}{2} \right) + \sum_{j=0}^{k} \sum_{i=0}^{j-1} 2^{j-i} a_i a_j . \tag{1}$$

Proof. This is obviously true for a single vertex ($k = 0$, $a_k = 0$), and it is easily seen that it also holds for the path (a_0 applications of Operation 1). Now the number of subtrees of $C(a_0, \dots, a_k)$ can be calculated from $C(a_0, \dots, a_{k-1})$ inductively:

After the application of our second operation, the pair becomes

$$2^k \left(1 + \sum_{j=0}^{k-1} 2^{-j} a_j \right) \text{ and } k + \sum_{j=0}^{k-1} \left(2^j a_j + \frac{a_j(a_j-1)}{2} \right) + \sum_{j=0}^{k-1} \sum_{i=0}^{j-1} 2^{j-i} a_i a_j ,$$

and a_k-fold application of Operation 1 yields

$$a_k + 2^k \left(1 + \sum_{j=0}^{k-1} 2^{-i} a_i \right)$$

and

$$k + \sum_{j=0}^{k-1} \left(2^j a_j + \frac{a_j(a_j-1)}{2} \right) + \sum_{j=0}^{k-1} \sum_{i=0}^{j-1} 2^{j-i} a_i a_j + a_k \cdot 2^k \left(1 + \sum_{j=0}^{k-1} 2^{-i} a_i \right) + \sum_{\ell=0}^{a_k-1} \ell ,$$

which completes the proof after some simplifications. ■

The formula for the number of subtrees $C(a_0, \dots, a_k)$ represents a quadratic form, but it is not very practical as it stands. However, a carefully chosen substitution simplifies matters decisively: we set $a_j = x_j - 2x_{j-1}$, where $x_{-1} = \frac{1}{2}$. Then, (1) reduces to

$$\begin{aligned}
&2^k + k + \sum_{j=0}^{k} \left(2^{k-j}(x_j - 2x_{j-1}) + 2^j (x_j - 2x_{j-1}) + \frac{(x_j - 2x_{j-1})(x_j - 2x_{j-1} - 1)}{2} \right) \\
&\quad + \sum_{j=0}^{k} \sum_{i=0}^{j-1} 2^{j-i} (x_i - 2x_{i-1})(x_j - 2x_{j-1}) \\
&= 2^k + k + \sum_{j=0}^{k} \left(2^{k-j} + 2^j \right) x_j - \sum_{j=-1}^{k-1} \left(2^{k-j} + 2^{j+2} \right) x_j + \sum_{j=0}^{k} \frac{x_j^2}{2} - \sum_{j=0}^{k} 2x_j x_{j-1} + \sum_{j=-1}^{k-1} 2x_j^2 - \sum_{j=0}^{k} \frac{x_j}{2} \\
&\quad + \sum_{j=-1}^{k-1} x_j + \sum_{j=0}^{k} \sum_{i=0}^{j-1} 2^{j-i} x_i x_j - \sum_{j=-1}^{k-1} \sum_{i=0}^{j} 2^{j-i+2} x_i x_j - \sum_{j=0}^{k} \sum_{i=-1}^{j-2} 2^{j-i} x_i x_j + \sum_{j=-1}^{k-1} \sum_{i=-1}^{j-1} 2^{j-i+2} x_i x_j \\
&= 2^k + k - 3 \sum_{j=0}^{k-1} 2^j x_j + (2^k + 1)(x_k - 1) + \sum_{j=0}^{k-1} \frac{5x_j^2 + x_j}{2} + \frac{x_k^2 - x_k}{2} + 1 - \sum_{j=0}^{k} 2^j x_j + \sum_{j=-1}^{k-1} 2^{j+2} x_j - 4 \sum_{j=-1}^{k-1} x_j^2 \\
&= k + \frac{x_k(x_k+1)}{2} - \sum_{j=0}^{k-1} \frac{3x_j^2 - x_j}{2} = \frac{x_k(x_k+1)}{2} - \sum_{j=0}^{k-1} \frac{3x_j^2 - x_j - 2}{2} .
\end{aligned}$$

Note that $\frac{3x^2-x}{2}$ is a so-called *pentagonal number*. It is part of a more general theorem (see [10]) that every positive integer is a sum of five pentagonal numbers. However, it is not possible to apply this theorem to our problem, since we have $a_j \geq 0$ for all j, which yields the restrictions $x_0 \geq 1$ and $x_j \geq 2x_{j-1}$ for $1 \leq j \leq k$. However, we can show the following, which will be sufficient for our purposes:

Proposition 3. *Every positive integer $n \geq 11034$ can be written as*

$$n = \sum_{j=0}^{k-1} \frac{3x_j^2 - x_j - 2}{2},$$

where $x_0 \geq 1$ and $x_j \geq 2x_{j-1}$ for $1 \leq j \leq k-1$.

Proof. We show that all integers within the closed interval $[11034, \frac{3(\ell+1)^2-(\ell+1)-2}{2} + 11033]$ can be written in the above form, where additionally $x_{k-1} \leq \ell$. This is done by induction on ℓ, the first case being $\ell = 160$, which can be checked directly by means of a computer. To proceed from $\ell - 1$ to ℓ, note that if n can be represented in the above form with $x_{k-1} \leq \ell/2$, then $n + \frac{3\ell^2-\ell-2}{2}$ can be written in this form as well, where now $k-1$ is replaced by k and $x_k = \ell$. Therefore, under the induction hypothesis, all numbers in the union

$$\left[11034 + \frac{3\ell^2 - \ell - 2}{2}, \frac{3(\lfloor \ell/2 \rfloor + 1)^2 - (\lfloor \ell/2 \rfloor + 1) - 2}{2} + 11033 + \frac{3\ell^2 - \ell - 2}{2}\right]$$
$$\cup \left[11034, \frac{3\ell^2 - \ell - 2}{2} + 11033\right] \supseteq \left[11034, \frac{3(\ell+1)^2 - (\ell+1) - 2}{2} + 11033\right]$$

(for the inclusion, simply note that $\frac{3(\lfloor \ell/2 \rfloor+1)^2-(\lfloor \ell/2 \rfloor+1)-2}{2} \geq 3\ell+1$ for $\ell \geq 160$) can be represented in the desired form, proving the claim. ■

An immediate consequence of the preceding proposition and its proof is that every integer within the interval

$$\left[\frac{2\ell(2\ell+1)}{2} - \left(\frac{3(\ell+1)^2 - (\ell+1) - 2}{2} + 11033\right), \frac{2\ell(2\ell+1)}{2} - 11034\right]$$

is the number of subtrees of some caterpillar tree, provided that $\ell \geq 160$ (simple take $x_k = 2\ell$ in the above formula). Taking the union over all $\ell \geq 160$ (where we note that subsequent intervals overlap, since

$$\frac{2\ell(2\ell+1)}{2} - \left(\frac{3(\ell+1)^2 - (\ell+1) - 2}{2} + 11033\right) \leq \frac{2\ell(2\ell-1)}{2} - 11034$$

holds for all $\ell \geq 1$), we find that all integers ≥ 1527 can be represented in this way. An additional computer search shows that in fact all positive integers, except for 2, 4, 5, 7, 8, 9, 12, 13, 14, 16, 18, 19, 22, 23, 26, 27, 29, 31, 33, 35, 38, 39, 42, 43, 46, 48, 50, 52, 54, 60, 61, 64, 65, 68, 72, 77, 79, 85, 89, 93, 96, 123, 157 and 183, are the number of subtrees of some caterpillar. Finally, it is not difficult to compute the number of subtrees of all trees with ≤ 18 vertices (by means of our recursions or directly in a brute-force manner) to obtain the final list. This is sufficient, since a tree with n vertices has at least $\binom{n+1}{2}$ subtrees (and thus at least 190 if the number of vertices is ≥ 19): consider any pair of two (possibly equal) vertices; then the unique path between them forms a subtree, proving that the total number of subtrees is indeed at least $\binom{n+1}{2}$.

4. Conclusion

It is tempting to conjecture that an approach as presented in the previous section works for many instances. Of course it can happen that only specific residue classes are covered, as in the case of chains. Generally, a heuristic argument is the following: if the number of trees with n vertices grows faster than the corresponding maximum value of a parameter (the number of non-isomorphic trees grows exponentially with an exponential base of 2.955765 —see [3]), then almost all integer values (or at least those in certain residue classes) should be covered. A Diophantine approach as shown in the preceding section for the number of subtrees is only likely to work if a sufficiently rich set of trees can be constructed for which the graph parameter can be expressed as a polynomial.

A general, but probably extremely hard problem can be stated as follows: given a parameter that is characterized by a certain set of polynomial maps on $\mathbb{R}^k$ (as in our examples in the introduction) and one or more initial values, are there criteria which guarantee that the inverse problem can be solved for all (almost all) positive integers? It might be possible to provide such criteria if all the involved maps are linear, but beyond that the problem is probably intractable.

References

[1] A. A. Dobrynin, R. Entringer, and I. Gutman. Wiener index of trees: theory and applications. *Acta Appl. Math.*, 66(3):211–249, 2001.

[2] R. C. Entringer, A. Meir, J. W. Moon, and L. A. Székely. The Wiener index of trees from certain families. *Australas. J. Combin.*, 10:211–224, 1994.

[3] F. Harary and E. M. Palmer. *Graphical enumeration*. Academic Press, New York, 1973.

[4] O. J. Heilmann and E. H. Lieb. Theory of monomer-dimer systems. *Commun. Math. Phys.*, 25(3):190–232, 1972.

[5] M. Klazar. Twelve countings with rooted plane trees. *European J. Combin.*, 18(2):195–210, 1997.

[6] M. Lepović and I. Gutman. A Collective Property of Trees and Chemical Trees. *J. Chem. Inf. Comput. Sci.*, 38:823–826, 1998.

[7] L. Z. Li, X. and L. Wang. The inverse problems for some topological indices in combinatorial chemistry. *J. Computational Biology*, 10:47–55, 2003.

[8] V. Linek. Bipartite graphs can have any number of independent sets. *Discrete Math.*, 76(2):131–136, 1989.

[9] R. E. Merrifield and H. E. Simmons. *Topological Methods in Chemistry*. Wiley, New York, 1989.

[10] M. B. Nathanson. *Additive number theory*, volume 164 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1996. The classical bases.

[11] L. A. Székely and H. Wang. On subtrees of trees. *Adv. in Appl. Math.*, 34(1):138–155, 2005.

[12] N. Trinajstić. *Chemical graph theory*. CRC Press, Boca Raton, FL., 1992.

[13] S. Wagner. A class of trees and its Wiener index. *Acta Appl. Math.*, 91(2):119–132, 2006.

[14] S. Wagner, H. Wang, and G. Yu. Molecular graphs and the inverse Wiener index problem. Preprint, submitted.

[15] H. Wang and G. Yu. All but 49 numbers are Wiener indices of trees. *Acta Appl. Math.*, 92(1):15–20, 2006.

[16] H. S. Wilf. The number of maximal independent sets in a tree. *SIAM J. Algebraic Discrete Methods*, 7(1):125–130, 1986.

Éva Czabarka, Department of Mathematics, University of South Carolina, Columbia, SC 29208, United States of America

E-mail address: czabarka@math.sc.edu

László Székely, Department of Mathematics, University of South Carolina, Columbia, SC 29208, United States of America

E-mail address: szekely@math.sc.edu

Stephan Wagner, Department of Mathematical Sciences, Stellenbosch University, 7602 Stellenbosch, South Africa

E-mail address: swagner@sun.ac.za