RAIRO-Inf. Theor. Appl. **39** (2005) 49-65
DOI: 10.1051/ita:2005003

HEREDITARY PROPERTIES OF WORDS *

József Balogh[1] and Béla Bollobás[2]

Abstract. Let $\mathcal{P}$ be a hereditary property of words, *i.e.*, an infinite class of finite words such that every subword (block) of a word belonging to $\mathcal{P}$ is also in $\mathcal{P}$. Extending the classical Morse-Hedlund theorem, we show that either $\mathcal{P}$ contains at least $n+1$ words of length n for every n or, for some N, it contains at most N words of length n for every n. More importantly, we prove the following quantitative extension of this result: if $\mathcal{P}$ has $m \leq n$ words of length n then, for every $k \geq n+m$, it contains at most $\lceil (m+1)/2 \rceil \lfloor (m+1)/2 \rfloor$ words of length k.

Mathematics Subject Classification. 05C.

1. Introduction

For a set A and a natural number n, a *word of length n over the alphabet A*, or, simply, an *n-word*, is a sequence $w = (w_1, w_2, \ldots, w_n) = (w_i)_{i=1}^n$ with $w_i \in A$ for every i. A *finite word* is an n-word for some n. We write $\mathcal{W}_n(A)$ for the set of all n-words over A, and $\mathcal{W}^*(A) = \cup_{n=1}^\infty \mathcal{W}_n(A)$ for the set of all finite words over A. The length of a word $w \in \mathcal{W}^*(A)$ is denoted by $|w|$. A *$\mathbb{Z}$-word* over A is a $\mathbb{Z}$-sequence

$$w = (\ldots, w_{-2}, w_{-1}, w_0, w_1, w_2, \ldots) = (w_i)_{-\infty}^{\infty}$$

with $w_i \in A$ for every i, and we denote by $\mathcal{W}_{\mathbb{Z}}(A)$ the set of $\mathbb{Z}$-words over A. Similarly, an *$\mathbb{N}$-word* is an $\mathbb{N}$-sequence $w = (w_1, w_2, \ldots) = (w_i)_{i=1}^{\infty}$, and $\mathcal{W}_{\mathbb{N}}(A)$ is the set of $\mathbb{N}$-words. An *infinite word* is a $\mathbb{Z}$-word or an $\mathbb{N}$-word.

When there is no danger of confusion, we frequently suppress various parameters like A, n, $\mathbb{Z}$ and $\mathbb{N}$. Thus a word may mean an n-word, a $\mathbb{Z}$-word or an $\mathbb{N}$-word.

Keywords and phrases. Graph properties, monotone, hereditary, speed, size.

* *Research of the first author was partly supported by NSF grant DMS-0302804.*

[1] Department of Mathematical Sciences, The Ohio State University, Columbus, OH 43210, USA; jobal@math.ohio-state.edu

[2] Department of Mathematical Sciences, The University of Memphis, Memphis, TN 38152, and Trinity College, Cambridge CB2 1TQ, England; bollobas@msci.memphis.edu

Also, we shall frequently omit the brackets in our notation; thus $w_1w_2\dots w_n$ means the same as $(w_1,\dots,w_n)$. Given a word $w=\dots w_i\dots$, we call the w_i the *letters* or *digits* of w.

An *n-block* of a word $w=(w_i)$ is an n-word of the form $w_{j+1}w_{j+2}\dots w_{j+n}$ for some j. For simplicity, we shall frequently say that a word u is a *subword* of w, or u is in w, or w contains u, if u is an n-block of w, where n is the length of u.[1] Note that a word of length N has $N-n+1$ n-blocks, *i.e.*, subwords of length n; in particular, a word of length $n+1$ has two subwords of length n. The set of n-blocks of a word w is denoted by $\mathcal{P}^n(w)$, and the function $n\mapsto|\mathcal{P}^n(w)|$ is the *speed* or *complexity* of the word w. For example, if $w=\dots0101010\dots$, then $\mathcal{P}^3(w)=\{010,101\}$, and $|\mathcal{P}^n(w)|=2$ for $n\ge1$. Similarly, for a set W of words and a natural number n, we put $\mathcal{P}^n(W)=\cup_{w\in W}\mathcal{P}^n(w)$, and define $n\mapsto|\mathcal{P}^n(W)|$ to be the *speed* or *complexity* of the set W. Thus, if $W=\{\dots0001000\dots,\ \dots1110111\dots,\ \dots00001111\dots\}$ then $|\mathcal{P}^n(W)|=3n-1$ for $n\ge3$. A $\mathbb{Z}$-word $w=(w_i)$ is *n-periodic* if for every $t\in\mathbb{Z}$, we have $w_t=w_{n+t}$. Similarly an $\mathbb{N}$-word $w=(w_i)$ is *eventually n-periodic* if there is an $N\in\mathbb{N}$ such that for every $t>N$ we have $w_t=w_{t+n}$.

The basic result concerning the complexity of a word is the classical theorem of Morse and Hedlund [6] stating that if w is a $\mathbb{Z}$-word then either it is periodic and so $|\mathcal{P}^n(w)|$ is constant for n large enough, or $|\mathcal{P}^n(w)|\ge n+1$ for every n (see for the precise statement Th. 2). There are words such that $|\mathcal{P}^n(w)|=n+1$ for every n, for example, the $\mathbb{Z}$-word $\dots0001000\dots$ and the Fibonacci $\mathbb{N}$-word 0 1 0 01 010 01001 01001010 .... The Fibonacci word is constructed from the sequence of finite words $a_1,a_2,\dots$ defined as follows: $a_1=0$, $a_2=01$ and for $k\ge3$ the word a_{k+1} is the concatenation of a_k and a_{k-1} (in this order): $a_{k+1}=a_ka_{k-1}$. Equivalently, the Fibonacci sequence is obtained from 0 by repeatedly substituting 01 for 0 and 0 for 1. Thus $a_3=01\ 0$, $a_4=010\ 01$, $a_5=01001\ 010$, and so on. The words $a_1,a_2,\dots$ are 'nested': if the kth digit of a_n is i then for $m>n$ the kth digit of a_m is also i. The Fibonacci $\mathbb{N}$-sequence is the 'union' of the finite sequences $a_1,a_2,\dots$

The complexity of infinite sequences has been studied in great detail in many papers; see, *e.g.*, Ferenczi [2], Tijdeman [7] and Heinis [4]. Here we shall consider some analogous problems concerning the complexity of a hereditary property of finite words. In keeping with the terminology applied to other combinatorial structures, we call an infinite set of finite words over a fixed alphabet a *property* of words. A property $\mathcal{P}$ of words is *hereditary* if every subword of a word in $\mathcal{P}$ is also in $\mathcal{P}$. Equivalently, a hereditary property is the collection $\mathcal{P}$ of all subwords of a set of words, *i.e.*, $\mathcal{P}=\mathcal{P}(W)$ for some set of words W, as defined above.

Also, a property $\mathcal{P}\subset\mathcal{W}^*(A)$ is hereditary if it is the collection of all words containing no subword of a given family $\mathcal{F}$ of words (usually called a family of *forbidden* words):

$$\mathcal{P}=\mathcal{P}(\neg\mathcal{F})=\{w\in\mathcal{W}^*(A):\ \text{no block of } w \text{ belongs to } \mathcal{F}\}.$$

[1] Note that we use the term "subword" in a different way from the usual one; usually 111 is a subword of 0101010.

As before, writing $\mathcal{P}^n$ for the set of n-words in $\mathcal{P}$, the function $n \mapsto |\mathcal{P}^n|$ is the *speed* or *complexity* of $\mathcal{P}$.

In case $\mathcal{P}$ is defined by forbidding finitely many words, *i.e.*, $\mathcal{P} = \mathcal{P}(\neg\mathcal{F})$ for a finite set $\mathcal{F}$, the complexity function $n \mapsto \mathcal{P}(\neg\mathcal{F})$ can be calculated exactly using recurrence equations. However, in the general case, *i.e.*, when $\mathcal{F}$ is infinite, the complexity function is rather mysterious.

Our main aim in this paper is to investigate what slow-growing functions may arise as complexities of hereditary properties of words. In the next section we shall prepare the ground for this; in particular, we shall prove the classical Morse-Hedlund theorem and one of the tools we shall use, the Fine-Wilf theorem. In Section 3 we shall extend the Morse-Hedlund Theorem for hereditary properties of words, and present our main result, a quantitative form of this extension, which is shown to be best possible. In Section 4 we construct a hereditary property whose speed oscillates rather wildly, and in Section 5 we make some concluding remarks.

2. Basic observations

In preparation for the results in Section 3, we introduce a little more terminology and make some observations.

A *word graph* over an alphabet A is a *directed graph* with loops whose edges (including loops) are *decorated* or *coloured* with elements of A such that no two edges with the same initial vertex have the same decoration, but all the edges with the same terminal vertex have the same decoration. (In particular, there is at most one loop at every vertex.)

Given a word w, let $G_k(w)$ be the word graph whose vertices are the k-words in w, and a vertex u sends an edge of decoration i to a vertex v if w contains a $(k+1)$-word ending in i whose first k-word is u and second (and last) k-word is v. Equivalently, v is obtained from u by omitting its first letter and adding i as its last letter. We call $G_k(w)$ the *k-de Bruijn graph* of w or simply the *k-word graph* of w. Similarly, given a set W of words, $G_k(W) = \cup_{w\in W} G_k(w)$ is the k-word graph of W.

Let us collect some observations concerning k-word graphs into the following lemma.

Lemma 1. *(i) A word graph G is a k-word graph iff any two walks of length at most k ending in the same vertex have the same sequence of decorations, and any two walks of length k with the same sequence of decorations end in the same vertex. If every vertex is the terminal vertex of a walk of length k then the alphabet of the k-word graph is the set formed by the decorations of the edges.*

(ii) A k-word graph is of the form $G_k(w)$ for some n-word w iff it has a (directed) walk of length $n-k+1$ passing through all edges.

Proof. We shall prove the only assertion that is not entirely trivial, namely the sufficiency of the conditions in (i). Suppose G satisfies the conditions; our aim is to assign a word $w(x)$ to each vertex x and define a suitable set W of $(k+1)$-words.

Given a vertex x of G, assign a k-word $w(x)$ to x as follows. If there is a walk of length k ending in x, set $w(x) = i_1 i_2 \dots i_k$, where $i_1, i_2, \dots, i_k$ are the decorations of the edges of this walk. Otherwise, pick a walk of maximal length ℓ ending in x, and set $w(x) = j_1 j_2 \dots j_{k-\ell} i_1 i_2 \dots i_k$, where $i_1, i_2, \dots, i_k$ are the decorations of the edges of the path (ending in an edge decorated with i_ℓ) and $j_1, j_2, \dots, j_{k-\ell}$ are letters chosen to ensure that different vertices get assigned different words. The conditions on G imply that different words are assigned to different vertices. Finally, for every edge $e_h = xy$, let w_h be the $(k+1)$-word in which the first k-word is $w(x)$ and the second (and last) is $w(y)$. The way we assigned words to the vertices ensures the existence of w_h, completing the proof of the sufficiency of the conditions for a word graph to be a k-word graph. □

We shall write uv for the concatenation of the finite words u and v; we also call uv the *product* of u and v. Thus, if $u = (u_i)_1^n$ and $v = (v_j)_1^m$ then uv is the $(n+m)$-word $u_1 u_2 \dots u_n v_1 v_2 \dots v_m$. Similarly, if $u \in \mathcal{W}_n(A)$ and $a \in A$ then ua denotes the $(n+1)$-word $u_1 u_2 \dots u_n a$. The symbol w_k^* denotes a word of length k, and w^* stands for a finite word. Thus, the equation $uw_p^* = w_p^* u$ means that the word u is *periodic* with period p: if $u = (u_i)_1^n$ then $u_{p+j} = u_j$ for $1 \le j \le n-p$. Note that in the equation $uw_p^* = w_p^* u$ the two occurrences of w_p^* need not denote the same p-letter word. Also, the equation $uw_p^* = w^* u$ is trivially equivalent to $uw_p^* = w_p^* u$ since it implies that $|uw_p^*| = |u| + p = |w^*| + |u| = |w^*| + |u|$, so $|w^*| = p$.

In this notation, the word graph $G_k(w)$ of $w \in \mathcal{W}_k(A)$ contains an edge from $u \in \mathcal{W}_k(A)$ to $v \in \mathcal{W}_k(A)$, decorated by $a \in A$ if $ua = bv$ for some $b \in A$, or $ua = w^* v$.

Let us recall that with the aid of k-word graphs one can give a very simple proof of the following strong form of the Morse-Hedlund theorem.

Theorem 2. *(i) Let w be a $\mathbb{Z}$-word such that $|\mathcal{P}^k(w)| \le k$ for some k. Then there is an n such that w is n-periodic and $|\mathcal{P}^m(w)| = n$ for every $m \ge n$.*
(ii) Let w be an $\mathbb{N}$-word such that $|\mathcal{P}^k(w)| \le k$ for some k. Then there is an n such that w is eventually n-periodic and $|\mathcal{P}^m(w)| = n$ for every $m \ge n$.

Proof. (i) The statement is trivial, if $|\mathcal{P}^1(w)| = 1$, hence we shall assume that $|\mathcal{P}'(w)| \ge 2$. If two words in $\mathcal{P}^{k+1}(w)$ have different initial k-words then they are themselves different. Consequently, the complexity $|\mathcal{P}^k(w)|$ is a monotone increasing sequence, and so if $|\mathcal{P}^k(w)| \le k$ for some k then $|\mathcal{P}^{n-1}(w)| = |\mathcal{P}^n(w)| = n$ for some n, *i.e.*, the $(n-1)$-word graph $G = G_{n-1}(w)$ has n vertices and n edges. Since G has a walk containing all the edges, and every vertex in G has indegree at least 1 and outdegree at least 1, the graph G is an (oriented) n-cycle. Therefore, the word w is n-periodic and $|\mathcal{P}^m(w)| = n$ for every $m \ge n$, as claimed.

(ii) Proceeding as in (i), we find that in $G_{n-1}(v)$ every vertex has outdegree at least 1 and, with the exception of at most one vertex (the word $v_1 \dots v_{n-1}$), all vertices have indegree at least 1. This implies that $G_{n-1}(v)$ is a cycle together with a path ending on the cycle. (This path may have length 0.) Consequently, v

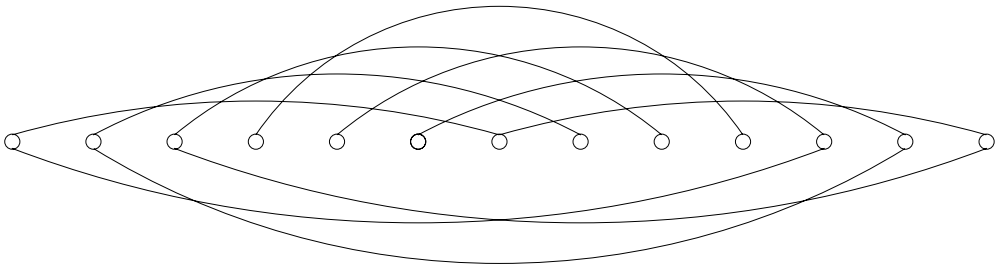

FIGURE 1. The graph $G_{13;6,10}$.

becomes periodic if we omit its initial segment formed by the blocks corresponding to the vertices of $G_{n-1}(v)$ not on the cycle, and so $|\mathcal{P}^m(w)| = n$ for every $m \geq n$. □

One of the tools in our investigation of words of low complexity is the *Fine-Wilf theorem* [3] which, for the sake of completeness, we prove next.

For integers $1 \leq p \leq n$, let $G_{n;p}$ be the graph with vertex set $[n]$ in which ij is an edge if $|i-j| = p$. Also, for $1 \leq p < q < n$, let $G_{n;p,q} = G_{n;p} \cup G_{n;q}$; thus, if $1 \leq i < j \leq n$ then ij is an edge of $G_{n;p,q}$ iff $j-i$ is p or q, as in Figure 1. Clearly, a word $w = w_1 w_2 \dots w_n$ has periods p and q iff $w_i = w_j$ for every edge ij of $G_{n;p,q}$. Equivalently, p and q are periods of w if the w_i are constant on the components of $G_{n;p,q}$, *i.e.*, $w_i = w_j$ whenever i and j are vertices of the same component of $G_{n;p,q}$. Note also that if i and j belong to the same component of $G_{n;p,q}$ then $i-j$ is a multiple of (p,q), the greatest common divisor of p and q. In particular, $G_{n;p,q}$ has at least (p,q) components, no matter what n is. If n is fairly small then the structure of $G_{n;p,q}$ is especially simple, as shown by the lemma below.

Lemma 3. *For $n \leq p+q-(p,q)$ the graph $G_{n;p,q}$ has $p+q-n$ components, each of which is a path.*

Proof. Let us start by showing that every component is a path. For $1 \leq i \leq (p,q)$ the subgraph of $G_{n;p,q}$ induced by the set of vertices congruent to i modulo (p,q) is a union of some components of $G_{n;p,q}$ and is isomorphic to $G_{n';p',q'}$, where $p' = p/(p,q)$, $q' = q/(p,q)$, so that $(p',q') = 1$, and $n' = \lfloor (n-i)/(p,q) \rfloor + 1 \leq p'+q'-1$. Hence in proving that every component of $G_{n;p,q}$ is a path, we may assume that p and q are relatively prime.

Let then $1 \leq p < q < n \leq p+q-1$, with p and q relatively prime. We claim that $G_{n;p,q}$ is a forest. Suppose that this is not the case, and let $C = i_1 i_2 \dots i_\ell$ be a cycle in $G_{n;p,q}$, with the convention that $i_{\ell+1} = i_1$. Note that if ab and bc are distinct edges of $G_{n;p,q}$ and $c-b = q$ then $a-b = p$. Similarly, if bc and cd are distinct edges and $c-b = q$ then $c-d = p$. (In particular, every vertex of $G_{n;p,q}$ has degree at most two.) Consequently, assuming, as we may, that $i_2 - i_1 = q$, the sequence $i-1, i_2, \dots, i_\ell, i_{\ell+1} = i_1$ is such that every increase is by q and

every decrease is by p, *i.e.*, $i_{j+1} - i_j = q$ whenever $i_{j+1} > i_j$ and $i_{j+1} - i_j = -p$ whenever $i_{j+1} < i_j$. Since the total increase equals the total decrease, there are positive integers a and b with $a + b = \ell \leq n \leq p + q - 1$ such that $ap = bq$. (Clearly, a is the number of 'p-edges' of the cycle C and b is the number of "q-edges".) Since p and q are relatively prime, $q|a$ and $p|b$; in particular, $a \geq q$ and $p \geq p$, so that $\ell = a + b \geq p + q$, contradicting that $\ell \leq n \leq p + q - 1$. Hence $G_{n;p,q}$ is indeed a forest.

To complete the proof of our lemma, set $s = p + q - n$. Note that each of the $2s$ vertices $p - s + 1, p - s + 2, \ldots, p$ and $q - s + 1, q - s + 2, \ldots, q$ has degree one, and every other vertex of $G_{n;p,q}$ has degree two. As $G_{n;p,q}$ is a forest, this implies that there are exactly s components, each of which is a path. □

The Fine-Wilf theorem is an immediate consequence of this simple lemma.

Theorem 4. *Let w be a word of length n with periods p and q. If (p, q) is not a period of w then $n \leq p + q - (p, q) - 1$, and this inequality is best possible.*

Proof. To simplify the notation, we set $r = (p, q)$.

(i) Let $w = w_1 w_2 \ldots w_n$ be a word with periods p and $q > p$, where $n = p+q-r$. We claim that r is also a period of w. To see this, note that by Lemma 3 the graph $G_{n;p,q}$ has precisely r components, so each subgraph induced by the vertices congruent to some number i modulo $r = (p, q)$ is connected. Hence, w is constant on congruence classes modulo r, *i.e.*, $r = (p, q)$ is also a period of w.

(ii) Let $n = p + q - r - 1$. Then, by Lemma 3, the graph $G_{n;p,q}$ has $r + 1$ components. (In fact, all we need is that there are at least $r + 1$ components: this follows from the fact that $G_{n;p,q}$ has $2r + 2$ vertices of degree one and all other vertices have degree two.) Then the word $w = w_1 w_2 \ldots w_n$, where w_i is (the label of) the component of i in $G_{n;p,q}$, has periods p and q but not $r = (p, q)$. □

After this preparation, we are ready to turn to the new results of this paper.

3. Hereditary properties of low complexity

What can we say about the complexity of a hereditary property $\mathcal{P}$ of finite words? If it is bounded, is it eventually constant? If it is unbounded, can it be smaller than a rather slowly-growing function, or does it have to "jump" above a certain level? As these questions concern low speeds, in answering them the size of the alphabet A will not matter beyond the trivial condition that it is at least two, so we shall take $A = \{0, 1\}$. Clearly, given $n_0 > m \geq 2$ we may have $|\mathcal{P}^n| = m$ for every $n \geq n_0$ and $|\mathcal{P}^{n_0-1}| > m$; indeed, this is the case if for $n < n_0$ the set $\mathcal{P}^n$ consists of all 2^n words of length n, and for $n \geq n_0$ a word of length n belongs to $\mathcal{P}^n$ iff it has at most one digit 1, and that digit is among the first $m-1$ digits of the word. Also, we may have $|\mathcal{P}^n| = n + 1$ for $n \geq n_0$, as shown by a property $\mathcal{P}$ such that for $n \geq n_0$ a word of length n is in $\mathcal{P}^n$ iff it has at most one digit 1. Our aim is to prove an analogue of the Morse-Hedlund theorem showing that these are the slowest speeds of hereditary properties of finite words. To prove our results

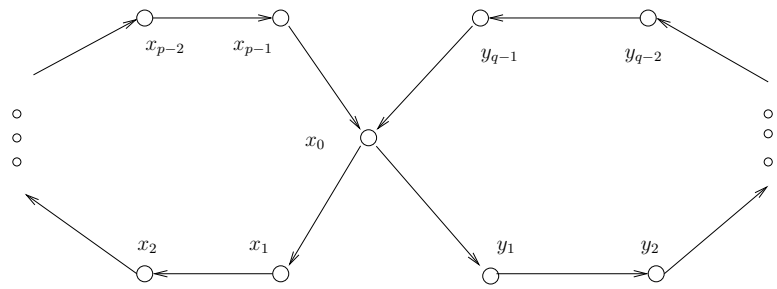

FIGURE 2. The word graph H_1.

we shall need the following key observation, for which we need to define some new graphs:

Let H_1 be the word graph shown in Figure 2. Thus H_1 is the union of a p-cycle and a q-cycle sharing a single vertex x_0:

$$x_0 \to x_1 \to \cdots \to x_{p-1} \to x_0,$$

$$x_0 \to y_1 \to \cdots \to y_{q-1} \to x_0,$$

with a_1 and $a_2 \neq a_1$ the decorations of the edges x_0x_1 and x_0y_1. Similarly, let H_2 be the word graph

$$x_0 \to x_1 \to \cdots \to x_{p-1} \to z_1 \to z_2 \to \ldots$$

$$\cdots \to z_r \to x_0 \to y_1 \to \cdots \to y_{q-1} \to z_1$$

shown in Figure 3, with a_1 and $a_2 \neq a_1$ decorating the edges x_0x_1 and x_0y_1. Note that H_2 is obtained by gluing together a $(p+r)$-cycle and a $(q+r)$-cycle along a path of length r. The graph H_3 is shown in Figure 4: it consists of a p-cycle and a q-cycle joined together by a path of length r. Let H_4 denote the graph shown in Figure 5 consisting of two disjoint cycles, a p-cycle $x_0 \to x_1 \to \cdots \to x_{p-1} \to x_0$ and a q-cycle $y_0 \to y_1 \to \cdots \to y_{q-1} \to y_0$, and a vertex u with a path to each of the cycles, $u \to z_1 \to \cdots \to z_r \to x_0$ and $u \to v_1 \to \cdots \to v_s \to y_0$. Let us define H_5 analogously, reversing the orientations of the paths from u to the cycles (see Fig. 6). Let H_4' be obtained from H_4 by identifying the two cycles; thus H_4 consists of a cycle and two vertex disjoint paths to the cycle starting from a vertex not on the cycle. Similarly, let H_5' be obtained from H_5 by identifying the two cycles in H_5.

Lemma 5. *Suppose G_n is an n-word graph of order at most n. Then G_n contains none of $H_1, H_2, H_3, H_4, H_5, H_4'$ and H_5' as a subgraph.*

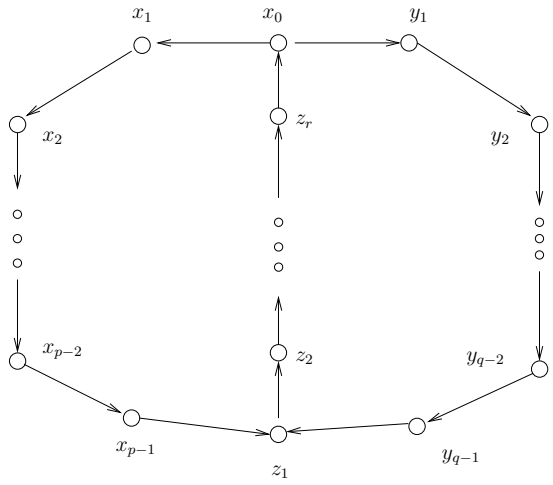

FIGURE 3. The word graph H_2.

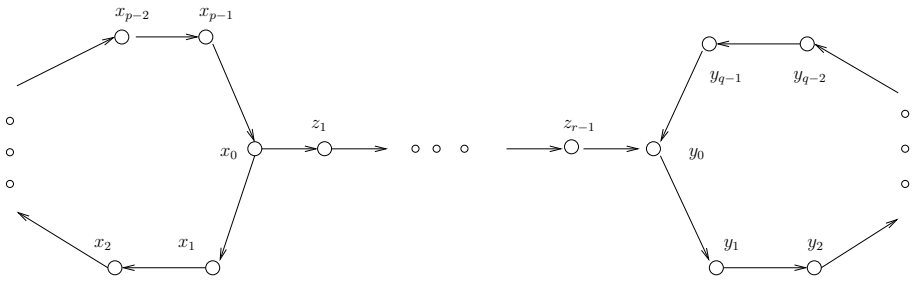

FIGURE 4. The word graph H_3.

Proof. (i) Suppose that H_1 is a subgraph of G. Our aim is to derive the contradiction that

$$n \le p + q - 2 < p + q - 1 = |V(H_1)|. \tag{1}$$

Clearly, we need to prove only the first inequality of (1). Let u_1 be the p-word formed by the decorations of the edges on the p-cycle $x_0x_1 \dots x_{p-1}x_0$, so that $x_0u_1 = w_p^*x_0$, and let u_2 be the corresponding q-word defined by the decorations on the q-cycle $x_0y_1 \dots y_{q-1}x_0$, so that $x_0u_2 = w_q^*x_0$. Recall that $a_1 \neq a_2$, where a_i is the first letter of u_i.

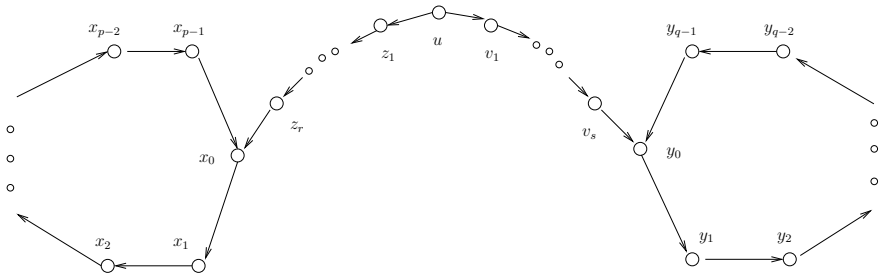

FIGURE 5. The word graph H_4.

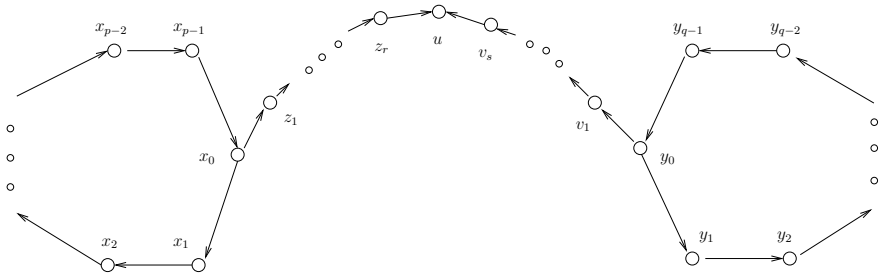

FIGURE 6. The word graph H_5.

Since $x_0u_1 = w_p^*x_0$, the word x_0 has period p; also, $x_0u_2 = w_q^*x_0$ implies that q is a period as well. Furthermore, the former relation tells us that the letter of x_0 in position $n-p+1$ is a_1, and the latter tells us that the letter of x_0 in position $n-q+1$ is a_2. Consequently, $|(n-p+1)-(n-q+1)| = |p-q|$ is not a period of x_0, and so (p,q) is not a period either. Hence, by Theorem 4,

$$n \leq p+q-2,$$

implying (1).

(ii) Secondly, suppose that G contains H_2. As in (i), we shall show that n is too small to make it possible for G to contain H_2, namely that

$$n \leq p+q+r-2 < p+q+r-1 = |V(H_2)|. \tag{2}$$

To see this, let u_1 and u_2 be the products of the decorations on the two $x_0 - z_1$ paths, and let v be given by the decorations on the $z_1 - x_0$ path. Recall that

$a_1 \neq a_2$, where a_i is the first letter of u_i. Recall that $|u_1| = p$, $|u_2| = q$, $|v| = r$, $x_0u_1 = w_p^*z_1$, $x_0u_2 = w_q^*z_1$ and $z_1v = w_r^*x_0$.

Set $w = z_1v$. Then

$$wu_1v = z_1vu_1v = w_r^*x_0u_1v = w_{p+r}^*z_1v = w_{p+r}^*w,$$

so w is $(p+r)$-periodic, and its letter in position $n-(p+r)+1$ is a_1; similarly, w is $(q+r)$-periodic and its letter in position $n-(q+r)+1$ is a_2. Since $a_1 \neq a_2$ and $|(n-(p+r)+1)-(n-(q+r)+1)| = |p-q|$, we see that $p \neq q$ and $|p-q|$ is not a period of w. Hence, $(p+r, q+r)$ is not a period of w. Consequently, by Theorem 4,

$$|w| = n + r \leq (p+r) + (q+r) - 2,$$

implying (2).

(iii) Suppose that G contains H_3. This time our aim is to prove that

$$n \leq p + q + r - 2 < p + q + r - 1 = |V(H_3)|. \tag{3}$$

Proceeding as in the previous cases, let u_1 be the product of the decorations of the cycle $x_0x_1 \dots x_{p-1}x_0$, and let u_2 be defined by $y_0y_1 \dots y_{q-1}y_0$. Also, let v be given by the path $x_0z_1 \dots z_{r-1}y_0$. Let us write a for the first letter in u_1 and b for the first letter in v. Then a and b decorate edges leaving the same vertex, x_0, so $a \neq b$. Note that $|u_1| = p$, $|u_2| = q$ and $|v| = r$.

Define the $(n-r)$-word w by

$$wv = y_0, \quad \textit{i.e.}, \quad x_0 = w_r^*w.$$

Then

$$x_0u_1 = w_p^*x_0 = w_p^*w_r^*w = w_{p+r}^*w$$

and

$$x_0u_1 = w_r^*wu_1.$$

Consequently,

$$wu_1 = w_p^*w,$$

so w has period p, and in position $|w| - |u_1| + 1 = |w| - p + 1$ it has the letter a.

Also,

$$wvu_2 = y_0u_2 = w_q^*y_0 = w_q^*wv,$$

so q is also a period of w, and b is the letter in position $|w| - q + 1$ in w. Since $a \neq b$, this implies that $p \neq q$ and $p - q$ is not a period of w. In particular, (p, q) is not a period of w either, so

$$|w| = n - r \leq p + q - 2,$$

implying (3).

(iv) Suppose now that G contains H_4, with $r \geq s$, say. Now, for a contradiction we shall prove that

$$n \leq p + q + \max\{r, s\} < p + q + r + s + 1 = |V(H_4)|. \tag{4}$$

Let a denote the decoration of the edge uz_1 and b that of uv_1. Define the $(n-r-1)$-word w by setting

$$w^*_{r+1} w = u.$$

This implies that

$$ww^*_{r+1} = x_0 \quad \text{and} \quad w^*_{r-s} w w^*_{s+1} = y_0. \tag{5}$$

As there is a $u - x_0$ walk of length $r + 1$ whose first edge is decorated by a, the $(n - r)$th digit of x_0 is a. Since the word x_0 is p-periodic, for every i, the $(n-r-ip)$th digit of the word x_0 (whenever it exists), is a; by (5), the same holds for the word w as well.

Similarly, there is a $u - y_0$ walk of length $s + 1$ whose first edge is decorated by b, and so the $(n - s)$th digit of y_0 is b. The word y_0 is q-periodic, and so for every j, the $(n - s - jq)$th digit of y_0 (whenever it exists) is b. Using (5), this implies that for every j, the $n-s-jq-(r-s) = (n-r-jq)$th digit of w (whenever it exists) is b. If $n \geq p + q + \max\{r, s\}$ then that both a and b appear in w, and their distance from each other is a multiple of (p, q). This implies that (p, q) is not a period of w. Consequently, by Theorem 4,

$$|w| = n - r - 1 \leq p + q - 2,$$

proving (4).

(v) Suppose that G contains H_5, with $r \geq s$, say. Now we shall derive a contradiction by proving that

$$n \leq p + q + \max\{r, s\} - 1 < p + q + r + s + 1 = |V(H_5)|. \tag{6}$$

Define the $(n - r - 1)$-word w by

$$ww^*_{r+1} = u.$$

Then we also have

$$w^*_{r+1} w = x_0 \quad \text{and} \quad w^*_{s+1} w w^*_{r-s} = y_0.$$

The p-periodicity of x_0 and the q-periodicity of y_0 imply that w has periods both p and q.

First, we consider the case when $r = s$. Since the vertices x_0 and y_0 are different, hence there is a position t where these two words differ. If the first inequality in (6) was false, then $|w| \geq \max\{p, q\}$; therefore there are integers i and j satisfying $r + 1 < t + ip$ and $r + 1 < t + iq$. Then in w there were different

digits in the positions $t+ip-r-1$ and $t+jq-r-1$. This implies that (p,q) is not a period of w, and so, by Theorem 4,

$$|w| = n-r-1 \le p+q-2,$$

implying (6).

Let us turn to the case when $r > s$. Let a be the decoration of the edge x_0x_1 and b that of x_0z_1. Since the initial vertices of the two edges are same, we have $a \neq b$. There is an i, satisfying $s-r \equiv i \pmod q$, such that there is an $(r+1)$-walk from y_i to u, where the walk may go around the y-cycle several times. This means that the edge y_iy_{i+1}, with $i+1$ taken modulo q, has decoration b. Similarly, there is a j satisfying $s-r-p \equiv j \pmod q$, such that there is an $(r+p+1)$-walk from y_j to u. Observe that there is a $(p+r+1)$-walk from x_0 to u (going first around the x-cycle), where the decoration of the first edge of the walk is a. This implies that the decoration of the edge y_jy_{j+1} is also a, with $j+1$ taken modulo q. Since the edges y_iy_{i+1} and y_jy_{j+1} have different colours, we find that (p,q) is not a period of w. Hence, by Theorem 4, we see that $|w| = n-r-1 \le p+q-2$, proving (6).

(vi) The proof of (iv) can be applied to the graph H_4'. The only difference is that now $p=q$, and $|V(H_4')| = p+r+s+1$.

(vii) The proof of (v) can be applied to the graph H_5'. Assume G contains an H_5'. Our aim is now to show that

$$n \le p + \max\{r,s\} < p+r+s+1 = |V(H_5')|. \tag{7}$$

Following the proof of part (v), observe that there cannot exist two x_0-u paths of the same lengths under the assumption (7). The rest of the argument is the same as in (v); here $p=q$ and for every i we have $x_i = y_{i+t}$, where the indices are understood modulo t. □

We are now ready to prove the first of our main results.

Theorem 6. *Let $\mathcal{P}$ be a hereditary property of finite words over an alphabet A. Then $|\mathcal{P}^n|$ is either bounded, or at least $n+1$ for every n.*

Proof. Suppose that, contrary to the assertion of the theorem, $|\mathcal{P}^n|$ is unbounded and $|\mathcal{P}^m| \le m$ for some integer m. Our aim is to arrive at a contradiction.

Since $|\mathcal{P}^n|$ is unbounded, the alphabet A contains at least two letters, and there is an integer $M > 3m$ such that $|\mathcal{P}^M| \ge |A|^{4m}$.

Let $\mathcal{R}^{M,m}$ be the set of M-words (over A, as always) that become periodic with period at most m after the deletion of the first m and last m digits. Clearly

$$|\mathcal{R}^{M,m}| \le |A|^{2m} \sum_{i=1}^{m} |A|^i < |A|^{3m+1},$$

so $\mathcal{P}^M \setminus \mathcal{R}^{M,m} \neq \emptyset$. Let $w_M \in \mathcal{P}^M \setminus \mathcal{R}^{M,m}$.

Let $G = G_m(w_M)$ be the m-word graph of w_M. Note that w_M has $M-m$ blocks of length $m+1$, and the sequence of these $(m+1)$-blocks gives an oriented

walk $T = u_1u_2 \dots u_{M-m+1}$ in G going through every edge uv as many times as the number of $(m+1)$-blocks isomorphic to $[u, v]$, the $(m+1)$-word in which the first m-word is u and the second (and last) is v. Also, as $|\mathcal{P}^m| \leq m$, the graph G has at most m vertices, and as $|T| = M - m + 1 > m$, some vertices are repeated in T. Let s be the minimal index such that $u_s = u_j$ for some $j > s$, and t be the maximal index such that $u_t = u_j$ for some $j < t$. Then the vertices $u_s, u_{s+1}, \dots, u_t$ span the 2-*core* H of G, the maximal subgraph in which every vertex has indegree at least 1 and outdegree at least 1. The existence of the walk T implies that the graph G consists of H, together with a path leading to H and a path leaving H. (A path may have only one vertex and no edge.) Furthermore, since $w_M \notin \mathcal{R}^{M,m}$, each of these paths has fewer than m vertices outside H, so H is non-empty. Again, since $w_M \notin \mathcal{R}^{M,m}$, the number of vertices of H is at most $m < M - 2m$, hence the 2-core H cannot be a single cycle. Consequently, H contains at least two cycles, so it contains either two cycles sharing an oriented path or two cycles joined by a path.

To be precise, H (and so G) contains a subgraph of the form H_1, H_2 or H_3, with some p, q and r natural numbers. By Lemma 5, the existence of H_i in G for $i = 1, 2, 3$ leads to a contradiction. □

It is tempting to conjecture that more is true than claimed by Theorem 6, namely that, just as in the Hedlund-Morse theorem, if the speed of a hereditary property is bounded then it is eventually constant. In fact, this is not the case.

Theorem 7. (i) *For $s \geq 1$ there is a hereditary property $\mathcal{P}$ such that* $\limsup |\mathcal{P}^n| = s^2$ *and* $\liminf |\mathcal{P}^n| = 2s - 1$*; also, $|\mathcal{P}^{4rs}| = s^2$ and $|\mathcal{P}^{(4r-2)s}| = 2s - 1$ for every $r \geq 1$.*

(ii) *Similarly, for $s \geq 1$ there is a hereditary property $\mathcal{P}$ such that* $\limsup |\mathcal{P}^n| = s(s+1)$ *and* $\liminf |\mathcal{P}^n| = 2s$*; also, $|\mathcal{P}^{4rs}| = s(s+1)$ and $|\mathcal{P}^{(4r-2)s}| = 2s$ for every $r \geq 1$.*

Proof. (i) As usual, we take our alphabet to be $A = \{0, 1\}$. For $i \geq 0$, write $[0]_i$ for the word of length i in which every letter is 0, and define $[1]$ analogously. (Thus $[0]_0 = [1]_0$ is the empty word.) For $n \geq 2s$, set

$$W_n = \{w = [1]_i[0]_{n-i-j}[1]_j,\ 0 \leq i, j \leq s - 1\},$$

so that $W_n \subset \mathcal{W}_n$ and $|W_n| = s^2$. Let $\mathcal{P}$ be the hereditary property consisting of all subwords of words in $\cup_{r=1}^{\infty} W_{4rs}$.

By construction, $|\mathcal{P}^{4rs}| = s^2$ and $|\mathcal{P}^n| \leq s^2$ for every n. Also, if $w \in \mathcal{P}^{(4r-2)s}$ then the 1s in w form a block of length at most $s - 1$, which is either at the beginning of w or at its end. Consequently, $|\mathcal{P}^{(4r-2)s}| = 2s - 1$. Furthermore, for every n, the set $\mathcal{P}^n$ contains all words of length n that either start or end with a block of 1s of length at most $s - 1$, so $|\mathcal{P}^n| \geq 2s - 1$.

(ii) This time, define W_n as follows:

$$W_n = \{w = [1]_i[0]_{n-i-j}[1]_j,\ 0 \le i \le s-1 \text{ and } 0 \le j \le s\},$$

and proceed as in (i). □

Clearly, each block of 1s in the definition of W_n in the proof above could be replaced by a single 1 at an appropriate distance from the end, say, in the first case W_n consists of all n-words that contain at most two 1s, each of which is either preceded or followed by at most $s-2$ 0 digits and nothing else. In fact, any *fixed* pattern of 1s would do in the initial block of length $s-1$, provided the $(s-1)$st digit is 1, and any *fixed* pattern of 1s would do in the terminal block of length $s-1$, provided the $(n-s)$th digit (the beginning of the block) is 1.

Let us turn to the main theorem of this paper. This result is not only a quantitative extension of Theorem 6 (and so of Th. 2, the Morse-Hedlund theorem), but also shows that the examples in Theorem 7 are best possible, *even if we do not demand oscillation*: if $|\mathcal{P}^n|$ is no more than n for some n then for no N can $|\mathcal{P}^N|$ be larger than indicated in Theorem 7.

Theorem 8. *Let $\mathcal{P}$ be a hereditary property of finite words over an alphabet A such that $|\mathcal{P}^n| = m \le n$. Then for all $k \ge n+m$ we have $|\mathcal{P}^k| \le \lfloor (m+1)/2 \rfloor \cdot \lceil (m+1)/2 \rceil$. For $n < k < n+m$ we have a weaker bound $\binom{m}{2}$. Furthermore, the inequality for $k \ge n+m$ is sharp.*

Proof. Let $G = G_n$ be the n-word graph of $\mathcal{P}^k$, *i.e.* $\cup_{w_k \in \mathcal{P}^k} G_n(w_k)$. By Lemma 5 G does not contain H_1, H_2, H_3, H_4, H_5, H_4' and H_5' as subgraphs.

Let us summarize what this implies so far about the structure of G.

Let $C_1, \dots, C_t$ denote the directed cycles of G. Since G contains neither H_1 nor H_2, these cycles are vertex-disjoint. Since G does not contain H_3, there is no walk from a cycle to another. Finally, as G does not contain any of the graphs H_4, H_4', H_5 and H_5', for every vertex u not on a cycle there is at most one cycle joined to u by a directed path of either orientation. Consequently, we can partition the vertex set of G as

$$V(G) = V(C_1) \cup V_{out}(C_1) \cup V_{in}(C_1) \cup \dots \cup V(C_t) \cup V_{out}(C_t) \cup V_{in}(C_t) \cup V_0,$$

where $V_{out}(C_i)$ is the set of vertices to which a path is leading from some vertex of C_i, $V_{in}(C_i)$ is the collection of vertices from where there is a path to C_i, and V_0, which is acyclic, is the rest of the vertices.

First, assume that $k \ge n+m$. Then, as $k-n+1 > m = |V(G)|$, for every k-word in $\mathcal{P}^k$ there is a $(k-n+1)$-walk in G with initial vertex in $V_{in}(C_i) \cup V(C_i)$ for some i, and terminal vertex in $V_{out}(C_i) \cup V(C_i)$.

For each i, we can count the number of walks as follows. If a walk starts in $u \in V_{in}(C_i)$, then it must reach $V(C_i)$ in a vertex depending only on u, and its terminal vertex is either a unique vertex in $V(C_i)$ or is in $V_{out}(C_i)$. If a walk starts in $v \in V(C_i)$ then its terminal vertex is either a unique vertex of $V(C_i)$, or is in $V_{out}(C_i)$. There is no $(k-n+1)$-walk with initial vertex in $V_{out}(C_i)$.

Furthermore there is no $(k-n+1)$-walk in V_0. The structure of G implies that there is no walk between vertex classes indexed with different numbers.

This implies that the number of $(k-n+1)$-walks in G is at most

$$\sum_{i=1}^{t}(|V_{in}(C_i)|+|C_i|)\cdot(|V_{out}(C_i)|+1)\le\lfloor(m+1)/2\rfloor\cdot\lceil(m+1)/2\rceil. \tag{8}$$

If equality holds in (8) then G has a unique cycle and V_0 is empty; a further case analysis shows that this cycle is a loop.

For $k\le n+m-1$ a little more work is needed, since there need not be cycle for a $(k-n+1\le m)$-walk in G.

Observe that between two vertices cannot exist two vertex disjoint walks of the same length $r<n$. The graph spanned by V_0 is acyclic, hence the number of $(k-n+1)$-walks in V_0 is at most $\binom{|V_0|}{2}$. In a component $V_{in}(C_i)\cup V(C_i)\cup V_{out}(C_i)$ there may be $\binom{|V_{in}(C_i)|}{2}+\binom{|V_{out}(C_i)|}{2}$ more $(k-n+1)$-walks. In conclusion, the total number of $(k-n+1)$-walks in G is at most

$$\binom{|V_0|}{2}+\sum_{i=1}^{t}\left((|V_{in}(C_i)|+|C_i|)\cdot(|V_{out}(C_i)|+1)+\binom{|V_{in}(C_i)|}{2}+\binom{|V_{out}(C_i)|}{2}\right)\le\binom{m}{2}, \tag{9}$$

completing our proof of the main part of Theorem 8.

Theorem 7 shows that for $k\ge n+m$ the bound is indeed best possible. □

4. Oscillation of complexities of linear order

Theorem 2 describes $\mathbb{N}$- and $\mathbb{Z}$-words w with $\lim_{n\to\infty}|\mathcal{P}^n(w)|/n\le 1$. It would be interesting to determine what numbers may arise as limits $\lim_{n\to\infty}|\mathcal{P}^n(w)|/n$, where w is an $\mathbb{N}$-word (or a $\mathbb{Z}$-word).

It is known that every integer is such a limit, and it was conjectured that only integers arise as limits (for survey of related results, see [7]). Heinis [4] proved that the open interval $(1,2)$ contains no limit points. On the other hand, Heinis also constructed a $\mathbb{Z}$-word w such that $\liminf|\mathcal{P}^n(w)|/n=3/2$ and $\limsup|\mathcal{P}^n(w)|/n=5/3$. Ferenczi [1] described an $\mathbb{N}$-sequence w whose speed is such that $\liminf|\mathcal{P}^n(w)|/n=2$ and for every constant β, $\limsup|\mathcal{P}^n(w)|/n^\beta=\infty$. Here we show that the complexity may oscillate rather wildly.

Theorem 9. *Let $\alpha(n)=o(\log n)$ be a function monotone increasing to ∞. Then there exist a $\mathbb{Z}$-sequence w over the alphabet $\{0,1\}$ and a monotone increasing sequence $\{n_k\}_{k=1}^{\infty}$ such that for k odd $|\mathcal{P}^{n_k}(w)|<2n_k+\alpha(n_k)$ and for k even $|\mathcal{P}^{n_k}(w)|>2^{n_k/\alpha(n_k)}$. Furthermore, there is an $\mathbb{N}$-word w such that for k odd $|\mathcal{P}^{n_k}(w)|<n_k+\alpha(n_k)$ and for k even $|\mathcal{P}^{n_k}(w)|>2^{n_k/\alpha(n_k)}$.*

Proof. We shall define recursively the $\mathbb{Z}$-word w and the sequence $\{n_k\}_{k=1}^{\infty}$. Let n_1 be an integer satisfying $\alpha(n_1) > 1$. Furthermore, place 0 at every position but 0 of the interval $[-n_1, n_1]$ in w. To construct the rest of the word w, we position the digits in such a way that the distance between any two 1s in w is at least n_1. This will guarantee that $\mathcal{P}^{n_1}(w)$ consists of words which have at most one 1 digit, hence $|\mathcal{P}^{n_1}(w)| = n_1 + 1 < 2n_1 + \alpha(n_1)$.

Assume that for some k we have constructed the sequence $n_1, n_2, \ldots, n_{k-1}$, and for some function f, the digits of w in the interval $[-f(k-1), f(k-1)]$ have been fixed. We shall consider two cases according to the parity of k.

First, we assume that k is even. Denote by $A_t(m, w)$ the set of m-words of w in which the distance of any two 1 digits is at least t. We shall choose the rest of w in such a way that for a properly chosen n_k the set $\mathcal{P}^{n_k}(w)$ contains the words intersecting w in one of the positions $[-f(k-1), f(k-1)]$, and the set $A_{n_{k-1}}(n_k, w)$. The recursive inequality $|A_{n_{k-1}}(n, w)| \le 2 \cdot |A_{n_{k-1}}(n + n_{k-1}, w)|$ implies that $|A_{n_{k-1}}(n, w)| \ge 2^{n/n_{k-1}}$. Hence, choosing n_k satisfying $\alpha(n_k) > n_{k-1}$, we shall have $|\mathcal{P}^{n_k}(w)| \ge |A_{n_{k-1}}(n_k, w)| \ge 2^{n_k/n_{k-1}}$. To ensure the containment $A_{n_{k-1}}(n_k, w) \subset \mathcal{P}^{n_k}(w)$, we put every words in $A_{n_{k-1}}(n_k, w)$ in w around the interval $[-f(k-1), f(k-1)]$, putting each of them on both sides, and separate them from each other with n_{k-1} 0 digits. These words can be placed in the interval $[-f(k), f(k)]$ where $f(k) = f(k-1) + (n_k + n_{k-1})|A_{n_{k-1}}(n_k, w)|$.

Whenever k is odd, our aim is to achieve that, in addition to the words intersecting w in one of the positions of $[-f(k-1), f(k-1)]$, the set $\mathcal{P}^{n_k}(w)$ should contain only words having at most one 1 digit. As in the previous case we have seen, this is easily achieved. This implies that $|\mathcal{P}^{n_k}(w)| \le 2n_k + 1 + 2f(k-1) + 1$. Consequently, choosing n_k such that $\alpha(n_k) < 2f(k-1) + 2$, we have $|\mathcal{P}^{n_k}(w)| \le 2n_k + \alpha(n_k)$.

To prove the second part of the theorem, simply consider the same sequence $\{n_k\}$ and the $\mathbb{N}$-sequence spanned by w. □

5. Concluding remarks

In addition to the complexity function studied above, there are various other ways of measuring the wealth of patterns that can be found in words. For example, Kamae and Zamboni [5] studied a complexity function defined by the patterns found at various fixed spaces in a word w, rather then in blocks. To be precise, for $k_1 < k_2 < \cdots < k_n$, consider the set $\mathcal{P}^*(k_1, k_2, \ldots, k_n, w)$ of the n-sub words for all m located at $k_1 + m, k_2 + m, \ldots, k_n + m$ in w. The *pattern complexity* of w is $\mathcal{P}^*(n, w) = \sup_{k_1 < k_2 < \cdots < k_n} |\mathcal{P}^*(k_1, k_2, \ldots, k_n, w)|$. It was shown in [5] that w is either (essentially) periodic or for every n, $\mathcal{P}^*(n, w) \ge 2n$. This result motivates the following question: Given a sequence $K = \{k_1 < k_2 < \ldots\}$, let $\mathcal{P}^*(n, K, w)$ be the set of n-subwords in w in the form of $k_1 + m, k_2 + m, \ldots, k_n + m$. The K-complexity of w is $|\mathcal{P}^*(n, K, w)|$. This coincides with the block complexity of w in case $K = \{1, 2, 3, \ldots\}$. It would be interesting to characterize the sequences K for which there is a word w whose K-complexity is unbounded but sublinear.

Acknowledgements. The authors thank the anonymous referee for numerous helpful suggestions which have contributed greatly to the improvements in exposition in the final version.

References

[1] S. Ferenczi, Rank and symbolic complexity. *Ergodic Theory Dyn. Syst.* **16** (1996) 663–682.
[2] S. Ferenczi, Complexity of sequences and dynamical systems. *Discrete Math.* **206** (1999) 145–154.
[3] N.J. Fine and H.S. Wilf, Uniqueness theorems for periodic functions. *Proc. Amer. Math. Soc.* **16** (1965) 109–114.
[4] A. Heinis, The $P(n)/n$-function for bi-infinite words. *Theoret. Comput. Sci.* **273** (2002) 35–46.
[5] T. Kamae and L. Zamboni, Sequence entropy and the maximal pattern complexity of infinite words. *Ergodic Theory Dynam. Syst.* **22** (2002) 1191–1199.
[6] M. Morse and A.G. Hedlund, Symbolic dynamics. *Amer. J. Math* **60** (1938) 815–866.
[7] R. Tijdeman, *Periodicity and almost periodicity.* Preprint, www.math.leidenuniv/∼tijdeman