Available online at www.sciencedirect.com
ScienceDirect
Information and Computation 205 (2007) 1652–1670
Information and Computation
www.elsevier.com/locate/ic

Magic numbers in the state hierarchy of finite automata☆

Viliam Geffert

Department of Computer Science, P.J. Šafárik University, Jesenná 5, 040 01 Košice, Slovakia

Received 2 March 2007
Available online 7 August 2007

Abstract

A number d is magic for n, if there is no regular language for which an optimal nondeterministic finite state automaton (nfa) uses exactly n states and, at the same time, the optimal deterministic finite state automaton (dfa) uses exactly d states. We show that, in the case of unary regular languages, the state hierarchy of dfa's, for the family of languages accepted by n-state nfa's, is not contiguous. There are some "holes" in the hierarchy, i.e., magic numbers in between values that are not magic. This solves, for automata with a single letter input alphabet, an open problem of existence of magic numbers. Actually, most of the numbers is magic in the unary case. As an additional bonus, we also get a new universal lower bound for the conversion of unary d-state dfa's into equivalent nfa's: nondeterminism does not reduce the number of states below $\log^2 d$, not even in the best case.

Keywords: Descriptional complexity; Finite-state automata; Regular languages

1. Introduction

Automata theory is one of the oldest topics in theoretical computer science, and also a popular first step to study this field. In spite of that, some important problems concerning regular languages are still open. The most famous problem is whether a two-way nondeterministic finite state automaton with n states can be converted into an equivalent two-way deterministic automaton using only a polynomial number of states [20]. (See also [8].)

At first glance, the situation is clear for one-way automata, the fundamental standard model for regular languages. By the classical subset construction [19], one of the oldest results, we know that a nondeterministic finite state automaton (nfa) with n states can be replaced by an equivalent deterministic finite state automaton (dfa) with d states, such that $n \leqslant d \leqslant 2^n$. It is also known that the exponential blow-up cannot be improved.In the

☆ This work was supported by the Slovak Grant Agency for Science (VEGA) under contract "Combinatorial Structures and Complexity of Algorithms," and by the Science and Technology Assistance Agency under contract APVT-20-004104. A preliminary version appeared in: Proc. Math. Found. Comput. Sci., Lect. Notes Comput. Sci., vol. 4162, Springer-Verlag, 2006, pp. 412–423.

E-mail address: viliam.geffert@upjs.sk (V. Geffert).

0890-5401/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.ic.2007.07.001

worst case, $d = 2^n$. (For examples of such languages, see [16,18,21].) On the other hand, we also know languages for which nondeterminism does not help at all, that is, $d = n$.

Thus, a natural question, raised for the first time by Iwama et al. [12], is the following:

> Is it possible, for a given number n, to find a number d satisfying $n \leqslant d \leqslant 2^n$, such that *no* optimal dfa using exactly d states can be simulated by *any* optimal nfa using exactly n states?

In [13], such numbers were named "**magic**," as numbers for which nondeterminism is especially weak. Adopting this kind of terminology, we say that d is a **muggle** number for n, if it is not magic for n, i.e., if there exists at least one optimal dfa using exactly d states which corresponds to an optimal nfa using exactly n states.

A negative answer to the question raised above would show that all numbers between n and 2^n are muggle numbers, which would give a complete and tight state hierarchy for the relation between dfa's and n-state nfa's, with no magic numbers leaving some "holes" in the hierarchy.

It turns out that the problem is easier to solve, if the size of the input alphabet is allowed to grow in n. Then there are no magic numbers at all [14]: For each n and d, such that $n \leqslant d \leqslant 2^n$, there exists an optimal n-state nfa for which the optimal dfa uses exactly d states. However, the size of the input alphabet for these automata is very large, namely, $2^{n-1}+1$. In the second part of [14], the input alphabet is reduced to $2n$. However, this result is shown by a "non-constructive" argument, proving the mere existence without giving an explicit exhibition of the witness regular languages. Finally, in [7], the complete state hierarchy is shown by a simpler "constructive" proof, displaying explicitly the witness automata and, at the same time, reducing the input alphabet size to $n+2$.

In chronological order, the first work concerning this problem [12] was devoted to binary regular languages, i.e., a family of languages with a fixed input alphabet. It was shown, for each n, that the values in the form $d = 2^n - 2^k$ or $d = 2^n - 2^k - 1$, where $0 \leqslant k \leqslant n/2-2$, are not magic. Later, in [13], this result has been extended for some new values of d, in the form $d = 2^n - k$, where $5 \leqslant k \leqslant 2n-2$, such that k satisfies also an additional coprimality condition. In [14], new muggle numbers have been found at the opposite end, namely, we have a binary witness language for each d satisfying $n \leqslant d \leqslant 1 + n\cdot(n+1)/2$. In addition, the paper extends the result for values in the form $d = n + 2^k - k$, where $1 \leqslant k \leqslant n$. Finally, in [7], a superpolynomial number of muggle numbers has been presented for the binary alphabet, namely, for each sufficiently large n, each d satisfying $n \leqslant d \leqslant e^{\Theta(n^{1/3}\cdot \ln^{2/3} n)}$ is muggle.

Summing up, for input alphabets growing in n, we have the complete state hierarchy, all numbers between n and 2^n are muggle numbers. On the other hand, there are many muggle numbers with binary witness languages, but no number has been shown to be magic in the binary case. The completeness of the state hierarchy for the regular languages over the binary alphabet (or any fixed input alphabet) is thus an open problem.

In this paper, we shall focus our attention on the state hierarchy of *unary* regular languages, i.e., on automata with a single letter input alphabet. Unary (tally) languages play an important role in theoretical computer science, as languages with a very low information content. Many of their properties are quite different from the general or binary case. (See, e.g., [3,4,5,6,8,17].)

For unary regular languages, we first present an almost exact approximation of $G_{\max}(n)$ and $G_{\min}(n)$, the largest and the smallest muggle numbers for n, respectively. Then we shall prove the existence of magic numbers between $G_{\min}(n)$ and $G_{\max}(n)$. Thus, in the unary case, the state hierarchy of dfa's, for the family of languages accepted by n-state nfa's, is not contiguous. We shall actually show a much stronger result, namely, that *most of the numbers* between $G_{\min}(n)$ and $G_{\max}(n)$ are magic. (A typical structure of the state hierarchy is shown in Fig. 2.)

In order to prove the existence of magic numbers for unary automata, we need to revise some of their properties first. In 1986, Chrobak [4] introduced a new normal form for unary nfa's. It was shown that each unary nfa M using at most n states can be replaced by an equivalent nfa M' consisting of an initial deterministic segment of length $O(n^2)$ and some m disjoint deterministic loops of lengths $\tilde{\ell}_1, \ldots, \tilde{\ell}_m$, with the total number of states in loops bounded by $\tilde{\ell}_1 + \cdots + \tilde{\ell}_m \leqslant n$. M' makes only a single nondeterministic decision, after passing through the initial segment, when it chooses one of the m loops.

This normal form reduced the cost of eliminating nondeterminism in the unary case, to $F(n) + O(n^2) \leqslant e^{(1+o(1))\cdot\sqrt{n\cdot\ln n}}$ states, where $F(n)$ denotes the Landau's function. (For exact definition of this function, see (1) in

Section 2.) This is far better than 2^n states required by the standard subset construction.[1] Chrobak presented also a lower bound, namely, $F(n-1)$, however, without a proof showing that the witness regular language cannot be replaced by a better language, and hence without a proof that the lower bound cannot be lifted up.

We shall need to introduce a more refined version of the Chrobak normal form, namely, we are going to reduce the length of the initial segment to n^2-2 states and, at the same time, the number of states in loops to $n-1$. This will reduce the cost of eliminating nondeterminism from $F(n)+O(n^2)$ states to $F(n-1)+(n^2-2)$.

Such improvement seems to be marginal at first glance. However, in Section 5, we shall present an *optimal* nfa using *exactly n* states, such that its *optimal* deterministic counterpart uses *exactly* $F(n-1)+k_n$ states, for some $k_n \in \{0,\ldots,n^2-2\}$. Thus, the new upper bound presented here is almost equal to the actually existing optimum, and hence the cost of eliminating nondeterminism is determined almost exactly. A potential difference of at most n^2-2 states is negligible, compared with the growth rate of $F(n)$, which is $e^{(1+o(1))\cdot\sqrt{n\cdot\ln n}}$. (The difference between $F(n)$ and $F(n-1)$ is much more significant.)

This improvement is required, among others, to obtain a sufficiently precise approximation for the largest muggle number, namely, $G_{\max}(n) = F(n-1)+k'_n$, for some $k'_n \in \{0,\ldots,n^2-2\}$.

Then we shall derive some properties of unary automata that are deterministic. Among others, an optimal dfa consisting of an initial segment of length s and a loop of length ℓ cannot be simulated by a dfa using a shorter initial segment or a loop of length ℓ' that factorizes into a "simpler" product of prime powers than does ℓ, even if the new machine uses far more states in total.

After that, utilizing the conflict between the properties of optimal dfa's and the revised Chrobak normal form for nfa's, we can prove the existence of magic numbers. As a simple consequence, we shall also get a new universal lower bound for the conversion of unary d-state dfa's into equivalent nfa's: nondeterminism does not reduce the number of states below $\log^2 d$, not even in the best case.

2. Preliminaries

We first briefly recall some basic definitions and notation on finite state automata. For more details, we refer the reader to [10,11], or any other standard textbook.

A *nondeterministic finite automaton* (*nfa*) is a quintuple $M=(Q,\Sigma,\delta,q_S,F)$, where Q denotes a finite set of states, Σ is a finite set of input symbols, $\delta: Q\times\Sigma\to 2^Q$ a transition function, $q_S\in Q$ an initial state, and $F\subseteq Q$ a set of final (accepting) states. The language accepted by M will be denoted, as usual, by $L(M)$.

The automaton M is *deterministic* (*dfa*), if $\|\delta(q,s)\|=1$, for each $q\in Q$ and each $s\in\Sigma$. Throughout the paper, $\|X\|$ denotes the cardinality of the set X.

Two automata are *equivalent*, if they accept the same language. An nfa (dfa) M is *optimal*, if no nfa (dfa, respectively) with fewer than $\|Q\|$ states is equivalent to M. (It is well known that the optimal dfa is unique, for each regular language, but we may potentially have several different optimal nfa's for the same language.)

All automata in this paper are unary, so we can fix the input alphabet to $\Sigma=\{1\}$. We also simplify the notation for transitions, a single-step transition $q'\in\delta(q,1)$ is presented in the form of an edge $q\to q'$. A path beginning in q, ending in q', and reading a string 1^u from the input (thus, consisting of u consecutive edges) can be displayed in a more compact form $q\xrightarrow{1^u}q'$. Finally, $q\rightsquigarrow q'$ indicates reachability by a path of any length (including zero, for $q=q'$).

The factorization of integers will also be important in the subsequent considerations. For a more detailed exposition concerning number theory, the reader is referred to [9,23].

Let X be a finite multiset of positive integers, with possible repetition of elements. Then $\operatorname{lcm} X$ denotes the least common multiple of all elements in X, $\gcd X$ the greatest common divisor of these elements, $\max X$ the largest element in X, and $\min X$ the smallest element.

The Fundamental Theorem of Arithmetic states that each $\ell>1$ can be uniquely *factorized* in the form $\ell = p_{i_1}^{\alpha_1}\cdot p_{i_2}^{\alpha_2}\cdot\ldots\cdot p_{i_e}^{\alpha_e}$, where $p_{i_1}^{\alpha_1}, p_{i_2}^{\alpha_2},\ldots,p_{i_e}^{\alpha_e}$ are some prime powers, with $\alpha_i\geqslant 1$ and $p_{i_f}\neq p_{i_g}$ for $f\neq g$. The set of

[1] In chronological order, this normal form was first presented by Ljubič in [15], together with upper and lower bounds for eliminating nondeterminism, of order $e^{O(\sqrt{n\cdot\ln n})}$ and $e^{\Omega(\sqrt{n\cdot\ln n})}$, respectively. However, the normal form presented by Chrobak in [4] produces tighter upper and lower bounds, both for nfa's and their deterministic counterparts.

prime powers used in the factorization of ℓ will be denoted by $\varphi(\ell) = \{p_{i_1}^{\alpha_1}, p_{i_2}^{\alpha_2}, \ldots, p_{i_e}^{\alpha_e}\}$. Clearly, $\ell = \prod_{p^\alpha \in \varphi(\ell)} p^\alpha$. We shall also need a *cost of factorization*, defined by

$$\Phi(\ell) = \sum_{p^\alpha \in \varphi(\ell)} p^\alpha .$$

By definition, $\varphi(1) = \emptyset$, which gives $\Phi(1) = 0$. The following simple properties of the factorization cost will be required later.

Lemma 2.1. *For each* $\ell \geqslant 1$, $\Phi(\ell) \leqslant \ell$. *If, moreover,* ℓ *divides some* ℓ', *then* $\Phi(\ell) \leqslant \Phi(\ell')$.

Proof. The statement is trivial for $\ell = 1$. If $\ell > 1$, then $\Phi(\ell) = \sum_{p^\alpha \in \varphi(\ell)} p^\alpha \leqslant \prod_{p^\alpha \in \varphi(\ell)} p^\alpha = \ell$, using the fact that prime powers are numbers greater than or equal to 2, and hence their sum cannot be larger than their product.

Second, if ℓ divides some ℓ', then, for each $p^\alpha \in \varphi(\ell)$, there must exist some $p^{\alpha'} \in \varphi(\ell')$, such that $\alpha' \geqslant \alpha$. Therefore, $\Phi(\ell) = \sum_{p^\alpha \in \varphi(\ell)} p^\alpha \leqslant \sum_{p^{\alpha'} \in \varphi(\ell')} p^{\alpha'} = \Phi(\ell')$. □

Lemma 2.2. *For each* $\ell_1, \ell_2, \ldots, \ell_m$, $\Phi(\mathrm{lcm}\{\ell_1, \ell_2, \ldots, \ell_m\}) \leqslant \sum_{i=1}^m \Phi(\ell_i)$.

Proof. Clearly, if a prime power $p^\alpha \in \varphi(\mathrm{lcm}\{\ell_1, \ldots, \ell_m\})$, it must appear in a factorization of at least one of the numbers $\ell_1, \ldots, \ell_m$. That is, $p^\alpha \in \varphi(\ell_i)$, for some $i \in \{1, \ldots, m\}$. Therefore, $\varphi(\mathrm{lcm}\{\ell_1, \ldots, \ell_m\}) \subseteq \varphi(\ell_1) \cup \ldots \cup \varphi(\ell_m)$. But then

$$\Phi(\mathrm{lcm}\{\ell_1, \ldots, \ell_m\}) = \sum_{p^\alpha \in \varphi(\mathrm{lcm}\{\ell_1, \ldots, \ell_m\})} p^\alpha \leqslant \sum_{p^\alpha \in \varphi(\ell_1) \cup \ldots \cup \varphi(\ell_m)} p^\alpha$$

$$\leqslant \sum_{i=1}^m \sum_{p^\alpha \in \varphi(\ell_i)} p^\alpha = \sum_{i=1}^m \Phi(\ell_i) . \quad \square$$

The following function will play a crucial role in our considerations. Let

$$F(n) = \max\{\mathrm{lcm}\{\ell_1, \ell_2, \ldots, \ell_m\} : \ell_1 + \ell_2 + \cdots + \ell_m = n\} . \tag{1}$$

This function, giving the largest least common multiple among all partitions of n, is known as Landau's function. It already plays an important role in the group theory. The best known approximation of $F(n)$ is due to Szalay [22]: $F(n) = \mathrm{e}^{\sqrt{n \cdot (\ln n + \ln\ln n - 1 + (\ln\ln n - 2 + o(1))/\ln n)}}$. For our purposes, this approximation can be simplified as follows:

$$F(n) = \mathrm{e}^{(1+o(1)) \cdot \sqrt{n \cdot \ln n}} . \tag{2}$$

The asymptotic notation in the formula above (and, in the same way, throughout the paper) is interpreted as follows. The exact value of $F(n)$ can be expressed in the form $F(n) = \mathrm{e}^{(1+r(n)) \cdot \sqrt{n \cdot \ln n}}$, where $r(n)$ is a real function satisfying $\lim_{n \to \infty} r(n) = 0$ and $r(n) \geqslant 0$, for each n. If we did not guarantee the condition $r(n) \geqslant 0$, we should use $\mathrm{e}^{(1 \pm o(1)) \cdot \sqrt{n \cdot \ln n}}$ instead. The meaning of the expression $\mathrm{e}^{(1-o(1)) \cdot \sqrt{n \cdot \ln n}}$ should be obvious.

3. Unary nondeterministic automata

This section is devoted to a more refined version of the so called Chrobak normal form for unary nfa's.

We shall begin with an auxiliary lemma stating that each computation path passing through a state q can be made to execute a loop beginning and ending in this state, provided that such a loop does exist, and that the remaining part of the computation is sufficiently long. (The lemma is a variant of the Dominant Loop Theorem, presented in [5].)

Lemma 3.1. *Let M be a unary nfa with at most n states. Then, if there exists a computation path $q_1 \xrightarrow{1^\alpha} q \xrightarrow{1^\beta} q_2$ for some states q_1, q, q_2 in M, and if there also exists a loop $q \xrightarrow{1^\ell} q$, such that $\beta \geqslant n \cdot \ell$, the path $q_1 \xrightarrow{1^\alpha} q \xrightarrow{1^\beta} q_2$ can be replaced by an equivalent path $q_1 \xrightarrow{1^\alpha} q \xrightarrow{1^\ell} q \xrightarrow{1^{\beta-\ell}} q_2$.*

Proof. Since $\beta \geqslant n \cdot \ell$, the string 1^β can be decomposed into at least n segments, all of equal length ℓ, except for some residual part of length ψ at the end, where $0 \leqslant \psi < \ell$. Thus, the path from q_1 to q_2 can be represented in the form

$$q_1 \xrightarrow{1^\alpha} q = r_0 \xrightarrow{1^\ell} r_1 \xrightarrow{1^\ell} r_2 \quad \ldots \quad r_{a-1} \xrightarrow{1^\ell} r_a \xrightarrow{1^\psi} q_2 \,,$$

where $a \geqslant n$ and $r_0, r_1, \ldots, r_a$ are some states. Hence, the segment between r_0 and q_2 contains more than n states, placed exactly ℓ positions apart. Using a simple pigeonhole argument, some state must be repeated, i.e., we have $r_i = r_j$, for some $0 \leqslant i < j \leqslant a$. Therefore, we must pass through a loop of length $(j-i) \cdot \ell$, beginning and ending in $r_i = r_j$. This allows us to divide the segment 1^β into three segments $1^{\beta_1}, 1^{(j-i)\cdot\ell}, 1^{\beta_2}$, corresponding, respectively, to the computations executed before this loop, the loop itself, and the part following this loop:

$$q_1 \xrightarrow{1^\alpha} q = r_0 \xrightarrow{1^{\beta_1}} r_i \xrightarrow{1^{(j-i)\cdot\ell}} r_j \xrightarrow{1^{\beta_2}} q_2 \,.$$

But then the original path can be replaced by an equivalent computation — beginning and ending in the same states and reading the same portion of the input — in which the loop of length $(j-i) \cdot \ell$ is replaced by the loop $q \xrightarrow{1^\ell} q$, iterated $j-i$ times:

$$q_1 \xrightarrow{1^\alpha} \underbrace{q \xrightarrow{1^\ell} q \xrightarrow{1^\ell} q \quad \ldots \quad q \xrightarrow{1^\ell} q}_{j-i \text{ times}} = r_0 \xrightarrow{1^{\beta_1}} r_i = r_j \xrightarrow{1^{\beta_2}} q_2 \,.$$

Since $i < j$, the loop $q \xrightarrow{1^\ell} q$ is iterated at least once. Thus, we have obtained a computation path $q_1 \xrightarrow{1^\alpha} q \xrightarrow{1^\ell} q \xrightarrow{1^{\beta'}} q_2$, where $\beta' = (j-i-1) \cdot \ell + \beta_1 + \beta_2 = \beta - \ell$, which completes the proof. □

Definition 3.2 (*Cardinal loops and states*)**.** Let M be a unary nfa with at most n states. Fix some loops in M, together with some states along these loops, as *cardinal*, in the following way. (There are several cases to consider, illustrated by Fig. 1.)

Case (a). There is no path $q_S \xrightarrow{1^\alpha} q_S$ beginning and ending in the initial state q_S, for no $\alpha > 0$. That is, no loop passes through q_S, and hence q_S does not belong to any strongly connected component. Partition the state set into $Q = Q_0 \cup Q_1 \cup \ldots \cup Q_m \cup Q_\infty$ as follows:

- Q_0 contains all states that are reachable from q_S, but not reachable by computation paths passing through some loops. This implies, among others, that the states in Q_0 do not belong to any strongly connected component, and that $q_S \in Q_0$.
- Q_i, for $i = 1, \ldots, m$, contains all states forming the ith strongly connected component in Q, reachable directly from Q_0. That is, if $q \in Q_i$, for some i, then *(i)* all states q' with paths $q \rightsquigarrow q' \rightsquigarrow q$ are included in Q_i, and *(ii)* there must exist a path $q_S \rightsquigarrow q$ consisting only of states in $Q_0 \cup Q_i$.
- Q_∞ contains all remaining states in Q. That is, the states that are either not reachable at all, or the states reachable only by computation paths passing through some states in $\bigcup_{i=1}^m Q_i$, but not belonging to $Q_0 \cup \bigcup_{i=1}^m Q_i$.

Now, for each $i = 1, \ldots, m$, let $\tilde{\ell}_i$ denote the length of the shortest loop in Q_i. For each Q_i, fix one such loop (there may potentially be more than one), and also a state $\tilde{q}_i \in Q_i$ along this loop. The fixed loops of lengths $\tilde{\ell}_1, \tilde{\ell}_2, \ldots, \tilde{\ell}_m$ will be *cardinal loops*, the states $\tilde{q}_1, \tilde{q}_2, \ldots, \tilde{q}_m$ fixing some positions along these loops will be *cardinal states*. (If Q does not contain any reachable strongly connected component, which can happen only if M accepts a finite language, there are no cardinal loops or states, i.e., $m = 0$ and $Q = Q_0 \cup Q_\infty$.)

It is easy to see that all these loops are *elementary*, i.e., no states are repeated in the course of a single iteration in any of them. Moreover, since these loops are in pairwise disjoint components and there is no loop passing through q_S,

$$\tilde{\ell}_1 + \tilde{\ell}_2 + \cdots + \tilde{\ell}_m \leqslant n-1 \,. \tag{3}$$

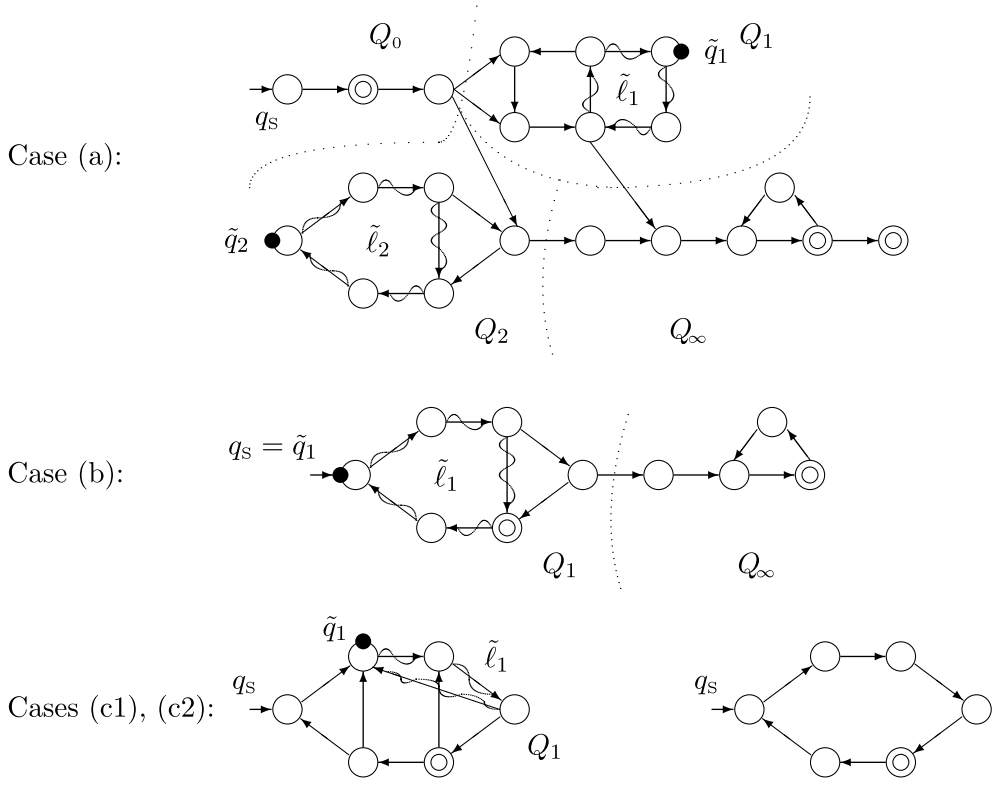

Fig. 1. Fixing cardinal loops and cardinal states. Partitioning of the state set into $Q = Q_0 \cup Q_1 \cup \ldots \cup Q_m \cup Q_\infty$ is represented by dotted territorial boundaries, cardinal loops are marked by wavy lines twining around the edges, and cardinal states by filled bullets.

Case (b). There exists a loop $q_S \xrightarrow{1^\alpha} q_S$ beginning and ending in the initial state q_S, for some $1 \leqslant \alpha \leqslant n-1$. That is, q_S belongs to some strongly connected component. Here, we get a partition in the form $Q = Q_1 \cup Q_\infty$, that is, $Q_0 = \emptyset$, $q_S \in Q_1$, and hence $m = 1$. Therefore, we fix some shortest loop beginning and ending in q_S as cardinal, of length $\tilde{\ell}_1$, and take $\tilde{q}_1 = q_S$ as the only cardinal state. It is easy to see that the condition (3) is satisfied again, since $\sum_{i=1}^{m} \tilde{\ell}_i = \tilde{\ell}_1 \leqslant \alpha \leqslant n-1$.

Case (c). There exists a loop beginning and ending in q_S, but the length of any such loop is at least n. Since the shortest loop does not repeat the same state twice, we have a path $q_S \xrightarrow{1^n} q_S$, visiting all states in Q. Therefore, the entire state set Q forms a single strongly connected component Q_1. By enumerating all states in order in which they appear in the loop $q_S \xrightarrow{1^n} q_S$, we get

$$q_S = q_0 \to q_1 \to q_2 \quad \ldots \quad q_{n-2} \to q_{n-1} \to q_0 \,. \tag{4}$$

There are now two subcases:

Case (c1). Besides the transitions displayed in (4), there exists at least one more edge in M. Such edge must be of the form $q_e \leftarrow q_f$, with $1 \leqslant e \leqslant f \leqslant n-1$. (Otherwise, we get either an edge already displayed in (4), or a loop $q_S \rightsquigarrow q_S$ shorter than n. Both cases lead to contradictions.)

Now, let $\bar{e}$ be the smallest $e \geqslant 1$ such that there exists a "backward" edge $q_{\bar{e}} \leftarrow q_f$, for some $f \geqslant \bar{e}$. Then let $\bar{f}$ be the smallest $f \geqslant \bar{e}$ with a backward edge $q_{\bar{e}} \leftarrow q_{\bar{f}}$. Finally, fix the loop $q_{\bar{e}} \to q_{\bar{e}+1} \quad \ldots \quad q_{\bar{f}-1} \to q_{\bar{f}} \to q_{\bar{e}}$ as cardinal, and fix $\tilde{q}_1 = q_{\bar{e}}$ as the only cardinal state. Clearly, the condition (3) is satisfied even in this case, since the only cardinal loop is of length $\tilde{\ell}_1 = \bar{f} - \bar{e} + 1 \leqslant n-1$.

Case (c2). The automaton M does not have any transitions except for those displayed in (4). In this case, M is already *deterministic*. We shall call such automaton a *trivial loop of length n*, and handle this special case separately. (Among others, here we do not try to fix any cardinal loops or states.)

The following technical theorem serves as a tool for converting nondeterministic automata into a normal form.

Theorem 3.3. *Let M be a unary nfa with at most $n > 1$ states, different from the trivial loop of length n, with cardinal loops of lengths $\tilde{\ell}_1, \tilde{\ell}_2, \ldots, \tilde{\ell}_m$ and cardinal states $\tilde{q}_1, \tilde{q}_2, \ldots, \tilde{q}_m$, as introduced by Def. 3.2. Then, for each $u \geqslant n^2-2$, the string 1^u is accepted by M if and only if*

(i) *there exists an $i \in \{1, \ldots, m\}$ (that is, also a strongly connected component Q_i with a cardinal loop $\tilde{\ell}_i$ and a cardinal state $\tilde{q}_i$),*
(ii) *there exists an $r \in \{0, \ldots, \tilde{\ell}_i - 1\}$ (a number modulo $\tilde{\ell}_i$),*
(iii) *there exists a $q' \in F$ (a final state of M),*
(iv) *there exists an $\alpha \leqslant n-1$, with a computation path $q_S \xrightarrow{1^\alpha} \tilde{q}_i$,*
(v) *there exists a $\beta \leqslant n^2-n-1$, with a computation path $\tilde{q}_i \xrightarrow{1^\beta} q'$,*
(vi) *such that $(\alpha+\beta) \bmod \tilde{\ell}_i = r$,*
(vii) *and $u \bmod \tilde{\ell}_i = r$.*

Proof. The argument for the "$\Longrightarrow$" part follows the cases (a), (b), (c1), and (c2), introduced by Definition 3.2 for fixing the cardinal loops and states.

Case (a), part I. Here we assume that there is no loop beginning and ending in the initial state q_S, that is, $q_S \in Q_0$. If 1^u is accepted by M, there must exist an accepting path, in the form $q_S \xrightarrow{1^u} q'$, for some $q' \in F$. This gives *a final state of the item (iii)*.

Since $u \geqslant n^2-2 \geqslant n$, some states must be repeated along the path $q_S \xrightarrow{1^u} q'$, consisting of at least $n+1$ states. Thus, the path must pass through at least one state not belonging to Q_0. (See also Definition 3.2 and Fig. 1.) On the other hand, by assumption, $q_S \in Q_0$. Now, take the first state $p \notin Q_0$ along this path, which gives

$$\underbrace{q_S \xrightarrow{1^{\alpha_0}}}_{\text{in } Q_0} p \xrightarrow{1^{\beta_0}} q', \tag{5}$$

for some α_0, β_0 satisfying $\alpha_0+\beta_0 = u$, such that all states in the segment $q_S \xrightarrow{1^{\alpha_0}} p$ belong to Q_0, except for the state p at the very end. Since $p \notin Q_0$ is reachable directly from Q_0, we have that $p \in Q_i$, for some strongly connected component Q_i. (That is, $p \notin Q_\infty$.) This gives *an $i \in \{1, \ldots, m\}$ of the item (i)*.

Note that both p and the cardinal state $\tilde{q}_i$ are in the same strongly connected component Q_i. Thus, M must also have a path from p to $\tilde{q}_i$, in the form

$$p = r_0 \to r_1 \to r_2 \quad \ldots \quad r_{a-1} \to r_a = \tilde{q}_i\,, \tag{6}$$

where $r_0, r_1, r_2, \ldots, r_a$ are states along the shortest path $p \rightsquigarrow \tilde{q}_i$. Combining this with (5), we shall show, by induction on $j = 0, 1, \ldots, a$, the existence of the path

$$q_S \underbrace{\rightsquigarrow p = r_0 \to r_1 \quad \ldots \quad r_{j-1} \to}_{1^{\alpha_j}} r_j \underbrace{\rightsquigarrow}_{1^{\beta_j}} q', \tag{7}$$

for some α_j, β_j satisfying $\alpha_j+\beta_j = u$. That is, the new path begins and ends in the same states as the original $q_S \xrightarrow{1^u} q'$, reading the same portion of the input. So far, the induction hypothesis has been shown for $j=0$. We are now going to extend it from j to $j+1$.

First, recall that all states in the segment $q_S \rightsquigarrow p$, except for p, are in Q_0, and hence they are all different, since the states in Q_0 do not belong to any strongly connected component. Second, all states in the segment $r_0 \rightsquigarrow r_j$ are in Q_i, and they are also all different, since (6) represents the shortest path connecting p with $\tilde{q}_i$. Third, $Q_0 \cap Q_i = \emptyset$, and hence the states in the path $q_S \xrightarrow{1^{\alpha_j}} r_j$ must be all different. Thus, the length of this path, measured in the number of states, is at most n. Moreover, if $j < a$, this length does not exceed $n-1$, since the state r_a is excluded. Expressing this length in the number of edges, we get

$$\begin{aligned} &\alpha_j \leqslant n-2\,, \quad \text{if } j < a\,, \\ &\alpha_j \leqslant n-1\,, \quad \text{if } j = a\,. \end{aligned} \tag{8}$$

Consider now $j < a$. Here, we can utilize the fact that both r_j and r_{j+1} belong to the same strongly connected component Q_i. Thus, there exists a path

$$r_j \to r_{j+1} \rightsquigarrow r_j \,.$$

Using the shortest path of this kind, we get that the segment $r_{j+1} \rightsquigarrow r_j$ does not repeat the same state twice, nor does it contain r_j or r_{j+1}, except for the very beginning and very end. But then $r_j \to r_{j+1} \rightsquigarrow r_j$ is an elementary loop, of length $\psi \leqslant n-1$. (Recall that q_S does not lie along any loop, which saves one state.)

Using (8), we then get $\beta_j = u - \alpha_j \geqslant (n^2-2) - (n-2) = n \cdot (n-1) \geqslant n \cdot \psi$. This allows us to use Lemma 3.1 and replace the path in (7) by an equivalent path iterating, at least once, the loop $r_j \to r_{j+1} \xrightarrow{1^{\psi-1}} r_j$:

$$q_S \underbrace{\rightsquigarrow p = r_0 \to r_1 \ \ldots \ r_{j-1} \to r_j \to}_{1^{\alpha_{j+1}}} r_{j+1} \underbrace{\xrightarrow{1^{\psi-1}} r_j \rightsquigarrow}_{1^{\beta_{j+1}}} q' .$$

Thus, we have $\alpha_{j+1}, \beta_{j+1}$ satisfying $\alpha_{j+1} + \beta_{j+1} = u$. This argument can be repeated for $j = 0, 1, \ldots, a-1$, which gives

$$q_S \underbrace{\rightsquigarrow p = r_0 \to r_1 \ \ldots \ r_{a-1} \to}_{1^{\alpha_a}} r_a = \tilde{q}_i \underbrace{\rightsquigarrow}_{1^{\beta_a}} q' ,$$

with $\alpha_a + \beta_a = u$ and $\alpha_a \leqslant n-1$, using (7) and (8), for $j = a$.

Now we can take $\alpha = \alpha_a$, which gives an $\alpha \leqslant n-1$ *of the item (iv)*, together with a path $q_S \xrightarrow{1^\alpha} \tilde{q}_i$. Finally, let $\beta'_0 = \beta_a$.

Case (a), part II. Now the input 1^u is accepted by a new computation path, passing through a cardinal state, in the form $q_S \xrightarrow{1^\alpha} \tilde{q}_i \xrightarrow{1^{\beta'_0}} q'$, with $\alpha + \beta'_0 = u$.

Recall that the state $\tilde{q}_i$ fixes some position along the cardinal loop of length $\tilde{\ell}_i$. Thus, we have also $\tilde{q}_i \xrightarrow{1^{\tilde{\ell}_i}} \tilde{q}_i$. But then the accepting computation can be expressed in the form

$$q_S \xrightarrow{1^\alpha} \underbrace{\tilde{q}_i \xrightarrow{1^{\tilde{\ell}_i}} \tilde{q}_i \ \ldots \ \tilde{q}_i \xrightarrow{1^{\tilde{\ell}_i}} \tilde{q}_i}_{k \text{ times}} \xrightarrow{1^{\beta'_k}} q' , \tag{9}$$

where the loop $\tilde{q}_i \xrightarrow{1^{\tilde{\ell}_i}} \tilde{q}_i$ is iterated k times, for some $k \geqslant 0$, and for some residual path, of length β'_k, following this iteration. So far, this statement has been shown only for $k = 0$.

However, if $\beta'_k \geqslant n \cdot \tilde{\ell}_i$, we can use Lemma 3.1 and replace the path $\tilde{q}_i \xrightarrow{1^{\beta'_k}} q'$ by an equivalent path iterating, at least once, the loop $\tilde{q}_i \xrightarrow{1^{\ell_i}} \tilde{q}_i$:

$$q_S \xrightarrow{1^\alpha} \underbrace{\tilde{q}_i \xrightarrow{1^{\tilde{\ell}_i}} \tilde{q}_i \ \ldots \ \tilde{q}_i \xrightarrow{1^{\tilde{\ell}_i}} \tilde{q}_i}_{k \text{ times}} \xrightarrow{1^{\tilde{\ell}_i}} \tilde{q}_i \xrightarrow{1^{\beta'_{k+1}}} q' .$$

Now the loop of length $\tilde{\ell}_i$ is iterated $k+1$ times, followed by a residual computation path of length β'_{k+1}, such that $\tilde{\ell}_i + \beta'_{k+1} = \beta'_k$.

This argument can be repeated for $k = 0, 1, 2, \ldots$, until we get, for some k, a value $\beta'_k < n \cdot \tilde{\ell}_i$. But then, using (3), we get that $\beta'_k \leqslant n \cdot \tilde{\ell}_i - 1 \leqslant n \cdot (n-1) - 1 = n^2 - n - 1$. The final value $\beta = \beta'_k$ gives a $\beta \leqslant n^2 - n - 1$ *of the item (v)*, together with a path $\tilde{q}_i \xrightarrow{1^\beta} q'$.

Finally, let $r = u \bmod \tilde{\ell}_i$. This gives *an* $r \in \{0, \ldots, \tilde{\ell}_i - 1\}$ *of the item* (*ii*), satisfying also *the condition of the item* (*vii*). It only remains to show that *the condition of the item* (*vi*) is valid as well. However, using $\beta = \beta'_k$ together with (9), we get that u can be expressed in the form $u = \alpha + k \cdot \tilde{\ell}_i + \beta$, for some $k \geqslant 0$, and hence $(\alpha + \beta) \bmod \tilde{\ell}_i = u \bmod \tilde{\ell}_i = r$.

This completes the argument for the "$\Longrightarrow$" part, Case (a).

Case (b). Assume now that there exists a loop beginning and ending in q_S, of length at most $n-1$. Here, we have only one strongly connected component that can be reached directly from q_S, namely, Q_1, with $\tilde{q}_1 = q_S \in Q_1$, and $\tilde{\ell}_1 \leqslant n-1$. But then the accepting computation path starts already in the cardinal state $\tilde{q}_1 = q_S$. Therefore,

it can be expressed in the form $q_S \xrightarrow{1^\alpha} \tilde{q}_i \xrightarrow{1^{\beta'_0}} q'$, where $\alpha = 0 \leqslant n-1$, $i = 1$, q' is a final state, and $1^{\beta'_0}$ is the entire input string 1^u. The rest of the argument is the same as in Case (a), part II.

Case (c1). Here we assume that the shortest loop $q_S \rightsquigarrow q_S$ is of length n (i.e., all states are in a single component Q_1), but there exists a shorter loop not passing through q_S. Recall that the cardinal loop of length $\tilde{\ell}_1 \leqslant n-1$ has been fixed by using a backward edge $q_{\bar{e}} \leftarrow q_{\bar{f}}$ with the smallest possible value of $\bar{e} \geqslant 1$, and that $\tilde{q}_1 = q_{\bar{e}}$.

Clearly, any path $q_S \xrightarrow{1^u} q'$ accepting the input of length $u \geqslant n^2-2$ must repeat some states. Therefore, it must also pass through at least one backward edge $q_f \to q_e$, with $f \geqslant e$, that is, through an edge *not increasing* the index of the state. (Otherwise, the path length is bounded by n.) All edges preceding the first backward edge along the path $q_S \xrightarrow{1^u} q'$ are forward edges, incrementing the state index by one. Thus, the path is in the form

$$q_S = q_0 \to q_1 \to q_2 \ \ \dots \ \ \dots \ \ q_{f-1} \to \underbrace{q_f \to q_e}_{\text{backwards for the first time}} \rightsquigarrow q'.$$

If $e \geqslant 1$, the state q_e is a target state of a backward edge in M that does not point to q_0. But the state $q_{\bar{e}}$ is the state with the smallest index among all such states. Therefore, $\bar{e} \leqslant e \leqslant f$. If $e = 0$, then $f = n-1$, the largest index value. (Smaller values of f are covered by Case (b) above.) Therefore, $\bar{e} \leqslant n-1 = f$. In both cases, $\bar{e} \leqslant f$. But then the sequence $q_0, q_1, q_2, \dots, q_f$ must contain the state $q_{\bar{e}} = \tilde{q}_1$, which gives

$$q_S = q_0 \to q_1 \to q_2 \ \ \dots \ \ q_{\bar{e}} = \tilde{q}_1 \ \ \dots \ \ q_f \to q_e \rightsquigarrow q'.$$

By taking $\alpha = \bar{e} \leqslant f \leqslant n-1$, $i = 1$, and $\beta'_0 = u-\alpha$, we get $q_S \xrightarrow{1^\alpha} \tilde{q}_i \xrightarrow{1^{\beta'_0}} q'$. The rest of the argument proceeds in the same way as in Case (a), part II.

Case (c2). This is here only for completeness. By assumption of the theorem, M is different from the trivial loop of length n, and hence this case has been excluded.

The "$\Longleftarrow$" part. Suppose that some $u \geqslant n^2-2$ satisfies the conditions listed in the items (i)–(vii). First, $u \geqslant n^2-2 = (n-1) + (n^2-n-1) \geqslant \alpha+\beta$, by items (iv) and (v). Second, $u \bmod \tilde{\ell}_i = r = (\alpha+\beta) \bmod \tilde{\ell}_i$, using (vii) and (vi), for i and r introduced by (i) and (ii). Therefore, u can be expressed in the form $u = \alpha + \beta + k \cdot \tilde{\ell}_i$, for some $k \geqslant 0$.

But then, by the use of the computation paths introduced by (iv) and (v), together with the cardinal loop $\tilde{q}_i \xrightarrow{1^{\tilde{\ell}_i}} \tilde{q}_i$, iterated k times, we can compose the path

$$q_S \xrightarrow{1^\alpha} \underbrace{\tilde{q}_i \xrightarrow{1^{\tilde{\ell}_i}} \tilde{q}_i \ \ \dots \ \ \tilde{q}_i \xrightarrow{1^{\tilde{\ell}_i}} \tilde{q}_i}_{k \text{ times}} \xrightarrow{1^\beta} q',$$

where q' is a final state, by (iii). Thus, the input 1^u is accepted by M, which completes the proof. □

Using the above theorem, we can fix some "significant" parts of nondeterministic computations so that they can be "precomputed" in advance.

Definition 3.4. Let M be a unary nfa with at most n states, different from the trivial loop of length n, with cardinal loops of lengths $\tilde{\ell}_1, \tilde{\ell}_2, \dots, \tilde{\ell}_m$ and cardinal states $\tilde{q}_1, \tilde{q}_2, \dots, \tilde{q}_m$. For each $i \in \{1, \dots, m\}$ and each $r \in \{0, \dots, \tilde{\ell}_i - 1\}$, define

- a boolean predicate $P_{i,r}$ = *true*/*false*, depending on whether
  - (iii) there exists a $q' \in F$,
  - (iv) there exists an $\alpha \leqslant n-1$, with a computation path $q_S \xrightarrow{1^\alpha} \tilde{q}_i$,
  - (v) there exists a $\beta \leqslant n^2-n-1$, with a computation path $\tilde{q}_i \xrightarrow{1^\beta} q'$,
  - (vi) such that $(\alpha+\beta) \bmod \tilde{\ell}_i = r$,
- a set of indices $R_i = \{r' \in \{0, \dots, \tilde{\ell}_i - 1\} : P_{i,r'} = \textit{true}\}$,
- a language $L_{i,r} = \{1^u : u \geqslant n^2-2 \text{ and } u \bmod \tilde{\ell}_i = r\}$,
- a language $L_0 = \{1^u : u < n^2-2 \text{ and } u \in L(M)\}$.

It is easy to see that, for each given $i \in \{1, \dots, m\}$ and each given $r \in \{0, \dots, \tilde{\ell}_i - 1\}$, the truth of the predicate $P_{i,r}$ can be precomputed without knowing u, the length of the input. This only requires to iterate over all possible

values of $q' \in F$, $\alpha = 0, \ldots, n-1$, and $\beta = 0, \ldots, n^2-n-1$, and verify the existence of the paths $q_S \xrightarrow{1^\alpha} \tilde{q}_i$ and $\tilde{q}_i \xrightarrow{1^\beta} q'$, for any combination of these such that $(\alpha+\beta) \bmod \tilde{\ell}_i = r$. (This, in turn, only requires to know the transition table of M, together with the allocation of cardinal loops and states.) Having computed the truth of all predicates $P_{i,r}$, we can easily precompute the sets R_i as well. Now we are ready for converting nfa's into the normal form.

Theorem 3.5 (*Chrobak normal form revised*)**.** *Let M be a unary nfa with at most $n > 1$ states, different from the trivial loop of length n. Then M can be replaced by an equivalent nfa M' consisting of an initial deterministic path of length $\tilde{s} \leqslant n^2-2$, and some m disjoint deterministic loops of lengths $\tilde{\ell}_1, \ldots, \tilde{\ell}_m$, with the total number of states in loops bounded by $\tilde{\ell}_1+\cdots+\tilde{\ell}_m \leqslant n-1$. M' makes a single nondeterministic decision (if any), after passing through the initial path, when it chooses one of the m loops (if $m > 1$).*

Proof. By Theorem 3.3, the string 1^u of length $u \geqslant n^2-2$ is in $L(M)$ if and only if it satisfies the statement of the items (i)–(vii). Using notation of Definition 3.4, this holds if and only if there exist an $i \in \{1, \ldots, m\}$ and an $r \in \{0, \ldots, \tilde{\ell}_i-1\}$, such that $P_{i,r} = true$ and $1^u \in L_{i,r}$. This, in turn, holds if and only if there exists an $i \in \{1, \ldots, m\}$ and an $r \in R_i$, such that $1^u \in L_{i,r}$. But then, using the fact that no string shorter than n^2-2 is in any $L_{i,r}$, but all accepted short strings are in L_0, the language $L(M)$ can be expressed in the form

$$L(M) = L_0 \cup \bigcup_{i=1}^{m} \bigcup_{r \in R_i} L_{i,r} \,.$$

It is easy to construct a machine M' for $L_0 \cup \bigcup_{i=1}^{m} \bigcup_{r \in R_i} L_{i,r}$. It consists of

- an initial deterministic segment, made up of some states $p_1, p_2, \ldots, p_{n^2-2}$, connected by edges $p_k \to p_{k+1}$, for $k = 1, \ldots, n^2-3$, with p_1 as the initial state,
- a separate deterministic loop of length $\tilde{\ell}_i$, for each $i \in \{1, \ldots, m\}$, made up of some states $q_{i,0}, q_{i,1}, \ldots, q_{i,\tilde{\ell}_i-1}$, connected by edges $q_{i,k} \to q_{i,(k+1) \bmod \tilde{\ell}_i}$, for $k = 0, \ldots, \tilde{\ell}_i-1$,
- edges $p_{n^2-2} \to q_{i,0}$, for $i \in \{1, \ldots, m\}$, connecting the initial segment to each of the loops. This is the only nondeterministic decision, ever made.
- Finally, mark as accepting each state $q_{i,k}$ in the loop of length $\tilde{\ell}_i$ such that $r_{i,k} = (n^2-2+k) \bmod \tilde{\ell}_i \in R_i$, and
- mark as accepting each state p_k in the initial segment such that $1^{k-1} \in L(M)$.

Note that $p_1 \xrightarrow{1^u} q_{i,k}$ if and only if $u = (n^2-2+k) + j \cdot \tilde{\ell}_i$, for some $j \geqslant 0$. This, in turn, holds if and only if $u \geqslant n^2-2$ and $u \bmod \tilde{\ell}_i = r_{i,k}$, which holds if and only if $1^u \in L_{i,r_{i,k}}$. Since $q_{i,k}$ is an accepting state if and only if $r_{i,k} \in R_i$, the paths ending in the loop $\tilde{\ell}_i$ accept exactly the strings belonging to $\bigcup_{r \in R_i} L_{i,r}$. The combination of $\tilde{\ell}_1, \tilde{\ell}_2, \ldots, \tilde{\ell}_m$ gives $\bigcup_{i=1}^{m} \bigcup_{r \in R_i} L_{i,r}$, while the initial segment is responsible for short inputs in L_0.

Clearly, M' uses n^2-2 states in the initial segment, and $\tilde{\ell}_1+\cdots+\tilde{\ell}_m \leqslant n-1$ states are in the loops, by (3). □

Now we can make the above automaton deterministic.

Theorem 3.6. *Let M be a unary nfa with at most $n > 1$ states, different from the trivial loop of length n. Then M can be replaced by an equivalent dfa M'' consisting of an initial segment of length $\tilde{s} \leqslant n^2-2$, and a loop of length $\tilde{\ell}$ satisfying $\Phi(\tilde{\ell}) \leqslant n-1$.*

Proof. The deterministic automaton M'' is obtained by the standard subset construction [19], from the nfa M' constructed in Theorem 3.5.

First, this gives an initial segment of states $\{p_1\}, \{p_2\}, \ldots, \{p_{n^2-2}\}$, connected by edges $\{p_k\} \to \{p_{k+1}\}$, for $k = 1, \ldots, n^2-3$, since the initial segment in M' is deterministic. Second, we have the edge $\{p_{n^2-2}\} \to \{q_{1,0}, q_{2,0}, \ldots, q_{m,0}\}$, since M' nondeterministically chooses one of the m loops after passing through the initial segment.

The remaining edges are in the form $\{q_{1,e_1}, \ldots, q_{m,e_m}\} \to \{q_{1,f_1}, \ldots, q_{m,f_m}\}$, where $f_i = (e_i+1) \bmod \tilde{\ell}_i$, for each $i \in \{1, \ldots, m\}$, since $q_{1,e_1}, \ldots, q_{m,e_m}$ lie along the deterministic loops, of respective lengths $\tilde{\ell}_1, \ldots, \tilde{\ell}_m$. Thus, after passing through the initial segment, M'' enters a loop of length $\tilde{\ell} = \mathrm{lcm}\{\tilde{\ell}_1, \ldots, \tilde{\ell}_m\}$.

In general, the subset construction produces 2^t states, if the original nfa consists of t states. However, if we reduce the state set of M'' to the subset that is actually reachable from the initial state $\{p_1\}$, the total number of states is $(n^2-2) + \tilde{\ell}$.

Moreover, by Lemmas 2.2 and 2.1, the length $\tilde{\ell}$ has a very low factorization cost, bounded by $\Phi(\tilde{\ell}) = \Phi(\mathrm{lcm}\{\tilde{\ell}_1,\ldots,\tilde{\ell}_m\}) \leqslant \sum_{i=1}^{m}\Phi(\tilde{\ell}_i) \leqslant \sum_{i=1}^{m}\tilde{\ell}_i \leqslant n-1$. □

There is one special case that should not be forgotten. If the original nfa M does not contain any reachable strongly connected component (which can happen only if M accepts a finite language), the construction in Case (a) of Definition 3.2 does not fix any cardinal loops or states, i.e., $m = 0$. As a consequence, Theorem 3.5 gives an nfa M' with $m = 0$, that is, just an initial path consisting of n^2-2 states, with no loops.[2] But then, in the subset construction of Theorem 3.6, the edge $\{p_{n^2-2}\} \to \{q_{1,0}, q_{2,0}, \ldots, q_{m,0}\}$ degenerates into $\{p_{n^2-2}\} \to \emptyset$, where $\emptyset$ denotes the empty set, which is a rejecting state in M''. Consequently, we get also the edge $\emptyset \to \emptyset$, that is, a loop of length $\tilde{\ell} = 1$. But then $\Phi(\tilde{\ell}) = \Phi(1) = 0 \leqslant n-1$. Note also that here we have $\mathrm{lcm}\{\tilde{\ell}_1, \tilde{\ell}_2, \ldots, \tilde{\ell}_m\} = \mathrm{lcm}\,\emptyset = 1 = \tilde{\ell}$, taking, by definition, $\mathrm{lcm}\,\emptyset = 1$.

Theorem 3.7. *Let M be a unary nfa with at most $n > 1$ states. Then M can be replaced by an equivalent dfa M'' using at most $d \leqslant F(n-1) + (n^2-2) \leqslant e^{(1+o(1))\cdot\sqrt{n\cdot\ln n}}$ states.*

Proof. First, if M is different from the trivial loop of length n, M'' is obtained by the use of Theorem 3.6. This automaton uses $d = \tilde{\ell} + (n^2-2)$ states, with $\Phi(\tilde{\ell}) \leqslant n-1$. This allows us to express $\tilde{\ell}$ in the form $\tilde{\ell} = \mathrm{lcm}\{1,\ldots,1, p_{i_1}^{\alpha_1},\ldots,p_{i_e}^{\alpha_e}\}$, where "1" is repeated $(n-1) - \Phi(\tilde{\ell})$ times and $\{p_{i_1}^{\alpha_1},\ldots,p_{i_e}^{\alpha_e}\}$ is the set of prime powers forming the factorization of $\tilde{\ell}$. The sum of these numbers is exactly $1+\cdots+1+p_{i_1}^{\alpha_1}+\cdots+p_{i_e}^{\alpha_e} = (n-1) - \Phi(\tilde{\ell}) + \Phi(\tilde{\ell}) = n-1$. But then $\tilde{\ell} \leqslant \max\{\mathrm{lcm}\{\ell_1,\ldots,\ell_f\} : \ell_1+\cdots+\ell_f = n-1\} = F(n-1)$. Thus, using (2), the number of states in M'' can be bounded by $d = \tilde{\ell} + (n^2-2) \leqslant F(n-1) + (n^2-2) \leqslant e^{(1+o(1))\cdot\sqrt{n\cdot\ln n}}$.

For completeness, if M turns out to be the trivial loop of length n, then M is already deterministic, so we take $M'' = M$, with $d = n$. But then $d \leqslant F(n-1) + (n^2-2)$, if $n > 1$. □

We shall conclude this section by a simple by-product that will not be required later, but which we consider to be interesting by itself. It shows that a superpolynomial gap between the size of unary nfa's and dfa's can be obtained only by machines *not* having the initial state in any strongly connected component.

Corollary 3.8. *Let M be a unary nfa with at most n states, with a loop passing through its initial state. Then M can be replaced by an equivalent dfa with at most $O(n^2)$ states.*

Proof. Because of the loop beginning and ending in the initial state, Case (a) is excluded in the construction of Definition 3.2. This either fixes exactly one cardinal loop, i.e., $m = 1$, or M turns out to be the trivial loop of length n. In the first case, Theorem 3.6 gives an M'' using n^2-2 states in the initial segment, and with a loop of length $\tilde{\ell} = \mathrm{lcm}\{\tilde{\ell}_1,\ldots,\tilde{\ell}_m\} = \mathrm{lcm}\{\tilde{\ell}_1\} = \tilde{\ell}_1 \leqslant n-1$. Thus, M'' uses $O(n^2)$ states. The same upper bound holds even if M is a trivial loop of length n. □

4. Unary deterministic automata

Now we are going to derive some properties of unary automata that are deterministic. Before passing further, we need to prove the following technical lemma.

Lemma 4.1. *Let m_1, m_2, and g be some positive integers with $\gcd\{m_1, m_2\} = 1$. Then, for each $k \in \{0,\ldots,g-1\}$,*

$$\{(k + i\cdot m_2 g) \bmod (m_1 g) :\ i = 0,\ldots,m_1-1\} = \{k + j\cdot g :\ j = 0,\ldots,m_1-1\}\,.$$

Proof. Consider the sequence $\mu_0, \mu_1, \ldots, \mu_{m_1-1}$, where

$$\mu_i = (k + i\cdot m_2 g) \bmod (m_1 g)\,,\quad \text{for } i = 0,\ldots,m_1-1\,.$$

First, for each i, the value of μ_i is a remainder after integer division by $m_1 g$, and hence $\mu_i \in \{0,\ldots,m_1 g-1\}$. Second, it can be expressed in the form $\mu_i = k + i\cdot m_2 g - M_i\cdot m_1 g$, for a suitable integer $M_i \geqslant 0$. Using $k \in \{0,\ldots,g-1\}$, this gives $\mu_i \bmod g = k \bmod g = k$.

[2] Theorem 3.3 remains also valid, the condition $u \geqslant n^2-2$ simply excludes all accepted inputs. Similarly, Definition 3.4 does not produce any predicates $P_{i,r}$, index sets R_i, or languages $L_{i,r}$.

On the other hand, there are exactly m_1 values such that $x \bmod g = k$ and, at the same time, $x \in \{0, \ldots, m_1 g - 1\}$. More exactly, these values form the set

$$X = \{k + j \cdot g : \; j = 0, \ldots, m_1 - 1\} .$$

By the argument above, we have shown that $\{\mu_0, \mu_1, \ldots, \mu_{m_1-1}\} \subseteq X$. Since $\|X\| = m_1$, the equality of these two sets can be proved by showing that the values $\mu_0, \mu_1, \ldots, \mu_{m_1-1}$ are all distinct.

So suppose, for contradiction, that $\mu_{i_1} = \mu_{i_2}$, for some $0 \leqslant i_1 < i_2 \leqslant m_1 - 1$. But then $(k + i_1 \cdot m_2 g) \bmod (m_1 g) = (k + i_2 \cdot m_2 g) \bmod (m_1 g)$, and hence $((i_2 - i_1) \cdot m_2 g) \bmod (m_1 g) = 0$. That is, $(i_2 - i_1) \cdot m_2 g$ is an integer multiple of $m_1 g$, which implies that $(i_2 - i_1) \cdot m_2$ is an integer multiple of m_1. But, by assumption, $\gcd\{m_1, m_2\} = 1$, and hence the factorization of m_2 into a product of primes does not contain a single prime p used in the factorization of m_1. Therefore, $i_2 - i_1$ must be an integer multiple of m_1. But this contradicts the fact that $1 \leqslant i_2 - i_1 \leqslant m_1 - 1$. Thus, $\mu_0, \mu_1, \ldots, \mu_{m_1-1}$ are all distinct, and therefore $\{\mu_0, \mu_1, \ldots, \mu_{m_1-1}\} = X$, which completes the proof. □

It is obvious that the transition function of an optimal unary dfa is determined by two quantities, the length of the initial segment and the length of the subsequent loop. If already the initial state is a part of the loop, we say that the segment length is zero. However, two dfa's of the same segment and loop lengths can still differ in the distribution of their final states.

Theorem 4.2. *Let M_1, M_2 be two unary dfa's accepting the same language L, consisting of initial segments of lengths s_1, s_2 and loops of lengths ℓ_1, ℓ_2, respectively. Then L can also be accepted by a dfa M consisting of an initial segment of length $s = \min\{s_1, s_2\}$ and a loop of length $\ell = \gcd\{\ell_1, \ell_2\}$.*

Proof. Without loss of generality, we can assume that $s_1 \leqslant s_2$. Then the input of length s_2 is sufficiently long so that both M_1 and M_2 pass through the initial segments and get into some states that are already parts of their loops. This gives two states, p_0 for M_1 and r_0 for M_2, fixing some positions along the loops of M_1 and M_2.

Now we can enumerate the ℓ_1 states in the loop of M_1 in order in which they appear along the loop, starting from the state p_0. Similarly, we can enumerate the ℓ_2 states in the loop of M_2, starting from r_0. This way we get a sequence $p_0, p_1, \ldots, p_{\ell_1-1}$ in M_1 and $r_0, r_1, \ldots, r_{\ell_2-1}$ in M_2, such that

$$q_{\mathrm{S},1} \xrightarrow{1^{s_2}} p_0 \xrightarrow{1^{e}} p_e , \quad \text{for } e = 0, \ldots, \ell_1 - 1 ,$$

$$q_{\mathrm{S},2} \xrightarrow{1^{s_2}} r_0 \xrightarrow{1^{f}} r_f , \quad \text{for } f = 0, \ldots, \ell_2 - 1 ,$$

where $q_{\mathrm{S},1}, q_{\mathrm{S},2}$ are the initial states of M_1, M_2, respectively.

Let $g = \gcd\{\ell_1, \ell_2\}$. Then ℓ_1, ℓ_2 can be expressed in the form $\ell_1 = m_1 g$ and $\ell_2 = m_2 g$, for some m_1, m_2 such that $\gcd\{m_1, m_2\} = 1$. For each $k \in \{0, \ldots, g - 1\}$, we can now consider the inputs of lengths $u_{k,0}, u_{k,1}, \ldots, u_{k,m_1-1}$, where

$$u_{k,i} = s_2 + k + i \cdot m_2 g , \quad \text{for } i = 0, \ldots, m_1 - 1 .$$

First, using $k \leqslant g - 1 \leqslant \ell_2 - 1$ and the fact that the loop in M_2 is of length $\ell_2 = m_2 g$, we get the path $q_{\mathrm{S},2} \xrightarrow{1^{u_{k,i}}} r_k$, for each i. Thus, in M_2, the computation paths for all inputs $u_{k,0}, u_{k,1}, \ldots, u_{k,m_1-1}$ end up in the same state. This implies that either *(i)* these strings are all in L, or *(ii)* none of these strings is in L. This holds for each $k \in \{0, \ldots, g - 1\}$.

Consider now the behaviour of M_1, for the same inputs $u_{k,0}, u_{k,1}, \ldots, u_{k,m_1-1}$. Here, the loop is of length $\ell_1 = m_1 g$, and hence, for each $i = 0, \ldots, m_1 - 1$, we get the path $q_{\mathrm{S},1} \xrightarrow{1^{u_{k,i}}} p_{(k+i \cdot m_2 g) \bmod (m_1 g)}$. Recall that M_1 and M_2 accept the same language L. Therefore, either *(i)* the states in the set $Q_k = \{p_{(k+i \cdot m_2 g) \bmod (m_1 g)} : \; i = 0, \ldots, m_1 - 1\}$ are all accepting, or *(ii)* none of these states is accepting. But, by Lemma 4.1,

$$Q_k = \{p_{k+j \cdot g} : \; j = 0, \ldots, m_1 - 1\} .$$

Summing up, we were able to partition the loop of M_1 into the state sets $Q_0, Q_1, \ldots, Q_{g-1}$, such that, for each $k \in \{0, \ldots, g-1\}$, the set Q_k consists of exactly m_1 states, distributed evenly g positions apart along the loop. Moreover, for each $k \in \{0, \ldots, g-1\}$, either *(i)* the states in Q_k are all accepting, or *(ii)* none of them is accepting.

This implies that if we take any two states p_a, p_b along the loop, such that the distance between them is an integer multiple of g, then p_a is accepting if and only if p_b is accepting. But then p_a, p_b must be equivalent, since, for each string t^μ, the states $p_{a'}, p_{b'}$, obtained by the paths $p_a \xrightarrow{1^{lt}} p_{a'}$ and $p_b \xrightarrow{1^{lt}} p_{b'}$, are again an integer multiple of g positions apart. Thus, the states in Q_k are all equivalent, for each $k \in \{0, \ldots, g-1\}$.

Therefore, M_1 can be replaced by an equivalent dfa M with an initial segment of length $s_1 = \min\{s_1, s_2\}$ and a loop of length $g = \gcd\{\ell_1, \ell_2\}$. □

We do not claim that M constructed in the above theorem is optimal and cannot be improved. Nevertheless, the theorem yields some consequences for automata that are optimal. The first application of this kind says that we cannot simulate an optimal dfa by a new automaton using a shorter initial segment or a loop length with a smaller factorization cost, even if the new machine uses far more states in total.

Theorem 4.3. *Let M be an optimal unary dfa consisting of an initial segment of length s and a loop of length ℓ. Then each dfa M', equivalent to M, must use an initial segment of length $s' \geqslant s$ and a loop of length ℓ' satisfying $\Phi(\ell') \geqslant \Phi(\ell)$.*

Proof. Let M and M' be two machines satisfying the assumptions of the theorem. By Theorem 4.2, we can replace M and M' by an equivalent dfa M'' with an initial segment of length $s'' = \min\{s, s'\}$ and a loop of length $\ell'' = \gcd\{\ell, \ell'\}$.

Suppose, for contradiction, that $\ell'' < \ell$. Then the total number of states in M'' can be bounded by $s''+\ell'' < s''+\ell = \min\{s, s'\}+\ell \leqslant s+\ell$. Thus, M'' uses fewer states than does M. But this is a contradiction, since M is optimal. Therefore, $\ell'' \geqslant \ell$. On the other hand, we have that $\ell'' = \gcd\{\ell, \ell'\}$ divides ℓ, and hence $\ell'' \leqslant \ell$. Summing up, $\ell'' = \ell$.

Second, $\ell'' = \gcd\{\ell, \ell'\}$ divides also ℓ' and hence, by Lemma 2.1, we get that $\Phi(\ell'') \leqslant \Phi(\ell')$. Therefore, $\Phi(\ell') \geqslant \Phi(\ell'') = \Phi(\ell)$.

Finally, if $s' < s$, then $s''+\ell'' = \min\{s, s'\}+\ell'' < s+\ell'' = s+\ell$, which again contradicts the fact that M is optimal. Therefore, $s' \geqslant s$. □

The next theorem gives a lower bound for nondeterministic simulation.

Theorem 4.4. *Let M be an optimal unary dfa consisting of an initial segment of length s and a loop of length ℓ, such that $s+\ell > n$, for some $n > 1$. If, moreover, either $s \geqslant n^2-1$ or $\Phi(\ell) \geqslant n$, then each nfa M', equivalent to M, must use more than n states.*

Proof. Suppose, for contradiction, that M can be replaced by an equivalent nfa M' with at most n states.

If M' is different from the trivial loop of length n, we can use Theorem 3.6 to obtain an equivalent dfa M'', consisting of an initial segment of length $s'' \leqslant n^2-2$, and a loop of length ℓ'' satisfying $\Phi(\ell'') \leqslant n-1$.

But, by Theorem 4.3, the dfa M'', equivalent to M, must use the initial segment of length $s'' \geqslant s$ and the loop of length ℓ'' satisfying $\Phi(\ell'') \geqslant \Phi(\ell)$.

Combining these facts together, we get $s \leqslant s'' \leqslant n^2-2$, and also $\Phi(\ell) \leqslant \Phi(\ell'') \leqslant n-1$. But this contradicts the assumption that either $s \geqslant n^2-1$ or $\Phi(\ell) \geqslant n$.

Finally, if M' turns out to be the trivial loop of length n, then M' is already deterministic, with $n < s+\ell$ states, which contradicts the assumption that M is optimal.

In either case, we can conclude that M' must use more than n states. □

5. Magic and muggle numbers

We are now ready to present the state hierarchy. Originally, in [12], magic numbers were introduced for regular languages over arbitrary input alphabets. In this paper, the definition has been adapted for the unary case.

Definition 5.1 (*Magic and muggle numbers*)**.** Let n and d be two positive integers. The number d is *magic* for n, if there is no unary regular language for which an optimal nfa uses exactly n states and, at the same time, the optimal dfa uses exactly d states.

Conversely, d is a *muggle* number for n, if it is not magic, i.e., if there exists at least one unary regular language for which an optimal nfa uses exactly n states and, at the same time, the optimal dfa exactly d states.

Theorem 5.2. *Let $G_{\max}(n)$ and $G_{\min}(n)$ denote, respectively, the largest and the smallest muggle numbers for $n > 1$. Then $G_{\max}(n) = F(n-1) + k_n$, for some $k_n \in \{0, \ldots, n^2-2\}$, which can be approximated by $G_{\max}(n) = \mathrm{e}^{(1\pm o(1))\cdot\sqrt{n\cdot\ln n}}$, and $G_{\min}(n) = n$.*

Proof. For each n, consider the sequence of languages $L_0, L_1, \ldots, L_{n^2-1}$, where

$$L_k = \{1^{k+u} \colon u \bmod F(n-1) \neq 0\}, \quad \text{for } k = 0, \ldots, n^2-1.$$

The construction of a deterministic automaton M_k for L_k is straightforward. It consists of

- an initial segment $p_1 \to p_2 \to p_3 \ \ldots \ p_k \to q_0$, skipping the first k symbols, where p_1 is the initial state, and q_0 the first state of the subsequent loop,
- a loop $q_0 \to q_1 \to q_2 \ \ldots \ q_{F(n-1)-1} \to q_0$, counting modulo $F(n-1)$.
- Finally, all states in the loop, except for q_0, are marked as accepting.

It is easy to see that M_k is optimal. First, there is no pair of equivalent states among the accepting states $q_1, \ldots, q_{F(n-1)-1}$, because of the path $q_1 \to q_2 \to q_3 \ldots \ q_{F(n-1)-1} \to q_0$, ending in the rejecting state q_0: For any pair $j > i$, take the string of length $v = F(n-1) - j$. This gives the path $q_j \xrightarrow{1^v} q_0$ ending in the rejecting q_0, together with $q_i \xrightarrow{1^v} q_{i'}$ ending in some accepting state $q_{i'}$. By a similar argument, there is no pair of equivalent states among the rejecting states $p_1, \ldots, p_k, q_0$, because of the path $p_1 \to p_2 \to p_3 \ \ldots \ p_k \to q_0 \to q_1$, ending in the accepting state q_1.

Thus, for each $k = 0, \ldots, n^2-1$, the number of states used by the optimal dfa M_k is exactly $F(n-1) + k$.

Now, for $k = 0, \ldots, n^2-1$, let f_k denote the exact number of states used in an optimal nfa for L_k.

First, we shall show that $f_0 \leqslant n$. Let $F(n-1)$ factorize into $F(n-1) = p_{i_1}^{\alpha_1} \cdot p_{i_2}^{\alpha_2} \cdot \ldots \cdot p_{i_e}^{\alpha_e}$, where $p_{i_1}^{\alpha_1}, \ldots, p_{i_e}^{\alpha_e}$ are some prime powers. Then $1^v \in L_0$ if and only if v is not divisible by $F(n-1)$ which, in turn, holds if and only if there exists some $j \in \{1, \ldots, e\}$ such that v is not divisible by $p_{i_j}^{\alpha_j}$. Therefore, L_0 can be accepted by an nfa M that, in the initial state, nondeterministically chooses one of e loops, of lengths $p_{i_1}^{\alpha_1}, \ldots, p_{i_e}^{\alpha_e}$, and then verifies, for some j, whether the length of the input is not an integer multiple of $p_{i_j}^{\alpha_j}$. Clearly, such automaton M uses $1 + \sum_{j=1}^{e} p_{i_j}^{\alpha_j} = 1 + \Phi(F(n-1))$ states. Using (1), the value of $F(n-1)$ can also be expressed in the form $F(n-1) = \mathrm{lcm}\{\ell_1, \ldots, \ell_m\}$, for some $\ell_1, \ldots, \ell_m$ satisfying $\ell_1 + \cdots + \ell_m = n-1$. But then, using Lemmas 2.2 and 2.1, we get that $1 + \Phi(F(n-1)) = 1 + \Phi(\mathrm{lcm}\{\ell_1, \ldots, \ell_m\}) \leqslant 1 + \sum_{i=1}^{m} \Phi(\ell_i) \leqslant 1 + \sum_{i=1}^{m} \ell_i = n$. Thus, we have an nfa M with at most n states, accepting L_0. We do not claim that M is optimal. For our purposes, it is sufficient to conclude that an optimal nfa for L_0 cannot use more states than does M, and hence $f_0 \leqslant n$.

Second, we shall show that $f_{n^2-1} > n$. This follows from the fact that the optimal deterministic automaton M_{n^2-1} for L_{n^2-1}, described above, contains the initial segment of length $k = n^2-1$. But then, by Theorem 4.4, each nfa accepting L_{n^2-1} must use more than n states. This must hold for optimal nfa's as well, and hence $f_{n^2-1} > n$.

Third, it is easy to see that $f_{k+1} \leqslant f_k + 1$, for each $k = 0, \ldots, n^2-2$. Let M'_k be an optimal nfa for L_k, with f_k states. To obtain an nfa M'' (not necessarily optimal) for the language L_{k+1}, we need only a new initial state q''_S, connected by a new edge $q''_S \to q'_S$ to the original initial state of M'_k. The rest of the computation is a direct simulation of M'_k. Clearly, for each v, M'' accepts the input 1^{1+v} if and only if M'_k accepts 1^v. This way we have obtained an nfa M'' for L_{k+1}, using only $f_k + 1$ states. But an optimal nfa M'_{k+1} accepting L_{k+1} cannot use more states than does M'', and hence $f_{k+1} \leqslant f_k + 1$.

Summing up, we have obtained an integer sequence $f_0, f_1, \ldots, f_{n^2-1}$, such that $f_0 \leqslant n$, $f_{n^2-1} > n$, and $f_{k+1} \leqslant f_k + 1$, for each $k = 0, \ldots, n^2-2$. This sequence is not necessarily monotone, nevertheless, it is easy to see that such sequence must contain an element equal to n, that is, $f_{k'} = n$, for some $k' \in \{0, \ldots, n^2-2\}$. But then $L_{k'}$ is a language for which the optimal nfa $M'_{k'}$ uses exactly $f_{k'} = n$ states and, at the same time, the optimal dfa $M_{k'}$ exactly $F(n-1) + k'$ states.

Therefore, the value $F(n-1)+k'$ is a muggle number for n. But then the largest muggle number for n is at least $G_{\max}(n) \geqslant F(n-1)+k' \geqslant F(n-1)$. On the other hand, by Theorem 3.7, each unary nfa with n states can be replaced by an equivalent dfa, not necessarily optimal, using at most $F(n-1)+(n^2-2)$ states. But then an optimal dfa does not use more states either. This gives that $F(n-1) \leqslant G_{\max}(n) \leqslant F(n-1)+(n^2-2)$. Using (2), we then get $G_{\max}(n) = e^{(1+o(1))\cdot\sqrt{(n-1)\cdot\ln(n-1)}} = e^{(1\pm o(1))\cdot\sqrt{n\cdot\ln n}}$.

For completeness, it is trivial to see that $G_{\min}(n) = n$. First, no optimal nfa with n states can be replaced by an equivalent dfa with a smaller number of states. Second, the language $L = \{1^u : u \bmod n = 0\}$ requires exactly n states, both in deterministic and nondeterministic case. □

The above theorem gives a corresponding lower bound for the simulation presented by Theorem 3.7.

Clearly, each number $d < G_{\min}(n)$ or $d > G_{\max}(n)$ is trivially magic for n. We are now going to prove the existence of *nontrivial* magic numbers, between $G_{\min}(n)$ and $G_{\max}(n)$, i.e., "holes" in the state hierarchy.

Definition 5.3 (*Darkly magic numbers*)**.** A number $d \geqslant 1$ is *darkly magic* for n, if, for each positive integer $\ell \in \{d-n^2+2, \ldots, d-1, d\}$, $\Phi(\ell) \geqslant n$.

Informally, a darkly magic number must be preceded by a sufficiently long contiguous sequence of integers, such that all of them have sufficiently high factorization costs.

Theorem 5.4. *Let d be a darkly magic number for $n > 1$. Then, for each optimal unary dfa M using exactly d states, an optimal nfa M', equivalent to M, must use more than n states. Therefore, if d is darkly magic for n, then it is magic for each $n' \leqslant n$.*

Proof. We begin with a small technical detail, showing that $d > n$. If $d \in \{2, \ldots, n\}$, we get $\Phi(d-1) \leqslant d-1 \leqslant n-1$, by the use of Lemma 2.1. If $d = 1$, then $\Phi(d) = 0 \leqslant n-1$. Thus, using a suitable $\ell' \in \{d-1, d\}$, we are able to obtain $\Phi(\ell') \leqslant n-1$, for each $d \leqslant n$. But this contradicts the assumption that d is darkly magic for n. Therefore, $d > n$.

Now, let M be an arbitrary optimal unary dfa, using exactly d states. Clearly, M consists of an initial segment of length $s \geqslant 0$ and a loop of length $\ell \geqslant 1$, such that $s+\ell = d > n$. Further, if the initial segment is of length $s \leqslant n^2-2$, the loop length is at least $\ell = d-s \geqslant d-n^2+2$. But then $\Phi(\ell) \geqslant n$, since d is darkly magic for n.

Summing up, $s+\ell > n$, and either $s \geqslant n^2-1$ or $\Phi(\ell) \geqslant n$. But then, by Theorem 4.4, each nfa M', equivalent to M, must use more than n states. □

Thus, to show the existence of nontrivial magic numbers, it is sufficient to prove the existence of nontrivial darkly magic numbers. This requires some more facts about the cost of factorization.

Lemma 5.5. *Let $F_{\#}(n)$ denote the number of different values d satisfying $\Phi(d) \leqslant n$. Then $F_{\#}(n) \leqslant e^{(1+o(1))\cdot 2\sqrt{\ln 2}\cdot\sqrt{n}}$.*

Proof. For the given value of n, consider the following sets A, B, and C:

- A consists of prime powers $p^\alpha \leqslant a$, where $a = \sqrt{n}\cdot\ln n/(2\sqrt{\ln 2})$,
- B consists of prime powers satisfying $a < p^\alpha \leqslant b$, where $b = \sqrt{n}\cdot\ln^2 n/\ln\ln n$,
- C consists of prime powers satisfying $b < p^\alpha \leqslant n$.

First, A is the set of all prime powers not exceeding a. But, as shown in [3], there are $\pi^*(a) \leqslant (1+o(1))\cdot a/\ln a$ prime powers[3] smaller than or equal to a. Therefore, the number of different integers d factorizing into $d = \prod_{p^\alpha \in \varphi(d)} p^\alpha$, such that $\varphi(d) \subseteq A$, does not exceed

$$A_{\#} \leqslant 2^{\pi^*(a)} \leqslant 2^{(1+o(1))\cdot a/\ln a} = 2^{(1+o(1))\cdot(\sqrt{n}\cdot\ln n/(2\sqrt{\ln 2}))/\ln(\sqrt{n}\cdot\ln n/(2\sqrt{\ln 2}))}$$
$$\leqslant 2^{(1+o(1))\cdot(\sqrt{n}\cdot\ln n/(2\sqrt{\ln 2}))/(1/2\cdot\ln n)} = 2^{(1+o(1))\cdot\sqrt{n}/\sqrt{\ln 2}} = e^{(1+o(1))\cdot\sqrt{\ln 2}\cdot\sqrt{n}}.$$

Second, the set B consists of prime powers satisfying $a < p^\alpha \leqslant b$. Consider now an integer d factorizing into $d = \prod_{p^\alpha \in \varphi(d)} p^\alpha$, such that $\varphi(d) \subseteq B$, with $\Phi(d) = \sum_{p^\alpha \in \varphi(d)} p^\alpha \leqslant n$. Clearly, the set $\varphi(d)$ contains at most n/a

[3] The proof in [3] actually shows that $\lim_{a\to\infty} \pi^*(a)\cdot\ln(a)/a = 1$. However, we need only an upper bound.

elements, taken from B. Therefore, $\varphi(d)$ can be represented by a unique monotonically decreasing sequence of integers, of length $\lfloor n/a \rfloor$, such that all elements in the sequence are in the range $\{1,\ldots,\lfloor b \rfloor\}$. (If $\varphi(d)$ contains fewer than $\lfloor n/a \rfloor$ elements, the integer sequence is "padded" by appending a suitable number of 1's at the end, so that the length of the sequence is exactly $\lfloor n/a \rfloor$.) Thus, the total number of different integers d factorizing into $d = \prod_{p^\alpha \in \varphi(d)} p^\alpha$, such that $\varphi(d) \subseteq B$ and $\Phi(d) \leqslant n$, does not exceed the total number of such integer sequences, which can be bounded by

$$\begin{aligned} B_\# &\leqslant \lfloor b \rfloor^{\lfloor n/a \rfloor} \leqslant (\sqrt{n} \cdot \ln^2 n / \ln\ln n)^{n/(\sqrt{n}\cdot \ln n/(2\sqrt{\ln 2}))} \\ &\leqslant (\sqrt{n} \cdot \ln^2 n)^{2\sqrt{\ln 2}\cdot\sqrt{n}/\ln n} = \mathrm{e}^{(1/2\cdot\ln n + 2\cdot\ln\ln n)\cdot 2\sqrt{\ln 2}\cdot\sqrt{n}/\ln n} \\ &= \mathrm{e}^{(1+4\cdot\ln\ln n/\ln n)\cdot\sqrt{\ln 2}\cdot\sqrt{n}} = \mathrm{e}^{(1+o(1))\cdot\sqrt{\ln 2}\cdot\sqrt{n}}. \end{aligned}$$

Third, the set C consists of prime powers satisfying $b < p^\alpha \leqslant n$. Consider now an integer d factorizing into $d = \prod_{p^\alpha \in \varphi(d)} p^\alpha$, such that $\varphi(d) \subseteq C$, with $\Phi(d) = \sum_{p^\alpha \in \varphi(d)} p^\alpha \leqslant n$. By the same reasoning as for the set B, we get that each such integer can be represented as a different integer sequence of length $\lfloor n/b \rfloor$, with all elements in the range $\{1,\ldots,n\}$. Therefore, the total number of different integers d factorizing this way is at most

$$C_\# \leqslant n^{\lfloor n/b \rfloor} \leqslant n^{n/(\sqrt{n}\cdot\ln^2 n/\ln\ln n)} = \mathrm{e}^{\ln n\cdot(\sqrt{n}\cdot\ln\ln n/\ln^2 n)} = \mathrm{e}^{\sqrt{n}\cdot\ln\ln n/\ln n} = \mathrm{e}^{o(1)\cdot\sqrt{n}}.$$

Finally, consider an arbitrary positive integer $d = \prod_{p^\alpha \in \varphi(d)} p^\alpha$ such that $\Phi(d) = \sum_{p^\alpha \in \varphi(d)} p^\alpha \leqslant n$. It is easy to see that $\varphi(d) \subseteq A \cup B \cup C$. Therefore, d can be partitioned into $d = d_A \cdot d_B \cdot d_C$, so that these three integers have all their factors in the respective sets A, B, and C. But then the total number of different numbers d satisfying $\Phi(d) \leqslant n$ is

$$F_\#(n) \leqslant A_\# \cdot B_\# \cdot C_\# \leqslant \mathrm{e}^{(1+o(1))\cdot\sqrt{\ln 2}\cdot\sqrt{n}} \cdot \mathrm{e}^{(1+o(1))\cdot\sqrt{\ln 2}\cdot\sqrt{n}} \cdot \mathrm{e}^{o(1)\cdot\sqrt{n}} \leqslant \mathrm{e}^{(1+o(1))\cdot 2\sqrt{\ln 2}\cdot\sqrt{n}}. \quad \square$$

Alternatively, $F_\#(n)$ could be defined as the number of different lcm's of partitions of n.

Lemma 5.6. *There are at most* $F_\#(n-1)\cdot(n^2-1) \leqslant \mathrm{e}^{(1+o(1))\cdot 2\sqrt{\ln 2}\cdot\sqrt{n}}$ *different numbers that are not darkly magic for* $n>1$. *Consequently, each set containing more than* $F_\#(n-1)\cdot(n^2-1)$ *positive integers contains at least one number that is darkly magic for* n.

Proof. If d is not darkly magic for n, it can be expressed in the form $d = s+\ell$, for some ℓ satisfying $\Phi(\ell) \leqslant n-1$, and some $s \in \{0,\ldots,n^2-2\}$. But there are only $F_\#(n-1)$ different numbers ℓ with factorization cost $\Phi(\ell) \leqslant n-1$, and n^2-1 different values of s. Therefore, by Lemma 5.5, the number of different ways in which we can form a number that is not darkly magic is bounded by $F_\#(n-1)\cdot(n^2-1) \leqslant \mathrm{e}^{(1+o(1))\cdot 2\sqrt{\ln 2}\cdot\sqrt{n-1}} \cdot \mathrm{e}^{2\cdot\ln n}$. The rest of the argument is straightforward. $\square$

By combining the results presented above, we get:

Theorem 5.7. *Let* $M_{\min}(n), M_{\max}(n)$ *denote the smallest and the largest nontrivial magic numbers, and* $D_{\min}(n), D_{\max}(n)$ *the smallest and the largest nontrivial darkly magic numbers for* n, *respectively. Except for some finitely many* n's, *such numbers do exist, and* $G_{\min}(n) < M_{\min}(n) \leqslant D_{\min}(n) < D_{\max}(n) \leqslant M_{\max}(n) < G_{\max}(n)$. *In addition,* $D_{\min}(n) \leqslant \mathrm{e}^{(1+o(1))\cdot 2\sqrt{\ln 2}\cdot\sqrt{n}}$, *and* $D_{\max}(n) = \mathrm{e}^{(1\pm o(1))\cdot\sqrt{n\cdot\ln n}}$.

Proof. Consider the set $X = \{1,\ldots,F_\#(n-1)\cdot(n^2-1)+1\}$. By Lemma 5.6, this set contains sufficiently many elements so that it contains at least one number that is darkly magic for n. Let $D_{\min}(n)$ be the smallest darkly magic number in X. First, $D_{\min}(n) > n$, since a number that is darkly magic for n must be larger than n, as shown in the proof of Theorem 5.4. Second, a darkly magic number larger than n is, by Theorem 5.4, also a magic number larger than n. Therefore, there must exist $M_{\min}(n) \leqslant D_{\min}(n)$, the smallest magic number larger than n. Using $G_{\min}(n) = n$, shown by Theorem 5.2, we thus get

$$G_{\min}(n) < M_{\min}(n) \leqslant D_{\min}(n) \leqslant F_\#(n-1)\cdot(n^2-1)+1\,. \tag{10}$$

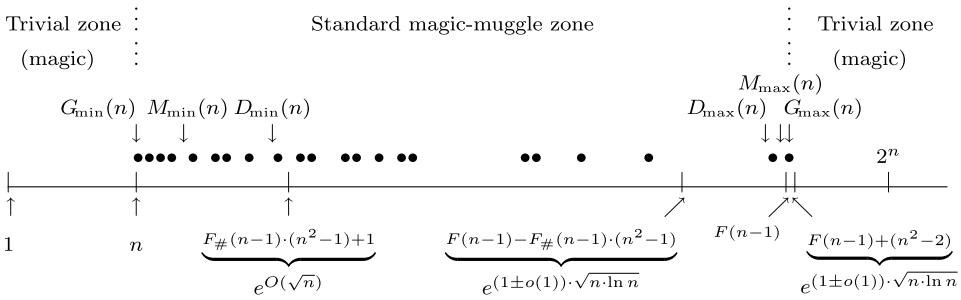

Fig. 2. An example of a typical distribution of muggle and magic numbers for n. Here, the "x-axis" grows in d, the number of states in dfa's. The filled bullets along this axis represent muggle numbers, while the "white space" surrounding the bullets represents magic numbers.

By the same reasoning for the set $Y = \{G_{\max}(n) - F_{\#}(n-1)\cdot(n^2-1), \ldots, G_{\max}(n)\}$, which also contains sufficiently many elements, we obtain

$$G_{\max}(n) - F_{\#}(n-1)\cdot(n^2-1) \leqslant D_{\max}(n) \leqslant M_{\max}(n) < G_{\max}(n)\,. \tag{11}$$

(The inequality $M_{\max}(n) \neq G_{\max}(n)$ follows from the trivial fact that no number can be, at the same time, magic and muggle for the same value of n.)

Finally, using the growth rates that were presented in Lemma 5.6 and Theorem 5.2, we obtain that $F_{\#}(n-1)\cdot(n^2-1)+1 \leqslant e^{(1+o(1))\cdot 2\sqrt{\ln 2}\cdot\sqrt{n}}$, together with $G_{\max}(n) - F_{\#}(n-1)\cdot(n^2-1) \geqslant e^{(1-o(1))\cdot\sqrt{n\cdot\ln n}} - e^{(1+o(1))\cdot 2\sqrt{\ln 2}\cdot\sqrt{n}} \geqslant e^{(1-o(1))\cdot\sqrt{n\cdot\ln n}}$. Combining this with (10) and (11), we get that $D_{\min}(n) < D_{\max}(n)$ for each sufficiently large n, together with the asymptotic bounds for these two values. □

Thus, the state hierarchy of dfa's, for the family of unary languages accepted by n-state nfa's, is not contiguous, there are some magic numbers between the smallest and the largest muggle numbers.

Now we can go farther and show that actually most of the numbers between $G_{\min}(n)$ and $G_{\max}(n)$ are magic. That is, quite surprisingly, muggle numbers are very sporadic. The structure of the state hierarchy and distribution of magic and muggle numbers is shown in Fig. 2.

Corollary 5.8. *Let $G_{\#}(n)$ denote the total number of different muggle numbers for n, and $M_{\#}(n)$ the total number of nontrivial magic numbers for n. Then $G_{\#}(n) \leqslant e^{(1+o(1))\cdot 2\sqrt{\ln 2}\cdot\sqrt{n}}$, and $M_{\#}(n) = e^{(1\pm o(1))\cdot\sqrt{n\cdot\ln n}}$. Consequently,* $\lim_{n\to\infty} G_{\#}(n)/M_{\#}(n) = 0$.

Proof. Clearly, a muggle number is not magic and hence, by Theorem 5.4, it is not darkly magic. But then, by Lemma 5.6, the total number of muggle numbers can be bounded by $G_{\#}(n) \leqslant F_{\#}(n-1)\cdot(n^2-1) \leqslant e^{(1+o(1))\cdot 2\sqrt{\ln 2}\cdot\sqrt{n}}$.

The number of nontrivial magic numbers between $G_{\min}(n)$ and $G_{\max}(n)$ can be expressed in the form $M_{\#}(n) = G_{\max}(n) - (G_{\min}(n)-1) - G_{\#}(n)$. But $G_{\#}(n) \leqslant e^{O(\sqrt{n})}$ and, by Theorem 5.2, we have $G_{\max}(n) = e^{(1\pm o(1))\cdot\sqrt{n\cdot\ln n}}$ and $G_{\min}(n) = n$. This gives $M_{\#}(n) = e^{(1\pm o(1))\cdot\sqrt{n\cdot\ln n}}$, and also $\lim_{n\to\infty} G_{\#}(n)/M_{\#}(n) = 0$. □

As an additional bonus, we get the following universal lower bounds.

Corollary 5.9. (*a*) *For each $d > 1$, no optimal unary dfa using d states can be simulated by an nfa using fewer than* $(1-o(1))/2 \cdot \ln^2 d/\ln\ln d \geqslant \Omega(\ln^2 d/\ln\ln d)$ *states*.

(*b*) *For infinitely many d's, no optimal unary dfa using d states can be simulated by an nfa using fewer than* $(1-o(1))/(4\ln 2)\cdot\ln^2 d \not\leqslant o(\ln^2 d)$ *states*.

Proof. (a) By Theorem 3.7, no optimal dfa with d states can be simulated by an n-state nfa, if $e^{(1+o(1))\cdot\sqrt{n\cdot\ln n}} < d$. This gives that $n \geqslant 1/(1+o(1))^2\cdot\ln^2 d/(2\cdot\ln\ln d)$, which can be simplified into the form $n \geqslant (1-o(1))/2\cdot\ln^2 d/\ln\ln d$.

(b) Consider the sequence $D_{\min}(n_0), D_{\min}(n_0+1), D_{\min}(n_0+2), \ldots$, starting from a sufficiently large n_0. Since $D_{\min}(n) > G_{\min}(n) = n$, by Theorems 5.7 and 5.2, it is obvious that this sequence must contain infinitely many different integers. By Theorem 5.7, we also have, for each $n \geqslant n_0$, that $D_{\min}(n) \leqslant e^{(1+o(1))\cdot 2\sqrt{\ln 2}\cdot\sqrt{n}}$. This gives that $n \geqslant (1-o(1))/(4\ln 2)\cdot\ln^2(D_{\min}(n))$. Moreover, $D_{\min}(n)$ is darkly magic for n and hence, by Theorem 5.4, no optimal dfa using exactly $D_{\min}(n)$ states can be simulated by an nfa using fewer than $n+1$ states. □

6. Concluding remarks

We have shown that, in the unary case, the state hierarchy of deterministic automata, for the family of languages accepted by nondeterministic automata using n states, is not contiguous. There are some "holes" in the hierarchy, i.e., magic numbers between the smallest muggle number $G_{\min}(n) = n$ and the largest muggle number $G_{\max}(n) = e^{(1\pm o(1))\cdot\sqrt{n\cdot\ln n}}$.

We have actually obtained a much stronger result, namely, that *most of the numbers* between $G_{\min}(n)$ and $G_{\max}(n)$ are magic. More precisely, if $G_{\#}(n)$ is the total number of different muggle numbers for n, and $M_{\#}(n)$ the number of nontrivial magic numbers, then $\lim_{n\to\infty} G_{\#}(n)/M_{\#}(n) = 0$. In addition, most numbers between $G_{\min}(n)$ and $G_{\max}(n)$ are magic not only for n itself, but also for each $n' \leqslant n$.

Using the growth rates for $G_{\max}(n) - G_{\min}(n)$ and $G_{\#}(n)$, it is also easy to see that there must exist two muggle numbers d_1, d_2, with $d_2 - d_1 \geqslant e^{\Omega(\sqrt{n\cdot\ln n})}$, such that all values between d_1 and d_2 are magic. This illustrates that some holes in the hierarchy, consisting of consecutive magic numbers, are quite spacious.

The above results required to revise the Chrobak normal form for unary nfa's, which reduced the cost of eliminating nondeterminism almost exactly to the actually existing optimum. As a by-product of this conversion, presented in Section 3, we have obtained that a superpolynomial gap between the size of unary nfa's and dfa's can be obtained only by nfa's without any loops passing through the initial state. Otherwise, by Corollary 3.8, the new conversion uses only $O(n^2)$ states.

We also have a new *universal* lower bound for the conversion of unary dfa's into equivalent nfa's. Clearly, using $L = \{t^u : u \bmod d = 0\}$, we get a dfa with d states that cannot be simulated by an nfa with a smaller number of states. This gives an "existential" lower bound $\Omega(d)$, showing that, for a carefully chosen worst case example, nondeterminism does not help at all.

On the other hand, Corollary 5.9 shows that nondeterminism *never* reduces the number of states below $\Omega(\ln^2 d/\ln\ln d)$, for no d, and no optimal unary dfa M using d states, not even in the best case. Moreover, because of infinitely many "critical" values of d, nondeterminism does not reduce the number of states to $o(\ln^2 d)$. This is, to the best of the author's knowledge, the highest known universal lower bound for the unary dfa-to-nfa conversion. It should be pointed out that universal lower bounds are very rare in general. (For some other examples, see [1,2].)

Some problems concerning the state hierarchy of regular languages are still open. We do not have a sufficient upper bound for $M_{\min}(n)$, the smallest nontrivial magic number for n, except for $M_{\min}(n) \leqslant e^{O(\sqrt{n})}$, given by Theorem 5.7. A better upper bound for the growth rate of $M_{\min}(n)$ would result in a better universal lower bound in Corollary 5.9. But the most important problem in this field is the completeness of the state hierarchy for the regular languages over the binary alphabet, or any other fixed input alphabet. Very little is known about the state hierarchy of two-way automata.

Acknowledgments

The author thanks Matúš Harminc for several useful discussions concerning the number theory.

References

[1] H. Alt, Lower bounds on space complexity for context-free recognition, Acta Inform. 12 (1979) 33–61.
[2] H. Alt, V. Geffert, K. Mehlhorn, A lower bound for the nondeterministic space complexity of context-free recognition, Inform. Process. Lett. 42 (1992) 25–27.
[3] A. Bertoni, C. Mereghetti, G. Pighizzini, An optimal lower bound for nonregular languages, Inform. Process. Lett. 50 (1994) 289–292. (Corrigendum: 1994, ibid. 52, 339.)
[4] M. Chrobak, Finite automata and unary languages, Theor. Comput. Sci. 47 (1986) 149–158 (Corrigendum: 2003, ibid. 302, 497–498).
[5] V. Geffert, Nondeterministic computations in sublogarithmic space and space constructibility, SIAM J. Comput. 20 (1991) 484–498.
[6] V. Geffert, Space hierarchy theorem revised, Theor. Comput. Sci. 295 (2003) 171–187.
[7] V. Geffert, (Non)determinism and the size of one-way finite automata, in: Proc. Descr. Compl. Formal Systems, IFIP & University Milano, 2005, pp. 23–37.

[8] V. Geffert, C. Mereghetti, G. Pighizzini, Converting two-way nondeterministic unary automata into simpler automata, Theor. Comput. Sci. 295 (2003) 189–203.
[9] G. Hardy, E. Wright, An Introduction to the Theory of Numbers, fifth ed., Oxford University Press, 1979.
[10] J. Hopcroft, R. Motwani, J. Ullman, Introduction to Automata Theory, Languages, and Computation, Addison-Wesley, 2001.
[11] J.E. Hopcroft, J.D. Ullman, Introduction to Automata Theory, Languages, and Computation, Addison-Wesley, 1979.
[12] K. Iwama, Y. Kambayashi, K. Takaki, Tight bounds on the number of states of DFA's that are equivalent to n-state NFA's, Theor. Comput. Sci. 237 (2000) 485–494.
[13] K. Iwama, A. Matsuura, M. Paterson, A family of NFA's which need $2^n - \alpha$ deterministic states, Theor. Comput. Sci. 301 (2003) 451–462.
[14] G. Jirásková, Note on minimal finite automata, in: Proc. Math. Found. Comput. Sci., Lect. Notes Comput. Sci., 2136, Springer-Verlag, 2001, pp. 421–431.
[15] Ju.I. Ljubič, Ocenki dlja optimaĺnoj determinizacii nedeterminirovannyh avtonomnyh avtomatov, Sibirsk. Mat. Zh. V/2 (1964) 337–355. (in Russian)
[16] O.B. Lupanov, Uber den Vergleich zweier Typen endlicher Quellen, Probleme der Kybernetik 6 (1966) 329–335. (Akademie-Verlag, Berlin, in German)
[17] C. Mereghetti, G. Pighizzini, Optimal simulations between unary automata, SIAM J. Comput. 30 (2001) 1976–1992.
[18] F. Moore, On the bounds for state-set size in the proofs of equivalence between deterministic, nondeterministic, and two-way finite automata, IEEE Trans. Comput. C-20 (1971) 1211–1214.
[19] M. Rabin, D. Scott, Finite automata and their decision problems, IBM J. Res. Develop. 3 (1959) 114–125.
[20] W. Sakoda, M. Sipser, Nondeterminism and the size of two-way finite automata, in Proc. ACM Symp. Theory of Comput., 1978, pp. 275–286.
[21] A. Salomaa, D. Wood, S. Yu, On the state complexity of reversals of regular languages, Theor. Comput. Sci. 320 (2004) 315–329.
[22] M. Szalay, On the maximal order in S_n and S_n^*, Acta Arith. 37 (1980) 321–331.
[23] S.Y. Yau, Number Theory for Computing, Springer-Verlag, 2002.