Characterizing Regular Languages with Polynomial Densities *

Andrew Szilard, Sheng Yu, Kaizhong Zhang
Department of Computer Science, the University of Western Ontario
London, Ontario, CANADA N6A 5B7 (syu@csd.uwo.ca)

Jeffrey Shallit
Department of Computer Science, University of Waterloo
Waterloo, Ontario, CANADA N2L 3G1

Abstract

A language L is said to have a polynomial density if the function $p_L(n) = |L \cap \Sigma^n|$ of L is bounded by a polynomial. We show that the function $p_R(n)$ of a regular language R is $O(n^k)$, for some $k \geq 0$, if and only if R can be represented as a finite union of the regular expressions of the form $xy_1^* z_1 \ldots y_t^* z_t$ with a nonnegative integer $t \leq k+1$, where $x, y_1, z_1, \ldots, y_t, z_t$ are all strings in Σ^*.

We prove a characterization for the (restricted) starheight-one languages. We show that a regular language is starheight one if and only if it is the image of a regular language of polynomial density under a finite substitution. We also show that the set of starheight-one languages includes all the regular languages with polynomial densities and their complements.

Keywords: Regular languages, population functions of languages, density functions, polynomial densities, starheight-one languages.

1 Introduction

Given a regular language, it is often useful to know how many words of a certain length are in the language. In this paper, we study problems related to this question.

For each language $L \subseteq \Sigma^*$, we define a function $p_L(n) = |L \cap \Sigma^n|$, where $|S|$ denotes the cardinality of a set S. In other words, $p_L(n)$ counts the number of words of length n in L. The density function of a language L is usually defined as $d_L(n) = |L \cap \Sigma^n|/|\Sigma^n|$, and $p_L(n)$, as defined above, is often called the *population function* of L. However, $d_L(n)$ is not always a good description of the sparsity of L, particularly when $d_L(n) \to 0$. Hence, in this paper, we use the function $p_L(n)$ rather than the function $d_L(n)$ to describe the density of a language L. For example, we say that L has a constant density if $p_L(n)$ is $O(1)$, and L has a polynomial density if $p_L(n)$ is $O(n^k)$ for some nonnegative integer k. Languages of polynomial density have been called *sparse languages* by many authors [1]. However, the term *sparse* has also been used to mean that $\lim_{n\to\infty} \inf(d_L(n)) = 0$ for a language L [12].

We link the densities of regular languages to the forms of regular expressions representing them. One of the results of this paper is that a regular language is of

*This research is supported in part by the Natural Sciences and Engineering Research Council of Canada grants OGP0041630 and OGP0046373

polynomial density (of degree k) *if and only if* it can be represented as a finite union of the regular expressions of the form

$$xy_1^*z_1 \dots y_t^*z_t$$

(with each $t \leq k+1$), where $x, y_1, z_1, \dots, y_t, z_t$ are all strings in Σ^*. In particular, a regular language has a constant density if and only if it can be represented as a finite union of the regular expressions of the form xy^*z. This latter result has also appeared in [10] with a different proof. The above results can be derived from some results in, e.g., [2, ?, 11] in the frame of rational formal series. However, this paper study the structural properties of finite automata and derive the above results using directly those structural properties.

We show that, for any polynomial function $f(n)$, there exists a regular language R such that the density function $p_R(n)$ is $\Theta(f(n))$. This means that there is a gap between the density functions of order n^k and n^{k+1} for each nonnegative integer k, and between the polynomial functions and the exponential functions with a linear exponent. The density function $p_R(n)$ of a regular language R can only be of the order of a polynomial or 2^{cn} for some constant c. For example, there exist regular languages that have a density function of the order $\Theta(1)$, $\Theta(n^5)$, or 2^{7n}. But, there are no regular languages having a density function of the order $\Theta(n^{1.3})$, $\Theta(\log n)$, $\Theta(n \log n)$, or $2^{\Theta(\sqrt{n})}$.

By our characterization of regular languages with polynomial densities, it is easy to see that all regular languages with polynomial densities are of starheight one. By starheight, we mean the restricted starheight. See [3] for details. Note that not all starheight-one languages are of polynomial density. For example, $(a+b)^*$ is of density 2^n. However, as a consequence of our characterization theorem for the regular languages with polynomial densities, that a language is of starheight one if and only if it is the image, under a finite substitution [9], of a regular language of a polynomial density. We also show that the starheight-one languages include all the regular languages with polynomial densities and their complements.

At the end of the paper, we list several closure properties of the regular languages with polynomial densities. The results include that if a regular language L has a density function $p_L(n) = O(n^k)$ for some integer $k \geq 0$, then the language that is the set of all prefixes of words in L has a density function also in $O(n^k)$.

2 Basic definitions and notations

In this paper, we use Σ to denote the alphabet of a language. Σ^* is the free monoid (with the identity ϵ) generated by Σ. For a set S, $|S|$ denotes the cardinality of S and 2^S the power set, i.e. the set of all subsets, of S. For a string $x \in \Sigma^*$, $|x|$ denotes the length of x. Let $u, v \in \Sigma^*$. We say that u is a prefix of v if $uw = v$ for some $w \in \Sigma^*$, and u is a nontrivial prefix of v if u is a prefix of v and $u \neq \epsilon$ and $u \neq v$.

As mentioned above, for a language $L \subseteq \Sigma^*$, we call the function

$$p_L(n) = |L \cap \Sigma^n|$$

the density function of L. We say that a language L has a polynomial density if $p_L(n)$ is $O(n^k)$ for some integer $k \geq 0$.

A deterministic finite automaton (DFA) A is denoted by a 5-tuple $(Q, \Sigma, \delta, q_0, F)$ where Q is the finite set of states, Σ is the input alphabet, $\delta : Q \times \Sigma \rightarrow Q$ is a partial function, q_0 is the initial state, and F is the set of final states. A is said to be *reduced* if every state in Q is reachable from q_0 and every state reaches a final state. A is said to be *complete* if δ is a total function. Note that a reduced DFA is not necessarily a complete DFA. $L(A)$ denotes the language accepted by A.

Let $A = (Q, \Sigma, \delta, q_0, F)$ be a DFA. For each word $w = a_1 \ldots a_n$ in Σ^*, the *state transition sequence* of A on w, denoted $STS_A(w)$, is the sequence of states $q_{i_0}, q_{i_1}, \ldots, q_{i_n}$ where $q_{i_0} = q_0$ and $\delta(q_{i_k}, a_{k+1}) = q_{i_{k+1}}$ for all $0 \leq k < n$.

Given two functions $f(n)$ and $g(n)$, we say that $f(n)$ is $O(g(n))$ or $f(n) = O(g(n))$ if there exist positive constants c and n_0 such that $0 \leq f(n) \leq cg(n)$ for all $n \geq n_0$; $f(n)$ is $\Omega(g(n))$ or $f(n) = \Omega(g(n))$ if there is a constant $c > 0$ and an infinite sequence $n_1, n_2, \ldots, n_k, \ldots$ such that $f(n_i) \geq cg(n_i)$ for all $i \geq 1$; and $f(n)$ is $\Theta(g(n))$ or $f(n) = \Theta(g(n))$ if both $f(n) = O(g(n))$ and $f(n) = \Omega(g(n))$.

Let Σ and Δ be two alphabets. A finite substitution [9] is a mapping $\sigma : \Sigma^* \rightarrow 2^{\Delta^*}$ such that

(*i*) $\sigma(\epsilon) = \{\epsilon\}$,

(*ii*) $\sigma(xy) = \sigma(x)\sigma(y)$ for any $x, y \in \Sigma^*$, and

(*iii*) $\sigma(a)$ is finite for each $a \in \Sigma$.

The (restricted) starheight [3] h of a regular expression is defined inductively as

(*a*) $h(\emptyset) = 0$, $h(\epsilon) = 0$, and $h(a) = 0$ for each $a \in \Sigma$;

(*b*) $h(e_1 + e_2) = h(e_1 e_2) = max(h(e_1), h(e_2))$;

(*c*) $h(e^*) = h(e) + 1$.

The starheight of a regular language R is the least height of a regular expression denoting R.

3 Regular languages with polynomial densities

First, we characterize the regular languages of density $O(n^k)$, integer $k \geq 0$, with certain forms of their regular-expression presentations. For example, this characterization shows that a regular language is of density $O(n^2)$ if and only if it can be represented as a finite union of regular expressions of the form uv^*wx^*y where u, v, w, x, y are words over the alphabet. Then, as a new characterization for the starheight-one languages, we show that a language is starheight one if and only if it is the image, under a finite substitution, of a regular language with a polynomial density.

Definition 1 *Let $A = (Q, \Sigma, \delta, s, F)$ be a DFA. A word w in Σ^* is said to be* t-tiered, *$t \geq 0$, with respect to A if the state transition sequence of w is given by*

$$STS_A(w) = \alpha \beta_1^{d_1} \gamma_1 \ldots \beta_t^{d_t} \gamma_t$$

where

(1) $\alpha = p_1 \ldots p_l$, $0 \leq l \leq |Q|$, *where the p's are states in Q,*

and for each i, $1 \leq i \leq t$,

(2) $\beta_i = q_{i,0} \ldots q_{i,k_i}$ *and* $\gamma_i = q_{i,0} r_{i,1} \ldots r_{i,l_i}$, $0 \leq k_i$, $l_i \leq |Q|$, *where the q's and r's are in Q,*

(3) $q_{i,0}$ *appears only as the first state in* β_i *and in* γ_i,

(4) $d_i > 0$.

Lemma 1 *Let* L *be accepted by a DFA* $A = (Q, \Sigma, \delta, s, F)$. *If there exists a word* $w \in L$ *such that* w *is* k-tiered *with respect to* A, *then the density of* L *is* $\Omega(n^{k-1})$.

Proof. According to the structure of $STS_A(w)$, w can be decomposed into $w = xy_1^{d_1}z_1 \ldots y_k^{d_k}z_k$ such that

$$\delta^*(p_1, x) = q_{1,0},$$

$\delta^+(q_{i,0}, y_i) = q_{i,0}$, which goes through the state sequence β_i exactly once, and

$$\delta^+(q_{i,0}, z_i) = q_{i+1,0}, \text{ for } 1 \leq i < k,$$

$$\delta^+(q_{k,0}, z_k) = r_{k,l_k}.$$

Let $C = |y_1||y_2| \ldots |y_k|$ and $C_i = C/|y_i|$, $1 \leq i \leq k$. For an arbitrary integer $t > 0$, let $n = |xz_1 \ldots z_k| + tC$. It is clear that for any k arbitrary nonnegative integers $t_1, \ldots, t_k$ such that $t_1 + \ldots t_k = t$, the word $w = xy_1^{C_1t_1}z_1 \ldots y_k^{C_kt_k}z_k$ is in L and of length n. Let $N^k(t)$ denote the set of all such k-tuples of nonnegative integers, i.e.,

$$N^k(t) = \{(t_1, \ldots, t_k) \mid t_1, \ldots, t_k \geq 0 \;\&\; t_1 + \ldots + t_k = t\}$$

It can be shown (page 45-47 of [4]) that the cardinality of $N^k(t)$ is

$$\binom{t+k-1}{t} = \binom{t+k-1}{k-1} = \Omega(t^{k-1}) = \Omega(n^{k-1})$$

since k is a constant and n is a linear function of t.

Let $(s_1, \ldots, s_k)$ and $(t_1, \ldots, t_k)$ be two arbitrary elements in $N^k(t)$ such that $(s_1, \ldots, s_k) \neq (t_1, \ldots, t_k)$. Consider the two words $u = xy_1^{C_1s_1}z_1 \ldots y_k^{C_ks_k}z_k$ and $v = xy_1^{C_1t_1}z_1 \ldots y_k^{C_kt_k}z_k$. Note that both u and v are in L. Now, we show that $u \neq v$. Let i be the smallest number such that $s_i \neq t_i$. Without any loss of generality, we assume that $s_i < t_i$. Then u and v have a common prefix $u_0 = xy_1^{C_1s_1}z_1 \ldots y_i^{C_is_i}$, i.e., $STS_A(u)$ and $STS_A(v)$ have a common prefix $STS_A(u_0)$. It is clear that $STS_A(u)$ does not contain the state $q_{i,0}$ after the prefix $STS_A(u_0)$ according to (3) of Definition 1. But $STS_A(v)$ contains at least one further appearance of $q_{i,0}$. Thus, $u \neq v$ since A is deterministic.

Therefore the number of distinct words of length n in L is $\Omega(n^{k-1})$. □

Lemma 2 *Let* L *be an arbitrary regular language with* $L = L(A)$, *for some DFA* A. *If* $p_L(n)$ *is* $O(n^k)$ *for some integer* $k \geq 0$, *then each word* $w \in L$ *is* t-tiered *with respect to* A *for some nonnegative integer* $t \leq k+1$.

Proof. Let $A = (Q, \Sigma, \delta, s, F)$. We prove this lemma by induction on the length of a word w in L.

For any word $|w| < |Q|$, the lemma holds trivially. We hypothesize that the lemma holds for all words w in L with $|w| < n$, for some $n \geq |Q|$.

Consider an arbitrary word $w \in L$ with $|w| = n$. We look at $STS_A(w)$ in reverse and find the first state that repeats. Let this state be q. Denote by $q^{(i)}$ the ith (from the left to right) appearance of q in $STS_A(w)$. Let $q^{(1)}, \ldots, q^{(h+1)}$ be the sequence of all appearances of q in $STS(w)$. Clearly, $h \geq 1$ and there are fewer than $|Q|$ states between $q^{(h)}$ and $q^{(h+1)}$. We write $w = w_0 v_1 v_2 \ldots v_h w_1$ such that $\delta^+(q^{(i)}, v_i) = q^{(i+1)}$, for $1 \leq i \leq h$. Then, clearly, $w_1 v_1^{n_1} \ldots v_h^{n_h} w_1$ is in L for any nonnegative integers $n_1, \ldots, n_h$.

We now prove that $v_1 = \ldots = v_h$. *Assume* that $v_i \neq v_j$ for some $i \neq j$. Then v_i is not a prefix of v_j because if it were so, there would be another appearance of q between $q^{(j)}$ and $q^{(j+1)}$ and this is impossible. Similarly, v_j is not a prefix of v_i. The monoid generated by $\{v_i v_j, v_j v_i\}$ is isomorphic to the free monoid on two generators. For any nonnegative integer g, the regular expression $(v_i v_j + v_j v_i)^g$ represents 2^g distinct words, each is of length $|v_i v_j|g$. Then all the 2^g words represented by $w_0(v_i v_j + v_j v_i)^g w_1$ have the same length, namely $|w_0 w_1| + |v_i v_j|g$, and are all in L. This contradicts the fact that L has a polynomial density. Therefore, $v_1 = \ldots = v_h$ and $STS_A(v_1) = \ldots = STS_A(v_h)$.

Since $w' = w_0 w_1$ is in L and $|w'| < n$, then, by our induction hypothesis, w' is t-tiered with respect to A for some $t \leq k+1$, i.e., $STS_A(w') = \alpha\beta_1^{d_1}\gamma_1 \ldots \beta_t^{d_t}\gamma_t$ satisfies the conditions (1) to (4) of Definition 1. Note that q must appear only in γ_t and no $q_{i,0}$ may appear in $STS_A(w_1)$ due to a similar argument as above. So, all $q_{i,0}$, $1 \leq i \leq t$, appear only before the first appearance of q. If $t < k+1$, then the proof is complete. Assume that $t = k+1$. Then w is $(k+2)$-tiered. By Lemma 1, the density of L is in $\Omega(n^{k+1})$. This is a contradiction and, thus, $t = k+1$ is impossible. □

Lemma 3 *Let L be accepted by a DFA $A = (Q, \Sigma, \delta, s, F)$. If there is an integer $k \geq 0$ such that each word w in L is t-*tiered *with respect to A, for some $t \leq k$, then L can be represented as a finite union of the regular expressions of the following form*

$$xy_1^* z_1 \ldots y_t^* z_t$$

with $0 \leq t \leq k$, where $x, y_1, \ldots, y_t, z_1, \ldots, z_t \in \Sigma^*$, $|x|, |z_1|, \ldots, |z_t| < |Q|$, and $|y_1|, \ldots, |y_t| \leq |Q|$.

Proof. Let w be an arbitrary word in L. Since w is t-tiered with respect to A for some $t \leq k$, according to the structure of $STS_A(w)$, w can be decomposed into $w = xy_1^{n_1} z_1 \ldots y_t^{n_t} z_t$ such that $|x|, |z_1|, \ldots, |z_t| < |Q|$ and $0 < |y_1|, \ldots, |y_t| \leq |Q|$, and

$$w \in xy_1^* z_1 \ldots y_t^* z_t \subseteq L \ .$$

For each $w \in L$, we denote this regular expression derived from w as $E_d(w)$. Then

$$L = \bigcup_{w \in L} E_d(w) \ .$$

Since Q is finite, there are a finite number of choices for x's, y's, and z's of length less than or equal to $|Q|$. Also because each t is bounded by the constant k, the number of distinct expressions on the righthand side of the equation is bounded by $|\Sigma|^{2k|Q|}$. Therefore, the lemma holds. □

Lemma 4 *Let E be a regular expression*

$$E = xy_1^* z_1 \ldots y_t^* z_t$$

where $t > 0$ and $x, y_1, z_1, \ldots, y_t, z_t$ are all strings in Σ^*, and let $L = L(E)$. Then the function $p_L(n)$ is $O(n^{t-1})$.

Proof. Note that if this lemma holds for all y's being nonempty, then it also holds for some y's being empty. So, we assume that all y's are nonempty in the following proof. Let n be an arbitrary nonnegative integer. Consider all the possible words of length n in L. Let w be such a word. Then $w = xy_1^{n_1}z_1 \dots y_t^{n_t}z_t$ for some nonnegative integers $n_1, \dots, n_t$ such that $n_1|y_1| + \dots + n_t|y_t| + |xz_1 \dots z_t| = n$. Let C denote $|xz_1 \dots z_t|$ and $m = n - C$, i.e., $m = n_1|y_1| + \dots + n_t|y_t|$. We know that the cardinality of the set

$$N^t(m) = \{(m_1, \dots, m_t) \mid 0 \le m_i, \text{ for } 1 \le i \le t, \text{ and } m_1 + \dots m_t = m\}$$

is

$$\binom{m+t-1}{m} = \binom{m+t-1}{t-1} = \binom{n-C+t-1}{t-1} = O(n^{t-1})$$

since C and t are constants. Define a mapping $\tau(n_1, \dots, n_t) = (n_1|y_1|, \dots, n_t|y_t|)$ over the set of t-tuples of nonnegative integers. Then for each t-tuple $(n_1, \dots, n_t)$ such that $xy_1^{n_1}z_1 \dots y_t^{n_t}z_t$ is of length n, $\tau(n_1, \dots, n_t)$ is in $N^t(m)$. Since τ is injective, the number of such distinct t-tuples is no more than the cardinality of $N^t(m)$, i.e., $O(n^{t-1})$. □

By the previous three lemmas, we obtain the subsequent three theorems.

Theorem 1 *Let R be a regular language and $A = (Q, \Sigma, \delta, s, F)$ be a DFA accepting R. Then the function $p_R(n)$ is $O(n^k)$, $k \ge 0$, if and only if each word of R is* t-tiered *with respect to A for some $t \le k + 1$.*

Theorem 2 *Let R be a regular language and $A = (Q, \Sigma, \delta, s, F)$ be an arbitrary DFA accepting R. Then the function $p_R(n)$ is $O(n^k)$, $k \ge 0$, if and only if R can be represented as a finite union of regular expressions of the following form:*

$$xy_1^*z_1 \dots y_t^*z_t, \text{ where } x, y_1, z_1, \dots, y_t, z_t \in \Sigma^*,$$

with $|x|, |y_1|, |z_1|, \dots, |y_t|, |z_t| \le |Q|$ and $0 \le t \le k + 1$.

A weaker form of the previous theorem is stated below.

Theorem 3 *A regular language R over Σ has a density in $O(n^k)$, $k \ge 0$, if and only if R can be represented as a finite union of regular expressions of the following form:*

$$xy_1^*z_1 \dots y_t^*z_t \text{ where } x, y_1, z_1, \dots, y_t, z_t \in \Sigma^*,$$

with $0 \le t \le k + 1$.

As a special case of the above theorem, we state a corollary for $k = 0$, which was first stated by Shallit [10] using a different method, and also proved by Yu and Fisch independently.

Corollary 1 *Let R be a regular language. The number of words of the same length in R is bounded by a constant C if and only if R can be represented as a finite union of the regular expressions of the form*

$$xy^*z \text{ where } x, y, z \in \Sigma^*.$$

In particular, if a language can be expressed by the union of C expressions in this form, then the number of words of the same length in the language is bounded by the constant C.

In what follows we assume that a regular language is given by a finite automaton or a regular expression.

Lemma 5 *Let R be a regular language accepted by a DFA of C states. If R has a polynomial density, then the function $p_R(n)$ is $O(n^{C-1})$.*

Proof. Notice that in Definition 1, each $q_{i,0}$, $1 \leq i \leq t$, is distinct. Since there are C states in A, each word in R is at most C-tiered with respect to A. Thus, $p_R(n)$ is $O(n^{C-1})$ by Theorem 1. □

With the previous lemma, we can easily prove the following theorem:

It is clear that all regular languages with polynomial densities are of starheight one. But, not all starheight-one languages are of polynomial density. For example, the language $(ab+b)^*(a+\epsilon)$ is of exponential density. However, there is a relation between these two subclasses of regular languages, which is stated in the following theorem.

Theorem 4 *A regular language is of starheight one if and only if it is the image of a regular language of a polynomial density under a finite substitution.*

Proof. The *if* part is obvious. For the *Only if* part, let E be the regular expression of starheight one. We define a finite substitution π such that for each $a \in \Sigma$, $a \rightarrow \{a\}$ and for each e^* in E, there is a distinct $\sigma \notin \Sigma$ such that $\sigma \rightarrow L(e)$. Let E' be the regular expression such that the image of E' under π is E. Then, clearly, $L(E')$ is a language of a polynomial density according to Theorem 3. □

4 Gap theorems on the densities of regular languages

In this section, we show that only the functions that are of the order of polynomial functions and exponential functions with linear exponents can be the densities functions of regular sets. This means that, for the densities of regular languages, there is a gap between $\Theta(n^k)$ and $\Theta(n^{k+1})$, for each integer $k \geq 0$, and between polynomial functions and exponential functions of $2^{\Theta(n)}$. For example, there is no regular language that has a density of the order $\sqrt{n}$, $n \log n$, or $2^{\sqrt{n}}$. We also show that for any integer $k \geq 0$, there exists a regular language R such that $p_R(n)$ is exactly n^k.

Theorem 5 *For any integer $k \geq 0$, there does not exist a regular language R such that the function $p_R(n)$ is neither $O(n^k)$ nor $\Omega(n^{k+1})$.*

Proof. Assume that there exists a regular language R such that $p_R(n)$ is neither $O(n^k)$ nor $\Omega(n^{k+1})$. Then $p_R(n)$ is $O(n^{k+1})$. Let $R = L(A)$ for some DFA A. By Lemma 2, for each word w in R, w is t-tiered with respect to A, for some nonnegative integer $t \leq k+2$. If there exists a word w in R such that w is $k+2$-tiered with respect to A, then, by Lemma 1, $p_R(n)$ is $\Omega(n^{k+1})$. It is a contradiction. Otherwise, $p_R(n)$ is $O(n^k)$ by Theorem 1. Again a contradiction. □

Lemma 6 *Let R be accepted by some reduced DFA $A = (Q, \Sigma, \delta, q_0, F)$. Then each word in R is t-tiered with respect to A for some nonnegative integer $t \leq |Q|$ if and only if there does not exist a state q of A such that* *(i)* $\delta^+(q,x) = q$ *and* $\delta^+(q,y) = q$,

(ii) $x \neq y$, *and*

(iii) there is no nontrivial prefix x' *of* x *or* y' *of* y *such that* $\delta^+(q,x') = q$ *or* $\delta^+(q,y') = q$.

Proof. Assume that there is no state of A that satisfies (i), (ii), and (iii) above. Then every state p of A satisfies the condition that there is at most one state-sequence that takes A from p to p (without passing through p). Since the number of states $k = |Q|$ is finite, it is easy to see that, for each $w \in R$, w is t-tiered with respect to A for some nonnegative t that is at most k.

Assume that each word in R is t-tiered with respect to A for some nonnegative t bounded by a constant k. Then $p_R(n)$ is $O(n^{k-1})$ by Theorem 1. Suppose that there exists a state p of A that satisfies (i), (ii), and (iii) above. Let $u, v \in \Sigma^*$ such that $\delta^*(q_0, u) = p$ and $\delta^*(p, v) \in F$. Let $R' = u(xy + yx)^*v$. Obviously, $R' \subseteq R$ and R' has a density of $2^{\Omega(n)}$. A contradiction. □

Theorem 6 *There does not exist a regular language* R *such that* $p_R(n)$ *is neither* $O(n^k)$, *for some integer* $k \geq 0$, *nor* $2^{\Omega(n)}$.

Proof. Suppose R is a regular language and $p_R(n)$ is neither $O(n^k)$, for some integer $k \geq 0$, nor $2^{\Omega(n)}$. Let A be a DFA that accepts R. Then either there exists a state q in A that satisfies (i), (ii), and (iii) of Lemma 6, or there does not. If there does not, then $p_R(n)$ is $O(n^k)$, for some $k \geq 0$, by Lemma 6 and Theorem 1. A contradiction. If there does, we can show that $p_R(n)$ is $2^{\Omega(n)}$. Again a contradiction. Therefore, such a language R does not exist. □

Thus, for example, there is no regular language that has a density of the order $2^{\sqrt[k]{n}}$ for any $k > 1$. The following corollary holds trivially.

Corollary 2 *If a regular language* R *is of starheight* $k \geq 2$, *then* $p_R(n)$ *is* $2^{\Omega(n)}$.

For each polynomial function $f(n)$, there exist a regular language R such that $p_R(n)$ is $\Theta(f(n))$, and for each regular language R, either there exists a polynomial function $f(n)$ such that $p_R(n) = \Theta(f(n))$, or $p_R(n) = 2^{\Theta(n)}$. It is not difficult to see that, for each nonnegative integer k, we can construct a regular language R such that $p_R(n)$ is exactly n^k.

5 Closure properties of regular languages with polynomial densities

In this section, we list several closure properties of the regular languages with polynomial densities. Since most of the properties can be derived directly from the results in Section 3, we omit the proofs.

Theorem 7 *Let* L_1 *and* L_2 *be regular languages over* Σ *with* $p_{L_1}(n) = \Theta(n^k)$ *and* $p_{L_2}(n) = \Theta(n^l)$. *Then the following statements hold:*

(a) If $L = prefix(L_1) = \{x \mid xy \in L_1 \text{ for some } y \in \Sigma^*\}$, *then* $p_L(n) = \Theta(n^k)$.

(b) If $L = infix(L_1) = \{y \mid xyz \in L_1 \text{ for some } x, y \in \Sigma^*\}$, *then* $p_L(n) = \Theta(n^k)$.

(c) *If* $L = suffix(L_1) = \{z \mid xz \in L_1 \text{ for some } x \in \Sigma^*\}$, *then* $p_L(n) = \Theta(n^k)$.

(d) *If* $L = L_1 \cup L_2$, *then* $p_L(n) = \Theta(n^{\max(k,l)})$.

(e) *If* $L = L_1 \cap L_2$, *then* $p_L(n) = O(n^{\min(k,l)})$.

(f) *If* $L = L_1 L_2$, *then* $p_L(n) = O(n^{k+l})$.

(g) *If* $L = h(L_1)$ *where* h *is an arbitrary morphism[5], then* $p_L(n) = O(n^k)$.

(h) *If* $L = \frac{1}{m}(L_1) = \{x \mid xy_1 \cdots y_{m-1} \in L_1, \text{ for } x, y_1, \ldots, y_{m-1} \in \Sigma^*, \text{ and } |x| = |y_1| = \ldots = |y_{m-1}|\}$, *then* $p_L(n) = \Theta(n^k)$.

It is clear that the set of regular languages with polynomial densities is not closed under complementation. However, we have the following result.

Theorem 8 *The complement of a regular language with a polynomial density is a starheight-one language.*

The idea of a proof is the following. Let A be a DFA accepting a regular language of a polynomial density and there is an integer $k \geq 0$ such that each word in $L(A)$ is t-tiered for some $t \leq k$. We assume that A is reduced. To make A complete, we add a new state, say, q, the so called *sink state*, such that all undefined transitions are directed to this state and $\delta(qa) = q$ for each $a \in \Sigma$. Then the complement of $L(A)$ is accepted by a DFA $\overline{A}$ which is exactly the same as the completed A except that the final states and non-final states are exchanged. Let B be another DFA which is the same as $\overline{A}$ except that the only transition for q is $\delta(q, \#) = q$ for some $\# \notin \Sigma$. It is not difficult to show that each word in $L(B)$ is t-tiered for some $t \leq k + 1$. Define a finite substitution σ such that $\sigma(a) = \{a\}$ for each $a \in \Sigma$ and $\sigma(\#) = \Sigma$. Then $\sigma(L(B)) = L(\overline{A})$. By Theorem 4, $L(\overline{A})$ is of starheight one.

Thus, the set of starheight-one languages includes all the regular languages with polynomial densities and their complements.

References

[1] L. Berman and J. Hartmanis, "On Isomorphisms and Density of NP and Other Complete Sets", *SIAM J. of Comput.* 6 (1977) 305-322.

[2] J. Berstel and C. Reutenauer, *Rational Series and Their Languages*, EATCS Monographs on Theoretical Computer Science, edited by Brauer, Rozenberg, and Salomaa, Springer-Verlag, 1988.

[3] J. Brzozowski, "Open Problems about Regular Languages", *Formal Language Theorem — Perspectives and Open Problems*, pp. 23-48, edited by R. V. Book, 1980.

[4] D. I. A. Cohen, *Basic Techniques of Combinatorial Theory*, John Wiley & Sons, New York, 1978.

[5] K. Culik II, F.E. Fich and A. Salomaa, "A Homomorphic Characterization of Regular Languages", *Discrete Applied Mathematics* 4, (1982)149-152.

[6] S. Eilenberg, *Automata, Languages and Machines*, vol. A, Academic Press, 1974.

[7] J.E. Hopcroft and J.D. Ullman, *Introduction to Automata Theory, Languages, and Computation*, Addison Wesley (1979), Reading, Mass.

[8] W. Kuich and A. Salomaa, *Semirings, Automata, Languages*, Springer-Verlag, 1986.

[9] B. Ravikumar and O. H. Ibarra, "Relating the type of ambiguity of finite automata to the succintness of their representation", *SIAM J. Comput.* vol. 18, no. 6 (1989) 1263-1282.

[10] A. Salomaa, *Computation and Automata*, Encyclopedia of Mathematics and Its Applications, vol. 25. Cambridge University Press, 1985.

[11] A. Salomaa and M. Soittola, *Automata-Theoretic Aspects of Formal Power Series*, Springer-Verlag, 1978.

[12] J. Shallit, "Numeration Systems, Linear Recurrences, and Regular Sets", *Research Report* CS-91-32, Dept. of Computer Science, Univ. of Waterloo, 1991.

[13] M. P. Schützenberger, Finite Counting Automata, *Information and Control*, 5 (1962) 91-107.

[14] S. Yu, "Can the Catenation of Two Weakly Sparse Languages be Dense?", *Discrete Applied Mathematics* 20 (1988) 265-267.