Theoretical Computer Science 732 (2018) 1–13

Contents lists available at ScienceDirect

Theoretical Computer Science

www.elsevier.com/locate/tcs

Leaf realization problem, caterpillar graphs and prefix normal words

Alexandre Blondin Massé [a,*,1], Julien de Carufel [b], Alain Goupil [b], Mélodie Lapointe [a,2], Émile Nadeau [a,3], Élise Vandomme [a]

[a] *Laboratoire de Combinatoire et d'Informatique Mathématique, Université du Québec à Montréal, Canada*
[b] *Laboratoire Interdisciplinaire de Recherche en Imagerie et en Combinatoire, Université du Québec à Trois-Rivières, Canada*

ARTICLE INFO

Article history:
Received 9 December 2017
Received in revised form 3 April 2018
Accepted 10 April 2018
Communicated by D. Perrin

Keywords:
Graph theory
Combinatorics on words
Induced subtrees
Leaf
Prefix normal words
Prefix normal form

ABSTRACT

Given a simple graph G with n vertices and a natural number $i \leq n$, let $L_G(i)$ be the maximum number of leaves that can be realized by an induced subtree T of G with i vertices. We introduce a problem that we call the *leaf realization problem*, which consists in deciding whether, for a given sequence of $n+1$ natural numbers $(\ell_0, \ell_1, \ldots, \ell_n)$, there exists a simple graph G with n vertices such that $\ell_i = L_G(i)$ for $i = 0, 1, \ldots, n$. We present basic observations on the structure of these sequences for general graphs and trees. In the particular case where G is a caterpillar graph, we exhibit a bijection between the set of the discrete derivatives of the form $(\Delta L_G(i))_{1 \leq i \leq n-3}$ and the set of prefix normal words.

1. Introduction

Subtrees of graphs are the object of intensive investigation from several communities. For instance, frequent subtrees mining [1] were studied in the data mining community; induced subtrees [2] are useful in information retrieval [3], and maximum leaf spanning subtrees [4,5] aroused the interest of the telecommunication network community [6,7]. In this paper, we focus on induced subtrees that have a maximal number of leaves. More precisely, for a graph G, let $L_G(i)$ be the maximum number of leaves that can be realized by an induced subtree of G with exactly i vertices. The induced subtrees having the maximum number of leaves are called *fully leafed* and already appear in [8,9]. Remarkable structures on regular lattices are presented in [8]. In [9], the decision problem called *Leafed Induced Subtree problem* and its associated optimization problem were introduced:

Problem 1.1 *(*Leafed Induced Subtree problem *(*LIS*)).* Given a graph G and two nonnegative integers i and ℓ, does there exist an induced subtree of G with i vertices and ℓ leaves?

* Corresponding author.
E-mail address: blondin_masse.alexandre@uqam.ca (A. Blondin Massé).

1 Supported by a grant from the National Sciences and Engineering Research Council of Canada (NSERC) through Individual Discovery Grant RGPIN-417269-2013.
2 Supported by a scholarship from the NSERC through Canada Graduate Scholarships-Doctoral CGSD3-488894-2016.
3 Supported by a scholarship from the NSERC.

https://doi.org/10.1016/j.tcs.2018.04.019
0304-3975/

Problem 1.2 *(*Maximum Leafed Induced Subtree problem *(*MLIS*)).* Given a graph G on n vertices, what is the maximum number of leaves, $L_G(i)$, that can be realized by an induced subtree of G with i vertices, for $i \in \{0, 1, \ldots, n\}$?

First, the authors of [9] proved that the problem LIS is NP-complete and fixed parameter intractable.[4] To tackle the MLIS problem, they provided a branch and bound algorithm with experimental results supporting its efficiency. When restricted to the case of trees, they showed that the MLIS problem is polynomial using a dynamic programming strategy.

In this paper, we are interested in studying the properties of the finite sequence $L_G(i)_{i=0,1,\ldots,n}$, called the *leaf sequence of* G, where n is the number of vertices of G. Naturally, we introduce the *Leaf Realization Problem*:

Problem 1.3 *(Leaf Realization Problem).* Given a sequence of $n+1$ natural numbers $(\ell_0, \ell_1, \ldots, \ell_n)$, does there exist a graph G with n vertices such that

$$(L_G(0), L_G(1), \ldots, L_G(n)) = (\ell_0, \ell_1, \ldots, \ell_n)?$$

Problem 1.3 shows many similarities with other famous realization problems investigated more than 50 years ago. For instance, considerable attention was devoted to the *graph realization problem* [11], which consists in deciding whether a finite sequence of natural numbers $(d_1, \ldots, d_n)$ is the degree sequence of some labeled simple graph. In the case where the answer is positive, the sequence $(d_1, \ldots, d_n)$ is called a *graphic sequence*. The problem was proven solvable in polynomial time [11–13]. In particular, it amounts to verifying n inequalities and whether the sum of degrees is even [11]. Several variations of the graph realization problem have been investigated, such as the *bipartite realization problem* [14,15] and the *digraph realization problem* [16–20].

Although we do not succeed, in this paper, in providing a complete answer to Problem 1.3, we solve the following subproblem, which casts some light on the structure of leaf sequences:

Problem 1.4 *(Leaf Realization Problem for Caterpillar Graphs).* Given a sequence of $n+1$ natural numbers $(\ell_0, \ell_1, \ldots, \ell_n)$, does there exist a caterpillar graph C of n vertices such that

$$(L_C(0), L_C(1), \ldots, L_C(n)) = (\ell_0, \ell_1, \ldots, \ell_n)?$$

It is worth mentioning that, contrary to the approaches used in [11–20], our solution relies on concepts studied in combinatorics on words. In particular, it turns out that the leaf sequences of caterpillar graphs have a surprising link with *prefix normal words* that were introduced by Fici and Lipták [21] and then further investigated by Burcsi et al. [22,23]. The defining property of these binary words is that their prefixes contain at least as many 1's as any of their factors of the same length. Moreover, for any binary word w, there is a prefix normal word w' of the same length such that for any integer k, the maximal number of 1's in a factor of length k coincide for w and w'. Such w' is called the *prefix normal form* of w.

If we consider the sequences of differences between consecutive elements of leaf sequences, also called the discrete derivatives of the leaf sequences, then we prove that, for caterpillar graphs, the set

$$\Delta L_{\mathcal{C}} = \{\Delta L_C : C \text{ is a caterpillar}\}$$

of discrete derivatives of leaf sequences of caterpillar graphs is precisely the set of prefix normal words. To prove this result, we introduce the notion of *reading caterpillars* of a word. The link between binary words and their prefix normal forms is then mirrored in terms of their reading caterpillars. Two words with the same prefix normal form are such that their reading caterpillars have the same leaf functions. This is yet another example of the fruitful interaction between graph theory and combinatorics on words (see for example [24,25]).

The manuscript is organized as follows. Preliminary notions are recalled in Section 2. We introduce the notions of leaf functions, leaf sequences and their discrete derivatives in Section 3. We develop some tools in Section 4 that are useful for the proof of our main theorems. Section 5 and 6 are devoted to our main theorems which address the relationship between caterpillar graphs, prefix normal words and prefix normal forms. We conclude with some perspectives on future work in Section 7.

2. Preliminaries

We start by recalling basics definition of graphs. We refer the reader to [26] for an introduction to this subject. All graphs considered in this text are simple and undirected. Let $G = (V, E)$ be a graph with vertex set V and edge set E. The *degree* of a vertex u is the number of vertices that are adjacent to u and is denoted by $\deg(u)$. We denote by $|G|$ or $|V|$ the total number of vertices of G, which is called the *size* of G. For $U \subseteq V$, the *subgraph of* G *induced by* U, denoted by $G[U]$, is the

[4] Assuming that the class of fixed parameter problems is distinct from the first complexity class $W[1]$ of the W-hierarchy introduced by Downey and Fellows [10].