ON THE NUMBER OF DISTINCT LANGUAGES ACCEPTED BY FINITE AUTOMATA WITH n STATES

MICHAEL DOMARATZKI
Department of Computer Science, Queen's University
Kingston, ON K7L 3N6, Canada
e-mail: domaratz@cs.queensu.ca

DEREK KISMAN
Department of Mathematics, University of Toronto
Toronto, ON M5S 1A1, Canada
e-mail: dkisman@acm.org

and

JEFFREY SHALLIT[1]
Department of Computer Science, University of Waterloo
Waterloo, ON N2L 3G1, Canada
e-mail: shallit@uwaterloo.ca

ABSTRACT

We give asymptotic estimates and some explicit computations for both the number of distinct languages and the number of distinct finite languages over a k-letter alphabet that are accepted by deterministic finite automata (resp. nondeterministic finite automata) with n states.

Keywords: enumeration, finite automata, minimal automaton, nondeterministic finite automaton

1. Introduction

The problem of enumeration of finite automata according to various criteria (with or without distinguished initial state, initially connected[2], strongly connected, non-isomorphic, etc.) was considered as early as 1959, when V. A. Vyssotsky apparently wrote a Bell Laboratories memorandum on this subject [33]. (We have not been able to obtain a copy.) Counting finite automata was problem 19 in Harary's 1960 list of unsolved problems in graph enumeration [3, pp. 75, 87]. (Also see Harary [4] and Harary and Palmer [6].) In 1962, Ginsburg proposed the problem of enumerating non-isomorphic sequential machines [1, p. 18].

Since then many authors examined these questions, particularly in the former Soviet Union. For example, see the papers of Livshits [22]; Korshunov [12, 13, 14, 15, 16, 17, 18]; Liskovets [19, 20, 21]; and Popov and Skibenko [27]. These papers are all in Russian and an English translation is not available for most of them.

For contributions of Western authors, see Harrison [7, 8]; Radke [28]; Harary and Palmer [5]; and Robinson [30].

[1] Research supported in part by a grant from NSERC.

[2] By *initially connected* we mean that for each state q there exists a directed path from the distinguished start state to q.

However, it appears that none of these authors have examined the problems that interest us in this paper: namely, counting the number of distinct *languages* (or distinct *finite* languages) accepted by DFA's (or NFA's) with n states. Indeed, many of the papers above deal with enumeration of *automata* rather than languages, and their automata often do *not* have a distinguished initial state or set of final states. For example, although the title of the paper of Livshits [22] suggests he was enumerating unary finite automata, he actually counted the number of non-isomorphic "functional digraphs", which is a very different quantity (and which had been previously studied by Harary [2], Read [29], and others). Korshunov [13, 15, 17] counted minimal automata, but it turns out he worked with Mealy machines that are not necessarily initially connected; in fact, his machines lacked a distinguished initial state. Therefore, his enumeration results are quite different from ours and apparently not trivially related.

Furthermore, it seems that the question of enumerating unary regular languages and languages accepted by NFA's has received little or no attention.

We define a DFA to be a 5-tuple, $M = (Q, \Sigma, \delta, q_0, F)$ where Q is a finite nonempty set of states, Σ is a finite nonempty input alphabet, $\delta : Q \times \Sigma \rightarrow Q$ is the transition function, $q_0 \in Q$ is the distinguished initial state, and $F \subseteq Q$ is the set of final states. The domain of δ is extended to $Q \times \Sigma^*$ in the obvious manner. An NFA is also a 5-tuple $M = (Q, \Sigma, \delta, q_0, F)$, but the transition function δ maps $Q \times \Sigma \rightarrow 2^Q$. In this paper we assume our NFA's do not have ϵ-transitions.

Two states $p, q \in Q$ are said to be *equivalent* if for all $x \in \Sigma^*$ we have $\delta(p, x) \in F \iff \delta(q, x) \in F$. If a DFA M lacks equivalent states then it is minimal.

We say a DFA (resp. NFA) M accepting L is *minimal* if there is no DFA (resp. NFA) M' with fewer states accepting L. By a well-known result, there is a unique minimal DFA, up to isomorphism (renaming of the states). This is not true for NFA's.

We define the following functions:

$$\begin{aligned} f_k(n) &= \text{the number of pairwise non-isomorphic minimal DFA's with } n \text{ states} \\ &\quad\ \text{over a } k\text{-letter input alphabet;} \\ g_k(n) &= \text{the number of distinct languages accepted by DFA's with } n \text{ states} \\ &\quad\ \text{over a } k\text{-letter input alphabet; and} \\ G_k(n) &= \text{the number of distinct languages accepted by NFA's with } n \text{ states} \\ &\quad\ \text{over a } k\text{-letter input alphabet.} \end{aligned}$$

Note that $f_k(n)$ counts the number of non-isomorphic, initially connected automata with n states such that no two states are equivalent. Robinson [30] suggests computing $f_k(n)$ as an open enumeration problem.

The goal of this paper is to develop good upper and lower bounds for $f_k(n)$, $g_k(n)$, and $G_k(n)$. We are particularly interested in *constructive* lower bounds, i.e., lower bounds which are provided by explicit construction of DFA's or NFA's accepting different languages. We have the following trivial observation:

Proposition 1 We have $g_k(n) = f_k(1) + f_k(2) + \cdots + f_k(n)$.

Proof. If L is accepted by a DFA with n states, then it is accepted by a unique (up to isomorphism) minimal DFA with $\leq n$ states. If L is accepted by a minimal DFA with $\leq n$ states, then by adding unreachable states it can be accepted by a DFA with exactly n states. □

There are some applications for good estimates for $g_k(n)$ and $G_k(n)$. For example, in Shallit and Breitbart [32] and Pomerance, Robson, and Shallit [26], upper bounds on $g_k(n)$ and $G_k(n)$ were used to provide lower bounds on "automaticity".

2. Enumeration of DFA languages for $k = 1$

In this section we develop both exact and asymptotic formulas in the unary case. We estimate both $f_1(n)$ and $g_1(n)$.

Nicaud [25] observed the following:

Theorem 2 *An n-state unary DFA $M = (Q, \{a\}, \delta, q_0, F)$ where $Q = \{q_0, q_1, \ldots, q_{n-1}\}$ is minimal iff the following three conditions hold:*

(a) It is connected, i.e., there are no unreachable states. Thus, after renaming, its transition diagram consists of a "loop" and a "tail", i.e., $\delta(q_i, a) = q_{i+1}$ for $0 \le i \le n-2$ and $\delta(q_{n-1}, a) = q_j$ for some j, $0 \le j \le n-1$.

(b) The loop is "minimal", that is, it cannot be replaced by an equivalent smaller loop.

(c) If $j \neq 0$, then q_{j-1} and q_{n-1} are of opposite "finality", i.e., $q_{j-1} \in F$ and $q_{n-1} \notin F$ or $q_{j-1} \notin F$ and $q_{n-1} \in F$.

Note that the loop $q_j, q_{j+1}, \ldots, q_{n-1}$ is minimal if and only if the word $a_j a_{j+1} \cdots a_{n-1}$ defined by

$$a_i = \begin{cases} 1, & \text{if } q_i \in F; \\ 0, & \text{if } q_i \notin F; \end{cases}$$

is primitive. (A nonempty word w is *primitive* if it cannot be written in the form $w = x^k$ for some word x and integer $k \ge 2$.)

Let $\psi_k(n)$ denote the number of primitive words of length n over a k-letter alphabet. It is well-known (e.g., [23, p. 9]) that

$$\psi_k(n) = \sum_{d \,|\, n} \mu(d) k^{n/d}$$

where μ is the Möbius function, defined as follows:

$$\mu(n) = \begin{cases} 0, & \text{if } n \text{ is divisible by a square} > 1; \\ (-1)^s, & \text{if } n = p_1 p_2 \cdots p_s, \text{ where the } p_i \text{ are distinct primes.} \end{cases}$$

Theorem 3 *We have*

$$f_1(n) = \psi_2(n) + \sum_{1 \le j \le n-1} \psi_2(n-j) 2^{j-1}.$$

Proof. The result follows from Nicaud's theorem (Theorem 2). The 2^{j-1} factor comes from the fact that there are j states in the tail and if $j \ge 1$, then the type of one of the states (final or non-final) is fixed by condition (c). □

By Proposition 1 we have the following expression for $g_1(n)$, the number of distinct languages accepted by unary DFA's with n states:

$$g_1(n) = \sum_{1 \le t \le n} f_1(t).$$

We can now state the first result of the paper.

Theorem 4 *We have $g_1(n) = \sum_{1 \le t \le n} \psi_2(t) 2^{n-t}$.*

Proof. We have

$$g_1(n) = \sum_{1 \leq t \leq n} f_1(t) = \sum_{1 \leq t \leq n} \left(\psi_2(t) + \sum_{1 \leq j \leq t-1} \psi_2(t-j)2^{j-1} \right)$$
$$= \sum_{1 \leq t \leq n} \left(\psi_2(t) + \sum_{1 \leq i \leq t-1} \psi_2(i)2^{t-i-1} \right) = \sum_{1 \leq t \leq n} \psi_2(t)2^{n-t}.$$

□

We now give a good asymptotic estimate for $g_1(n)$:

Theorem 5 *We have*

$$g_1(n) = 2^n(n - \alpha + O(n2^{-n/2}))$$

where

$$\alpha = \sum_{d \geq 2} \frac{\mu(d)}{1 - 2^{d-1}} \doteq 1.38271445540239628547.$$

Proof. By Theorem 4 we get

$$g_1(n) = \sum_{1 \leq t \leq n} \psi_2(t)2^{n-t} = \sum_{1 \leq t \leq n} \left(\sum_{d \,|\, t} \mu(d)2^{t/d} \right) 2^{n-t}$$
$$= 2^n \sum_{1 \leq t \leq n} \sum_{d \,|\, t} \mu(d)2^{t/d-t} = 2^n \left(n + \sum_{1 \leq t \leq n} \sum_{\substack{d \,|\, t \\ d \neq 1}} \mu(d)2^{t/d-t} \right),$$

so it suffices to estimate $\sum_{1 \leq t \leq n} \sum_{\substack{d \,|\, t \\ d \neq 1}} \mu(d)2^{t/d-t}$. Let $t = kd$ and reverse the order of summation. We find

$$\sum_{1 \leq t \leq n} \sum_{\substack{d \,|\, t \\ d \neq 1}} \mu(d)2^{t/d-t} = \sum_{2 \leq d \leq n} \mu(d) \sum_{1 \leq k \leq \frac{n}{d}} 2^{k-kd}$$
$$= \sum_{2 \leq d \leq n} \mu(d) \left(O(2^{n/d-n}) + \sum_{k \geq 1} 2^{k-kd} \right)$$
$$= \sum_{2 \leq d \leq n} \mu(d) \left(O(2^{n/d-n}) + \frac{1}{2^{d-1}-1} \right)$$
$$= \left(\sum_{2 \leq d \leq n} \frac{\mu(d)}{2^{d-1}-1} \right) + O(n2^{-n/2})$$

and the result follows. □

Corollary 6 *We have*

$$f_1(n) = 2^{n-1}(n + 1 - \alpha + O(n2^{-n/2}))$$

where α is defined in Theorem 5.

Note: Nicaud [25] proved the weaker result that $f_1(n) \sim 2^{n-1}n$.

Proof. By Proposition 1 we have

$$\begin{aligned} f_1(n) &= g(n) - g(n-1) \\ &= 2^n(n - \alpha + O(n \cdot 2^{-n/2})) - 2^{n-1}(n - 1 - \alpha + O(n \cdot 2^{-n/2})) \\ &= 2^{n-1}(n + 1 - \alpha + O(n \cdot 2^{-n/2})). \end{aligned}$$

□

3. Enumeration of DFA languages for $k \geq 2$

In this section, we turn our attention to larger alphabet sizes. First, we give a constructive lower bound for $f_k(n)$, the number of pairwise non-isomorphic minimal n-state DFA's over a k-letter alphabet.

Suppose we are given an automaton $M = (Q, \Sigma, \delta, q_0, F)$ and a subset of the input alphabet $\Delta \subseteq \Sigma$. We can consider the restriction of M to Δ, written M_Δ, which is the automaton $(Q, \Delta, \delta_\Delta, q_0, F)$ where δ_Δ is the restriction of the function δ to the domain $Q \times \Delta$.

Theorem 7 *We have* $f_k(n) \geq f_1(n) n^{(k-1)n} \sim n 2^{n-1} n^{(k-1)n}$.

Proof. Fix $k \geq 2$ and $n \geq 1$. Consider the set $S_{k,n}$ of DFA's M over the k-letter alphabet $\{0, 1, \ldots, k-1\}$ defined by

(a) Letting $M_{\{0\}}$ be one of the $f_1(n)$ different minimal unary DFA's on n states; and

(b) Choosing any $k-1$ functions $h_i : Q \to Q$ for $1 \leq i < k$ and defining $\delta(q, i) = h_i(q)$ for $1 \leq i < k$ and $q \in Q$.

Then we claim that $S_{k,n}$ contains $f_1(n) n^{(k-1)n}$ different DFA's, each is minimal and no two DFA's accept the same language.

Each DFA M in $S_{k,n}$ is minimal because its restriction $M_{\{0\}}$ is minimal. To see that no two DFA's in $S_{k,n}$ accepts the same language, let $M = (Q, \Sigma, \delta, q_0, F)$ and $M' = (Q, \Sigma, \delta', q_0, F')$ be two distinct DFA's in $S_{k,n}$. If the restrictions $M_{\{0\}}$ and $M'_{\{0\}}$ represent different minimal unary DFA's, then by definition they accept different languages, so M and M' accept different languages, too.

Otherwise we may assume that the restrictions $M_{\{0\}}$ and $M'_{\{0\}}$ are identical; in particular $F = F'$. Without loss of generality, assume $Q = \{q_0, \ldots, q_{n-1}\}$, where $\delta(q_0, 0^s) = q_s$ for $0 \leq s < n-1$. Then the restrictions $M_{\{1,2,\ldots,k-1\}}$ and $M'_{\{1,2,\ldots,k-1\}}$ are different, and these correspond to sets of functions h_i and h'_i that must differ. Then there exists an index i, $1 \leq i < k$ and state $q \in Q$ such that $h_i(q) \neq h'_i(q)$. Without loss of generality assume $q = q_l$, $q_j = h_i(q)$, $q_k = h'_i(q)$, $j < k$. Then consider the string $x = 0^l i$. We then have $\delta(q_0, x) = q_j$ and $\delta'(q_0, x) = q_k$.

If there exists $t \geq 0$ such that either $\delta(q_j, 0^t) \in F$ and $\delta'(q_k, 0^t) \notin F$ or $\delta(q_j, 0^t) \notin F$ and $\delta'(q_k, 0^t) \in F$, then the string $x0^t = 0^l i 0^t$ distinguishes $L(M)$ from $L(M')$. Otherwise $\delta(q_j, 0^t) \in F \iff \delta'(q_k, 0^t) \in F$ for all $t \geq 0$. But the restrictions $M_{\{0\}}$ and $M'_{\{0\}}$ are identical, so q_j and q_k are equivalent states in the restriction $M_{\{0\}}$. But then $M_{\{0\}}$ is not minimal, a contradiction.

□

We may improve the lower bound in Theorem 7 slightly, as follows: instead of demanding that the restriction $M_{\{0\}}$ be minimal, we allow the restriction $M_{\{i\}}$ to *any* letter to be minimal. Of course, this double-counts those automata whose restriction are minimal on two or more letters. One way to remove this double counting is to remove all those automata which under some permutation of the states $\{1, 2, \ldots, n-1\}$ results in an initially-connected automaton on some lower-numbered letter. This gives the following theorem.

Theorem 8 *We have*

$$f_k(n) \geq f_1(n) \left(k n^{(k-1)n} - \frac{k(k-1)}{2}(n-1)! n^{(k-2)n} \right).$$

For fixed k this gives $f_k(n) \geq (k - o(1)) n 2^{n-1} n^{(k-1)n}$.

We now turn to upper bounds for $f_k(n)$ and $g_k(n)$. We may obtain the trivial bound $g_k(n) \leq 2^n n^{kn}/(n-1)!$ easily as follows. We can choose the final states in 2^n ways. The transition function maps $\delta : Q \times \Sigma \to Q$, and there are n^{kn} such functions. As observed by Robinson [30] we may now divide by $(n-1)!$, because after removing those automata with n states that are not initially connected, the names of all but the initial state in the remaining automata are irrelevant.

This upper bound can be improved somewhat by considering only initially connected finite automata. Let $C_k(n)$ be the number of pairwise non-isomorphic initially connected finite automata without any final states. Then the analysis above gives $f_k(n) \leq 2^n C_k(n)/(n-1)!$. As both Liskovets [19] and Robinson [30] have observed, a simple combinatorial argument gives

$$C_k(n) = n^{kn} - \sum_{1 \leq j < n} \binom{n-1}{j-1} C_k(j) n^{k(n-j)}. \tag{3.1}$$

Robinson stated that

$$C_k(n) = n^{kn} \gamma_k^{n(1+o(1))}$$

where

$$\gamma_k = \frac{(1-c_k)^{\frac{1-c_k}{c_k}}}{c_k^{k-1}}$$

and c_k is the unique positive root of the equation $c_k = 1 - e^{-kc_k}$. (We have corrected a misprint in the formula Robinson gave for γ_k.) For $k = 2$, we have $c_k \doteq .79681213002002$ and $\gamma_k \doteq .83590576750085$. For more precise results, see Korshunov [17, p. 50].

Our lower and upper bounds for $f_k(n)$ differ by a factor of approximately $(\gamma_k e)^n$, which for $k = 2$ is roughly 2.27^n. It seems likely to us that $\log f_k(n) \sim (k-1) n \log n + \beta_k n$ where $\beta_2 \doteq 1.5$.

4. Enumeration of NFA languages

In this section we consider the computation of $G_k(n)$, the number of distinct languages accepted by NFA's with n states over a k-letter alphabet. Other than the single result mentioned below, it appears that this question has not been examined previously.

First, the unary case. Pomerance, Robson, and Shallit [26] proved that there exists a constant c_1 such that $G_1(n) \leq (\frac{c_1 n}{\log n})^n$, which appears to be the best known upper bound.

Evidently $G_1(n) \geq 2^n$, since each subset $L \subseteq \{\epsilon, a, a^2, \ldots, a^{n-1}\}$ can be accepted by some NFA with n states. This lower bound can be slightly improved as follows:

Theorem 9 *We have $G_1(n) > 2^{n + (2.295 - o(1))\sqrt{\frac{n}{\log n}}}$.*

Proof. Let p_m denote the m'th prime (with $p_1 = 2$). Let C be a constant to be determined later, with $C \geq 2$. Given $n \geq 2$, define b such that $b p_{\lfloor Cb \rfloor} \leq n < (b+1) p_{\lfloor C(b+1) \rfloor}$. Then from the well-known approximation $p_m < m(\log m + \log\log m)$ for $m \geq 6$, we have $b > (\sqrt{2/C} - o(1))\sqrt{\frac{n}{\log n}}$.

For each of the $\binom{\lfloor Cb \rfloor}{b}$ ways to choose b distinct primes $r_1 < r_2 < \cdots < r_b$ from the set $\{p_1, p_2, \ldots, p_{\lfloor Cb \rfloor}\}$, construct a unary NFA with a tail of $s := n - (r_1 + r_2 + \cdots + r_b)$ states, with the last state branching nondeterministically into b distinct cycles, of lengths $r_1, r_2, \ldots, r_b$, respectively. See Figure 1.

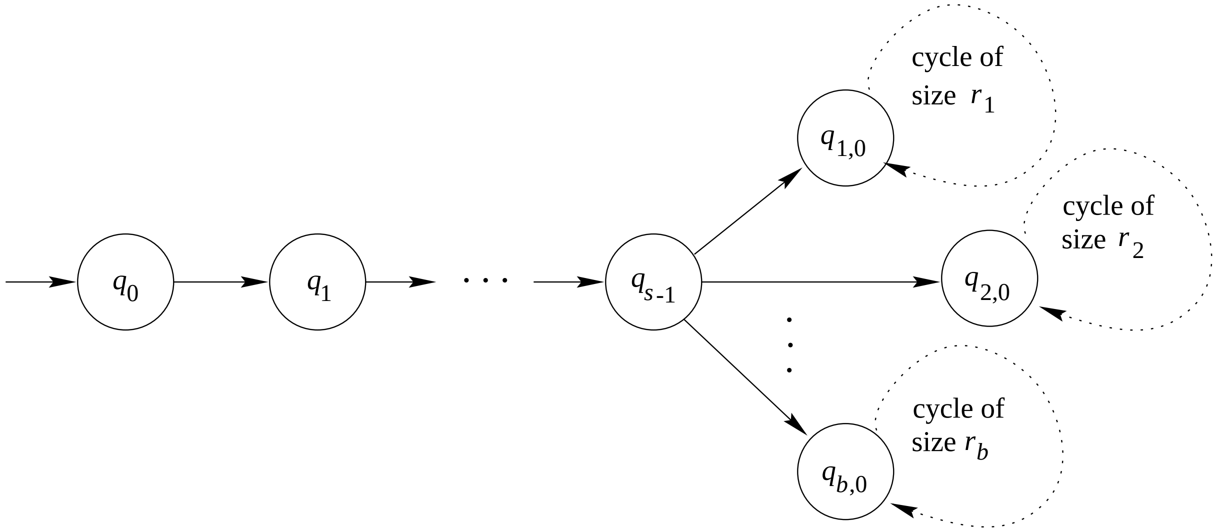

Figure 1: Constructing an NFA

More precisely, define

$$\begin{aligned} \delta(q_i, a) &= \{q_{i+1}\} \quad \text{for } 0 \le i < s-1; \\ \delta(q_{s-1}, a) &= \{q_{1,0}, q_{2,0}, \ldots, q_{b,0}\}; \\ \delta(q_{i,j}, a) &= \{q_{i,(j+1) \bmod r_i}\} \quad \text{for } 1 \le i \le b,\ \ 0 \le j < r_i. \end{aligned} \tag{4.2}$$

We have

- if $t < s$, then a^t is accepted iff q_t is final;
- if $t \ge s$, then a^t is accepted iff there exists i, $1 \le i \le b$ such that $t - s \equiv e \pmod{r_i}$ and $q_{i,e}$ is final.

Now choose the final states in the tail in all possible ways, and choose the final states in the cycles in all possible ways, subject to the restriction that not all the states in any given cycle are of the same type (i.e., not all final or all non-final).

We claim that no two of these NFA's accept the same language. Suppose the contrary, i.e., there exist two NFA's $M = (Q, \Sigma, \delta, q_0, F)$ and $M' = (Q', \Sigma, \delta', q_0', F')$ such that $L(M) = L(M')$. Suppose the states of M are given as in Eq. (4.2) (based on the set of primes $R = \{r_1, r_2, \ldots, r_a\}$) and the states of M' are similarly defined, based on the set of primes $R' = \{r_1', r_2', \ldots, r_a'\}$.

Case 1: $R = R'$. We may assume $r_i = r_i'$ for $1 \le i \le a$ and $Q = Q'$. Since $L(M) = L(M')$, it must be that M and M' are identical except that $F \ne F'$. If the final state that differs occurs in the tail, then the appropriate prefix is accepted by one NFA and not the other. Hence the difference occurs in some cycle, say the state which is congruent to $e \pmod{r_i}$. More precisely, assume $q_{i,e} \in F$ but $q_{i,e} \notin F'$. By our hypothesis, for all j, $1 \le j \le b$, there exists a state $q_{j,c_j} \notin F'$.

Then, by the Chinese remainder theorem, we can choose $t \ge s$ such that $t - s \equiv e \pmod{r_i}$, but $t - s \equiv c_j \pmod{r_j}$ for $1 \le j \le b$, $j \ne i$. Then a^t is accepted by M but not by M'.

Case 2: $R \ne R'$. Without loss of generality there is an $r \in R$ such that $r \notin R'$. By our hypothesis, if $r = r_l$, there exists a state q_{l,e_l} which is final in M. Also, for all j with $r_j' \in R' - R$ there exists a state q'_{j,c_j} which is not final in M'. Finally, for all k with $r_k \ne r$, there exists a state q_{k,d_k} which is not final in M.

Then, by the Chinese remainder theorem, we can choose $t \geq n$ such that $t - s \equiv l \pmod{r_l}$, $t - s' \equiv c_j \pmod{r'_j}$ for $r'_j \in R' - R$ and $t - s \equiv d_k \pmod{r_k}$ for $r_k \in R - \{r\}$. Then M accepts a^t, but M' does not.

How many different NFA's are there in our construction? There are $\binom{\lfloor Cb \rfloor}{b}$ ways to choose the subset of b primes, and for each subset $R = \{r_1, r_2, \ldots, r_b\}$ we can assign the final states in every possible way, except that we remove two possibilities for each cycle (all final or all non-final). This gives

$$2^s(2^{r_1} - 2)(2^{r_2} - 2) \cdots (2^{r_b} - 2) = 2^n(1 - 2^{1-r_1})(1 - 2^{1-r_2}) \cdots (1 - 2^{1-r_b})$$

ways to assign the final states. Now clearly

$$(1 - 2^{1-r_1})(1 - 2^{1-r_2}) \cdots (1 - 2^{1-r_b}) \geq \prod_{i \geq 1} (1 - 2^{1-p_i})$$

where p_i is the i'th prime, and this infinite product converges to the absolute constant $\beta \doteq .345640293832338$. It follows that there are $\geq \beta 2^n \binom{\lfloor Cb \rfloor}{b}$ different NFA's. Now from Stirling's approximation we have

$$\binom{\lfloor Cm \rfloor}{m} \sim C^{Cm}(C-1)^{m(1-C)}\sqrt{C/(2\pi(C-1)m)}$$

for fixed C and as $m \to \infty$, so we get an asymptotic lower bound on the number of different NFA's of $2^{n+(\gamma - o(1))\sqrt{\frac{n}{\log n}}}$, where $\gamma = \sqrt{2/C}(C \log_2 C - (C-1)\log_2(C-1))$. Now, choosing $C \doteq 4.141$ in order to maximize γ, we find we can take $\gamma \doteq 2.295$. This completes the proof. □

Now we turn to the case $k \geq 2$.

Proposition 10 For $k \geq 2$ we have $n2^{(k-1)n^2} \leq G_k(n) \leq (2n-1)2^{kn^2} + 1$.

Proof. For the upper bound, note that any NFA can be specified by specifying, for each pair (q, a) of state and symbol, which subset of Q equals $\delta(q, a)$. We may assign the final states as follows: either the initial state is final or not, and then since the names of states are unimportant we may assume the remaining final states are $\{1, 2, \ldots, k\}$ for $0 \leq k \leq n-1$. Finally, if we choose no final states, we obtain only the empty language $\emptyset$.

For the lower bound, we use the same technique as in Theorem 7. Define an NFA $M = (Q, \Sigma, \delta, q_0, F)$ over the k-letter alphabet $\Sigma = \{0, 1, \ldots, k-1\}$ as follows. Let $Q = \{q_0, q_1, \ldots, q_{n-1}\}$ and define

$$\delta(q_i, 0) = q_{(i+1) \bmod n} \quad \text{for } 0 \leq i < n;$$
$$\delta(q_i, j) = h_j(i) \quad \text{for } 0 \leq i < n,\ 1 \leq a < k;$$

where $h_j : \{1, 2, \ldots, n-1\} \to 2^Q$ is any set-valued function. Finally, let $F = \{q_i\}$ for any i, $0 \leq i \leq n$. There are $(2^n)^{(k-1)n}$ such functions and n ways to choose the set of final states. Using similar reasoning to that in Theorem 7, no two such NFA's accept the same language. □

The upper bound may be marginally improved by considering only NFA's that are initially connected, as follows: let $D_k(n)$ be the number of NFA's on n states, over a k-letter alphabet such that every state is reachable from the start state. Then we have $G_k(n) \leq (2n-1)D_k(n) + 1$. Unfortunately we have

Theorem 11 *We have* $D_k(n) \sim 2^{kn^2}$.

Proof. In analogy with Eq. (3.1), we have

$$2^{kn^2} = \sum_{1 \leq j \leq n} \binom{n-1}{j-1} D_k(j) 2^{kn(n-j)}.$$

(Here j is the size of the connected component involving the start state. There are $\binom{n-1}{j-1}$ ways to choose the other $j-1$ states, and then transitions from the remaining $n-j$ states can be chosen in all possible ways.) From this we obtain $D_k(1) = 2^k$ and

$$D_k(n) = 2^{kn^2} - \sum_{1 \le j \le n-1} \binom{n-1}{j-1} D_k(j) 2^{kn(n-j)}.$$

If $k \ge 2$, then

$$\binom{n-1}{j-1} D_k(j) 2^{kn(n-j)} \le 2^{n-1} 2^{kj^2} 2^{kn^2-knj} = 2^{k(n^2+j^2-nj)+n-1}.$$

Now over the range $1 \le j \le n-1$, the expression $n^2 + j^2 - nj$ is strictly decreasing for $1 \le j \le n/2$ and strictly increasing for $n/2 \le j \le n-1$. It follow that $n^2+j^2-nj \le n^2-n+1$ for $1 \le j \le n-1$. Thus, if $k \ge 2$ we have

$$\binom{n-1}{j-1} D_k(j) 2^{kn(n-j)} \le 2^{k(n^2-n+1)+n-1} \le 2^{kn^2-n+k}.$$

Hence $D_k(n) \ge 2^{kn^2}(1 - n2^{-n+k})$.

If $k = 1$, then

$$\begin{aligned}\binom{n-1}{j-1} D_1(j) 2^{n(n-j)} &\le 2 \cdot 2^{n^2-n+1} + \sum_{2 \le j \le n-2} \binom{n-1}{j-1} D_1(j) 2^{n(n-j)} \\ &\le 2^{n^2-n+2} + n \cdot 2^{n-1} \cdot 2^{n^2-2n+4} \\ &\le 2^{n^2-n+2}(1+2n).\end{aligned}$$

It follows that $D_1(n) \ge 2^{n^2}(1 - (2n+1)2^{-n+2})$. □

Let $E_k(n,r)$ denote the number of distinct languages L over a k-letter alphabet such that L can be accepted by an NFA with n states, but the minimal DFA accepting L has precisely r states. Currently it is not even known whether or not $E_2(n,j) > 0$ for every j with $n \le j \le 2^n$; see [9, 10, 11]. The distribution of $E_k(n,r)$ is an even harder question. We make a small amount of progress in this paper by showing $E_k(n, 2^n) \ge 2^{n-2}$ for $n \ge 2$:

Theorem 12 *Let $n \ge 2$ and $k = |\Sigma| \ge 2$. There are at least 2^{n-2} distinct languages $L \subseteq \Sigma^*$ such that*

(a) L can be accepted by an NFA with n states; and

(b) The minimal DFA accepting L has 2^n states.

Proof. Without loss of generality we may assume $\Sigma = \{a,b\}$. We construct 2^{n-2} different NFA's, as follows: for each subset $S \subseteq \{1,2,\ldots,n-1\}$ we create an NFA $M_S = (Q, \Sigma, \delta_S, q_0, F)$ where $Q = \{0,1,\ldots,n-1\}$, $q_0 = 0$, $F = \{n-1\}$, and δ_S is defined as follows:

$$\delta_S(i,a) = \begin{cases} \{i+1\}, & \text{if } 0 \le i < n-1; \\ \emptyset, & \text{if } i = n-1; \end{cases}$$

$$\delta_S(i,b) = \begin{cases} \{0,1\}, & \text{if } i = 0; \\ \{0,i+1\}, & \text{if } 1 \le i \le n-2 \text{ and } i \in S; \\ \{i+1\}, & \text{if } 1 \le i \le n-2 \text{ and } i \notin S; \\ \{0\}, & \text{if } i = n-1; \end{cases}$$

First, we show that every state-set in the corresponding DFA is reachable. We claim that each string of the form $b^{n-1}w$, where $|w| = n$, takes us to a different state-set. For upon reading

b^{n-1} we will be in $\{0, 1, \ldots, n-1\}$; and whenever after this we read a, the state numbered 0 is excluded and whenever we read b, the state numbered 0 is included. This gives us a bijection between the 2^n possible state sets and strings of the form $b^{n-1}|w|$.

Next, we show that no two of these state-sets is equivalent. Pick two such sets, say T and U, with $T \neq U$. Without loss of generality we may assume there exists $i \in T$ such that $i \notin U$. Then $\delta_S(T, a^{n-1-i})$ contains the final state $n-1$, but $\delta_S(U, a^{n-1-i})$ does not.

Finally, we show that if $S \neq S'$, then M_S and $M_{S'}$ accept different languages. Without loss of generality we may assume $j \in S$ and $j \notin S'$ for some j, $1 \leq j \leq n-2$. Then M_S accepts $a^j ba^{n-1}$ but $M_{S'}$ does not. □

5. Enumeration of finite languages

We now turn to consideration of finite languages. We define

$f'_k(n) =$ the number of non-isomorphic minimal DFA's with n states over a k-letter input alphabet that accept a *finite* language;

$g'_k(n) =$ the number of distinct *finite* languages accepted by DFA's with n states over a k-letter input alphabet; and

$G'_k(n) =$ the number of distinct *finite* languages accepted by NFA's with n states over a k-letter input alphabet.

Once again, it is clear that $g'_k(n) = f'_k(1) + f'_k(2) + \cdots + f'_k(n)$.

The following proposition is useful.

Proposition 13 Let M be a minimal n-state DFA with $L(M)$ finite. Then M is isomorphic (up to renaming of the states) to a DFA $M' = (Q, \Sigma, \delta, q_0, F)$ satisfying $Q = \{q_0, q_1, \ldots, q_{n-1}\}$ and the following conditions:

(a) $\delta(q_{n-1}, a) = q_{n-1}$ for all $a \in \Sigma$;

(b) If $n \geq 2$, then $\delta(q_{n-2}, a) = q_{n-1}$ for all $a \in \Sigma$;

(c) $q_{n-1} \notin F$;

(d) If $n \geq 2$, then $q_{n-2} \in F$;

(e) If $\delta(q_i, a) = q_j$ for $i < n-1$ then $i < j$.

Proof. If M is minimal, then it is initially connected. Now discard all states from which no final state is reachable. (We must discard at least one such state, for if a final state is reachable from every state, start at q_0 and follow a path by making transitions on some fixed input symbol until some state occurs for the second time. This gives a cycle from which a final state is reachable, so $L(M)$ is not finite, a contradiction.) The resulting graph G must be acyclic (for if there were a cycle $L(M)$ would not be finite). Hence we can impose an ordering on the remaining nodes (excluding q_0), say $q_1, q_2, \ldots, q_{n-2}$ such that $\delta(q_i, a) = q_j$ implies $i < j$. Now add back q_0, with edges to the appropriate q_i. Since G is acyclic we cannot have $\delta(q_0, a) = q_0$ for any $a \in \Sigma$, and so the ordering is preserved.

Finally, add a new sink state labeled q_{n-1} satisfying conditions (a) and (c) and additional transitions as necessary to q_{n-1} to make the DFA complete. This new DFA M' clearly accepts the same language as M. Since we removed at least one state and added back the sink state, the number of states in M' is no larger than the number of states in M. Hence M' is minimal.

It remains to verify conditions (b) and (d). By the ordering in condition (e), the transitions from q_{n-2} can only go to a higher numbered state. This proves (b). If $q_{n-2} \notin F$, then q_{n-2} and q_{n-1} would be equivalent states, and hence M' not minimal. This proves (d). □

Theorem 14 *We have*

(a) $f'_1(1) = 1$ *and* $f'_1(n) = 2^{n-2}$ *for* $n \geq 2$*;*

(b) $g'_1(n) = 2^{n-1}$*;*

(c) For $k \geq 2$ *we have* $f'_k(n) \geq 2^{n-2}((n-1)!)^{k-1}$.

Proof. We start by computing $f'_1(n)$. By Proposition 13 or by using Nicaud's theorem (Theorem 2) if M is a minimal unary DFA accepting a finite language, then the transition diagram of M must have a loop of size 1, and that state must be non-final. If there are ≥ 2 states, the state immediately preceding this sink state must be final. It follows that $f'_1(1) = 1$ and $f'_1(n) = 2^{n-2}$ for $n \geq 2$. This proves (a). Part (b) is an immediate consequence.

We now turn to estimating $f'_k(n)$ for $k \geq 2$. We can construct a DFA M with state set $Q = \{q_0, q_1, \ldots, q_{n-1}\}$ such that its restriction $M_{\{0\}}$ satisfies

$$\begin{aligned} \delta(q_i, 0) &= q_{i+1}, \quad 0 \leq i \leq n-2; \\ \delta(q_{n-1}, 0) &= q_{n-1}. \end{aligned}$$

Furthermore, we may choose the set of final states to be $S \cup \{q_{n-2}\}$, where S is any subset of $\{q_0, \ldots, q_{n-3}\}$. It is now easy to see that these $2^{n-2}((n-1)!)^{k-1}$ automata are all minimal and pairwise different. This proves (c). □

Now let us consider G'.

Theorem 15 *We have*

(a) $G'_1(n) = 2^n$*;*

(b) For $n \geq 2$ *we have* $2^{(k-1)n(n-1)/2} \leq G'_k(n) \leq 2^{n-1+kn(n-1)/2}$.

Proof. It is easy to see that $G'_1(n) = 2^n$. (If an initially-connected NFA M of n states accepts a finite language, then the longest string accepted is of length $< n$. For if a longer string is accepted, we would have a directed cycle in M's transition diagram, and hence $L(M)$ would be infinite.) It follows that $G'_1(n) \leq 2^n$. On the other hand, every subset of $\{\epsilon, a, a^2, \ldots a^{n-1}\}$ can be accepted, since we may form a linear chain of n states and assign the final states in 2^n ways. Each assignment gives a distinct language.

Now let us consider $G'_k(n)$ for $k \geq 2$. We may assume that the states are numbered $0, 1, \ldots, n-1$ in such a way that every transition goes from a lower-numbered state to a higher-numbered state. If $n > 1$, we may also assume vertex $n-1$ is final, for if not we could simply remove all edges leading into it, and renumber the resulting state to appear earlier in the ordering. This gives the upper bound $G'_k(n) \leq 2^{n-1+kn(n-1)/2}$. On the other hand, by a technique similar to that given in Proposition 10, we see that $G'_k(n) \geq 2^{(k-1)n(n-1)/2}$. (We let the states be $\{q_0, \ldots, q_{n-1}\}$ and define $\delta(q_i, 0) = q_{i+1}$ for $0 \leq i < n-1$. We choose the other transitions in all possible ways, provided they go from a lower-numbered to a higher-numbered state. We fix q_{n-1} as final.) □

6. Tables

In this section we report on some explicit computations.

The following table gives the first 10 values of $f_1(n)$ and $g_1(n)$:

n	1	2	3	4	5	6	7	8	9	10
$f_1(n)$	2	4	12	30	78	180	432	978	2220	4926
$g_1(n)$	2	6	18	48	126	306	738	1716	3936	8862

The following table gives $f_2(n)$ and $g_2(n)$ for $1 \le n \le 6$.

n	1	2	3	4	5	6
$f_1(n)n^n$	2	16	324	7680	243750	8398080
$f_2(n)$	2	24	1028	56014	3705306	286717796
$g_2(n)$	2	26	1054	57068	3762374	290480170
$2^n C_2(n)/(n-1)!$	2	48	1728	83968	5141600	379618560

The following table gives $f_3(n)$ and $g_3(n)$ for $1 \le n \le 4$.

n	1	2	3	4
$f_1(n)n^{2n}$	2	64	8748	1966080
$f_3(n)$	2	112	41928	26617614
$g_3(n)$	2	114	42042	26659656
$2^n C_3(n)/(n-1)!$	2	224	63720	34049024

We give some brief remarks about how the numbers in the tables above were computed. We considered each of the n^{kn} possible digraphs on vertices labeled $\{0, 1, \ldots, n-1\}$ where each node has out-degree k. Naively there would be an additional factor of 2^n for the possible choices of F, the set of final states, but we can reduce this to n by exploiting symmetries. First, if we fix 0 as the initial state, then either 0 is final or not, and then we can consider any subset of the form $\{1, 2, \ldots, k\}$ as additional members of F, for $0 \le k \le n$. This gives us $2n$ possible choices for F. For each resulting automaton, we determine if it is minimal, and if so, we count it with weight $\frac{\binom{n-1}{k}}{(n-1)!}$. Second, we can reduce the number of possible sets of final states from $2n$ to n by exploiting the symmetry that a DFA for L is minimal iff the the corresponding DFA obtained by changing final states to non-final and vice-versa is minimal.

The following table gives $G_1(n)$ for $1 \le n \le 5$:

n	1	2	3	4	5
$G_1(n)$	3	9	29	88	269

The distribution of state sizes is given below. The entry in row n and column j gives the number of distinct unary languages accepted by NFA's with n states which are accepted by a minimal DFA with j states.

$n \setminus j$	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	2	1																
2	2	4	3															
3	2	4	12	7	3	1												
4	2	4	12	30	16	11	8	2	1	1	1							
5	2	4	12	30	78	33	27	29	23	9	6	6	2	3	2	1	1	1

A complete listing of the distinct NFA languages for $1 \le n \le 5$ can be found at
`http://www.math.uwaterloo.ca/~shallit/papers.html`.

We give some brief remarks about how the numbers in the tables above were computed. We consider all 2^{n^2} different digraphs on n vertices. As above, we can restrict our attention to sets of final states that either contain state 0 not, and additionally contains the set $\{1, 2, \ldots, k\}$ for some k with $0 \le k \le n$. For each machine, we convert it to a DFA and then minimize the DFA using Theorem 2. Using hashing, we then build a table of distinct minimal DFA's and as each new machine is considered, we check to see if we have already enumerated it.

The following table gives $D_k(n)$ for $1 \le k \le 4$ and $1 \le n \le 4$:

$k \setminus n$	1	2	3	4
1	2	8	256	38912
2	4	192	221184	4020240384
3	8	3584	128450560	279166431789056
4	16	61440	67947724800	18428089759432704000

The following table gives $G_2(n)$ for $1 \le n \le 3$:

n	1	2	3
$G_2(n)$	5	213	45113

The distribution of state sizes is given below. The entry in row n and column j gives the number of distinct languages over $\{0, 1\}$ accepted by NFA's with n states which are accepted by a minimal DFA with j states.

$n \setminus j$	1	2	3	4	5	6	7	8
1	2	3						
2	2	24	117	70				
3	2	24	1028	5595	11211	14537	10580	2136

The following table gives $f_2'(n)$ and $g_2'(n)$ for $1 \le n \le 7$.

n	1	2	3	4	5	6	7
$f_2'(n)$	1	1	6	60	900	18480	487560
$g_2'(n)$	1	2	8	68	968	19448	507008

The following table gives $G_2'(n)$ for $1 \le n \le 5$:

n	1	2	3	4	5
$G_2'(n)$	2	8	80	1904	102848

The distribution of state sizes is given below. The entry in row n and column j gives the number of distinct finite languages over $\{0, 1\}$ accepted by NFA's with n states which are accepted by a minimal DFA with j states.

$n \setminus j$	1	2	3	4	5	6	7	8	9	10	11
1	1	1									
2	1	1	6								
3	1	1	6	60	12						
4	1	1	6	60	900	672	264				
5	1	1	6	60	900	18480	31720	30992	15920	4288	480

As Mandl [24] and Salomaa and Yu [31] have shown, the maximum number of states needed by a DFA to accept a finite language accepted by an NFA with n states over $\{0, 1\}$ is $2^{(n+2)/2} - 1$ if n is even and $3 \cdot 2^{(n-1)/2} - 1$ if n is odd.

References

[1] S. Ginsburg. *An Introduction to Mathematical Machine Theory.* Addison-Wesley, 1962.

[2] F. Harary. The number of functional digraphs. *Math. Annalen* **138** (1959), 203–210.

[3] F. Harary. Unsolved problems in the enumeration of graphs. *Magyar Tud. Akad. Math. Kutató Int. Közl.* **5** (1960), 63–95.

[4] F. Harary. Combinatorial problems in graphical enumeration. In E. Beckenbach, editor, *Applied Combinatorial Mathematics*, pp. 185–217. Wiley, 1964.

[5] F. Harary and E. Palmer. Enumeration of finite automata. *Inform. Control* **10** (1967), 499–508.

[6] F. Harary and E. M. Palmer. *Graphical Enumeration.* Academic Press, 1973.

[7] M. A. Harrison. A census of finite automata. In *Proc. 5th Annual Symposium on Switching Circuit Theory and Logical Design*, pp. 44–46. IEEE Press, 1964.

[8] M. A. Harrison. A census of finite automata. *Canad. J. Math.* **17** (1965), 100–113.

[9] K. Iwama, Y. Kambayashi, and K. Takaki. Tight bounds on the number of states of DFAs that are equivalent to n-state NFAs. *Theoret. Comput. Sci.* **237** (2000), 485–494.

[10] K. Iwama, A. Matsuura, and M. Paterson. A family of NFA's which need $2^n - \alpha$ deterministic states. In M. Nielsen and B. Rovan, editors, *Proc. 25th Symposium, Mathematical Foundations of Computer Science 2000*, Vol. 1893 of *Lecture Notes in Computer Science*, pp. 436–445. Springer-Verlag, 2000.

[11] G. Jirásková. Note on minimal finite automata. In J. Sgall, A. Pultr, and P. Kolman, editors, *Proc. 26th Symposium, Mathematical Foundations of Computer Science 2001*, Vol. 2136 of *Lecture Notes in Computer Science*, pp. 421–431. Springer-Verlag, 2001.

[12] A. D. Korshunov. On asymptotic estimates of the number of finite automata. *Diskretnyi Analiz*, No. 6, (1966), 35–50. In Russian.

[13] A. D. Korshunov. Asymptotic estimates of the number of finite automata. *Kibernetika* **3**(2) (1967), 12–19. In Russian. English translation in *Cybernetics* **3** (2) (1967), 9–14.

[14] A. D. Korshunov. The invariant properties of finite automata. *Diskretnyi Analiz*, No. 16, (1970), 51–76. In Russian.

[15] A. D. Korshunov. A survey of certain trends in automata theory. *Diskretnyi Analiz*, No. 25, (1974), 19–55, 62. In Russian.

[16] A. D. Korshunov. The number of automata and boundedly determined functions. Hereditary properties of automata. *Dokl. Akad. Nauk SSSR* **221** (1975), 1264–1267. In Russian. English translation in *Soviet Math. Doklady* **16** (1975), 515–518.

[17] A. D. Korshunov. Enumeration of finite automata. *Problemy Kibernetiki*, No. 34, (1978), 5–82, 272. In Russian.

[18] A. D. Korshunov. On the number of non-isomorphic strongly connected finite automata. *Elektronische Informationsverarbeitung und Kybernetik* **22** (1986), 459–462.

[19] V. A. Liskovets. The number of connected initial automata. *Kibernetika* **5**(3) (1969), 16–19. In Russian. English translation in *Cybernetics* **5** (1969), 259–262.

[20] V. A. Liskovets. The number of initially connected digraphs. *Doklady Akad. Nauk BSSR* **15** (1971), 293–294. In Russian.

[21] V. A. Liskovets. Enumeration of non-isomorphic strongly connected automata. *Vesci Akad. Navuk BSSR, Ser. Fiz.-Mat. Navuk*, No. 3, (1971), 26–30. In Russian.

[22] E. M. Livshits. Asymptotic formula for the number of classes of isomorphic autonomous automata with n states. *Ukrainskii Matematicheskii Zhurnal* **16** (1964), 245–246. In Russian.

[23] M. Lothaire. *Combinatorics on Words*, Vol. 17 of *Encyclopedia of Mathematics and Its Applications*. Addison-Wesley, 1983.

[24] R. Mandl. Precise bounds associated with the subset construction on various classes of nondeterministic finite automata. In *Proc. 7th Princeton Conference on Information and System Sciences*, pp. 263–267. 1973.

[25] C. Nicaud. Average state complexity of operations on unary automata. In M. Kutylowski, L. Pacholski, and T. Wierzbicki, editors, *Proc. 24th Symposium, Mathematical Foundations of Computer Science 1999*, Vol. 1672 of *Lecture Notes in Computer Science*, pp. 231–240. Springer-Verlag, 1999.

[26] C. Pomerance, J. M. Robson, and J. Shallit. Automaticity II: Descriptional complexity in the unary case. *Theoret. Comput. Sci.* **180** (1997), 181–201.

[27] V. A. Popov and I. T. Skibenko. Enumeration of abstract automata. *Avtomatika i Telemekhanika* **36** (1975), 143–148. In Russian. English translation in *Automation and Remote Control* **36** (1975), 129–133.

[28] C. E. Radke. Enumeration of strongly connected sequential machines. *Inform. Control* **8** (1965), 377–389.

[29] R. C. Read. A note on the number of functional digraphs. *Math. Annalen* **143** (1961), 109–110.

[30] R. W. Robinson. Counting strongly connected finite automata. In Y. Alavi, G. Chartrand, L. Lesniak, D. R. Lick, and C. E. Wall, editors, *Graph Theory with Applications to Algorithms and Computer Science*, pp. 671–685. Wiley, 1985.

[31] K. Salomaa and S. Yu. NFA to DFA transformation for finite languages over arbitrary alphabets. *J. Automata, Languages, and Combinatorics* **2** (1997), 177–186.

[32] J. Shallit and Y. Breitbart. Automaticity I: Properties of a measure of descriptional complexity. *J. Comput. System Sci.* **53** (1996), 10–25.

[33] V. A. Vyssotsky. A counting problem for finite automata. Technical report, Bell Telephone Laboratories, May 1959.