Complexity of Suffix-Free Regular Languages*

Janusz Brzozowski[1] and Marek Szykuła[2]

[1] David R. Cheriton School of Computer Science, University of Waterloo,
Waterloo, ON, Canada N2L 3G1
{brzozo@uwaterloo.ca}
[2] Institute of Computer Science, University of Wrocław,
Joliot-Curie 15, PL-50-383 Wrocław, Poland
{msz@cs.uni.wroc.pl}

Abstract. A sequence $(L_k, L_{k+1} \dots)$ of regular languages in some class $\mathcal{C}$, where n is the state complexity of L_n, is called a *stream*. A stream is *most complex* in class $\mathcal{C}$ if its languages together with their dialects (that is, languages that differ only very slightly from the languages in the stream) meet the state complexity bounds for boolean operations, product (concatenation), star, and reversal, have the largest syntactic semigroups, and have the maximal numbers of atoms, each of which has maximal state complexity. It is known that there exist such most complex streams in the class of regular languages, and also in the classes of right, left, and two-sided ideals. In contrast to this, we prove that there does not exist a most complex stream in the class of suffix-free regular languages. However, we do exhibit one ternary suffix-free stream that meets the bound for product and whose restrictions to binary alphabets meet the bounds for star and boolean operations. We also exhibit a quinary stream that meets the bounds for boolean operations, reversal, size of syntactic semigroup, and atom complexities. Moreover, we solve an open problem about the bound for the product of two languages of state complexities m and n in the binary case by showing that it can be met for infinitely many m and n.

Two transition semigroups play an important role for suffix-free languages: semigroup $\mathbf{T}^{\leqslant 5}(n)$ is the largest suffix-free semigroup for $n \leqslant 5$, while semigroup $\mathbf{T}^{\geqslant 6}(n)$ is largest for $n = 2, 3$ and $n \geqslant 6$. We prove that all witnesses meeting the bounds for the star and the second witness in a product must have transition semigroups in $\mathbf{T}^{\leqslant 5}(n)$. On the other hand, witnesses meeting the bounds for reversal, size of syntactic semigroup and the complexity of atoms must have semigroups in $\mathbf{T}^{\geqslant 6}(n)$.

Keywords: most complex, regular language, state complexity, suffix-free, syntactic complexity, transition semigroup

* This work was supported by the Natural Sciences and Engineering Research Council of Canada grant No. OGP000087, and by Polish NCN grant DEC-2013/09/N/ST6/01194.

1 Introduction

Suffix-Free Languages A language is *suffix-free* if no word in the language is a suffix of another word in the language. The languages ba^*, $\{a^nb^n \mid n \geqslant 1\}$, and Σ^n, where Σ is a finite alphabet and n is a positive integer, are all examples of suffix-free languages. Every suffix-free language (except that consisting of the empty word ε) is a suffix code. Suffix codes are an important subclass of general codes, which have numerous applications in cryptography, data compression and error correction. Codes have been studied extensively; see [2] for example. In addition to being codes, suffix-free languages are also a special subclass of suffix-convex languages [1], where a language is *suffix-convex* if, whenever a word w and its suffix u are in the language, then so is every suffix of w that has u as a suffix.

We study complexity properties of suffix-free regular languages.

Quotient Complexity A basic complexity measure of a regular language L over an alphabet Σ is the number n of distinct left quotients of L, where a *(left) quotient* of L by a word $w \in \Sigma^*$ is $w^{-1}L = \{x \mid wx \in L\}$. We denote the set of quotients of L by $K = \{K_0, \ldots, K_{n-1}\}$, where $K_0 = L = \varepsilon^{-1}L$ by convention. Each quotient K_i can be represented also as $w_i^{-1}L$, where $w_i \in \Sigma^*$ is such that $w_i^{-1}L = K_i$. The number of quotients of L is its *quotient complexity* [3] $\kappa(L)$. A concept equivalent to quotient complexity is *state complexity* [23] of L, which is the number of states in a minimal deterministic finite automaton (DFA) recognizing L.

Let L_n be a regular language of quotient complexity n. The *quotient complexity of a unary operation* $\circ$ on L_n is the maximum value of $\kappa(L_n^\circ)$ as a function of n. To establish the quotient complexity of L_n°, first we need to find an upper bound on this complexity. For example, 2^n is an upper bound on the reversal operation on regular languages [20, 22]. Second, we need a sequence $(L_n, n \geqslant k) = (L_k, L_{k+1}, \ldots)$, called a *stream*, of languages that meet this bound; here k is usually some small integer because the bound may not apply for $n < k$. A language L_n that meets the bound $\kappa(L_n^\circ)$ for the operation $\circ$ is a *witness (language)* for that operation. A stream in which every language meets the bound is called a *witness (stream)*. The languages in a stream are normally defined in the same way, differing only in the parameter n. For example, we might have the stream $(L_n = \{w \in \{a,b\}^* \mid$ the number of a's is 0 modulo $n\}$, $n \geqslant 2)$.

Similarly, $\kappa(K_m \circ L_n)$ is the *quotient complexity of a binary operation* $\circ$ on regular languages K_m and L_n of complexities m and n, respectively. Again, we need an upper bound on $\kappa(K_m \circ L_n)$. For example, an upper bound on product (concatenation) is $(m-1)2^n + 2^{n-1}$ [19, 22]. Then we have to find two streams $(K_m, m \geqslant h)$ and $(L_n, n \geqslant k)$ of languages meeting this bound. In general, the two streams are different, but there are many examples where K_n "differs only slightly" from L_n; such a language K_n has been called a *dialect* [4] of L_n. The notion "differs only slightly" will be made precise below. A pair (K_m, L_n) of languages that meets the bound $\kappa(K_m \circ L_n)$ for the operation $\circ$ is a *witness (pair)* for that operation. A stream in which every pair of languages meets the bound is called a *witness (stream)*.

The quotient/state complexity of an operation gives a worst-case lower bound on the time and space complexity of the operation. For this reason it has been studied extensively; see [3, 23] for additional references. The quotient complexity of suffix-free languages was examined in [14, 15, 17].

We also extend the notions of maximal complexity, stream, and witness to DFAs.

Syntactic Complexity A second measure of complexity of a regular language is its syntactic complexity. Let Σ^+ be the set of non-empty words of Σ^*. The *syntactic semigroup* of L is the set of equivalence classes of the Myhill congruence $\approx_L$ on Σ^+ defined by $x \approx_L y$ if and only if $uxv \in L \Leftrightarrow uyv \in L$ for all $u, v \in \Sigma^*$. The syntactic semigroup of L is isomorphic to the *transition semigroup* of a minimal DFA $\mathcal{D}$ recognizing L [21], which is the semigroup of transformations of the state set of $\mathcal{D}$ induced by non-empty words. The *syntactic complexity* of L is the cardinality of its syntactic/transition semigroup. It was pointed out in [13] that languages having the same quotient complexity can have vastly different syntactic complexities. Thus syntactic complexity can be a finer measure of complexity. Syntactic complexity of suffix-free languages was studied in [9, 11].

Complexities of Atoms A possible third measure of complexity of a regular language L is the number and quotient complexities, which we call simply complexities, of certain languages, called atoms, uniquely defined by L. Atoms arise from an equivalence on Σ^* which is a left congruence refined by the Myhill congruence, where two words x and y are equivalent if $ux \in L$ if and only if $uy \in L$ for all $u \in \Sigma^*$ [16]. Thus x and y are equivalent if $x \in u^{-1}L \Leftrightarrow y \in u^{-1}L$. An equivalence class of this relation is called an *atom* [12] of L. It follows that an atom is a non-empty intersection of complemented and uncomplemented quotients of L. The quotients of a language are unions of its atoms.

Most Complex (Streams of) Languages The concept of *most complex* languages in a class C of languages was introduced in [4]. Such languages, with some of their dialects, must meet the bounds in the class C on the quotient complexities of the unary operations reversal and (Kleene) star, and on the product and the binary boolean operations. Moreover, they must also have the largest possible syntactic semigroups and the most complex atoms. It is surprising that there exists a single stream of languages that meets all these conditions for maximal complexity [4]. Moreover, most complex right, left, and two-sided ideals also exist [7]. We show in this paper, however, that this is not the case for suffix-free languages.

Most complex languages are useful for testing the efficiency of systems. The complexity of operations on languages gives a measure of time and space requirements for these operations. Hence to check the maximal size of the objects that a system can handle, we can use most complex languages. It is certainly simpler to have just one or two universal worst-case examples.

Terminology and Notation A *deterministic finite automaton (DFA)* is a quintuple $\mathcal{D} = (Q, \Sigma, \delta, q_0, F)$, where Q is a finite non-empty set of *states*, Σ is a finite non-empty *alphabet*, $\delta\colon Q \times \Sigma \to Q$ is the *transition function*, $q_0 \in Q$ is the *initial* state, and $F \subseteq Q$ is the set of *final* states. We extend δ to a function

$\delta\colon Q \times \Sigma^* \to Q$ as usual. A DFA $\mathcal{D}$ *accepts* a word $w \in \Sigma^*$ if $\delta(q_0, w) \in F$. The language accepted by $\mathcal{D}$ is denoted by $L(\mathcal{D})$. If q is a state of $\mathcal{D}$, then the language L^q of q is the language accepted by the DFA $(Q, \Sigma, \delta, q, F)$. A state is *empty* if its language is empty. Two states p and q of $\mathcal{D}$ are *equivalent* if $L^p = L^q$. A state q is *reachable* if there exists $w \in \Sigma^*$ such that $\delta(q_0, w) = q$. A DFA is *minimal* if all of its states are reachable and no two states are equivalent. Usually DFAs are used to establish upper bounds on the quotient complexity of operations and also as witnesses that meet these bounds.

A *nondeterministic finite automaton (NFA)* is a quintuple $\mathcal{D} = (Q, \Sigma, \delta, I, F)$, where Q, Σ and F are defined as in a DFA, $\delta\colon Q \times \Sigma \to 2^Q$ is the *transition function*, and $I \subseteq Q$ is the *set of initial states*. An *ε-NFA* is an NFA in which transitions under the empty word ε are also permitted.

The *quotient DFA* of a regular language L with n quotients is defined by $\mathcal{D} = (K, \Sigma, \delta_{\mathcal{D}}, K_0, F_{\mathcal{D}})$, where $\delta_{\mathcal{D}}(K_i, w) = K_j$ if and only if $w^{-1}K_i = K_j$, and $F_{\mathcal{D}} = \{K_i \mid \varepsilon \in K_i\}$. To simplify the notation, without loss of generality we use the set $Q_n = \{0, \ldots, n-1\}$ of subscripts of quotients as the set of states of $\mathcal{D}$; then $\mathcal{D}$ is denoted by $\mathcal{D} = (Q_n, \Sigma, \delta, 0, F)$, where $\delta(p, w) = q$ if $\delta_{\mathcal{D}}(K_p, w) = K_q$, and F is the set of subscripts of quotients in $F_{\mathcal{D}}$. The quotient DFA of L is unique and it is isomorphic to each complete minimal DFA of L.

A *transformation* of Q_n is a mapping $t\colon Q_n \to Q_n$. The *image* of $q \in Q_n$ under t is denoted by qt. The *range* of t is $\mathrm{rng}(t) = \{q \in Q_n \mid pt = q \text{ for some } p \in Q_n\}$. In any DFA, each letter $a \in \Sigma$ induces a transformation δ_a of the set Q_n defined by $q\delta_a = \delta(q, a)$; we denote this by $a\colon \delta_a$. By a slight abuse of notation we use the letter a to denote the transformation it induces; thus we write qa instead of $q\delta_a$. We also extend the notation to sets of states: if $P \subseteq Q_n$, then $Pa = \{pa \mid p \in P\}$. If s, t are transformations of Q, their composition is denoted by $s \circ t$ and defined by $q(s \circ t) = (qs)t$; the $\circ$ is usually omitted. Let $\mathcal{T}_{Q_n}$ be the set of all n^n transformations of Q_n; then $\mathcal{T}_{Q_n}$ is a monoid under composition.

For $k \geqslant 2$, a transformation (permutation) t of a set $P = \{q_0, q_1, \ldots, q_{k-1}\} \subseteq Q$ is a *k-cycle* if $q_0t = q_1, q_1t = q_2, \ldots, q_{k-2}t = q_{k-1}, q_{k-1}t = q_0$. A k-cycle is denoted by $(q_0, q_1, \ldots, q_{k-1})$. A 2-cycle (q_0, q_1) is called a *transposition*. A transformation that changes only one state p to a state $q \neq p$ is denoted by $(p \to q)$. A transformation mapping a subset P of Q to a single state and acting as the identity on $Q \setminus P$ is denoted by $(P \to q)$. We also denote by $[q_0, \ldots, q_{n-1}]$ the transformation that maps $p \in \{0, \ldots, n-1\}$ to q_p.

In this paper we consider only three types of simple dialects of DFAs and languages. Let $\Sigma = \{a_1, \ldots, a_k\}$, and let π be a permutation of Σ:

1. DFA $\mathcal{D}_n(a_1, \ldots, a_k)$ is *permutationally equivalent* to DFA $\mathcal{D}'_n(a_1, \ldots, a_k)$ if $\mathcal{D}'_n(a_1, \ldots, a_k) = \mathcal{D}_n(\pi(a_1), \ldots, \pi(a_k))$. In other words, the transformation induced by a_i in $\mathcal{D}'_n$ is the transformation induced by $\pi(a_i)$ in $\mathcal{D}_n$.

2. Let $\Gamma \subseteq \Sigma$. DFA $\mathcal{D}'_n$ is the *restriction* of a DFA $\mathcal{D}_n$ to Γ if all the transitions induced by letters in $\Sigma \setminus \Gamma$ are deleted from $\mathcal{D}_n$. For example, $\mathcal{D}_n(-, b, -, d)$ is the DFA $\mathcal{D}_n(a, b, c, d)$ restricted to the alphabet $\{b, d\}$. However, if the dashes appear only at the end, they may be omitted.

3. DFA $\mathcal{D}'_n$ is the *permutational restriction* of a DFA $\mathcal{D}_n$ to Γ if the letters of Σ are first permuted and then the letters in $\Sigma \setminus \Gamma$ are deleted.

The same notational conventions are used for languages.

Contributions

1. We prove that a most complex stream of suffix-free languages does not exist. This is in contrast with the existence of streams of most complex regular languages [4], right ideals [5, 6], and left and two-sided ideals [6, 7].

2. We exhibit a single ternary witness that meets the bounds for star, product, and boolean operations.

3. We exhibit a single quinary witness that meets the bounds for boolean operations, reversal, number of atoms, syntactic complexity, and quotient complexities of atoms.

4. We show that when $m, n \geqslant 6$ and $m-2$ and $n-2$ are relatively prime, there are binary witnesses that meet the bound $(m-1)2^{n-2}+1$ for product.

5. We prove that any witness DFA for star and any second witness DFA for product must have transition semigroups that are subsemigroups of the suffix-free semigroup of transformations $\mathbf{T}^{\leqslant 5}(n)$ which is largest for $n \leqslant 5$; that the witness DFAs for reversal, syntactic complexity and quotient complexities of atoms must have transition semigroups that are subsemigroups of the suffix-free semigroup of transformations $\mathbf{T}^{\geqslant 6}(n)$ which is largest for $n = 2, 3$ and $n \geqslant 6$; and that the witness DFAs for boolean operations can have transition semigroups that are subsemigroups of $\mathbf{T}^{\leqslant 5} \cap \mathbf{T}^{\geqslant 6}$.

The full version of this paper is available at [10].

2 Suffix-Free Transformations

In this section we discuss some properties of suffix-free languages with emphasis on their syntactic semigroups as represented by the transition semigroups of their quotient DFAs. We assume that our basic set is always $Q_n = \{0, \ldots, n-1\}$.

Let $\mathcal{D}_n = (Q_n, \Sigma, \delta, 0, F)$ be the quotient DFA of a suffix-free language L, and let T_n be its transition semigroup. For any transformation t of Q_n, the sequence $(0, 0t, 0t^2, \ldots)$ is called the 0-*path* of t. Since Q_n is finite, there exist i, j such that $0, 0t, \ldots, 0t^i, 0t^{i+1}, \ldots, 0t^{j-1}$ are distinct but $0t^j = 0t^i$. The integer $j-i$ is the *period* of t and if $j-i=1$, t is *initially aperiodic*. The following properties of suffix-free languages are known [9, 15]:

Lemma 1. *If L is a suffix-free language, then*

1. *There exists $w \in \Sigma^*$ such that $w^{-1}L = \emptyset$; hence $\mathcal{D}_n$ has an empty state, which is state $n-1$ by convention.*
2. *For $w, x \in \Sigma^+$, if $w^{-1}L \neq \emptyset$, then $w^{-1}L \neq (xw)^{-1}L$.*
3. *If $L \neq \emptyset$ and $w^{-1}L = L$, then $w = \varepsilon$. This is known as the* non-returning *property* [15] *and also as* unique reachability [8].
4. *For any $t \in T_n$, the 0-path of t in $\mathcal{D}_n$ is aperiodic and ends in $n-1$.*

An (unordered) pair $\{i, j\}$ of distinct states in $Q_n \setminus \{0, n-1\}$ is *colliding* (or p *collides* with q) in T_n if there is a transformation $t \in T_n$ such that $0t = p$ and $rt = q$ for some $r \in Q_n \setminus \{0, n-1\}$. A pair of states is *focused* by a transformation u of Q_n if u maps both states of the pair to a single state $r \notin \{0, n-1\}$. We then say that $\{p, q\}$ is *focused to state* r. If L is a suffix-free language, then from Lemma 1 (2) it follows that if $\{p, q\}$ is colliding in T_n, there is no transformation $t' \in T_n$ that focuses $\{p, q\}$. So colliding states can be mapped to a single state by a transformation in T_n only if that state is the empty state $n-1$.

Following [9], for $n \geqslant 2$, we let

$$\mathbf{B}(n) = \{t \in \mathcal{T}_Q \mid 0 \notin \mathrm{rng}(t),\ (n-1)t = n-1,\ \text{and for all } j \geqslant 1,$$

$$0t^j = n-1 \text{ or } 0t^j \neq qt^j,\ \forall q \text{ such that } 0 < q < n-1\}.$$

Proposition 1 ([9]). *If L is a regular language having quotient DFA $\mathcal{D}_n = (Q_n, \Sigma, \delta, 0, F)$ and syntactic semigroup T_L, then the following hold:*

1. *If L is suffix-free, then T_L is a subset of $\mathbf{B}(n)$.*
2. *If L has the empty quotient, only one final quotient, and $T_L \subseteq \mathbf{B}(n)$, then L is suffix-free.*

Since the transition semigroup of a minimal DFA of a suffix-free language must be a subsemigroup of $\mathbf{B}(n)$, the cardinality of $\mathbf{B}(n)$ is an upper bound on the syntactic complexity of suffix-free regular languages with quotient complexity n. This upper bound, however, cannot be reached since $\mathbf{B}$ is not a semigroup for $n \geqslant 4$: We have $s = [1, 2, n-1, \ldots, n-1]$ and $t = [n-1, 2, 2, \ldots, 2, n-1]$ in $\mathbf{B}(n)$, but $st = [2, 2, n-1, \ldots, n-1]$ is not in $\mathbf{B}(n)$.

We now consider semigroups that are largest for $n \leqslant 5$. For $n \geqslant 2$, let $\mathbf{T}^{\leqslant 5}(n) = \{t \in \mathbf{B}(n) \mid \text{ for all } p, q \in Q_n \text{ where } p \neq q, \text{ either } pt = qt = n-1 \text{ or } pt \neq qt\}$.

Proposition 2. *For $n \geqslant 4$, semigroup $\mathbf{T}^{\leqslant 5}(n)$ is generated by the following set $\mathbf{H}^{\leqslant 5}(n)$ of transformations of Q:*

- $a\colon (0 \to n-1)(1, \ldots, n-2)$,
- $b\colon (0 \to n-1)(1, 2)$,
- *for* $1 \leqslant p \leqslant n-2$, $c_p\colon (p \to n-1)(0 \to p)$.

For $n = 4$, a and b coincide, and so $\mathbf{H}^{\leqslant 5}(4) = \{a, c_1, c_2\}$. Also, $\mathbf{H}^{\leqslant 5}(3) = \{a, c_1\} = \{[2, 1, 2], [1, 2, 2]\}$ and $\mathbf{H}^{\leqslant 5}(2) = \{c_1\} = \{[1, 1]\}$.

A DFA using these transformations is illustrated in Figure 1 for $n = 5$.

Proposition 3. *For $n \geqslant 2$, $\mathbf{T}^{\leqslant 5}(n)$ is the unique maximal semigroup of a suffix-free language in which all possible pairs of states are colliding.*

Proposition 4. *For $n \geqslant 5$, the number n of generators of $\mathbf{T}^{\leqslant 5}(n)$ cannot be reduced.*

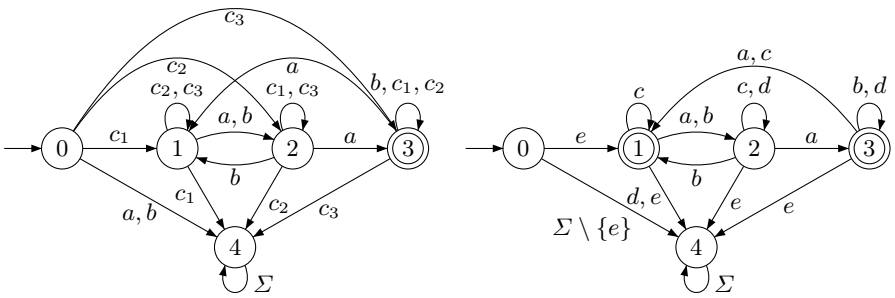

Fig. 1. DFAs with $\mathbf{T}^{\leqslant 5}(5)$ (left) and $\mathbf{T}^{\geqslant 6}(5)$ (right) as their transition semigroups

Next, we present semigroups that are largest for $n \geqslant 6$. For $n \geqslant 2$, let

$$\mathbf{T}^{\geqslant 6}(n) = \{t \in \mathbf{B}(n) \mid 0t = n-1 \text{ or } qt = n-1,\ \forall\ q \text{ such that } 1 \leqslant q \leqslant n-2\}.$$

Proposition 5 ([11]). *For $n \geqslant 4$, $\mathbf{T}^{\geqslant 6}(n)$ is a semigroup contained in $\mathbf{B}(n)$, its cardinality is $(n-1)^{n-2} + (n-2)$, and it is generated by the set $\mathbf{G}^{\geqslant 6}(n)$ of the following transformations:*

- a: $(0 \to n-1)(1, \ldots, n-2)$;
- b: $(0 \to n-1)(1, 2)$;
- c: $(0 \to n-1)(n-2 \to 1)$;
- d: $(\{0, 1\} \to n-1)$;
- e: $(Q \setminus \{0\} \to n-1)(0 \to 1)$.

For $n = 4$, a and b coincide, and so $\mathbf{G}^{\geqslant 6}(4) = \{a, c, d, e\}$. Also $\mathbf{G}^{\geqslant 6}(3) = \{a, e\} = \{[2, 1, 2], [1, 2, 2]\}$ and $\mathbf{G}^{\geqslant 6}(2) = \{e\} = \{[1, 1]\}$.

A DFA using the transformations of Proposition 5 is shown in Figure 1 for $n = 5$. Semigroups $\mathbf{T}^{\geqslant 6}(n)$ are the largest suffix-free semigroups for $n \geqslant 6$ [11].

3 Witnesses with Transition Semigroups in $\mathbf{T}^{\leqslant 5}(n)$

In this section we consider DFA witnesses whose transition semigroups are subsemigroups of $\mathbf{T}^{\leqslant 5}(n)$. We show that there is one witness that satisfies the bounds for star, product and boolean operations.

Definition 1. *For $n \geqslant 6$, we define the DFA $\mathcal{D}_n = (Q_n, \Sigma, \delta, 0, \{1\})$, where $Q_n = \{0, \ldots, n-1\}$, $\Sigma = \{a, b, c\}$, and δ is defined by the transformations*

- a: $(0 \to n-1)(1, 2, 3)(4, \ldots, n-2)$,
- b: $(2 \to n-1)(1 \to 2)(0 \to 1)(3, 4)$,
- c: $(0 \to n-1)(1, \ldots, n-2)$.

Theorem 1 (Star, Product, Boolean Operations). *Let $\mathcal{D}_n(a, b, c)$ be the DFA of Definition 1, and let the language it accepts be $L_n(a, b, c)$. For $n \geqslant 6$, L_n and its dialects meet the bounds for star, product and boolean operations as follows:*

1. $L_n^*(a,b,-)$ *meets the bound* $2^{n-2}+1$. [Cmorik and Jirásková [14]]
2. $L_m(a,b,c)\cdot L_n(b,c,a)$ *meets the bound* $(m-1)2^{n-2}+1$.
3. $L_m(a,b,-)$ *and* $L_n(c,b,-)$ *meet the bounds* $mn-(m+n-2)$ *for union and symmetric difference,* $mn-2(m+n-3)$ *for intersection, and* $mn-(m+2n-4)$ *for difference.*

The claim about the star operation was proved in [14]. We add a result about the transition semigroup of the star witness and prove the remaining two claims.

In 2009 Han and Salomaa [15] showed that the language of a DFA over a four-letter alphabet meets the bound $2^{n-2}+1$ for the star operation for $n \geqslant 4$. The transition semigroup of this DFA is a subsemigroup of $\mathbf{T}^{\leqslant 5}(n)$. In 2012 Cmorik and Jir'asková [14] showed that for $n \geqslant 6$ a binary alphabet $\{a,b\}$ suffices. The transition semigroup of this DFA is again a subsemigroup of $\mathbf{T}^{\leqslant 5}(n)$. We prove that these are special cases of the following general result:

Theorem 2. *For* $n \geqslant 4$*, the transition semigroup of the quotient DFA* $\mathcal{D}$ *of a suffix-free language* L *that meets the bound* $2^{n-2}+1$ *for the star operation is a subsemigroup of* $\mathbf{T}^{\leqslant 5}(n)$ *and is not a subsemigroup of* $\mathbf{T}^{\geqslant 6}(n)$*.*

For the product, to avoid confusing the states of the two DFAs, we label the states of the first DFA differently. Let $\mathcal{D}'_m = \mathcal{D}'_m(a,b,c) = (Q'_m, \Sigma, \delta', 0', \{1'\})$, where $Q'_m = \{0', \ldots, (m-1)'\}$, and $\delta'(q',x) = p'$ if $\delta(q,x) = p$, and let $\mathcal{D}_n = \mathcal{D}_n(b,c,a)$. We use the standard construction of the ε-NFA $\mathcal{N}$ for the product: the final state of $\mathcal{D}'_m$ becomes non-final, and an ε-transition is added from that state to the initial state of $\mathcal{D}_n$. This is illustrated in Figure 2 for $m = 9, n = 8$.

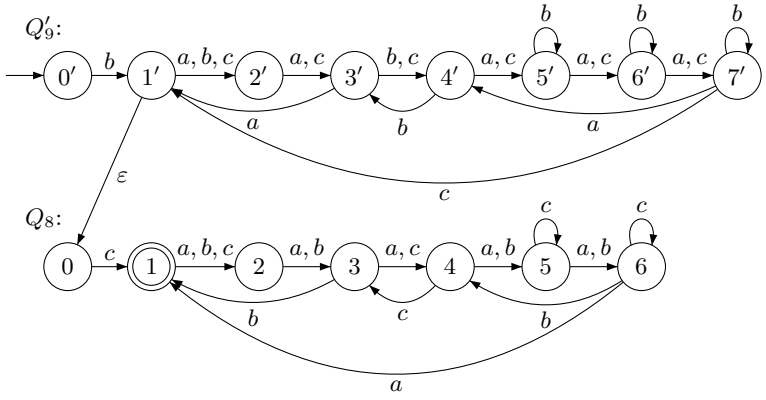

Fig. 2. The NFA $\mathcal{N}$ for product $L'_9(a,b,c)\cdot L_8(b,c,a)$. The empty states $8'$ and 7 and the transitions to them are omitted

We use the subset construction to determinize $\mathcal{N}$ to get a DFA $\mathcal{P}$ for the product. The states of $\mathcal{P}$ are subsets of $Q'_m \cup Q_n$ and have one of three forms: $\{0'\}$, $\{1',0\}\cup S$ or $\{p'\}\cup S$, where $p' = 2', \ldots, (m-1)'$ and $S \subseteq \{1,\ldots,n-1\}$.

Note that for each $x \in \Sigma$ every state $q \in Q_n \setminus \{0, n-1\}$ has a unique predecessor state $p \in Q_n \setminus \{n-1\}$ such that $px = q$. For $w \in \Sigma^*$, the w-predecessor of $S \subseteq Q_n \setminus \{0, n-1\}$ is denoted by Sw^{-1}.

Lemma 2. *For each $n \geqslant 6$ and each $q \in Q_n$ there exists a word $w_q \in c\{a,b\}^*$ such that $1'w_q = 3'$, $0w_q = q$, and each state of $Q_n \setminus \{0, q, n-1\}$ has a unique w_q-predecessor in $Q_n \setminus \{0, n-1\}$.*

Theorem 3 (Product: Ternary Case). *For $n \geqslant 6$, the product $L'_m(a,b,c) \cdot L_n(b,c,a)$ meets the bound $(m-1)2^{n-2}+1$.*

Cmorik and Jirásková [14, Theorem 5] also found binary witnesses that meet the bound $(m-1)2^{n-2}$ in the case where $m-2$ and $n-2$ are relatively prime. It remained unknown whether the bound $(m-1)2^{n-2}+1$ is reachable with a binary alphabet. We show that a slightly modified first witness of [14] meets the upper bound exactly. For $m \geqslant 6, n \geqslant 3$, let the first DFA be that of [14], except that the set of final states is changed to $\{2', 4'\}$; thus let $\Sigma = \{a,b\}$, $\mathcal{D}'_m(a,b) = (Q'_m, \Sigma, \delta', 0', \{2', 4'\})$, and let $\mathcal{D}_n(a,b) = (Q_n, \Sigma, \delta, 0, \{1\})$ be the second DFA [14]. Let $L'_m(a,b)$ and $L_n(a,b)$ be the corresponding languages.

Theorem 4 (Product: Binary Case). *For $m, n \geqslant 6$, $L'_m(a,b)$ is suffix-free and $L'_m(a,b) \cdot L_n(a,b)$ meets the bound $(m-1)2^{n-2}+1$ when $m-2$ and $n-2$ are relatively prime.*

Theorem 5. *Suppose $m, n \geqslant 4$ and $L'_m L_n$ meets the bound $2^{n-2}+1$. Then the transition semigroup T_n of a minimal DFA $\mathcal{D}_n$ of L_n is a subsemigroup of $\mathbf{T}^{\leqslant 5}(n)$ and is not a subsemigroup of $\mathbf{T}^{\geqslant 6}$.*

For boolean operations we use the witnesses $L'_m(a,b,-)$ and $L_n(c,b,-)$ and relabel them as $L'_m(a,b)$ and $L_n(a,b)$. See Figure 3.

Theorem 6. *For $m, n \geqslant 6$, $L'_m(a,b)$ and $L_n(a,b)$ meet the bounds for boolean operations.*

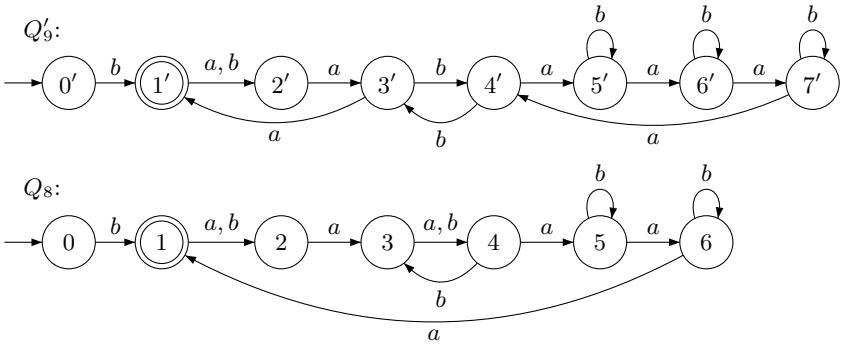

Fig. 3. Witnesses $\mathcal{D}'_9(a,b)$ and $D_8(a,b)$ for boolean operations. The empty states $8'$ and 7 and the transitions to them are omitted figure moved

4 Witnesses with Semigroups in $\mathbf{T}^{\geqslant 6}(n)$

We now turn to the operations which cannot have witnesses with transition semigroups in $\mathbf{T}^{\leqslant 5}$.

Definition 2. *For $n \geqslant 4$, we define the DFA $\mathcal{D}_n(a,b,c,d,e) = (Q_n, \Sigma, \delta, 0, F)$, where $Q_n = \{0,\ldots,n-1\}$, $\Sigma = \{a,b,c,d,e\}$, δ is defined by the transformations of Proposition 5, and $F = \{q \in Q_n \setminus \{0,n-1\} \mid q \text{ is odd}\}$. For $n = 4$, a and b coincide, and we can use $\Sigma = \{b,c,d,e\}$. The structure of $\mathcal{D}_5(a,b,c,d,e)$ is illustrated in Figure 1.*

Our main result in this section is the following theorem:

Theorem 7 (Boolean Operations, Reversal, Number and Complexities of Atoms, Syntactic Complexity). *Let $\mathcal{D}_n(a,b,c,d,e)$ be the DFA of Definition 2, and let the language it accepts be $L_n(a,b,c,d,e)$. Then $L_n(a,b,c,d,e)$ meets the following bounds:*

1. *For $n,m \geqslant 4$, $L_m(a,b,-,d,e)$ and $L_n(b,a,-,d,e)$ meet the bounds $mn-(m+n-2)$ for union and symmetric difference, $mn-2(m+n-3)$ for intersection, and $mn-(m+2n-4)$ for difference.*
2. *For $n \geqslant 4$, $L_n(a,-,c,-,e)$ meets the bound $2^{n-2}+1$ for reversal and number of atoms.*
3. *For $n \geqslant 6$, $L_m(a,b,c,d,e)$ meets the bound $(n-1)^{n-2}+n-2$ for syntactic complexity, and the bounds on the quotient complexities of atoms.*

The claim about syntactic complexity is known from [11]. It was shown in [12] that the number of atoms of a regular language L is equal to the quotient complexity of L^R. In the next subsections we prove the claim about boolean operations, reversal, and atom complexities. First we state some properties of $\mathcal{D}_n$.

Proposition 6. *For $n \geqslant 4$ the DFA of Definition 2 is minimal, accepts a suffix-free language, and its transition semigroup T_n has cardinality $(n-1)^{n-2}+n-2$. In particular, T_n contains (a) all $(n-1)^{n-2}$ transformations that send 0 and $n-1$ to $n-1$ and map $Q\setminus\{0,n-1\}$ to $Q\setminus\{0\}$, and (b) all $n-2$ transformations that send 0 to a state in $Q\setminus\{0,n-1\}$ and map all the other states to $n-1$. Also, T_n is generated by $\{a,b,c,d,e\}$ and cannot be generated by a smaller set of transformations.*

We now show that witness DFAs for boolean operations may have transition semigroups in $\mathbf{T}^{\geqslant 6}$.

Theorem 8. *For $n,m \geqslant 4$, $L_m(a,b,-,d,e)$ and $L_n(b,a,-,d,e)$ meet the bounds for boolean operations.*

Since $t_d = t_{c_1}t_{c_1}$ and $t_e = t_{c_1}t_{c_2}\cdots t_{c_{n-1}}$, where the c_i are from Proposition 2, the semigroup of $\mathcal{D}_n(a,b,-,d,e)$ is in $\mathbf{T}^{\leqslant 5}(n) \cap \mathbf{T}^{\geqslant 6}(n)$. In fact, one can verify that the semigroup of $\mathcal{D}_n(a,b,-,d,e)$ is $\mathbf{T}^{\leqslant 5}(n) \cap \mathbf{T}^{\geqslant 6}(n)$.

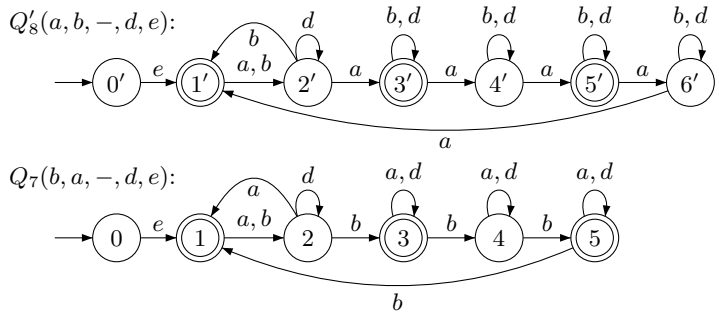

Fig. 4. The DFAs $\mathcal{D}'_8$ and $\mathcal{D}_7$ for boolean operations; empty states omitted

Han and Salomaa [15] showed that to meet the bound for reversal one can use the binary DFA of Leiss [18] and add a third input to get a suffix-free DFA. Cmorik and Jirásková [14] showed that a binary alphabet will not suffice. We show a different ternary witness below, and prove that any witness must have its transition semigroup in $\mathbf{T}^{\geqslant 6}$.

Theorem 9 (Semigroup of Reversal Witness). *For $n \geqslant 4$, the transition semigroup of a minimal DFA of a suffix-free language L_n that meets the bound $2^{n-2}+1$ for reversal is a subsemigroup of $\mathbf{T}^{\geqslant 6}(n)$ but not of $\mathbf{T}^{\leqslant 5}(n)$.*

Theorem 10 (Reversal Complexity). *If $n \geqslant 4$, then $L_n(a,-,c,-,e)$ of Definition 2 meets the bound $2^{n-2}+1$ for reversal.*

Although $L_n(a,-,c,-,e)$ meets the bound for the number of atoms, it does not meet the bounds on the quotient complexities of atoms; we now show that $L_n(a,b,c,d,e)$ does.

Let $Q_n = \{0,\ldots,n-1\}$ and let L be a non-empty regular language with quotients $K = \{K_0,\ldots,K_{n-1}\}$. Let $\mathcal{D} = (Q_n, \Sigma, \delta, 0, F)$ be the minimal DFA of L in which the language of state q is K_q.

Denote the complement of a language L by $\overline{L} = \Sigma^* \setminus L$. Each subset S of Q_n defines an *atomic intersection* $A_S = \bigcap_{i\in S} K_i \cap \bigcap_{i\in\overline{S}} \overline{K_i}$, where $\overline{S} = Q_n \setminus S$. An *atom* of L is a non-empty atomic intersection. Since atoms are pairwise disjoint, every atom A has a unique atomic intersection associated with it, and this atomic intersection has a unique subset S of K associated with it.

Let $A_S = \bigcap_{i\in S} K_i \cap \bigcap_{i\in\overline{S}} \overline{K_i}$ be an atom. For any $w \in \Sigma^*$ we have

$$w^{-1}A_S = \bigcap_{i\in S} w^{-1}K_i \cap \bigcap_{i\in\overline{S}} \overline{w^{-1}K_i}.$$

Since a quotient of a quotient of L is also a quotient of L, $w^{-1}A_S$ has the form:

$$w^{-1}A_S = \bigcap_{i\in X} K_i \cap \bigcap_{i\in Y} \overline{K_i},$$

where $|X| \leqslant |S|$ and $|Y| \leqslant n - |S|$, $X, Y \subseteq Q_n$.

Proposition 7. *Suppose L is a suffix-free language with $n \geqslant 4$ quotients. Then L has at most $2^{n-2}+1$ atoms. Also, the complexity $\kappa(A_S)$ of atom A_S satisfies*

$$\kappa(A_S) \begin{cases} \leqslant 2^{n-2}+1, & \text{if } S = \emptyset; \\ = n, & \text{if } S = \{0\}; \\ \leqslant 1 + \sum_{x=1}^{|S|} \sum_{y=0}^{n-2-|S|} \binom{n-2}{x}\binom{n-2-x}{y}, & \emptyset \neq S \subseteq \{1,\ldots,n-2\}. \end{cases} \tag{1}$$

Following Iván [16] we define a DFA for each atom:

Definition 3. *Suppose $\mathcal{D} = (Q, \Sigma, \delta, q_0, F)$ is a DFA and let $S \subseteq Q$. Define the DFA $\mathcal{D}_S = (Q_S, \Sigma, \Delta, (S, \overline{S}), F_S)$, where*

- $Q_S = \{(X,Y) \mid X, Y \subseteq Q, X \cap Y = \emptyset\} \cup \{\bot\}$.
- *For all $a \in \Sigma$, $(X,Y)a = (Xa, Y, a)$ if $Xa \cap Ya = \emptyset$, and $(X,Y)a = \bot$ otherwise; and $\bot a = \bot$.*
- $F_S = \{(X,Y) \mid X \subseteq F, Y \subseteq \overline{F}\}$.

DFA $\mathcal{D}_S$ recognizes the atomic intersection A_S of L. If $\mathcal{D}_S$ recognizes a non-empty language, then A_S is an atom.

Theorem 11. *For $n \geqslant 4$, the language $L_n(\mathcal{D}(a,b,c,d,e))$ of Definition 2 meets the bounds of Proposition 7 for the atoms.*

Remark 1. The complexity of atoms in left ideals [6] is

$$\kappa(A_S) \begin{cases} = n, & \text{if } S = Q_n; \\ \leqslant 2^{n-1}, & \text{if } S = \emptyset; \\ \leqslant 1 + \sum_{x=1}^{|S|} \sum_{y=1}^{n-|S|} \binom{n-1}{x}\binom{n-1-x}{y-1}, & \text{otherwise.} \end{cases} \tag{2}$$

The formula for $S \notin \{\emptyset, Q_n\}$ evaluated for $n-1$ and $S \subseteq \{1,\ldots,n-2\}$ becomes $1 + \sum_{x=1}^{|S|} \sum_{y=1}^{n-2-|S|} \binom{n-2}{x}\binom{n-2-x}{y-1}$, which is precisely the formula for suffix-free languages. ■

5 Conclusions

It may appear that semigroup $\mathbf{T}^{\leqslant 5}(n)$ should not be of great importance, since it exceeds $\mathbf{T}^{\geqslant 6}(n)$ only for $n = 4$ and $n = 5$, and therefore should not matter when n is large. However, our results show that this is not the case. We conclude with our result about the non-existence of single universal suffix-free witness.

Theorem 12. *There does not exist a most complex stream in the class of suffix-free languages.*

The first four studies of most complex languages were done for the classes of regular languages [4], right ideals [5, 6], left ideals [6, 7], and two-sided ideals [6, 7]. In those cases there exists a single stream witness of languages over a minimal alphabet which, together with their dialects, cover all the complexity measures. In the case of suffix-free languages such a stream does not exist. Our study is an example of a general problem: Given a class of regular languages, find the smallest number of streams over minimal alphabets that together cover all the measures. The witness of Definition 1 is conjectured to be over a minimal alphabet, unless the bound for product can be met by binary DFAs for every $n, m > c$, for some c; this is an open problem. The witness of Definition 2 is over a minimal alphabet, since five letters are required to meet then bound for syntactic complexity.

References

1. Ang, T., Brzozowski, J.: Languages convex with respect to binary relations, and their closure properties. Acta Cybernet. 19(2), 445–464 (2009)
2. Berstel, J., Perrin, D., Reutenauer, C.: Codes and Automata. Cambridge University Press (2009)
3. Brzozowski, J.: Quotient complexity of regular languages. J. Autom. Lang. Comb. 15(1/2), 71–89 (2010)
4. Brzozowski, J.: In search of the most complex regular languages. Int. J. Found. Comput. Sc. 24(6), 691–708 (2013)
5. Brzozowski, J., Davies, G.: Most complex regular right ideals. In: Jürgensen, H., et al. (eds.) DCFS. LNCS, vol. 8614, pp. 90–101. Springer (2014)
6. Brzozowski, J., Davies, S.: Quotient complexities of atoms in regular ideal languages (2015), `http://arxiv.org/abs/1503.02208`
7. Brzozowski, J., Davies, S., Liu, B.Y.V.: Most complex regular ideals (2015), in preparation
8. Brzozowski, J., Jirásková, G., Li, B., Smith, J.: Quotient complexity of bifix-, factor-, and subword-free regular languages. Acta Cybernet. 21, 507–527 (2014)
9. Brzozowski, J., Li, B., Ye, Y.: Syntactic complexity of prefix-, suffix-, bifix-, and factor-free regular languages. Theoret. Comput. Sci. 449, 37–53 (2012)
10. Brzozowski, J., Szykuła, M.: Complexity of suffix-free regular languages (2015), `http://arxiv.org/abs/1504.05159`
11. Brzozowski, J., Szykuła, M.: Upper bound for syntactic complexity of suffix-free languages. In: Okhotin, A., Shallit, J. (eds.) DCFS 2015. LNCS, vol. 9118, pp. 33–45. Springer (2015), full paper at `http://arxiv.org/abs/1412.2281`
12. Brzozowski, J., Tamm, H.: Theory of átomata. Theoret. Comput. Sci. 539, 13–27 (2014)
13. Brzozowski, J., Ye, Y.: Syntactic complexity of ideal and closed languages. In: Mauri, G., Leporati, A. (eds.) DLT. LNCS, vol. 6795, pp. 117–128. Springer (2011)
14. Cmorik, R., Jirásková, G.: Basic operations on binary suffix-free languages. In: Kotásek, Z., et al. (eds.) MEMICS. pp. 94–102 (2012)
15. Han, Y.S., Salomaa, K.: State complexity of basic operations on suffix-free regular languages. Theoret. Comput. Sci. 410(27-29), 2537–2548 (2009)
16. Iván, S.: Complexity of atoms, combinatorially (2015), `http://arxiv.org/abs/1404.6632`

17. Jirásková, G., Olejár, P.: State complexity of union and intersection of binary suffix-free languages. In: Bordihn, H., et al. (eds.) NMCA. pp. 151–166. Austrian Computer Society (2009)
18. Leiss, E.: Succinct representation of regular languages by boolean automata. Theoret. Comput. Sci. 13, 323–330 (2009)
19. Maslov, A.N.: Estimates of the number of states of finite automata. Dokl. Akad. Nauk SSSR 194, 1266–1268 (Russian) (1970), english translation: Soviet Math. Dokl. **11** (1970), 1373–1375
20. Mirkin, B.G.: On dual automata. Kibernetika (Kiev) 2, 7–10 (Russian) (1966), English translation: Cybernetics **2**, (1966) 6–9
21. Pin, J.E.: Syntactic semigroups. In: Handbook of Formal Languages, vol. 1: Word, Language, Grammar, pp. 679–746. Springer, New York, NY, USA (1997)
22. Yu, S., Zhuang, Q., Salomaa, K.: The state complexities of some basic operations on regular languages. Theoret. Comput. Sci. 125, 315–328 (1994)
23. Yu, S.: State complexity of regular languages. J. Autom. Lang. Comb. 6, 221–234 (2001)