Unary Prime Languages

Ismaël Jecker
Institute of Science and Technology, Austria
ismael.jecker@ist.ac.at

Orna Kupferman
School of Computer Science and Engineering, Hebrew University, Israel
orna@cs.huji.ac.il

Nicolas Mazzocchi
Université Libre de Bruxelles, Belgium
nicolas.mazzocchi@ulb.ac.be

Abstract

A regular language L of finite words is *composite* if there are regular languages $L_1, L_2, \ldots, L_t$ such that $L = \bigcap_{i=1}^{t} L_i$ and the index (number of states in a minimal DFA) of every language L_i is strictly smaller than the index of L. Otherwise, L is *prime*. Primality of regular languages was introduced and studied in [9], where the complexity of deciding the primality of the language of a given DFA was left open, with a doubly-exponential gap between the upper and lower bounds. We study primality for unary regular languages, namely regular languages with a singleton alphabet. A unary language corresponds to a subset of $\mathbb{N}$, making the study of unary prime languages closer to that of primality in number theory. We show that the setting of languages is richer. In particular, while every composite number is the product of two smaller numbers, the number t of languages necessary to decompose a composite unary language induces a strict hierarchy. In addition, a primality witness for a unary language L, namely a word that is not in L but is in all products of languages that contain L and have an index smaller than L's, may be of exponential length. Still, we are able to characterize compositionality by structural properties of a DFA for L, leading to a LogSpace algorithm for primality checking of unary DFAs.

2012 ACM Subject Classification Theory of computation → Formal languages and automata theory; Mathematics of computing → Mathematical analysis

Keywords and phrases Deterministic Finite Automata (DFA), Regular Languages, Primality

Digital Object Identifier 10.4230/LIPIcs.MFCS.2020.45

Funding *Ismaël Jecker*: This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Grant Agreement No. 754411.
Nicolas Mazzocchi: PhD fellowship FRIA from the F.R.S.-FNRS.

1 Introduction

Compositionality is a well motivated and studied notion in mathematics and computer science [2]. By decomposing a problem into several smaller problems, it is possible not only to increase parallelism, but also to sometimes handle inputs that are otherwise intractable. A major challenge is to identify problems and instances that can be decomposed. Motivated by practical barriers of the automata-theoretic approach to formal verification [8], Kupferman and Mosheiff introduced in [9] the notion of compositionality for regular languages. The algebraic approach to DFAs associates each DFA with a *monoid*, and is used in [7] in order to show that every DFA $\mathcal{A}$ can be presented as a wreath product of reset DFAs and permutation DFAs, whose algebraic structure is simpler than that of $\mathcal{A}$. The definition of decomposition in [9] is simpler, and is based on the right-congruence relation $\sim_L$ between words in Σ^*: given a regular language $L \subseteq \Sigma^*$, we have that two words $x, y \in \Sigma^*$ satisfy $x \sim_L y$, if for every

45th International Symposium on Mathematical Foundations of Computer Science (MFCS 2020).
Editors: Javier Esparza and Daniel Král'; Article No. 45; pp. 45:1–45:12
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

word $z \in \Sigma^*$, it holds that $x \cdot z \in L$ iff $y \cdot z \in L$. By the Myhill-Nerode theorem [10, 11], the equivalence classes of $\sim_L$ constitute the state space of a minimal canonical DFA for L. The number of equivalence classes is referred to as the *index* of L. Then, according to [9], a language $L \subseteq \Sigma^*$ is *composite* if there are languages $L_1, \ldots, L_t$ such that $L = \bigcap_{i=1}^{t} L_t$ and the index of L_i, for all $1 \leq i \leq t$, is strictly smaller than the index of L. Otherwise, L is *prime*[1]. The definitions apply also to DFAs, referring to the languages they recognize. Back to formal verification, by decomposing a specification automaton $\mathcal{A}$ to automata $\mathcal{A}_1, \mathcal{A}_2, \ldots, A_t$ such that $L(\mathcal{A}) = \bigcap_{i=1}^{t} L(\mathcal{A}_t)$, one can replace a language-containment problem $L(\mathcal{S}) \subseteq L(\mathcal{A})$ by a sequence of simpler problems, namely $L(\mathcal{S}) \subseteq L(\mathcal{A}_i)$, for the automata $\mathcal{A}_i$ in the decomposition.

Decompositions of width 2 were studied in [3]. For such decompositions, the question of deciding whether a given DFA $\mathcal{A}$ is composite is clearly in NP, as one can guess the two factors. It is shown in [9] that there are regular languages whose decomposition require width 3, which was generalized in [12] to languages whose decomposition require arbitrarily large widths. In fact, the only bound known for the required width is exponential in $|\mathcal{A}|$, which follows from the bound on the size of the underlying DFAs. Accordingly, the best upper bound known for the problem of deciding the compositionality of a given DFA is ExpSpace. This is quite surprising, especially given that the best lower bound for the problem is NLogSpace, making the gap between the upper and lower bounds doubly-exponential. For the class of *permutation* DFAs, whose monoid consists of permutations, compositionality can be decided in PSpace [9], making the gap exponentially less embarassing, but the general case is still open.

We study regular languages over a unary alphabet, thus $\Sigma = \{1\}$. Each word $1^i \in \Sigma^*$ can be identified with its length $i \in \mathbb{N} = \{0, 1, 2, \ldots\}$, and a language $L \subseteq 1^*$ can be viewed as a subset of $\mathbb{N}$. The association of words with natural numbers strengthens the relation between the notions of primality in number theory and regular languages. In particular, it is shown in [9] that for every $k \in \mathbb{N}$, we have that the language $(1^k)^*$ is composite iff k is not a prime power (see Example 1). The fact, however, that each DFA defines a set of numbers, makes the regular setting much richer [1]. We present two indications of this rich setting. The first concerns the *width* of a decomposition, namely the number t of languages in it. The width of decompositions in number theory is 2. Indeed, every composite number is the product of two smaller numbers. We show that for unary regular languages, the width is arbitrarily large. Specifically, if a language L is defined by a unary DFA of size n, then the width of a decomposition of L may be $\omega(n)$, namely the number of distinct prime divisors of n. This bound is tight.

An additional indication to the richness of the setting is the length of *primality witnesses*. Consider a DFA $\mathcal{A}$. It is not hard to see that $\mathcal{A}$ is prime iff there exists a word w that is rejected by $\mathcal{A}$ yet accepted by all DFAs $\mathcal{B}$ that are *potential decomposers* of $\mathcal{A}$, namely $L(\mathcal{A}) \subseteq L(\mathcal{B})$ and $|\mathcal{B}| < |\mathcal{A}|$. Indeed, such a word w indicates that every product of DFAs that attempts to decompose $\mathcal{A}$ would fail on w. Accordingly, w is termed a primality witness for $\mathcal{A}$, and a decision procedure for checking primality can be based on a search for a primality witness. In the general (non unary) case, the best known upper bound for the length of a primality witness is doubly exponential in $\mathcal{A}$, with no matching lower bound [9]. We study the length of primality witnesses for unary DFAs and show an exponential tight bound.

In spite of the above two hardness indications, we are able to describe a LogSpace

[1] We note that a different notion of primality, relative to the concatenation operator rather than to intersection, has been studied in [4].

algorithm for checking primality of unary DFAs. Our algorithm is based on the trivial observation that a unary DFA $\mathcal{A}$ is *lasso shaped*, and the not-at-all trivial observation that $\mathcal{A}$ is composite iff it can be decomposed to *clean quotients* – once quotients obtained by folding the cycle of length ℓ of $\mathcal{A}$'s lasso to a cycle of length d, for d that is a strict divisor of ℓ. All the clean quotients over-approximate the language of $\mathcal{A}$, and the algorithm essentially has to check whether each rejecting state q of $\mathcal{A}$ is *covered* by some clean quotient, in the sense that this clean quotient rejects all words that $\mathcal{A}$ rejects in a run that reaches q. As we show, the above condition can be checked in logarithmic space.

2 Preliminaries

A *deterministic finite automaton* (DFA hereafter) is a 5-tuple $\mathcal{A} = \langle \Sigma, Q, q_I, \delta, F\rangle$, where Q is the finite set of states, Σ is a finite non-empty alphabet, $\delta\colon Q \times \Sigma \to Q$ is a transition function, $q_I \in Q$ is an initial state, and $F \subseteq Q$ is a set of accepting states. For each state $q \in Q$, we use $\mathcal{A}^q$ to denote the DFA $\mathcal{A}$ with q as the initial state. That is, $\mathcal{A}^q = \langle \Sigma, Q, q, \delta, F\rangle$. We extend δ to words in the expected way, thus $\delta\colon Q \times \Sigma^* \to Q$ is defined recursively by $\delta(q, \epsilon) = q$ and $\delta(q, w_1 w_2 \cdots w_n) = \delta(\delta(q, w_1 w_2 \cdots w_{n-1}), w_n)$. We sometimes omit the initial state q_I as a parameter of δ and write $\delta(w)$ instead of $\delta(q_I, w)$ in order to refer to the state that $\mathcal{A}$ visits after reading w. The DFA $\mathcal{A}$ naturally induces an equivalence relation $\sim_{\mathcal{A}}$ over the set of words Σ^* defined by $v \sim_A w$ iff $\delta(v) = \delta(w)$.

The *run* of $\mathcal{A}$ on a word $w = w_1 \ldots w_n$ is the sequence of states $s_0, s_1 \ldots s_n$ such that $s_0 = q_I$ and for each $1 \leq i \leq n$ it holds that $\delta(s_{i-1}, w_i) = s_i$. Note that $s_n = \delta(q_I, w)$. The DFA $\mathcal{A}$ *accepts* w iff $\delta(q_I, w) \in F$. Otherwise, $\mathcal{A}$ *rejects* w. The set of words accepted by $\mathcal{A}$ is denoted $L(\mathcal{A})$ and is called the *language of* $\mathcal{A}$. We say that $\mathcal{A}$ *recognizes* $L(\mathcal{A})$. A language recognized by some DFA is called a *regular language*. Two DFAs $\mathcal{A}$ and $\mathcal{B}$ are *equivalent* if $L(\mathcal{A}) = L(\mathcal{B})$. The complement of a regular language L over Σ is $comp(L) = \Sigma^* \setminus L$.

We refer to the size of a DFA $\mathcal{A}$, denoted $|\mathcal{A}|$, as the number of states in $\mathcal{A}$. A DFA $\mathcal{A}$ is *minimal* if every DFA $\mathcal{B}$ equivalent to $\mathcal{A}$ satisfies $|\mathcal{B}| \geq |\mathcal{A}|$. Every regular language L has a single (up to isomorphism) minimal DFA $\mathcal{A}$ such that $L(\mathcal{A}) = L$. The index of L, denoted $ind(L)$, is the size of the minimal DFA recognizing L.

Quotient DFA Consider a DFA $\mathcal{A} = \langle \Sigma, Q, q_I, \delta, F\rangle$. We say that an equivalence relation $\sim \subseteq Q \times Q$ is *coherent* with δ if for every two states $p, q \in Q$, if $p \sim q$ then $\delta(p, a) \sim \delta(q, a)$ for all $a \in \Sigma$. Then, the *quotient* $\mathcal{A}'$ of $\mathcal{A}$ by $\sim$ is the DFA obtained by merging the states of $\mathcal{A}$ that are equivalent with respect to $\sim$. Formally, $\mathcal{A}' = \langle \Sigma, Q', [q_I], \delta', F'\rangle$, where Q' is the set of equivalence classes $[p]$ of the states $p \in Q$, the transition function δ' is such that for all $a \in \Sigma$ we have that $\delta'([p], a) = [\delta(p, a)]$, and F' is composed of the classes $[p]$ such that there is $q \in F$ such that $p \sim q$. Note that the coherency of $\sim$ with respect to δ guarantees that the definition of δ' is independent of the choice of the state p in $[p]$. On the other hand, we do not require states related by $\sim$ to agree on membership in F, and define F' so that the language of $\mathcal{A}'$ over-approximates that of $\mathcal{A}$. Formally, $L(\mathcal{A}) \subseteq L(\mathcal{A}')$, as every accepting run of $\mathcal{A}$ induces an accepting run of $\mathcal{A}'$.

Composite and Prime DFAs A DFA $\mathcal{A}$ is *composite* if there are $t \geq 1$ and DFAs $\mathcal{A}_1, \ldots, \mathcal{A}_t$ such that for all $1 \leq i \leq t$, it holds that $|\mathcal{A}_i| < |\mathcal{A}|$, and $\bigcap_{i=1}^{t} L(\mathcal{A}_i) = L(\mathcal{A})$. Thus, $L(\mathcal{A})$ can be described by means of an intersection of DFAs all strictly smaller than $\mathcal{A}$. Otherwise, $\mathcal{A}$ is *prime*. We refer to t as the *width* of the decomposition of $\mathcal{A}$. For $t \geq 2$, we say that $\mathcal{A}$ is t*-composite* if it has a decomposition of width t. Otherwise, $\mathcal{A}$ is t*-prime*. Then, the width

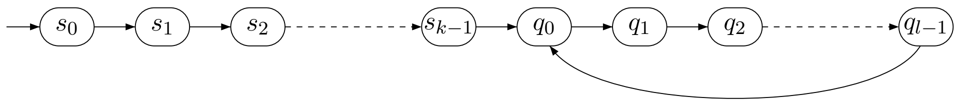

Figure 1 A lasso-shaped unary DFA

of a composite $\mathcal{A}$ is the minimal $t \geq 1$ such that $\mathcal{A}$ is t-composite. Note that non-minimal DFAs are 1-composite, and so compositionality is of interest mainly for minimal DFAs, where $|\mathcal{A}| = ind(L(\mathcal{A}))$. We identify a regular language with its minimal DFA. Thus, we talk also about a regular language being composite or prime, referring to its minimal DFA. The PRIME-DFA problem is to decide, given a DFA $\mathcal{A}$, whether $\mathcal{A}$ is prime.

A *primality witness* for a DFA $\mathcal{A}$ is a word $w \in \Sigma^*$ such that $w \notin L(\mathcal{A})$ and $w \in L(\mathcal{B})$ for all $\mathcal{B}$ with $L(\mathcal{A}) \subseteq L(\mathcal{B})$ and $|\mathcal{B}| < |\mathcal{A}|$. Note that such a word w indeed witnesses that $\mathcal{A}$ is prime, as w is a member in every intersection of DFAs that attempts to compose $\mathcal{A}$. Moreover, every prime DFA $\mathcal{A}$ admits at least one primality witness, as otherwise $L(\mathcal{A})$ would be equal to the intersection of the languages of all the DFAs $\mathcal{B}$ satisfying $L(\mathcal{A}) \subseteq L(\mathcal{B})$ and $|\mathcal{B}| < |\mathcal{A}|$.

2.1 Unary DFAs

A DFA $\mathcal{A} = \langle \Sigma, Q, q_I, \delta, F \rangle$ is *unary* if its alphabet Σ is of size 1. Discussing unary DFAs, we denote the alphabet by $\mathbb{1}$, its single letter by 1, and we identify a word $1^i \in \mathbb{1}^*$ with its length $i \in \mathbb{N} = \{0, 1, 2, \ldots\}$. Thus, the language of a unary DFA $\mathcal{A}$ is viewed as a subset of $\mathbb{N}$. Likewise, we refer to the transition function of a unary DFA as $\delta \colon \mathbb{N} \to Q$, where $\delta(i)$ is the state that $\mathcal{A}$ visits after reading 1^i. Clearly, $i \in L(\mathcal{A})$ iff $\delta(i) \in F$. Finally, note that a unary DFA must be *lasso shaped*. Indeed, as $|Q|$ is finite, there must be $k, j \in \mathbb{N}$ such that $k < j$ and $\delta(j) = \delta(k)$. Then, as $\mathcal{A}$ is deterministic, we have that $\delta(k+i) = \delta(j+i)$ for all $i \geq 0$. When j is minimal, we say that $\mathcal{A}$ is a (k, ℓ)-DFA, for $\ell = j - k$. Thus, $\mathcal{A}$ is lasso-shape with a prefix of lenght k and cycle of length ℓ. We refer to the states $\delta(0), \ldots, \delta(k)$ by $s_0, \ldots, s_{k-1}, q_0$, and to the states $\delta(k+1), \ldots, \delta(j-1)$ by $q_1, \ldots, q_{\ell-1}$ (see Figure 1). Note that since $k < j$, it must be that $\ell > 0$. When we want to fix only one of the two parameters of the lasso, we use the notations $(*, \ell)$-DFA, for fixing only the cycle, and $(k, *)$-DFA for fixing only the prefix. In particular, a $(0, *)$-DFA consists of a single cycle.

As demonstrated in Example 1 below, taken from [9], primality questions about unary languages are strongly related to primality questions in number theory.

▶ **Example 1.** Let $L_k = \{x : x \equiv 0 \bmod k\}$. Clearly, the minimal DFA that recognizes L_k is a $(0, k)$-DFA, and so $ind(L_k) = k$. We show that L_k is composite iff k is not a prime power.

Assume first that k is not a prime power. Thus, there exist co-prime integers $1 < p, q < k$ such that $p \cdot q = k$. It then holds that $L_k = L_p \cap L_q$. Since $ind(L_p) < k$ and $ind(L_q) < k$, it follows that L_k is composite.

For the other direction, assume that k is a prime power. Let $p, r \in \mathbb{N}$ be such that p is a prime and $k = p^r$. Let $x = (p+1)p^{r-1}$. Note that $x \notin L_k$. We claim that x is a primality witness for L_k, and conclude that L_k is prime.

Recall that $ind(L_k) = k$. Consider a language L' such that $L \subseteq L'$ and $ind(L') < ind(L)$. We show that $x \in L'$. Assume by contradiction that $x \notin L'$. Let $\mathcal{A}'$ be a DFA for L'. Since $ind(L') < ind(L_k) = k$ and $x > k$, the rejecting run of $\mathcal{A}'$ on x must traverse the cycle of $\mathcal{A}'$. Let ℓ be the length of this cycle, and note that $0 < \ell < k$. Note that for all $i \geq 0$, we have that $i\ell + (p+1)p^{r-1}$ is not accepted by $\mathcal{A}'$, and hence, $i\ell + (p+1)p^{r-1} \notin L'$. On the other hand, since $\ell \not\equiv 0 \bmod k$, there exists $i \geq 0$ such that $i\ell \equiv -p^{r-1} \bmod k$. For this value of i, we have that $i\ell + (p+1)p^{r-1} \in L \setminus L'$, and thus, $L \not\subseteq L'$, and we have reached a contradiction. Therefore, $x \in L$, and we are done. ◀

▶ Remark 2. Since DFAs can be complemented by dualizing the set of final states, we can dualize the definition of composite and prime DFAs and consider definitions that are based on union of DFAs. Specifically, L is $\cup$-composite if there are DFAs $\mathcal{A}_1, \ldots, \mathcal{A}_t$ such that for all $1 \leq i \leq t$, it holds that $|\mathcal{A}_i| < |\mathcal{A}|$, and $\bigcup_{i=1}^{t} L(\mathcal{A}_i) = L(\mathcal{A})$. Otherwise, $\mathcal{A}$ is $\cup$-*prime*. Clearly, L is $\cap$-composite iff $comp(L)$ is $\cup$-composite.

3 Decompositions of Unary DFAs

In this section we study decompositions of unary DFAs. We characterize these decompositions by means of *clean quotients*, which will become handy when we study the width of decompositions, the length of primality witnesses, and the complexity of the Prime-DFA problem for unary DFAs.

3.1 Clean quotients of unary DFA

Let $\mathcal{A} = \langle \mathbb{1}, Q, \Sigma, q_I, \delta, F \rangle$ be a unary (k, ℓ)-DFA. Recall (see Figure 1) that we refer to the states leading to the cycle of $\mathcal{A}$ by $s_0, s_1, \ldots, s_{k-1}$, and to the states in the cycle by $q_0, q_1, \ldots, q_{\ell-1}$. A *clean quotient* $\mathcal{A}_d$ of $\mathcal{A}$ is a (k, d)-DFA obtained by quotienting $\mathcal{A}$ by folding its cycle to a cycle of length d, for some strict divisor d of ℓ. Formally, $\mathcal{A}_d$ is induced by the equivalence relation $\sim_d$ defined by

$$\begin{array}{lll} s_i \sim_d s_j & \text{if and only if} & i = j; \\ q_i \sim_d q_j & \text{if and only if} & i \equiv j \bmod d. \end{array}$$

Note that $\sim_d$ is coherent with δ, and so $L(\mathcal{A}) \subseteq L(\mathcal{A}_d)$. As with general quotient DFAs, the latter containment may be strict.

▶ **Example 3.** In Figure 2, we describe a $(4, 6)$-DFA $\mathcal{A}$, and its two clean quotients: the $(4, 3)$-DFA $\mathcal{A}_3$ and the $(4, 2)$-DFA $\mathcal{A}_2$.

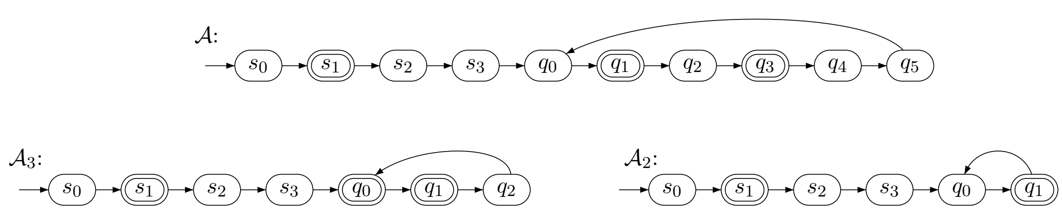

Figure 2 The DFA $\mathcal{A}$ and its clean quotients $\mathcal{A}_3$ and $\mathcal{A}_2$.

Omega function For $n \in \mathbb{N}$, the *omega function* $\omega(n)$ maps n to the number of distinct prime divisors of n. Formally, for every integer n, if the decomposition of n into prime factors is $n = p_1^{g_1} p_2^{g_2} \ldots p_t^{g_t}$, then $\omega(n) = t$. For example, as $45 = 3 \cdot 3 \cdot 5$, then $\omega(45) = 2$. The asymptotic behavior of ω is tricky, as it behaves irregularly. Indeed, if n is a prime number, then $\omega(n) = 1$. On the other hand, if n is a *primorial*, namely $n = p_1 p_2 \ldots p_t$ is the product of the first t prime numbers, then $\omega(n) \sim \frac{\ln(n)}{\ln(\ln(n))}$ [5]. Note that for every $t \in \mathbb{N}$, the primorial $n = p_1 p_2 \dots p_t$ is the smallest integer satisfying $\omega(n) \geq t$. Accordingly, $\frac{\ln(n)}{\ln(\ln(n))}$ serves as an upper asymptotical bound for $\omega(n)$.

In Subsection 3.2, we relate compositionality with compositionality by clean quotients. Here, we bound the width of such compositions:

▶ **Lemma 4.** *Every unary (k, ℓ)-DFA that has a decomposition into clean quotients is $\omega(\ell)$-composite.*

Proof. Let $\mathcal{A}$ be a unary (k, ℓ)-DFA, and assume that $L(\mathcal{A}) = \bigcap_{i=1}^{m} L(\mathcal{A}_{d_i})$ for some strict divisors d_i of ℓ. Let $p_1, p_2, \dots, p_{\omega(\ell)}$ be an enumeration of the prime strict divisors of ℓ, and for every $1 \leq i \leq \omega(\ell)$, let $\ell_i = \ell / p_i$. For all $1 \leq i \leq \omega(\ell)$, we get $L(\mathcal{A}) \subseteq L(\mathcal{A}_{\ell_i})$ since $\mathcal{A}_{\ell_i}$ is a quotient of $\mathcal{A}$, hence $L(\mathcal{A}) \subseteq \bigcap_{i=1}^{\omega(\ell)} \mathcal{A}_{\ell_i}$. Conversely, for every $1 \leq i \leq m$, there exists $1 \leq j \leq \omega(\ell)$ such that d_i divides ℓ_j, thus the DFA $\mathcal{A}_{d_i}$ is a subquotient of the DFA $\mathcal{A}_{\ell_j}$, which implies that $L(\mathcal{A}_{d_i}) \supseteq L(\mathcal{A}_{\ell_j})$. Since this is true for every $1 \leq i \leq m$, it follows that $L(\mathcal{A}) = \bigcap_{i=1}^{\omega(m)} \mathcal{A}_{d_i} \supseteq \bigcap_{i=1}^{\omega(\ell)} \mathcal{A}_{\ell_i}$. Hence $L(\mathcal{A}) = \bigcap_{i=1}^{\omega(\ell)} \mathcal{A}_{\ell_i}$, thus $\mathcal{A}$ is $\omega(\ell)$-composite. ◀

Bézout's Identity We use in our proofs a weaker version of Bézout's Identity, a well known theorem in number theory. For the sake of completeness, we state here the specific part of the result that we use, along with its proof.

▶ **Lemma 5.** *Consider an integer $b \in \mathbb{N}$. If b has a strict divisor, then for all $a < b$ we have that b has a strict divisor that can be expressed as a linear combination $\lambda a - \mu b$, for some $\lambda, \mu \in \mathbb{N}$.*

Proof. Let U be the set of integers definable as a linear combination $\lambda a - \mu b$ for some $\lambda, \mu \in \mathbb{N}$. We prove that the minimal strictly positive element d of U satisfies the statement. First, since $a \in U$, then $d \leq a < b$. Now, since $d \in U$, there exist $\lambda_0, \mu_0 \in \mathbb{N}$ satisfying $d = \lambda_0 a - \mu_0 b$. Let $\beta \in \mathbb{N}$ be the minimal integer satisfying $\beta d \geq b$. Then $0 \leq \beta d - b < d$, yet $\beta d - b = \beta\lambda_0 a - (\beta\mu_0 + 1)b$, is an element of U. Since we chose d as the minimal strictly positive integer of U, this implies that $\beta d - b = 0$. Hence, d divides b and we are done. ◀

Key Lemma Recall that every clean quotient $\mathcal{A}_d$ of $\mathcal{A}$ is such that $L(\mathcal{A}) \subseteq L(\mathcal{A}_d)$, and that the latter containment may be strict. We now prove the existence of clean quotients of $\mathcal{A}$ for which this strict containment is good enough for our decomposition goal. Intuitively, each clean quotient rejects large parts of the language rejected by $\mathcal{A}$. Formally, we have the following.

▶ **Lemma 6.** *Let $\mathcal{A}$ be a unary (k, ℓ)-DFA. For every unary $(k_{\mathcal{B}}, \ell_{\mathcal{B}})$-DFA $\mathcal{B}$ such that $\ell_{\mathcal{B}} < \ell$ and $L(\mathcal{A}) \subseteq L(\mathcal{B})$, there is a strict divisor d of ℓ such that the clean quotient $\mathcal{A}_d$ rejects all the words $w > k_{\mathcal{B}}$ rejected by $\mathcal{B}$.*

Proof. Let $\mathcal{A} = \langle \mathbb{1}, Q, q_I, \delta, F \rangle$, and let $\mathcal{B}$ be a unary $(k_{\mathcal{B}}, \ell_{\mathcal{B}})$-DFA such that $\ell_{\mathcal{B}} < \ell$ and $L(\mathcal{A}) \subseteq L(\mathcal{B})$. Since $\ell_{\mathcal{B}} < \ell$, then, by Lemma 5, there exists a strict divisor d of ℓ that can be expressed as a linear combination $d = \lambda \ell_{\mathcal{B}} - \mu \ell$ for some $\lambda, \mu \in \mathbb{N}$.

We prove that the clean quotient $\mathcal{A}_d$ of $\mathcal{A}$ rejects all the words $w > k_{\mathcal{B}}$ rejected by $\mathcal{B}$. Assume by way of contradiction that there is a word $w > k_{\mathcal{B}}$ accepted by $\mathcal{A}_d$. If $w < k$, $w \in L(\mathcal{A}_d)$ immediately implies that $w \in L(\mathcal{A})$. Then, as $L(\mathcal{A}) \subseteq L(\mathcal{B})$, we have that $w \in L(\mathcal{B})$, and we reach a contradiction. If $w \geq k$, then by definition of the quotient $\mathcal{A}_d$, the equivalence class of the state $\delta(w) \in Q$ in $\mathcal{A}_d$ contains an accepting state of $\mathcal{A}$ since $w \in L(\mathcal{A}_d)$. Therefore, there exists an integer $\alpha \in \mathbb{N}$ such that $w + \alpha d$ is accepted by $\mathcal{A}$. Since adding a multiple of ℓ to $w + \alpha d$ yields another element of $L(\mathcal{A})$, we obtain that

$x = w + \alpha d + \alpha\mu\ell \in L(\mathcal{A})$. Since $L(\mathcal{A}) \subseteq L(\mathcal{B})$, it follows that x is also accepted by $\mathcal{B}$. Now, by the definition of d, we have that

$$x = w + \alpha d + \alpha\mu\ell = w + \alpha\lambda\ell_{\mathcal{B}}.$$

Therefore, since $\mathcal{B}$ accepts x, and $w > k_{\mathcal{B}}$ by supposition, $\mathcal{B}$ also accepts the word w, and we reach a contradiction. ◀

3.2 Characterizing compositionality

Consider a unary (k,ℓ)-DFA $\mathcal{A} = \langle \mathbb{1}, Q, q_I, \delta, F\rangle$. We say that a rejecting state q of $\mathcal{A}$ is *covered* by a quotient $\mathcal{A}'$ of $\mathcal{A}$ if the state $[q]$ of $\mathcal{A}'$ is rejecting. That is, q is covered by $\mathcal{A}'$ iff $\mathcal{A}'$ rejects all the words w such that $\delta(w) = q$. We show that we can determine if $\mathcal{A}$ is composite by checking whether some of its rejecting states are covered by clean quotients. Our analysis distinguishes between several cases, as detailed below. We start with (k,ℓ)-DFAs satisfying $k = 0$.

▶ **Lemma 7.** *Consider a unary $(0,\ell)$-DFA $\mathcal{A}$. The following are equivalent:*
1. *$\mathcal{A}$ is $\omega(\ell)$-composite;*
2. *$\mathcal{A}$ is composite;*
3. *For every rejecting state q_i of $\mathcal{A}$, the word $\ell + i$ is not a primality witness of $\mathcal{A}$;*
4. *Every rejecting state of $\mathcal{A}$ is covered by a clean quotient.*

Proof. It is clear that Item 1 implies Item 2. Moreover, Item 2 implies Item 3: indeed, if $\mathcal{A}$ is composite, then it has no primality witness.

We now show that Item 3 implies Item 4. Consider a rejecting state q_i of $\mathcal{A}$. We argue that either the word $w_i = i + \ell \in \mathbb{N}$ is a primality witness, or there is a clean quotient of $\mathcal{A}$ covering q_i. Assume that w_i is not a primality witness for $\mathcal{A}$. Thus, there is a unary $(k_{\mathcal{B}}, \ell_{\mathcal{B}})$-DFA $\mathcal{B}_i$ such that $|\mathcal{B}_i| < |\mathcal{A}|$, $L(\mathcal{A}) \subseteq L(\mathcal{B}_i)$, and $w_i \notin L(\mathcal{B}_i)$.

As $k = 0$, we have that $|\mathcal{A}| = \ell$, and so $\ell_{\mathcal{B}} \leq |\mathcal{B}| < |\mathcal{A}| = \ell$. Hence, by Lemma 6, there is a clean quotient $\mathcal{A}_{d_i}$ of $\mathcal{A}$ that rejects all the words longer than $k_{\mathcal{B}}$ that are rejected by $\mathcal{B}$. In particular, since $k_{\mathcal{B}} \leq |\mathcal{B}| < |\mathcal{A}| \leq i + \ell$, the DFA $\mathcal{A}_{d_i}$ rejects w_i. However, as $\mathcal{A}_{d_i}$ is a quotient of $\mathcal{A}$, then $\mathcal{A}_{d_i}$ either accepts all words w with $\delta(w) = q_i$ or it rejects them all. Therefore, as $\delta(w_i) = q_i$ and $\mathcal{A}_{d_i}$ rejects w_i, we conclude that the clean quotient $\mathcal{A}_{d_i}$ rejects all words w with $\delta(w) = q_i$, implying it covers q_i.

To conclude, we show that Item 4 implies Item 1. Assume that every rejecting state q_i of $\mathcal{A}$ is covered by a clean quotient $\mathcal{A}_{d_i}$. Let $I \subseteq \{0, \ldots, \ell - 1\}$ be such that $i \in I$ iff q_i is rejecting. We show that $L(\mathcal{A}) = \bigcap_{i \in I} L(\mathcal{A}_{d_i})$, which implies that $\mathcal{A}$ is $\omega(\ell)$-composite by Lemma 4. First, by definition of a quotient DFA, we have that $L(\mathcal{A}) \subseteq L(\mathcal{A}_{d_i})$ for all $i \in I$, and thus $L(\mathcal{A}) \subseteq \bigcap_{i \in I} L(\mathcal{A}_{d_i})$. Second, each word w that $\mathcal{A}$ rejects reaches a rejecting state q_i of $\mathcal{A}$. Therefore, $\mathcal{A}_{d_i}$ also rejects w, and so $L(\mathcal{A}) \supseteq \bigcap_{i \in I} L(\mathcal{A}_{d_i})$. ◀

We continue to (k,ℓ)-DFAs with $k > 0$. Consider such a DFA $\mathcal{A}$, and consider the state s_{k-1}, namely the last state visited by $\mathcal{A}$ before entering the cycle, and the state $q_{\ell-1}$, namely the last state of the cycle. Let $\widetilde{A}$ be the quotient DFA of $\mathcal{A}$ induced by the equivalent $s_{k-1} \sim q_{\ell-1}$. Thus, $\widetilde{A}$ is obtained from $\mathcal{A}$ by merging s_{k-1} and q_{l-1}. Clearly, $|\widetilde{\mathcal{A}}| < |\mathcal{A}|$.

The following lemmas handle three possible cases.

▶ **Lemma 8.** *Consider a unary (k,ℓ)-DFA $\mathcal{A}$ with $k > 0$. If s_{k-1} and $q_{\ell-1}$ are both in F or are both not in F, then $\mathcal{A}$ is composite.*

Proof. The agreement of s_{k-1} and $q_{\ell-1}$ on membership in F guarantees that $L(\widetilde{A}) = L(\mathcal{A})$. Hence, $\mathcal{A}$ is not minimal, and is thus composite with $t = 1$. ◀

▶ **Lemma 9.** *Consider a unary (k, ℓ)-DFA $\mathcal{A}$ with $k > 0$. If $s_{k-1} \notin F$ and $q_{\ell-1} \in F$, then $\mathcal{A}$ is composite iff $\ell > 1$.*

Proof. If $\ell = 1$, then $\mathcal{A}$ is prime with primality witness $k-1$. If $\ell > 1$, then $\mathcal{A}$ is 2-composite. Indeed, consider the language $\mathbb{N} \setminus \{k-1\}$. Clearly, it can be accepted by a $(k-1, 1)$-DFA. Since $L(\mathcal{A})$ is the intersection of $L(\widetilde{\mathcal{A}})$ and $\mathbb{N} \setminus \{k-1\}$, we are done. ◀

▶ **Lemma 10.** *Consider a unary (k, ℓ)-DFA $\mathcal{A}$ with $k > 0$. If $s_{k-1} \in F$ and $q_{\ell-1} \notin F$, then the following assertions are equivalent:*

1. *$\mathcal{A}$ is 2-composite;*
2. *$\mathcal{A}$ is composite;*
3. *The word $k - 1 + (|\mathcal{A}| - 1)!$ is not a primality witness of $\mathcal{A}$;*
4. *The rejecting state $q_{\ell-1}$ of $\mathcal{A}$ is covered by a clean quotient.*

Proof. It is clear that Item 1 implies Item 2. Moreover, Item 2 implies Item 3: indeed, if $\mathcal{A}$ is composite, then it has no primality witness.

To prove that Item 3 implies Item 4, we argue that either the word $w = k - 1 + (|\mathcal{A}| - 1)!$ is a primality witness of $\mathcal{A}$, or there exists a clean quotient of $\mathcal{A}$ covering $q_{\ell-1}$. Assume that the word w is not a primality witnesses of $\mathcal{A}$. Thus, there exists a unary $(k_{\mathcal{B}}, \ell_{\mathcal{B}})$-DFA $\mathcal{B}$ such that $|\mathcal{B}| < |\mathcal{A}|$, $L(\mathcal{A}) \subseteq L(\mathcal{B})$, and $w \notin L(\mathcal{B})$. In order to use Lemma 6, we show that the cycle of $\mathcal{B}$ is strictly smaller than the cycle of $\mathcal{A}$. Assume by way of contradiction that $\ell_{\mathcal{B}} \geq \ell$. Since $k_{\mathcal{B}} + \ell_{\mathcal{B}} = |\mathcal{B}| < |\mathcal{A}| = k + \ell$, this implies that $k_{\mathcal{B}} < k$. Therefore, $\mathcal{B}$ reaches its cycle while reading the word $k - 1$. Since $s_{k-1} \in F$, the word $k - 1$ is accepted by $\mathcal{A}$. Since $L(\mathcal{A}) \subseteq L(\mathcal{B})$, the word $k - 1$ is also accepted by $\mathcal{B}$, which thus accepts all words in $(k - 1) + \mu\ell_B$. Indeed, the run of $\mathcal{B}$ on all of them reaches the same accepting state. In particular, $\mathcal{B}$ accepts the witness $w = k - 1 + (|\mathcal{A}| - 1)!$, and we have reached a contradiction.

Now that we have proven that $\ell_{\mathcal{B}} < \ell$, we can apply Lemma 6 to guarantee the existence of a clean quotient $\mathcal{A}_d$ of $\mathcal{A}$ that rejects (in particular) the word w. However, as $\mathcal{A}_d$ is a quotient of $\mathcal{A}$, then $\mathcal{A}_d$ either accepts all words w' with $\delta(w') = q_{\ell-1}$ or it rejects them all. Therefore, as $\delta(w) = q_{\ell-1}$ and $\mathcal{A}_d$ rejects w, we conclude that the clean quotient $\mathcal{A}_d$ rejects all words w' with $\delta(w') = q_{\ell-1}$, hence it covers $q_{\ell-1}$.

We conclude by showing that Item 4 implies Item 1. Assume that the rejecting state $q_{\ell-1}$ of $\mathcal{A}$ is covered by a clean quotient $\mathcal{A}_d$. We show that $L(\mathcal{A}) = L(\widetilde{\mathcal{A}}) \cap L(\mathcal{A}_d)$. Since both $\widetilde{\mathcal{A}}$ and $\mathcal{A}_d$ are quotients of $\mathcal{A}$, then $L(\mathcal{A}) \subseteq L(\widetilde{\mathcal{A}}) \cap L(\mathcal{A}_d)$. Now consider a word w rejected by $\mathcal{A}$. Then, either $\delta(w) = q_{\ell-1}$, in which case, as $\mathcal{A}_d$ covers $q_{\ell-1}$, the word w is also rejected by $\mathcal{A}_d$, or $\delta(w) \neq q_{\ell-1}$, in which case it is also rejected by $\widetilde{\mathcal{A}}$. Therefore, $L(\mathcal{A}) \supseteq L(\widetilde{\mathcal{A}}) \cap L(\mathcal{A}_d)$. ◀

4 The Width of Unary Languages

Recall that [9, 12] shows that in the general (non unary) case, the width of composite languages may be arbitrarily large. This is in contrast with composite numbers, which are always 2-composite. The languages used in [9, 12] for showing the strict hierarchy are over alphabets of size $O(t)$. In this section we show that the hierarchy is strict even for unary languages, which are closer to number theory. We show that the width of a unary language of index n is closely related to the omega function $\omega(n)$ that counts the number of distinct prime divisors of n.

First, our results from Section 3 provide an upper bound on the width of a composite (k,ℓ)-DFA $\mathcal{A}$: If $k=0$, then, by Lemma 7, we have that $\mathcal{A}$ is $\omega(\ell)$-composite, and if $k>0$, then, by Lemmas 8, 9, and 10, we have that $\mathcal{A}$ is 2-composite. We thus have the following.

▶ **Theorem 11.** *Every unary composite language of index n is $\max(2,\omega(n))$-composite.*

We prove that such a large width is sometimes required.

▶ **Theorem 12.** *For every $n \in \mathbb{N}$ with $\omega(n) \geq 2$, there is a composite unary language of index n and width $\omega(n)$.*

Proof. Let $n \in \mathbb{N}$, and consider the decomposition $n = p_1^{g_1} p_2^{g_2} \dots p_{\omega(n)}^{g_{\omega(n)}}$ of n into prime factors. Assume that $\omega(n) \geq 2$. For every $1 \leq i \leq \omega(n)$, let $\gamma_i = n/p_i^{g_i}$, and let $L_i = \{x : x \not\equiv 0 \bmod \gamma_i\}$. We set $L = \bigcap_{i=1}^{\omega(n)} L_i$, and prove that L is $\omega(n)$-composite and $(\omega(n)-1)$-prime.

It is easy to see that L can be recognized by a $(0,n)$-DFA, and that each L_i can be recognized by a $(0,\gamma_i)$-DFA. To conclude, we show that if L is expressed as the intersection of $m < \omega(n)$ languages, then at least one of these language has an index bigger or equal to n. This implies that the index of L is n (using the particular case $m=1$), and that $L = \bigcap_{i=1}^{\omega(n)} L_i$ has minimal width amongst the decompositions of L into languages of indices smaller than n. Formally, we prove the following:

▷ Claim. Let $m < \omega(n)$, and let $\mathcal{B}_1, \dots, \mathcal{B}_m$ be m unary DFAs satisfying $\bigcap_{i=1}^{m} L(\mathcal{B}_i) = L$. Then, there exists $1 \leq i \leq m$, such that $|\mathcal{B}_i| \geq n$.

Since $m < \omega(n)$ and $\bigcap_{i=1}^{m} L(\mathcal{B}_i) = L = \bigcap_{i=1}^{\omega(n)} L_i$, there exist $1 \leq i \leq m$ and $1 \leq j_1 < j_2 \leq \omega(n)$ such that $\mathcal{B}_i$ rejects both $n+\gamma_{j_1} \notin L_{j_1}$ and $n+\gamma_{j_2} \notin L_{j_2}$. We prove that $|\mathcal{B}_i| \geq n$.

Let $k,\ell \in \mathbb{N}$ be the integers such that $\mathcal{B}_i$ is a (k,ℓ)-DFA. If $k \geq n$, we are done. Otherwise, $\mathcal{B}$ reaches its cycle while reading the word $n+\gamma_j$ for both $j \in \{j_1,j_2\}$. As the cycle of $\mathcal{B}$ is of length ℓ, we also have that $n+\gamma_j+\ell\cdot p_j \notin L(\mathcal{B}_i)$. Therefore, as $L \subseteq L(\mathcal{B}_i)$, it must be that $n+\gamma_j+\ell\cdot p_j \notin L$. Thus, there exists $1 \leq j' \leq \omega(n)$ such that $n+\gamma_j+\ell\cdot p_j \notin L_{j'}$. Hence,

$$n+\gamma_j+\ell\cdot p_j \equiv 0 \bmod \gamma_{j'}. \tag{1}$$

As p_j divides both n and $\ell\cdot p_j$ but not γ_j, we get from Equation 1 that $\gamma_{j'}$ is not divisible by p_j, which is possible only if $j'=j$. Therefore, $\gamma_{j'}=\gamma_j$, and as n is divisible by γ_j, Equation 1 becomes $\ell\cdot p_j \equiv 0 \bmod \gamma_j$. Then, since p_j and γ_j are co-prime, it follows that $\ell \equiv 0 \bmod \gamma_j$. Finally, since this equation holds for both $j=j_1$ and $j=j_2$, and $j_1 \neq j_2$, it must be that $\ell \equiv 0 \bmod n$ by definition of γ_{j_1} and γ_{j_2}. This implies that $\ell \geq n$, hence $|\mathcal{B}_i| \geq n$. ◀

5 Primality Witnesses For Unary Languages

Recall that every prime DFA $\mathcal{A}$ has a primality witness: a word that is rejected by $\mathcal{A}$ yet accepted by all DFAs $\mathcal{B}$ that are *potential decomposers* of $\mathcal{A}$, namely $L(\mathcal{A}) \subseteq L(\mathcal{B})$ and $|\mathcal{B}| < |\mathcal{A}|$. Note that indeed $\mathcal{A}$ is prime iff it has a primality witness w: since w is accepted by all the potential decomposers of $\mathcal{A}$, then w is accepted by all products of potential decomposers, implying they strictly contain $\mathcal{A}$.

For general DFAs, [9] provides a doubly-exponential upper bound on the length of a minimal primality witnesses, with no lower bound. In this section we describe a tight exponential bound for unary DFAs, and we start with the lower bound:

▶ **Theorem 13.** *For every $n \geq 1$, there is a unary prime language L_n that is recognized by a DFA with $O(n)$ states, yet the shortest primality witness for L_n is of length exponential in n.*

Proof. For $n \in \mathbb{N}$, let $\mathcal{A}_n$ be the unary $(2n+1, 2)$-DFA whose language is the union of the odd numbers and the singleton $2n$. Thus, $L(\mathcal{A}_n) = \{2\lambda + 1 : \lambda \in \mathbb{N}\} \cup \{2n\}$. We define $L_n = L(\mathcal{A}_n)$. Clearly, $\mathcal{A}_n$ has $2n+3$ states, which is linear in n. We prove that L_n is prime, yet the size of its smallest primality witness is exponential in n.

Let $p_1, p_2, \ldots, p_m$ be an enumeration of the prime numbers smaller than or equal to $n+1$, moreover for every $1 \le i \le m$, let g_i be the highest power such that $p_i^{g_i} \le n+1$. Finally, let $P = p_1^{g_1} \cdot p_2^{g_2} \cdot \ldots \cdot p_m^{g_m}$. We prove that the word $2(n+P)$ is a primality witness for L_n. Since $2(n+P)$ is even and is different from $2n$, then it is rejected by $\mathcal{A}_n$. We show that $2(n+P)$ is accepted by every unary $(k_\mathcal{B}, \ell_B)$-DFA $\mathcal{B}$ that satisfies $|\mathcal{B}| < |\mathcal{A}_n|$ and $L_n \subseteq L(\mathcal{B})$.

We distinguish between the two cases, according to the parity of $\ell_\mathcal{B}$ – the length of the cycle of $\mathcal{B}$. If $\ell_\mathcal{B}$ is odd, then, in order to ensure that $L_n \subseteq L(\mathcal{B})$, all the states in the cycle of $\mathcal{B}$ have to be accepting. Therefore $\mathcal{B}$ accepts every word greater than $k_\mathcal{B} < |\mathcal{B}| < |\mathcal{A}_n| = 2n+3$. In particular, it accepts $2(n+P)$.

If $\ell_\mathcal{B}$ is even, let $\ell' \ge 1$ be such that $\ell_B = 2\ell'$. Then, since $k_\mathcal{B} + 2\ell' = |\mathcal{B}| < |\mathcal{A}_n| = 2n+3$, we obtain that $k_\mathcal{B} < 2n+1$, and $\ell' \le n+1$. Since $L_n \subseteq L(\mathcal{B})$, the run of $\mathcal{B}$ on the word $2n$ is accepting. Since $k_\mathcal{B} < 2n+1$, the accepting run of $\mathcal{B}$ on $2n$ reaches its cycle. Thus, $\mathcal{B}$ also accepts all words obtained by adding to $2n$ a multiple of $\ell_B = 2\ell'$. However, $2P$ is a multiple of $2\ell'$, as the definition of P ensures that every divisor of integers smaller than $n+1$, in particular ℓ', is also a divisor of P. Therefore, $\mathcal{B}$ accepts the word $2(n+P)$, and we are done.

Next, we prove that P is exponential in n. Recall that there are m prime numbers smaller than or equal to $n+1$. By the Prime Number Theorem, the integer m can be approximated with $(n+1)/\ln(n+1)$. Also, for every $1 \le i \le m$, the definition of g_i implies that $p_i^{g_i} \ge \sqrt{n+1}$. As a consequence, we get

$$P = p_1^{g_1} \cdot p_2^{g_2} \cdot \ldots \cdot p_m^{g_m} \ge \sqrt{n+1}^m = (n+1)^{\frac{m}{2}} \sim (n+1)^{\frac{n+1}{2\ln(n+1)}} = e^{\frac{n+1}{2}}.$$

Finally, we prove that every word smaller than $2(n+P)$ is not a primality witness for L_n. Let $x \in \mathbb{N}$ be such that $x < 2(n+P)$ and $x \notin L_n$. We prove that there is an NBW $\mathcal{B}$ such that $|\mathcal{B}| < |\mathcal{A}_n|$ and $x \notin L(\mathcal{B})$. Since $x \notin L_n$, then it is of the form $2(n+\lambda)$, for some $\lambda \in \mathbb{N}$ satisfying $0 < |\lambda| < P$. Therefore, there exists an index $1 \le i \le m$ such that the prime power $p_i^{g_i}$ does not divide $|\lambda|$. Let $\mathcal{B}$ be the unary $(0, 2p_i^{g_i})$-DFA whose language is the union of the odd words and the words equivalent to $2n$ modulo $2p_i^{g_i}$. That is, $L(\mathcal{B}) = \{2\kappa + 1 : \kappa \in \mathbb{N}\} \cup \{2p_i^{k^i} + 2n : \kappa \in \mathbb{N}\}$. Note that $L_n \subseteq L(\mathcal{B})$. Moreover, as $p_i^{g_i} \le n+1$, we have that $|\mathcal{B}| = 2p_i^{g_i} \le 2(n+1) < |\mathcal{A}_n|$. Finally, $x \notin L(\mathcal{B})$. Indeed, since x is even, it is not in $\{2\kappa + 1 : \kappa \in \mathbb{N}\}$. Also, as we chose i so that $p_i^{g_i}$ does not divide $|\lambda|$, we also have that $x \notin \{2p_i^{k^i} + 2n : \kappa \in \mathbb{N}\}$. Thus, x is not a primality witness for L_n, and we are done. ◀

We continue with a matching upper bound.

▶ **Theorem 14.** *Every prime DFA $\mathcal{A}$ has a primality witness of length at most exponential in $|\mathcal{A}|$.*

Proof. Consider a prime (k, ℓ)-DFA $\mathcal{A}$. If $k = 0$ then, by Lemma 7, there is a primality witness for $\mathcal{A}$ of length smaller than 2ℓ. If $k > 0$ then, by Lemmas 8, 9, and 10, there is a primality witness for $\mathcal{A}$ of length smaller than $|\mathcal{A}|!$. In order to reduce the $|\mathcal{A}|!$ bound to an exponential one, we do a more careful analysis of the length of the primality witness in Item 3 of Lemma 10, reducing it from $k - 1 + (|\mathcal{A}| - 1)!$ down to $k - 1 + P$, where P is the product of the maximal prime powers $p_i^{g_i}$ smaller than $|\mathcal{A}|$ (the same P used in the proof of Theorem 13). Essentially, this follows form the fact the DFA $\mathcal{B}$ in the proof of Lemma 10 accepts all words in $\{(k-1)+\mu\ell_{\mathcal{B}} : \mu \in \mathbb{N}\}$, in particular it accepts $k-1+P$, as $\ell_{\mathcal{B}}$ can be decomposed into prime factors in $\{p_1, \ldots, p_m\}$. ◀

6 Solving the Prime-DFA problem

The Prime-DFA problem is to decide, given a DFA $\mathcal{A}$, whether $\mathcal{A}$ is prime. As discussed in [9], the Prime-DFA problem for general DFAs is in ExpSpace and is hard for NLogSpace. In this section we show that for unary DFAs, the problem can be solved in deterministic logarithmic space.

▶ **Theorem 15.** *The* Prime-DFA *problem for unary DFAs is in* LogSpace.

Proof. We first describe a deterministic algorithm for the problem, and then explain its correctness and argue it uses logarithmic space.

Function IsComposite($\mathcal{A}$: unary $\langle k, \ell\rangle$-DFA)
  if $k \stackrel{?}{=} 0$ **then** /* by Lemma 7 */
    foreach $q_i \notin F$ **do**
      if not IsCleanlyCovered($\mathcal{A}, q_i$) **then** **return** false
    return true
  if $s_{k-1} \in F \Leftrightarrow q_{\ell-1} \in F$ **then** **return** true /* by Lemma 8 */
  if $s_{k-1} \notin F \wedge q_{\ell-1} \in F$ **then** **return** $\ell \stackrel{?}{\neq} 1$ /* by Lemma 9 */
  if $s_{k-1} \in F \wedge q_{\ell-1} \notin F$ **then** /* by Lemma 10 */
    return IsCleanlyCovered($\mathcal{A}, q_{\ell-1}$)

Function IsCleanlyCovered($\mathcal{A}$: unary $\langle k, \ell\rangle$-DFA, $q_i \notin F$)
  foreach $1 < d < \ell$ such that d divides ℓ **do**
    nb_final := 0
    foreach $0 \leq j < \ell$ with $j \equiv i \bmod d$ **do**
      if $q_j \in F$ **then** nb_final := nb_final + 1
    if nb_final $\stackrel{?}{=} 0$ **then** **return** true
  return false

Let $\mathcal{A} = \langle \Sigma, Q, q_I, \delta, F\rangle$ be a unary (k, ℓ)-DFA. The main decision procedure is straightforward from the cases considered in Subsection 3.2, and uses a constant local space. However, a call to the function "IsCleanlyCovered", which takes as input a DFA $\mathcal{A}$ and a rejecting state q_i from its cycle, requires a logarithmic space. We prove that "IsCleanlyCovered" return true iff there exists a strict divisor d of ℓ such that the clean quotient $\mathcal{A}_d$ of $\mathcal{A}$ covers q_i.

First, the function searches for divisors d by checking every decomposition $d \cdot m$ of ℓ with $d, m \in \{2, 3, \ldots \ell-1\}$. Then, given d, let $\mathcal{A}_d = \langle \Sigma, Q', [q_I], \delta', F'\rangle$. Recall that $\mathcal{A}_d$ has a cycle of length d. To perform in logarithmic space, the function cannot construct $\mathcal{A}_d$ explicitly and has to perform on-the-fly. It counts in nb_final how many accepting states belong to the equivalence class $[q_i]$ by increasing a counter on all states $q_j \in F$ for which $i \equiv j \bmod d$. By the definition of a quotient automaton, $\mathcal{A}_d$ rejects all the words w for which $\delta(w) = q_i$ iff $[q_i] \notin F'$; that is, iff $q_j \notin F$ for every $i \equiv j \bmod d$. Hence, q_i is covered by $\mathcal{A}_d$ iff the counter nb_final stays zero. These operations are doable within space logarithmic in $|\mathcal{A}|$ since all the numerical values are bounded by ℓ and thus representable in $O\big(\log_2(|\mathcal{A}|)\big)$ bits with a binary encoding. ◀

7 Discussion

We studied primality for unary regular languages, and showed that while the setting is richer than that of primality in number theory, we can decide primality of a given unary DFA in deterministic logarithmic space. Beyond the interest in unary languages and their relation to number theory, we believe that our results can contribute to an improved upper bound in the general (non unary) case, where the best known algorithm for the Prime-DFA problem requires exponential space. A promising direction for closing the doubly-exponential gap is to consider more special cases. Different semantic fragments of regular languages induce different structural properties of the their DFAs. For example, languages closed for letter-swapping are recognized by DFAs that are products of lassos, and bounded semilinear languages, namely languages L for which there exists $k > 0$ and words $u_1, \ldots, u_k \in \Sigma^*$ such that $L \subseteq u_1^* \ldots u_k^*$, are recognized by DFAs that are concatenation of lassos, as well as deterministic Parikh automata [6] – all are good candidates for a tighter analysis. Likewise, the considerations we made for lasso-shape DFAs may be extendible to DFAs that are trees with back edges. Another interesting direction is to allow richer compositions, in particular ones that allow both intersection and union.

References

1 A. Cobham. On the base-dependence of sets of numbers recognizable by finite automata. *Math. Systems Theory*, 3:186?192, 1969.

2 W-P. de Roever, H. Langmaack, and A. Pnueli, editors. *Compositionality: The Significant Difference. Proceedings of Compositionality Workshop*, volume 1536 of *Lecture Notes in Computer Science*. Springer, 1998.

3 P. Gazi. Parallel decompositions of finite automata. Master's thesis, Comenius University, Bratislava, Slovakia, 2006.

4 Y.-S. Han, A. Salomaa, K. Salomaa, D. Wood, and S. Yu. On the existence of prime decompositions. *Theoretical Computer Science*, 376:60–69, 2007.

5 G.H. Hardy and E.M. Wright. *An introduction to the theory of numbers*. Oxford university press, 1979.

6 F. Klaedtke and H. Rueß. Monadic second-order logics with cardinalities. In *Proc. 30th Int. Colloq. on Automata, Languages, and Programming*, volume 2719 of *Lecture Notes in Computer Science*, pages 681–696. Springer, 2003.

7 K. Krohn and J. Rhodes. Algebraic theory of machines. i. prime decomposition theorem for finite semigroups and machines. *Transactions of the American Mathematical Society*, 116:450–464, 1965.

8 O. Kupferman, R.P. Kurshan, and M. Yannakakis. Existance of reduction hierarchies. In *Proc. 6th Annual Conf. of the European Association for Computer Science Logic*, volume 1414 of *Lecture Notes in Computer Science*, pages 327–340. Springer, 1997.

9 O. Kupferman and J. Mosheiff. Prime languages. *Information and Computation*, 240:90–107, 2015.

10 J. Myhill. Finite automata and the representation of events. Technical Report WADD TR-57-624, pages 112–137, Wright Patterson AFB, Ohio, 1957.

11 A. Nerode. Linear automaton transformations. *Proceedings of the American Mathematical Society*, 9(4):541–544, 1958.

12 A. Netser. Decomposition of safe langauges. Amirim Research Prohect, The Hebrew University, 2018.