ON THE STAR-HEIGHT OF SUBWORD COUNTING LANGUAGES AND THEIR RELATIONSHIP TO REES ZERO-MATRIX SEMIGROUPS

TOM BOURNE AND NIK RUŠKUC

Abstract. Given a word w over a finite alphabet, we consider, in three special cases, the generalised star-height of the languages in which w occurs as a contiguous subword (factor) an exact number of times and of the languages in which w occurs as a contiguous subword modulo a fixed number, and prove that in each case it is at most one. We use these combinatorial results to show that any language recognised by a Rees (zero-)matrix semigroup over an abelian group is of generalised star-height at most one.

1. Introduction and preliminaries

The generalised star-height problem, which asks whether or not there exists an algorithm to compute the generalised star-height of a regular language, is a long-standing problem in the field of formal language theory. In particular, it is not yet known whether there exist languages of generalised star-height greater than one; see [10, Section I.6.4] and [9]. The aim of the present paper is to present some new contributions concerning this problem. In Section 2, we take a combinatorial approach and find the generalised star-height of languages where a fixed word w appears as a contiguous subword of the words in our language precisely k times or k modulo n times. In Section 3, we apply these results to prove that languages recognised by Rees (zero-)matrix semigroups over abelian groups are of generalised star-height at most one.

An *alphabet* A is a finite, non-empty set; its elements are *letters*. A finite sequence of letters is a *word (over A)*. The *length* of a word w, denoted by $|w|$, is the total number of letters appearing in w. The *empty word*, denoted by ε, is the unique word of length zero. The set of all words over A is denoted by A^*, and the set of all non-empty words over A is denoted by A^+. A *semigroup* (respectively, *monoid*) *language* is a subset of A^+ (respectively, A^*).

Given an alphabet A, we define the empty set, the empty word, and each of the letters in A to be *basic regular expressions*. Using these, we recursively define new *regular expressions* by using the (finite) union, concatenation product, (Kleene) star and complement operations; that is, if E and F are regular expressions then so too are $E \cup F$, EF, E^* and E^c. A language is *regular* if it can be represented by a regular expression.

The (generalised) *star-height* of a regular expression E, denoted by $h(E)$, is defined recursively as follows: for the basic regular expressions, $h(\emptyset) = h(\varepsilon) = h(a) = 0$, where a is a letter from A; for union and product, $h(E \cup F) = h(EF) = \max\{h(E), h(F)\}$; for the star operation, $h(E^*) = h(E) + 1$; and for complementation, $h(E^c) = h(E)$. The (generalised) *star-height* of a language L, denoted by $h(L)$, is

$$h(L) = \min\{h(E) \mid E \text{ is a regular expression representing } L\}.$$

Key words and phrases. Regular language, star-height, subword, Rees matrix semigroup.

Note that we can use De Morgan's laws to express intersection and set difference, and that

$$h(E \cap F) = h(E \setminus F) = \max\{h(E), h(F)\}.$$

It is well known that the class of regular languages remains unchanged if complementation is removed from the list of allowed operations. One can define the notion of (restricted) star-height with respect to this signature. In this context, the star-height problem has been solved: there exist languages of arbitrary (restricted) star-height [1], and the (restricted) star-height of a language is algorithmically computable [4].

From this point on, the phrase "star-height" will always refer to generalised star-height.

The following simple observation, which allows 'removal' of stars, will be used throughout the paper:

Observation 1.1 ([8])**.** *For any alphabet A and any subset B of A we have*

$$A^* = \emptyset^c \qquad \text{and} \qquad B^* = A^* \setminus (A^*(A \setminus B)A^*).$$

Hence,

$$h(A^*) = h(B^*) = 0.$$

Let u, w and x be elements of A^*. If $v = uwx$ then u is a *prefix* of v, w is a *contiguous subword* (or *factor*) of v, and x is a *suffix* of v. Throughout this paper, the phrase "subword" will always mean contiguous subword. A prefix of a word that is also a suffix of that word is a *border*, and the proper border of greatest length is said to be *maximal*.

For every word w in A^+ and every word v in A^*, we denote the number of times that w appears as a subword of v by $|v|_w$. When w is a letter, say $w = a$, the notation $|v|_a$ coincides with its usual meaning; that is, the number of times the letter a appears in the word v. For every word w in A^+ and every non-negative integer k, we define the language $\mathrm{Count}(w, k)$ by

$$\mathrm{Count}(w, k) = \{v \in A^* \mid |v|_w = k\};$$

that is, the set of words v over A such that w appears as a subword of v precisely k times. As such, we regard $\mathrm{Count}(w, 0)$ as the set of all words that do not feature w as a subword. From this characterisation, we note that

$$v \in \mathrm{Count}(w, 0) \Leftrightarrow v \in A^* \setminus A^*wA^* \Leftrightarrow v \in (A^*wA^*)^c \Leftrightarrow v \in (\emptyset^c w \emptyset^c)^c\,, \tag{1}$$

where the the final equivalence follows by Observation 1.1. Thus, for a fixed word w, the language $\mathrm{Count}(w, 0)$ is representable by a star-free expression and is therefore of star-height zero.

In a similar manner, for every word w in A^+, every integer n greater than or equal to 2 and every non-negative integer k with $0 \leq k < n$, we define the language $\mathrm{ModCount}(w, k, n)$ by

$$\mathrm{ModCount}(w, k, n) = \{v \in A^* \mid |v|_w \equiv k \pmod{n}\};$$

that is, the set of words v over A such that w appears as a subword of v precisely k modulo n times.

It should be noted that the languages $\mathrm{Count}(w, k)$ and $\mathrm{ModCount}(w, k, n)$ are regular. This can be proved directly; for example, by building a finite state automaton accepting the language and appealing to Kleene's Theorem (see, for example, [10, Theorem I.2.3]). For the languages under consideration in this paper, regularity also follows from the proofs in Section 2.

In Section 2 we prove the following result:

Proposition 1.2. *Let A be an alphabet. For any word w in A^+ with $|w| \leq 3$, the language* $\mathrm{Count}(w, k)$ *is of star-height zero, and the language* $\mathrm{ModCount}(w, k, n)$ *is of star-height at most one.*

In Section 3 we are interested in languages recognised by Rees (zero-)matrix semigroups over abelian groups. A language $L \subseteq A^+$ is *recognised* by a semigroup S if there exists a semigroup morphism $\varphi : A^+ \to S$ and a subset X of S such that $L = X\varphi^{-1}$. Again by Kleene's Theorem, a language is recognisable by a finite semigroup if and only if it is regular. We then prove:

Theorem 1.3. *A language recognised by a Rees (zero-)matrix semigroup over an abelian group is of star-height at most one.*

In order to prove this, we combine the results of Section 2 and a general result on Rees zero-matrix semigroups over semigroups with the following known results:

(AG1) A language L is recognised by a finite abelian group if and only if L is a boolean combination of languages of the form $\mathrm{ModCount}(a, k, n)$, where a is a letter from an alphabet A; see, for example, [7, Corollary 2.3.12].

(AG2) A language recognised by a finite abelian group is of star-height at most one; [5].

2. Counting subwords

Throughout this section we will consistently make use of the notation $\mathrm{Count}(w, k)$ and $\mathrm{ModCount}(w, k, n)$ as introduced in Section 1. We split our analysis into the following three special cases:

(1) Counting subwords over a unary alphabet;
(2) Counting subwords with maximal border ε over a non-unary alphabet;
(3) Counting subwords that are a power of a letter over a non-unary alphabet.

Case (1) is simple, but we consider it for the sake of establishing some equalities that will be useful subsequently. The substantive difference between cases (2) and (3) is that in (3) the letters of w may appear as components of multiple subwords. For example, if we are counting the number of occurrences of aa and we encounter the expression aaa then we have two occurrences of aa and the central a belongs to both. This is not a problem in the second case as having maximal border ε ensures that occurrences of w do not overlap.

2.1. *Case 1: a unary alphabet*

Let $A = \{a\}$ be a unary alphabet. A language L over A is regular if and only if L is of the form $X \cup Y(a^r)^*$, where X and Y are finite sets and r is an integer greater than or equal to 0; see [10, Exercise II.2.4]. Thus, every language over a unary alphabet is of star-height at most one. However, we want to find expressions of minimal star-height for the languages $\mathrm{Count}(a^r, k)$ and $\mathrm{ModCount}(a^r, k, n)$, where r is a natural number, to be used later in the non-unary cases.

We begin with finding an expression for $\mathrm{Count}(a^r, k)$. If we consider an arbitrary word a^s then each a appearing in it is the start of an occurrence of a^r, except for the final $r-1$ letters. It immediately follows that

$$\mathrm{Count}(a^r, 0) = \varepsilon \cup a \cup \cdots \cup a^{r-1}, \tag{2}$$

$$\mathrm{Count}(a^r, k) = a^{r+k-1} \ (k > 0). \tag{3}$$

Next we find an expression for $\mathrm{ModCount}(a^r, k, n)$. The approach taken is to first count k occurrences of the subword a^r and then repeat in multiples of n. Recalling the

expression for $\mathrm{Count}(a^r, k)$ in (3) we obtain

$$\mathrm{ModCount}(a^r, k, n) = a^{r+k-1}(a^n)^*.$$

An expression for the remaining language, namely $\mathrm{ModCount}(a^r, 0, n)$, is obtained by using similar reasoning, but keeping in mind the special nature of $\mathrm{Count}(a^r, 0)$ as in (2); it yields

$$\mathrm{ModCount}(a^r, 0, n) = \varepsilon \cup a \cup \cdots \cup a^{r-1} \cup a^{r+n-1}(a^n)^*.$$

A combination of the above constitutes a proof for the following lemma:

Lemma 2.1. *Let $A = \{a\}$ be a unary alphabet. For every natural number r, the language $\mathrm{Count}(a^r, k)$ is of star-height zero, and the language $\mathrm{ModCount}(a^r, k, n)$ is of star-height at most one.* □

2.2. *Case 2: a non-unary alphabet and maximal border ε*

We now consider the case where A is a non-unary alphabet and the subword w under consideration has maximal border ε, meaning that w does not overlap itself. As such, once we have started to read w we can continue reading it until it finishes without worrying that another occurrence of w may have already begun.

From (1), we know that the language $\mathrm{Count}(w, 0)$ can be represented by the star-free expression $(\emptyset^c w \emptyset^c)^c$. Knowing this, we can obtain an expression representing $\mathrm{Count}(w, k)$ which is star-free:

$$\mathrm{Count}(w, k) = [\mathrm{Count}(w, 0) \cdot w]^k \cdot \mathrm{Count}(w, 0).$$

As can be seen from this expression, we begin with a word from $\mathrm{Count}(w, 0)$, which may be empty, and then count the k occurrences of the subword w, with each pair of occurrences 'padded' by a word from $\mathrm{Count}(w, 0)$. We finish with a word from $\mathrm{Count}(w, 0)$, which, again, may be empty.

We now turn our attention to counting subwords modulo n. An expression for $\mathrm{ModCount}(w, k, n)$ which is of star-height one is given by

$$[\mathrm{Count}(w, 0) \cdot w]^k \big[[\mathrm{Count}(w, 0) \cdot w]^n\big]^* \cdot \mathrm{Count}(w, 0).$$

As can be seen from this expression, we begin with a word from $\mathrm{Count}(w, 0)$ and then count the first k occurrences of the subword w, with each pair of occurrences 'padded' by a word from $\mathrm{Count}(w, 0)$. After this, we allow the same expression to repeat in non-negative multiples of n before ending with a final word from $\mathrm{Count}(w, 0)$.

A combination of the above constitutes a proof for the following lemma:

Lemma 2.2. *Let A be a non-unary alphabet and let w have maximal border ε. Every language $\mathrm{Count}(w, k)$ is of star-height zero, and every language $\mathrm{ModCount}(w, k, n)$ is of star-height at most one.* □

2.3. *Case 3: a non-unary alphabet and powers of a letter*

We now analyse the third case where the subword under consideration consists of a power of a letter a from an alphabet A which contains at least two letters. Specifically, we are interested in finding generalised regular expressions for the languages $\mathrm{Count}(a^r, k)$ and $\mathrm{ModCount}(a^r, k, n)$, where r is a natural number.

From (1), we know that the language $\mathrm{Count}(a^r, 0)$ can be represented by the star-free expression $(\emptyset^c (a^r) \emptyset^c)^c$.

In the case where $k > 0$, we find an expression representing the language $\mathrm{Count}(a^r, k)$ by first considering only those words that have a^r as a border. We denote this language by $\mathrm{CountWB}(a^r, k)$. Let $B = A \setminus \{a\}$. We think of B as a set of 'buffers' that stop

us from 'accidentally' reading two as in a row. This is important as letters may appear as a component of more than one subword and the 'buffers' are used to mark the points where we stop reading powers of a. We also define the subset W of A^* by

$$W = B \cup (B \cdot \mathrm{Count}(a^r, 0) \cdot B),$$

which is the set of non-empty words that do not feature a^r as a subword and neither start nor end with a. It is useful to think of elements of W as 'wedges', separating the strings that feature a^r from one another. Note that the individual components of W are all star-free expressions which implies that W is a language of star-height zero.

A general formula for $\mathrm{CountWB}(a^r, k)$ is given by

$$\mathrm{CountWB}(a^r, k) = \bigcup_{j=1}^{k} \bigcup_{\substack{k_1,k_2,\ldots,k_j \geq r \\ k_1+k_2+\cdots+k_j = k+(r-1)j}} a^{k_1} W a^{k_2} W \ldots W a^{k_j},$$

where the right-hand side is a regular expression since both unions are finite. Note that the expression is star-free. To see that this equality is correct, consider an arbitrary word w in $\mathrm{CountWB}(a^r, k)$. Let $a^{k_1}, \ldots, a^{k_j}$ be the maximal subwords of w that are powers of a and have length greater than or equal to r. Note that a^{k_1} must be a prefix of w as w starts with a^r, and, likewise, a^{k_j} must be a suffix. Hence, we have a decomposition $w = a^{k_1} w_1 a^{k_2} w_2 \ldots w_{j-1} a^{k_j}$, where, necessarily, $w_1, \ldots, w_{j-1}$ belong to W. Furthermore, each a^{k_i} contains precisely $k_i - r + 1$ occurrences of a^r by (3). Since all of the occurrences of a^r appear as subwords of a^{k_i}, we must have

$$k = |w|_{a^r} = \sum_{i=1}^{j} (k_i - r + 1) = k_1 + \cdots + k_j - (r-1)j,$$

and so w belongs to the right-hand side. A similar analysis shows that, conversely, every element of the right-side side belongs to $\mathrm{CountWB}(a^r, k)$.

Now, a star-free expression representing the language consisting of *all* words that contain precisely k occurrences of a^r as a subword, namely $\mathrm{Count}(a^r, k)$, is given by

$$\big[\varepsilon \cup [\mathrm{Count}(a^r, 0) \cdot B]\big] \cdot \mathrm{CountWB}(a^r, k) \cdot \big[[B \cdot \mathrm{Count}(a^r, 0)] \cup \varepsilon\big].$$

To see this, note that the k occurrences of a^r all appear in the central term $\mathrm{CountWB}(a^r, k)$. This term can be preceded by either the empty word or a word that does not contain a^r as a subword; that is, a word from the language $\mathrm{Count}(a^r, 0)$. However, since words in $\mathrm{Count}(a^r, 0)$ have the potential to end with a power of a, we must utilise a 'buffer' from the set B. A dual argument deals with potential suffices. Since each of the components of the above expression are star-free, the language $\mathrm{Count}(a^r, k)$ must be of star-height zero.

We now turn our attention to counting occurrences of a^r modulo n. Our strategy here is to count the first k occurrences of a^r using the expression found above for $\mathrm{CountWB}(a^r, k)$, and then count occurrences of a^r in multiples of n before adding appropriate prefixes and suffices (as in the case of $\mathrm{Count}(a^r, k)$).

Having used $\mathrm{CountWB}(a^r, k)$ to count the first k occurrences of a^r, we note that the suffix a^{r-1} has the potential to be a component of a new occurrence of a^r if the part of the word immediately following a^{r-1} begins with an a. Similarly, the suffix a^{r-2} immediately followed by an a^2 leads to another occurrence of a^r. In order to take these possibilities into account, let $\mathrm{Mult}(a^r, n)$ denote the language whose words contain precisely n occurrences

of the subword a^r when left concatenated by a^{r-1} and also have suffix a^r:

$$\mathrm{Mult}(a^r, n) = \{w \in A^* \mid |a^{r-1}w|_{a^r} = n \text{ and } w \text{ has suffix } a^r\}.$$

The significance of the assumption about the suffix a^r is that every count stops precisely when the n-th occurrence of a^r is met, and that this suffix 'feeds into' the next group of occurrences of a^r.

A star-free expression for this language is given by

$$\mathrm{Mult}(a^r, n) = a^n \cup \bigcup_{i=0}^{n-1} a^i W \cdot \mathrm{CountWB}(a^r, n-i).$$

Note that the right-hand side is a regular expression since the union is finite. To see that this equality is correct, consider an arbitrary word w in $\mathrm{Mult}(a^r, n)$. If $w = a^k$ for some natural number k then

$$n = |a^{r-1}w|_{a^r} = |a^{r-1}a^k|_{a^r} = |a^{r+k-1}|_{a^r} = k$$

by (3), and hence $w = a^n$. Otherwise, we can decompose w as

$$w = a^{k_1} w_1 a^{k_2} w_2 \dots w_{j-1} a^{k_j},$$

where,

$$w_1, \dots, w_{j-1} \in W, \qquad k_1 \geq 0 \qquad \text{and} \qquad k_2, \dots, k_j \geq r.$$

The maximal subwords of $a^{r-1}w$ that are powers of a of exponent greater than or equal to r are $a^{r-1}a^{k_1} = a^{r+k_1-1}$ (provided that $k_1 > 0$) and $a^{k_2}, \dots, a^{k_j}$. Furthermore, our decomposition of w can be used to split $a^{r-1}w$ as $a^{r-1}w = xy$, where $x = a^{r+k_1-1}w_1$ and $y = a^{k_2}w_2 \dots w_{j-1}a^{k_j}$. Suppose that x contains i occurrences of a^r. Then

$$i = |a^{r+k_1-1}w_1|_{a^r} = |a^{r+k_1-1}|_{a^r} = k_1$$

by (3). Moreover, y must contain the remaining $n-i$ occurrences of a^r and has a^r as a border. Hence y belongs to $\mathrm{CountWB}(a^r, n-i)$. Thus, w belongs to $a^iW \cdot \mathrm{CountWB}(a^r, n-i)$ and hence belongs to the union on the right-hand side. A similar analysis shows that, conversely, every element of the right-hand side belongs to $\mathrm{Mult}(a^r, n)$.

Putting all of this together, we have that an expression representing $\mathrm{ModCount}(a^r, k, n)$, where $k > 0$, is given by

$$\big[\varepsilon \cup [\mathrm{Count}(a^r, 0) \cdot B]\big] \cdot \mathrm{CountWB}(a^r, k) \cdot \mathrm{Mult}(a^r, n)^* \cdot \big[[B \cdot \mathrm{Count}(a^r, 0)] \cup \varepsilon\big],$$

and an expression representing $\mathrm{ModCount}(a^r, 0, n)$ is given, with slight abuse of notation, by

$$\mathrm{Count}(a^r, 0) \cup \mathrm{ModCount}(a^r, n, n).$$

Both of these expressions are of star-height one, and so the language $\mathrm{ModCount}(a^r, k, n)$ is of star-height at most one.

A combination of the above constitutes a proof for the following lemma:

Lemma 2.3. *Let A be a non-unary alphabet. For every natural number r, the language* $\mathrm{Count}(a^r, k)$ *is of star-height zero, and the language* $\mathrm{ModCount}(a^r, k, n)$ *is of star-height at most one.* □

2.4. *Discussion*

Based on the results presented so far, it is natural to ask whether the language $\mathrm{Count}(w,k)$ is of star-height zero and whether the language $\mathrm{ModCount}(w,k,n)$ is of star-height at most one for all words w. As a consequence of the foregoing results, this is certainly the case for words of length ≤ 2: indeed every such word is either a power of a letter or has maximal border ε.

Proposition 2.4. *Let A be an alphabet. For any word w in A^+ with $|w| \leq 2$, the language $\mathrm{Count}(w,k)$ is of star-height zero, and the language $\mathrm{ModCount}(w,k,n)$ is of star-height at most one.* □

When the subword under consideration is of length three we are presented with a new hurdle to overcome. The possible types for words of length three are

$$aaa, \qquad aab, \qquad aba, \qquad abb, \qquad abc,$$

where a, b and c are distinct letters in A. Counting occurrences of the word aaa is covered by Lemmas 2.1 and 2.3, while the words aab, baa and abc are covered by Lemma 2.2.

With the final type, namely aba, we must be more careful as the maximal border in this case is a, meaning that the suffix a can act as a prefix a in a new occurrence of the subword. For example, the word $abababa$ contains three occurrences of the subword aba. However, we can proceed in a similar manner to that in Section 2.3 to resolve this issue.

Define W to be the set of words that are not b, do not have prefix ba, do not have suffix ab, and do not contain aba as a subword; that is,

$$W = (b \cup baA^* \cup A^*ab \cup A^*abaA^*)^c = (b \cup ba\emptyset^c \cup \emptyset^c ab \cup \emptyset^c aba\emptyset^c)^c.$$

Then, a general formula for $\mathrm{CountWB}(aba,k)$, where k is a natural number, is given by

$$\mathrm{CountWB}(aba,k) = \bigcup_{j=1}^{k} \bigcup_{\substack{k_1,k_2,\ldots,k_j \geq 1 \\ k_1+k_2+\cdots+k_j=k}} a(ba)^{k_1} W a(ba)^{k_2} W \ldots W a(ba)^{k_j},$$

which is star-free, and the language $\mathrm{Count}(aba,k)$, expressed by

$$(\emptyset^c aba\emptyset^c \cup \emptyset^c ab)^c \cdot \mathrm{CountWB}(aba,k) \cdot (ba\emptyset^c \cup \emptyset^c aba\emptyset^c)^c,$$

is of star-height zero.

To find an expression for $\mathrm{ModCount}(aba,k,n)$ we introduce the language

$$\mathrm{Mult}(aba,n) = \{w \in A^* \mid |aw|_{aba} = n \text{ and } w \text{ has suffix } aba\}.$$

A star-free expression representing $\mathrm{Mult}(aba,n)$ is given by

$$(ba)^n \cup \bigcup_{i=1}^{n-1} (ba)^i W \cdot \mathrm{CountWB}(aba, n-i).$$

Putting all of this together, an expression representing $\mathrm{ModCount}(aba,k,n)$, where $k > 0$, is given by

$$(\emptyset^c aba\emptyset^c \cup \emptyset^c ab)^c \cdot \mathrm{CountWB}(aba,k) \cdot \mathrm{Mult}(aba,n)^* \cdot (ba\emptyset^c \cup \emptyset^c aba\emptyset^c)^c,$$

and an expression representing $\mathrm{ModCount}(aba,0,n)$ is given, with slight abuse of notation, by

$$\mathrm{Count}(aba,0) \cup \mathrm{ModCount}(aba,n,n).$$

This establishes that the language $\mathrm{ModCount}(aba,k,n)$ is of star-height at most one.

Hence, we have proven the following result:

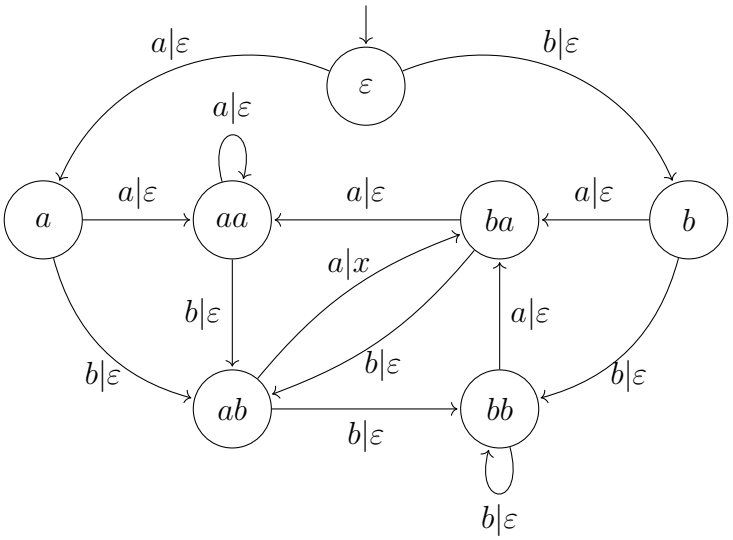


FIGURE 1. A finite state transducer realising f_{aba}.

Proposition 2.5. *Let A be an alphabet. For any word w in A^+ with $|w| \leq 3$, the language* $\mathrm{Count}(w, k)$ *is of star-height zero, and the language* $\mathrm{ModCount}(w, k, n)$ *is of star-height at most one.* □

It should be noted that Proposition 2.5 can also be proved using existing theoretical results. We briefly outline the proof strategy below.

Let A and $X = \{x\}$ be alphabets, and consider the languages $L = \mathrm{ModCount}(x, k, n) = x^k(x^n)^*$ over X and $K = \mathrm{ModCount}(w, k, n)$ over A. Define a function $f_w : A^* \to X^*$ by $f_w(v) = x^{|v|_w}$. It is easy to show that $K = Lf_w^{-1}$.

Note that for all words w with $|w| \leq 3$, f_w is a generalised sequential function (in the sense of Eilenberg [2, p. 299]). For example, a transducer realising f_{aba} is shown in Figure 1. In this diagram, edges labelled with $c|\varepsilon$, where $c \in A \setminus \{a, b\}$, have been removed for clarity, since all of these edges point directly to the initial state.

Standard calculations show that the transition monoid of each transducer realising f_w, where $|w| \leq 3$ is aperiodic. Moreover, the transition monoid of the automaton recognising L is an abelian group by (AG1) and (AG2). Hence, by Eilenberg [3, Proposition IX.1.1] (suitably modified to deal with generalised sequential functions), the transition monoid of K divides a wreath product of an abelian group by an aperiodic monoid. Since all languages that belong to the pseudovariety generated by wreath products of abelian groups by aperiodic monoids have star-height at most one [9, Theorem 7.8], we conclude that K is of star-height at most one.

3. Applications to Rees zero-matrix semigroups

In this section, we change tack and use our combinatorial results to prove new results in an algebraic setting. Specifically, we show that languages recognised by Rees zero-matrix semigroups over abelian groups are of star-height at most one.

Let S be a semigroup without zero. Let I and Λ be non-empty indexing sets and let P be a $|\Lambda| \times |I|$ matrix with entries from $S \cup \{\mathbf{0}\}$, where $\mathbf{0}$ is a new symbol not in S. The *Rees zero-matrix semigroup* $M^0[S; I, \Lambda; P]$ is the set $(I \times S \times \Lambda) \cup \{\mathbf{0}\}$ equipped with the

binary operation defined by

$$(i,s,\lambda)(j,t,\mu) = \begin{cases} (i, sp_{\lambda j}t, \mu) & \text{if } p_{\lambda j} \neq \mathbf{0}, \\ \mathbf{0} & \text{if } p_{\lambda j} = \mathbf{0}, \end{cases}$$

and $s\mathbf{0} = \mathbf{0} = \mathbf{0}s$ for all s in $S \cup \{\mathbf{0}\}$. If we disregard the new symbol $\mathbf{0}$ but leave everything else intact then the resulting semigroup, denoted by $M[S; I, \Lambda; P]$, is simply a *Rees matrix semigroup*. Throughout the rest of this section we work in full generality with Rees zero-matrix semigroups. The results of course remain true when restricted to Rees matrix semigroups.

We say that the matrix P is *regular* if each row and column contain a non-zero entry. Rees zero-matrix semigroups with finite underlying groups and regular matrices are precisely finite 0-simple semigroups according to Rees' Theorem [6, Theorem 3.2.3]. In turn, these semigroups together with zero semigroups completely exhaust principal factors of arbitrary finite semigroups.

We begin by exploring which languages are recognised by Rees zero-matrix semigroups over cyclic groups $\mathbb{Z}_n$, where n is a natural number, and then extend this to arbitrary abelian groups via the Fundamental Theorem for Finite Abelian Groups.

3.1. *Rees zero-matrix semigroups over cyclic groups*

Let $S = M^0[\mathbb{Z}_n; I, \Lambda; P]$ be a Rees zero-matrix semigroup, where the zero in S is denoted by $\mathbf{0}$ and the identity in $\mathbb{Z}_n$ is denoted by 0. Let A be an alphabet and define a map $\varphi : A \to S$ by either $a\varphi = \mathbf{0}$ or $a\varphi = (i_a, g_a, \lambda_a)$, where $0 \leq g_a < n$. Let

$$A_{(i,g,\lambda)} = (i,g,\lambda)\varphi^{-1} \qquad \text{and} \qquad A_{\mathbf{0}} = \mathbf{0}\varphi^{-1}.$$

Uniquely extend φ to a morphism $\bar{\varphi} : A^+ \to S$.

Now, consider the image of $w = a_1 a_2 \dots a_r$ under $\bar{\varphi}$. If $a_t\bar{\varphi} = \mathbf{0}$ for at least one $t \in \{1, 2, \dots, r\}$ then $w\bar{\varphi} = \mathbf{0}$. Likewise, if $p_{\lambda_{a_t} i_{a_{t+1}}} = \mathbf{0}$ for at least one $t \in \{1, 2, \dots, r-1\}$ then $w\bar{\varphi} = \mathbf{0}$. Otherwise, if $a_t\varphi \neq \mathbf{0}$ for all $t \in \{1, 2, \dots, r\}$ and $p_{\lambda_{a_t} i_{a_{t+1}}} \neq \mathbf{0}$ for all $t \in \{1, 2, \dots, r-1\}$, then

$$\begin{aligned} w\bar{\varphi} &= (i_{a_1}, g_{a_1}, \lambda_{a_1})(i_{a_2}, g_{a_2}, \lambda_{a_2}) \dots (i_{a_r}, g_{a_r}, \lambda_{a_r}) \\ &= (i_{a_1}, g_{a_1} + p_{\lambda_{a_1} i_{a_2}} + g_{a_2} + p_{\lambda_{a_2} i_{a_3}} + \dots + p_{\lambda_{a_{r-1}} i_{a_r}} + g_{a_r}, \lambda_{a_r}). \end{aligned}$$

We proceed by finding regular expressions for preimages of elements in S. We split into two cases: the preimage of the zero $\mathbf{0}$ and the preimage of an arbitrary non-zero element $s = (i, g, \lambda)$.

Lemma 3.1. *With the notation as above, $\mathbf{0}\bar{\varphi}^{-1}$ is of star-height zero.*

Proof. According to the analysis preceding the lemma, a word $w = a_1 a_2 \dots a_r$ belongs to the preimage of $\mathbf{0}$ if and only if at least one of the following holds:

(1) a_t lies in $A_{\mathbf{0}}$ for some t in $\{1, 2 \dots, r\}$; or
(2) $p_{\lambda j} = \mathbf{0}$, where $a_t \in A_{(i,g,\lambda)}$ and $a_{t+1} \in A_{(j,h,\mu)}$.

It follows that

$$\mathbf{0}\bar{\varphi}^{-1} = A^* A_{\mathbf{0}} A^* \cup \Big[\bigcup A^* A_{(i,g,\lambda)} A_{(j,h,\mu)} A^* \Big],$$

where the second union is taken over all $(i,g,\lambda), (j,h,\mu) \in S \setminus \{\mathbf{0}\}$ with $p_{\lambda j} = \mathbf{0}$, a language of star-height zero by Observation 1.1. □

Lemma 3.2. *For a non-zero element $s = (i, g, \lambda)$ in S, its preimage, $s\bar{\varphi}^{-1}$, is of star-height at most one.*

Proof. We begin by writing $s\bar{\varphi}^{-1}$ as the intersection of three regular languages as follows:

$$s\bar{\varphi}^{-1} = (\{i\} \times \mathbb{Z}_n \times \Lambda)\,\bar{\varphi}^{-1} \cap (I \times \{g\} \times \Lambda)\,\bar{\varphi}^{-1} \cap (I \times \mathbb{Z}_n \times \{\lambda\})\,\bar{\varphi}^{-1}. \tag{4}$$

Due to the nature of the multiplication on S, it is clear to see that

$$(\{i\} \times \mathbb{Z}_n \times \Lambda)\,\bar{\varphi}^{-1} = \left[\bigcup_{h\in\mathbb{Z}_n,\mu\in\Lambda} A_{(i,h,\mu)}\right] \cdot A^*,$$

$$(I \times \mathbb{Z}_n \times \{\lambda\})\,\bar{\varphi}^{-1} = A^* \cdot \left[\bigcup_{j\in I,h\in\mathbb{Z}_n} A_{(j,h,\lambda)}\right].$$

By Observation 1.1, these languages have star-height zero.

It remains to find an expression for $(I \times \{g\} \times \Lambda)\,\bar{\varphi}^{-1}$. Consider an arbitrary $w = a_1a_2\ldots a_r$ belonging to this language. Continuing to use the notation introduced before Lemma 3.1, we know that $p_{\lambda_{a_t} i_{a_{t+1}}} \neq \mathbf{0}$ for $t = 1, 2, \ldots, r-1$, and

$$g_{a_1} + p_{\lambda_{a_1} i_{a_2}} + g_{a_2} + p_{\lambda_{a_2} i_{a_3}} + \cdots + p_{\lambda_{a_{r-1}} i_{a_r}} + g_{a_r} \equiv g \pmod{n}.$$

We split the above sum into two:

$$\underbrace{g_{a_1} + g_{a_2} + \cdots + g_{a_r}}_{\equiv g_1 \pmod{n}} + \underbrace{p_{\lambda_{a_1} i_{a_2}} + p_{\lambda_{a_2} i_{a_3}} + \cdots + p_{\lambda_{a_{r-1}} i_{a_r}}}_{\equiv g_2 \pmod{n}} \equiv g \pmod{n},$$

and we examine them separately. The first sum corresponds to the contributions from 'group' summands, while the second is the contributions from 'matrix' summands.

For the group contribution, we consider the congruence given by

$$g_{a_1} + g_{a_2} + \cdots + g_{a_r} \equiv g_1 \pmod{n}.$$

Grouping together summands corresponding to the same letter, we see that the above congruence is equivalent to

$$\sum_{a\in A} g_a |w|_a \equiv g_1 \pmod{n},$$

which, in turn, is equivalent to

$$\sum_{a\in A} g_a(|w|_a \pmod{n}) \equiv g_1 \pmod{n}.$$

The point here is that while $|w|_a$ can take infinitely many values, the same is not true for $|w|_a \pmod{n}$. More formally, let T be the following set of tuples of elements $\{0, 1, \ldots, n-1\}$ indexed by A:

$$T = \{(k_a)_{a\in A} \mid \sum_{a\in A} g_a k_a \equiv g_1 \pmod{n}\}.$$

For any fixed tuple $(k_a)_{a\in A}$ in T, every word w such that $|w|_a \equiv k_a \pmod{n}$, where a lies in A, will have group contribution equal to $g_1 \bmod n$. The set of all such words is obtained by forming the finite intersection of the languages $\mathrm{ModCount}(a, k_a, n)$ for $a \in A$. Taking the finite union over all tuples in T results in the expression

$$\mathrm{GrpContrib}(g_1, n) = \bigcup_{(k_a)\in T} \bigcap_{a\in A} \mathrm{ModCount}(a, k_a, n),$$

which is of star-height at most one, since $\mathrm{ModCount}(a, k_a, n)$ is of star-height at most one by Lemmas 2.1 and 2.2.

In a similar fashion, we consider the contributions made by 'matrix' summands; that is, we consider the congruence given by

$$p_{\lambda_{a_1} i_{a_2}} + p_{\lambda_{a_2} i_{a_3}} + \cdots + p_{\lambda_{a_{r-1}} i_{a_r}} \equiv g_2 \pmod{n}.$$

Counting the contribution of each matrix entry separately, we see that the above congruence is equivalent to

$$\sum_{ab \in A^2} p_{\lambda_a i_b} |w|_{ab} \equiv g_2 \pmod{n},$$

which, in turn, is equivalent to

$$\sum_{ab \in A^2} p_{\lambda_a i_b} (|w|_{ab} \pmod{n}) \equiv g_2 \pmod{n}.$$

Consider the finite family U of tuples $(k_{ab})_{ab \in A^2}$ of elements$\{0, 1, \ldots, n-1\}$, indexed by A^2:

$$U = \{(k_{ab})_{ab \in A^2} \mid \sum_{ab \in A^2} p_{\lambda_a i_b} k_{ab} \equiv g_2 \pmod{n}\}.$$

For a fixed tuple in U, the set of all words w satisfying $|w|_{ab} \equiv k_{ab} \pmod{n}$, where ab lies in A^2, is obtained by taking the finite intersection of the languages $\mathrm{ModCount}(ab, k_{ab}, n)$. Taking the union over all tuples in U yields

$$\mathrm{MatContrib}(g_2, n) = \bigcup_{(k_{ab})_{ab \in A^2} \in U} \bigcap_{ab \in A^2} \mathrm{ModCount}(ab, k_{ab}, n),$$

which is of star-height at most one, since $\mathrm{ModCount}(ab, k_{ab}, n)$ is of star-height at most one by Proposition 2.4.

Combining the 'group' contribution and the 'matrix' contribution appropriately leads to

$$(I \times \{g\} \times \Lambda)\,\bar{\varphi}^{-1} = \bigcup_{\substack{(g_1, g_2) \in \mathbb{Z}_n^2 \\ g_1 + g_2 \equiv g \pmod{n}}} (\mathrm{GrpContrib}(g_1, n) \cap \mathrm{MatContrib}(g_2, n)),$$

and completes the proof. □

An immediate consequence of the above propositions is the following theorem:

Theorem 3.3. *A regular language recognised by a Rees zero-matrix semigroup over a cyclic group is of star-height at most one.*

Proof. Every language recognised by a Rees zero-matrix semigroup over a cyclic group can be expressed as a finite union of preimages of elements in the semigroup. Since each individual preimage is of star-height at most one and taking finite unions does not increase star-height, the result follows. □

3.2. *Extending to abelian groups*

We now extend Theorem 3.3 to Rees zero-matrix semigroups over abelian groups. In order to do this we make use of properties of homomorphisms and projection maps and appeal to the Fundamental Theorem of Finite Abelian Groups.

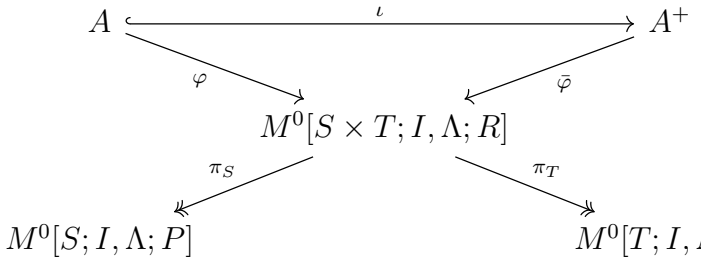
We begin with some general theory concerning Rees matrix semigroups over direct products of semigroups. Consider a Rees zero-matrix semigroup $M^0[S \times T; I, \Lambda; R]$, with $R = (r_{\lambda i})$, where $r_{\lambda i} = (p_{\lambda i}, q_{\lambda i})$ lies in $S \times T$ or $r_{\lambda i} = \mathbf{0}_{S \times T}$, the zero element. Define two further Rees matrix semigroups $M^0[S; I, \Lambda; P]$ and $M^0[T; I, \Lambda; Q]$, with zeros $\mathbf{0}_S$ and $\mathbf{0}_T$

respectively, and matrices P and Q defined by $P = (p_{\lambda i})$ and $Q = (q_{\lambda i})$, where we take $p_{\lambda i} = \mathbf{0}_S$ and $q_{\lambda i} = \mathbf{0}_T$ whenever $r_{\lambda i} = \mathbf{0}_{S\times T}$. We then have two natural projections:

$$\pi_S \,:\, M^0[S\times T; I, \Lambda; R] \to M^0[S; I, \Lambda; P] \,:\, (i,(s,t),\lambda) \mapsto (i,s,\lambda),$$
$$\pi_T \,:\, M^0[S\times T; I, \Lambda; R] \to M^0[T; I, \Lambda; Q] \,:\, (i,(s,t),\lambda) \mapsto (i,t,\lambda).$$

Proof that these are epimorphisms is routine and is left as an exercise.

Now suppose that we are given an alphabet A and a map $\varphi : A \to M^0[S\times T; I, \Lambda; R]$, which extends uniquely to a homomorphism $\bar{\varphi} : A^+ \to M^0[S\times T; I, \Lambda; R]$. Then the compositions $\bar{\varphi}\pi_S$ and $\bar{\varphi}\pi_T$ are homomorphisms from A^+ to $M^0[S; I, \Lambda; P]$ and $M^0[T; I, \Lambda; Q]$ respectively. The entire set-up is summarised in the following diagram:



In the following lemma we relate the preimage of a non-zero element in $M^0[S\times T; I, \Lambda; R]$ to the preimages of non-zero elements in $M^0[S; I, \Lambda; P]$ and $M^0[T; I, \Lambda; Q]$.

Lemma 3.4. *For any $(i,(s,t),\lambda)$ in $M^0[S\times T; I, \Lambda; R]$ we have*

$$(i,(s,t),\lambda)\,\bar{\varphi}^{-1} = (i,s,\lambda)(\bar{\varphi}\pi_S)^{-1} \cap (i,t,\lambda)(\bar{\varphi}\pi_T)^{-1},$$
$$\mathbf{0}_{S\times T}\bar{\varphi}^{-1} = \mathbf{0}_S(\bar{\varphi}\pi_S)^{-1} \cap \mathbf{0}_T(\bar{\varphi}\pi_T)^{-1}.$$

Proof. First, suppose that $w \in (i,(s,t),\lambda)\,\bar{\varphi}^{-1}$; that is, $w\bar{\varphi} = (i,(s,t),\lambda)$. Then

$$w(\bar{\varphi}\pi_S) = (i,(s,t),\lambda)\,\pi_S = (i,s,\lambda)$$

Hence $w \in (i,s,\lambda)(\bar{\varphi}\pi_S)^{-1}$, and, analogously, $w \in (i,t,\lambda)(\bar{\varphi}\pi_T)^{-1}$.

Conversely, suppose that $w \in (i,s,\lambda)(\bar{\varphi}\pi_S)^{-1} \cap (i,t,\lambda)(\bar{\varphi}\pi_T)^{-1}$, so that $w(\bar{\varphi}\pi_S) = (i,s,\lambda)$ and $w(\bar{\varphi}\pi_T) = (i,t,\lambda)$. Note that $w\bar{\varphi} \neq \mathbf{0}_{S\times T}$, so we must have that $w\bar{\varphi} = (i_w,(s_w,t_w),\lambda_w)$ for some $i_w \in I$, $(s_w,t_w) \in S\times T$ and $\lambda_w \in \Lambda$. Now,

$$(i,s,\lambda) = (w\bar{\varphi})\pi_S = (i_w,(s_w,t_w),\lambda_w)\,\pi_S = (i_w,s_w,\lambda_w)$$

and, similarly, $(i,t,\lambda) = (i_w,t_w,\lambda_w)$. Hence, $i_w = i$, $s_w = s$, $t_w = t$ and $\lambda_w = \lambda$. Therefore, $w\bar{\varphi} = (i_w,(s_w,t_w),\lambda_w) = (i,(s,t),\lambda)$ and $w \in (i,(s,t),\lambda)\,\bar{\varphi}^{-1}$, as required.

The second equality is proved in essentially the same way. □

We can now prove the following:

Theorem 3.5. *Let S and T be finite semigroups. If languages recognised by finite Rees zero-matrix semigroups over S or T all have star-height $\leq h$, then all the languages recognised by the finite Rees zero-matrix semigroups over the direct product $S\times T$ also have star-height $\leq h$.*

Proof. Lemma 3.4 allows us to express the preimage of an element in the Rees zero-matrix semigroup over the direct product as the intersection of two preimages of elements in Rees zero-matrix semigroups over the factors. Since the preimage of any subset is a finite union of preimages of elements, the result follows. □

By combining the above results we can now extend Theorem 3.3 to Rees zero-matrix semigroups over abelian groups.

Theorem 3.6. *A regular language recognised by a Rees zero-matrix semigroup over an abelian group is of star-height at most one.*

Proof. Invoking the Fundamental Theorem of Finite Abelian Groups and applying Corollary 3.5 a finite number of times to Rees zero-matrix semigroups over cyclic groups yields the result. □

Theorem 3.6 can also be deduced from existing theoretical results when attention is restricted to the basic Rees matrix construction (without zero). Indeed, let S be a Rees matrix semigroup over an abelian group G. By [3, Proposition XI.3.1], S divides a wreath product of G by an aperiodic monoid. However, by [9, Theorem 7.8], every language recognised by $\mathbf{Gcom} * \mathbf{A}$ is of star-height at most one, where $\mathbf{Gcom} * \mathbf{A}$ is the pseudovariety generated by wreath products of commutative groups by aperiodic monoids. Hence, S belongs to the pseudovariety $\mathbf{Gcom} * \mathbf{A}$ and every language recognised by S is of star-height at most one.

Acknowledgement. The authors would like to thank the anonymous referee for their suggestions concerning alternative proof strategies for some of the results.

References

[1] L. C. Eggan. Transition graphs and the star-height of regular events. *Michigan Math. J.*, 10:385–397, 1963.
[2] S. Eilenberg. *Automata, Languages and Machines; Volume A*. Academic Press, 1974.
[3] S. Eilenberg. *Automata, Languages and Machines; Volume B*. Academic Press, 1976.
[4] K. Hashiguchi. Representation theorems on regular languages. *J. Comput. System Sci.*, 27:101–115, 1983.
[5] W. H. Henneman. *Algebraic theory of automata*. PhD thesis, MIT, 1971.
[6] J.M. Howie. *Fundamentals of semigroup theory*, volume 12 of *London Mathematical Society Monographs. New Series*. The Clarendon Press, Oxford University Press, New York, 1995. Oxford Science Publications.
[7] J.-E. Pin. *Varieties of Formal Languages*. North Oxford Academic, 1986.
[8] J.-E. Pin, H. Straubing, and D. Therien. New results on the generalized star-height problem. *STACS 89, Lecture Notes in Computer Science*, 349:458–467, 1989.
[9] J.-E. Pin, H. Straubing, and D. Thérien. Some results on the generalized star-height problem. *Inform. and Comput.*, 101(2):219–250, 1992.
[10] J. Sakarovitch. *Elements of automata theory*. Cambridge University Press, Cambridge, 2009.

School of Mathematics and Statistics, University of St Andrews, St Andrews, Scotland, U.K.
E-mail address: tom.bourne@st-andrews.ac.uk

School of Mathematics and Statistics, University of St Andrews, St Andrews, Scotland, U.K.
E-mail address: nik.ruskuc@st-andrews.ac.uk