Theoretical Computer Science 652 (2016) 102–108

Contents lists available at ScienceDirect

Theoretical Computer Science

www.elsevier.com/locate/tcs

The state complexity of random DFAs

Daniel Berend, Aryeh Kontorovich *

Ben-Gurion University, Beer Sheva, Israel

ARTICLE INFO

Article history:
Received 5 May 2015
Received in revised form 15 August 2016
Accepted 17 September 2016
Available online 26 September 2016
Communicated by D. Perrin

Keywords:
Random
Deterministic finite-state automaton
DFA
Minimal

ABSTRACT

The state complexity of a Deterministic Finite-state automaton (DFA) is the number of states in its minimal equivalent DFA. We study the state complexity of random n-state DFAs over a k-symbol alphabet, drawn uniformly from the set $[n]^{[n]\times[k]} \times 2^{[n]}$ of all such automata. We show that, with high probability, the latter is $\alpha_k n + O(\sqrt{n}\log n)$ for a certain explicit constant α_k.

1. Introduction

A randomly generated deterministic finite automaton (DFA) on n states and k symbols is drawn as follows: for each state and each of the k symbols in the alphabet, the transition arrow's destination is chosen uniformly at random among the n states; the nk random choices are independent.[1] Then each state is chosen to be accepting (or not) independently with probability 1/2. This natural model for a "typical" DFA goes back to [15] and was considered in [1,12] in the context of learning theory. In particular, in [1] it is shown (perhaps surprisingly) that random DFAs possess sufficient complexity to embed nontrivial parity problems.

Let us define the *state complexity* of a DFA M as the number of states in the canonical (minimal) DFA equivalent to M, and denote it by $\|M\|$. In this paper, we study the state complexity of random DFAs in the model defined above.

Related work. The somewhat related problem of enumerating finite automata according to various criteria has been extensively studied; see [7] and the references therein. Some recent results include generation of random complete DFAs [5], enumeration and generation of accessible DFAs [4], and enumeration of minimal automata [3]. (In particular, Ref. [3] studies a question that is very much related to one of those we study here; see Remark 8 below.)

In a different line of enquiry (whose connection to the present work is explained in Remark 2 below), Pittel investigated the distributions induced by transitive closures [13] and rumor spreading [14]. The uniformly random DFA model seems to have first been proposed in [15]. Working in this model, Grusho [9] considered the accessibility spectrum of a random DFA and gave a central limit theorem, which is analogous to our large deviation result in Theorem 4. A recent result of Balle [2] analyzes the mixing properties of random DFA graphs.

* Corresponding author.
E-mail address: karyeh@cs.bgu.ac.il (A. Kontorovich).

[1] By symmetry, we may always take the state $q = 0$ to be the starting state.

http://dx.doi.org/10.1016/j.tcs.2016.09.012
0304-3975/

2. Background and notation

We use standard automata-theoretic notation throughout; the reader is referred to [11,16] for background. We put $[n] = \{0, \ldots, n-1\}$. Thus, $[k]$ is a k-ary alphabet and $[k]^*$ is the set of all finite words (strings) over this alphabet. The notation $|\cdot|$ is used for both word length and set cardinality. The floor and ceiling functions, $\lfloor r \rfloor$ and $\lceil r \rceil$, are used, respectively, to denote the greatest (smallest) integer majorized (minorized) by $r \in \mathbb{R}$. When obvious from context (such as in summations), we will occasionally omit the floor function, leaving it as implicit. Standard order-of-magnitude notation $o(\cdot)$ and $O(\cdot)$ is used, as well as their "with high probability" variants $o_P(\cdot)$ and $O_P(\cdot)$. The $\tilde{O}(\cdot)$ notation ignores polylog factors.

An n-state k-ary Deterministic Finite-state Automaton is a tuple $M = (Q, q_0, A, \delta)$ where

- $Q = [n]$ is the set of states;
- $q_0 = 0$ is the starting state;
- $A \subseteq [n]$ is the set of accepting states;
- $\delta : [n] \times [k] \to [n]$ is the transition function.

The transition function δ may be extended to $[n] \times [k]^*$ via the recursion

$$\delta(q, u_1 u_2 \cdots u_n) = \delta(\delta(q, u_1), u_2 \cdots u_n). \tag{1}$$

If the accepting states are unspecified, the transition function δ induces a directed multigraph on n nodes with regular outdegree k, called a k-ary *semiautomaton*.

We recall the standard equivalence relation over the states of a DFA: a word $x \in [k]^*$ *distinguishes* between the states $p, q \in [n]$ if exactly one of the states $\delta(p, x)$, $\delta(q, x)$ is accepting. If no $x \in [k]^*$ distinguishes between p and q, these states are *equivalent*, denoted by $p \equiv q$.

A standard high-level algorithm[2] for minimizing a DFA proceeds in two stages:

- `REMOVE-UNREACHABLE`: Remove all states q such that there is no directed path from the starting state q_0 to q.
- `COLLAPSE-EQUIVALENT`: Collapse each set of mutually equivalent states into a single state.

3. Main results

Our main result is the following estimate on the state complexity of random DFAs:

Theorem 1. *Let $M_n^{(k)}$ be a random DFA on n states and k symbols, drawn uniformly from $[n]^{[n]\times[k]} \times 2^{[n]}$. Then, for any fixed $k \geq 2$ and sufficiently large n,*

$$\mathbb{P}\left(\left|\left\|M_n^{(k)}\right\| - \alpha_k n\right| > \sqrt{n}\log n\right) = \Theta\left(\frac{1}{n^k}\right), \tag{2}$$

where α_k is the unique positive root[3] of $x = 1 - e^{-kx}$. In particular,

$$\mathbb{E}\left\|M_n^{(k)}\right\| = \alpha_k n + O(\sqrt{n}\log n).$$

Remark 2. Observe that $0.7968 \approx \alpha_2 < \alpha_3 < \ldots < \alpha_\infty = 1$. For $k = 1$, the behavior of $\left\|M_n^{(k)}\right\|$ is qualitatively different than described in Theorem 1. The equation $x = 1 - e^{-x}$ has no positive solution and $\mathbb{E}\left\|M_n^{(1)}\right\| = \Theta(\sqrt{n})$, which follows from the analysis in [13].

Remark 3. The lower bound of $\Omega(1/n^k)$ in (2) is trivial since, with probability $1/n^k$, all of state 1's arrows point back to itself and $\left\|M_n^{(k)}\right\| = 1$.

Our proof of Theorem 1 proceeds in two principal stages. First we show that, in our model, a random semiautomaton has roughly $\alpha_k n$ reachable states with high probability. As in [15], we refer to the number of reachable states as the *accessibility spectrum* of the semiautomaton.

[2] Hopcroft's celebrated algorithm [10] for minimizing a DFA has runtime complexity $O(n \log n)$.

[3] A closed-form expression for α_k is possible via the Lambert W function [6]: $\alpha_k = 1 + W(-ke^{-k})/k$. This constant also appears (as ω_k) in [3] and seems to be intimately related to average reachability properties of semiautomata in $[n]^{[n]\times[k]}$ under the uniform measure.

Theorem 4. *Let $R_n^{(k)}$ be the accessibility spectrum of a random semiautomaton on n states and k symbols, drawn uniformly from $[n]^{[n]\times[k]}$. Then, for every fixed $k \geq 2$ and as $n \to \infty$,*

$$\mathbb{P}\left(\left|R_n^{(k)} - \alpha_k n\right| > \sqrt{n}\log n\right) = O\left(\frac{1}{n^k}\right). \tag{3}$$

Remark 5. Note that the fact that $\mathbb{E}R_n^{(k)}$ is approximately $\alpha_k n$ follows in particular from Grusho [9]. However, whereas that result deals with the behavior of $R_n^{(k)}$ near the expectation, our result concerns large deviations.

Second, we show that with high probability, very few states are lost when equivalent ones are merged. Define $E_n^{(k)}$ to be the number of "excess" reachable states:

$$E_n^{(k)} = R_n^{(k)} - \left\|M_n^{(k)}\right\|.$$

Theorem 6. *For every fixed $k \geq 2$,*

$$\mathbb{P}\left(E_n^{(k)} > C_k \frac{\log n}{\log\log n}\right) = O\left(\frac{1}{n^k}\right)$$

for an appropriate constant C_k.

Remark 7. Note that in principle we need only show that the number of states lost due to merging is small *after* the unreachables are removed, but we will actually show that this is true even without removing them. Theorem 6 continues to hold when each state is accepting with probability $0 < p < 1$ instead of $1/2$; only the constants C_k and those implicit in $O(\cdot)$ will change.

Remark 8. The main quantity studied in [3] is the probability that a random automaton will be minimal. However, the sample space there consists of all n-state automata in which all states are reachable. It is proved there that, for $k \geq 2$, such an automaton is minimal (i.e., satisfies $E_n^{(k)} = 0$) except on a set of asymptotic probability $O(1/n^{k-2})$. Theorem 6 says that the probability that $E_n^{(k)}$ is larger than $C_k \log n/\log\log n$ drops to $O(1/n^k)$. Note, though, that the sample space in [3] is very different from ours.

4. Proofs

Lemma 9. *Define the function*

$$F(t) = n(1-(1-1/n)^t) - (t-1)/2, \qquad 0 \leq t \leq 2n.$$

Then, for sufficiently large n,

$$\frac{F^2(t)}{t} \geq \begin{cases} 0.01t, & t \leq n/2, \\ \Omega(\log^2 n), & t \in [n/2, 2\alpha_2 n - 2\sqrt{n}\log n] \cup [2\alpha_2 n + 2\sqrt{n}\log n, 2n]. \end{cases}$$

Proof. We have $F(0) = 1/2$, and for $t \leq n/2$

$$\begin{aligned} F'(t) &= -\frac{1}{2} + n\log\left(1+\frac{1}{n-1}\right)\cdot\left(1-\frac{1}{n}\right)^t \\ &\geq -\frac{1}{2} + n\left(\frac{1}{n-1} - \frac{1}{2(n-1)^2}\right)\cdot\left(1-\frac{1}{n}\right)^{n/2} \\ &\geq -\frac{1}{2} + \frac{n}{n-1}\cdot\frac{2n-3}{2n-2}\cdot\left(\frac{1}{\sqrt{e}} - o(1)\right) \\ &\geq \frac{1}{\sqrt{e}} - \frac{1}{2} - o(1) > 0.1. \end{aligned}$$

This proves the estimate on $F^2(t)/t$ in the range $[1, n/2]$.

Now consider $t \in [n/2, 2\alpha_2 n - 2\sqrt{n}\log n]$, and observe that $F(t) = H(t) + O(1)$, where

$$H(t) = n - t/2 - n\exp(-t/n).$$

By the definition of α_2, we have $H(2\alpha_2 n) = H(0) = 0$. Furthermore, $H''(t) = -\exp(-t)/n < 0$, and so H is concave with $H'(n\log 2) = 0$, and therefore increasing on $[0, n\log 2]$ and decreasing on $[n\log 2, \infty)$. Hence, to lower-bound $H^2(t)/t$ in the given range, it suffices to estimate H at its right endpoint:

$$H(2\alpha_2 n - 2\sqrt{n}\log n) = \frac{1}{2}\sqrt{n}\log n - e^{-2\alpha_2}\sqrt{n}\log n + O(\log^2 n) = \Omega(\sqrt{n}\log n).$$

Since $H'(t) < -1/2 + e^{-2\alpha_2}$ for $t > 2\alpha_2 n$, we have $H(2\alpha_2 n + x) = \Omega(x)$ for $x > 0$, which completes the proof. □

Proof of Theorem 4. We will prove the theorem for $k = 2$; the general case is completely analogous — only the constants implicit in $O(1/n^k)$ will vary with k. For readability, we will write $\alpha = \alpha_2$ and $R_n = R_n^{(2)}$. It will be convenient to embed R_n in a slightly more general random process. Fix $n \geq 1$, and define the sequence of random variables $(\nu_t)_{t=1}^{\infty}$ as follows:

$$\nu_1 = 1,$$
$$\nu_{t+1} = \begin{cases} \nu_t, & \text{with probability } \nu_t/n, \\ \nu_t + 1, & \text{with probability } 1 - \nu_t/n. \end{cases}$$

Clearly, ν_t is with probability 1 nondecreasing, upper-bounded by n, and reaches n after a finite number of steps. Let us also put

$$\omega_t = 2\nu_t + 1 - t, \qquad t \geq 1. \tag{4}$$

Now consider the following process for generating random directed multigraphs with regular outdegree 2. For time steps $t = 1, 2, \ldots$, we will maintain the set of nodes N_t, reached from $q_0 = 0$ by time t, and two sets of edges: *open edges* O_t and *closed edges* C_t. A closed edge c is an ordinary directed arrow from a source node p to a destination node q marked with a $\sigma \in [k]$ and denoted by $c = (p \xrightarrow{\sigma} q)$. An open edge o has a specified source p but an as yet unspecified destination; such an edge will be denoted by $o = (p \xrightarrow{\sigma} \star)$. We initialize $N_1 = \{0\}$, $C_1 = \emptyset$ and $O_1 = \left\{(0 \xrightarrow{0} \star), (0 \xrightarrow{1} \star)\right\}$. If, at any time t, we have $O_t = \emptyset$, then the process stops. Otherwise, at time $t+1$, some (arbitrarily chosen[4]) open edge in $o \in O_t$ (if one exists) chooses a destination node q as follows:

(i) with probability $|N_t|/n$, we choose $q \in N_t$ uniformly at random (thus, o will point to a previously reached node);
(ii) with probability $1 - |N_t|/n$, we choose $q \in [n] \setminus N_t$ uniformly at random.

In event (i), $O_{t+1} = O_t \setminus \{o\}$, while in event (ii), $O_{t+1} = (O_t \setminus \{o\}) \cup \left\{(q \xrightarrow{0} \star), (q \xrightarrow{1} \star)\right\}$; in both cases, $N_{t+1} = N_t \cup \{q\}$ and $C_{t+1} = C_t \cup \{o\}$.

The random semiautomaton corresponds to the process (ν_t, ω_t) via the following natural mapping: $\nu_t = |N_t|$ and $\omega_t = |O_t|$ as long as N_t and O_t are defined. Let τ be the smallest t for which $\omega_t = 0$ — i.e., the first time there are no longer any open edges to choose from. Then the pair (N_τ, C_τ) defines[5] a semiautomaton with accessibility spectrum $R_n = \nu_\tau$, drawn uniformly from $[n]^{[n]\times\{0,1\}}$. We observe that

$$\begin{aligned} |\nu_\tau - \alpha n| > \sqrt{n}\log n \quad &\Longleftrightarrow \quad |2\nu_\tau - 2\alpha n| > 2\sqrt{n}\log n \\ &\Longleftrightarrow \quad |\omega_\tau - 1 + \tau - 2\alpha n| > 2\sqrt{n}\log n \\ &\Longleftrightarrow \quad |\tau - 2\alpha n - 1| > 2\sqrt{n}\log n \end{aligned}$$

and hence, proving (3) amounts to showing that

$$\mathbb{P}\left(|\tau - 2\alpha n - 1| > 2\sqrt{n}\log n\right) = O\left(\frac{1}{n^2}\right). \tag{5}$$

Indeed, (5) implies that $\tau = (2 + o_P(1))\alpha n$ and $\nu_\tau = (1 + o_P(1))\nu_{\alpha n}$. Since, by definition, τ is the smallest t for which $\nu_t = (t-1)/2$, we have

$$\begin{aligned} \mathbb{P}(\tau \in [a, b]) \quad &\leq \quad \mathbb{P}(\exists t \in [a, b] : \nu_t = (t-1)/2) \\ &\leq \quad \mathbb{P}(\exists t \in [a, b] : \nu_t \leq (t-1)/2) \\ &\leq \quad \textstyle\sum_{t=a}^{b} \mathbb{P}(\nu_t \leq (t-1)/2)\,. \end{aligned} \tag{6}$$

We estimate the left tail of τ as follows:

[4] It is easy to see that the distribution of N_t, C_t, O_t is unaffected by the order in which the open edges are selected.
[5] Since the quantity of interest is the accessibility spectrum, it is unnecessary to define transitions out of unreachable states.

$$\mathbb{P}\left(\tau \le 2\alpha n + 1 - 2\sqrt{n}\log n\right) \le P_0 + P_1 + P_2,$$

where

$$P_0 = \mathbb{P}(\tau \in [1, 150\log n]),$$
$$P_1 = \mathbb{P}(\tau \in [150\log n, n/2]),$$
$$P_2 = \mathbb{P}\left(\tau \in [n/2, 2\alpha n + 1 - 2\sqrt{n}\log n]\right).$$

To bound P_0, we first argue, by elementary combinatorics, that $\mathbb{P}(\omega_4 < 3) = O(1/n^2)$. Now we condition on the high-probability event that there are at least 3 open arrows available at time $t = 4$. If all of the open arrows have been exhausted between time $t = 4$ and $t = T$, then certainly at least three of these arrows must point back to the $O(T)$ previously reached states. Thus, for $T = 150\log n$,

$$P_0 \in O\left(\frac{1}{n^2} + \binom{T}{3}\left(\frac{T}{n}\right)^3\right) \subset O\left(\frac{1}{n^2}\right).$$

Clearly,

$$P_1 \le \sum_{t=150\log n}^{n/2} \mathbb{P}(\nu_t - \mathbb{E}\nu_t \le -F(t)), \tag{7}$$

$$P_2 \le \sum_{t=n/2}^{2\alpha n - 2\sqrt{n}\log n + 1} \mathbb{P}(\nu_t - \mathbb{E}\nu_t \le -F(t)), \tag{8}$$

where $F(t)$ is as in Lemma 9.

To bound P_1 and P_2, we observe that an alternate interpretation is possible for ν_t. Namely, when t balls are thrown into n bins uniformly at random, the number of non-empty bins is distributed as ν_t. We also observe that

$$\mathbb{E}\nu_t = n(1 - (1 - 1/n)^t), \qquad t \ge 1. \tag{9}$$

Let us recall the notion of *negatively associated* random variables [8]: the random variables X_i, $i \in I$, are negatively associated if for all disjoint $J, J' \subseteq I$ and all $f : \mathbb{R}^J \to \mathbb{R}$, $g : \mathbb{R}^{J'} \to \mathbb{R}$ that are both monotonic non-increasing or non-decreasing,

$$\mathbb{E}[f(X_j, j \in J)g(X_{j'}, j' \in J')] \le \mathbb{E}[f(X_j, j \in J)]\mathbb{E}[g(X_{j'}, j' \in J')].$$

A straightforward adaptation of the balls and bins analysis in [8, Thm. 13] shows that the indicator variables of the empty bins are negatively associated, and hence the corresponding version of the Chernoff bound [8, Prop. 5] applies:

$$\mathbb{P}(\nu_t - \mathbb{E}\nu_t \le -\Delta) \le \exp(-2\Delta^2/t), \qquad \Delta > 0. \tag{10}$$

By (7), (8) and (10):

$$P_1 \le \frac{n}{2}\exp(-3\log n) \in O\left(\frac{1}{n^2}\right)$$

and

$$P_2 \in O(n)\exp(-\Omega(\log^2 n)) \subset O\left(\frac{1}{n^2}\right).$$

This proves the left-tail estimate in (5). To prove the corresponding right-tail estimate, we observe that, analogously to (6),

$$\mathbb{P}(\tau \in [a, b]) \le \mathbb{P}(\exists t \in [a, b] : \nu_t \ge (t - 1)/2).$$

The deviation probability is bounded as in (10):

$$\mathbb{P}(\nu_t - \mathbb{E}\nu_t \ge \Delta) \le \exp(-2\Delta^2/t), \qquad \Delta > 0.$$

Hence

$$\mathbb{P}\left(\tau > 2\alpha n + 2\sqrt{n}\log n + 1\right) \le \sum_{t=2\alpha n + 2\sqrt{n}\log n + 1}^{2n} \mathbb{P}(\nu_t - \mathbb{E}\nu_t \ge G(t)),$$

where $G(t) = -F(t)$. Invoking again Lemma 9, we have

$$\mathbb{P}\left(\tau > 2\alpha n + 2\sqrt{n}\log n + 1\right) \in O(n)\exp(-\Omega(\log^2 n)) \subset O\left(\frac{1}{n^2}\right). \qquad \square$$

Proof of Theorem 6. Again, for ease of exposition, we only prove the claim for $k = 2$. We start by explaining the idea of the proof. We need to show that there are usually "few" pairs of equivalent states. Let us start by describing two "typical" situations in which equivalent states emerge.

The first is where a state is mapped into itself by every member of $\{0, 1\}$. Two such states are equivalent if and only if both are accepting or both are rejecting, which happens with a probability of $1/2$. More generally, if from each of the two states one can reach very few states, then there is a non-negligible probability that the states are equivalent. Thus, we will have to show that there are few states with small accessibility spectra. In the preceding sentence, "few" means (with high probability) "at most 2", while "small" means "less than $4\log_2 n$".

The second principal reason for two states q, q' to be equivalent is that $\delta(q, 0) = \delta(q', 0)$ and $\delta(q, 1) = \delta(q', 1)$. Again, q and q' are equivalent in this case with probability $1/2$. Thus, we will need to show that there are few pairs of states q, q' for which there are few words in $\{0, 1\}^*$ taking q and q' to distinct states. Here, the first "few" means "at most $C \log n / \log\log n$" and the second means "up to $4\log_2 n$".

Let us now consider the above scenarios in more detail. The (random) set of states reachable from q is $\{q, \delta(q, 0), \delta(q, 1), \delta(q, 00), \delta(q, 01), \ldots\}$. Thus, the states reachable from q reside on a binary tree whose edges are marked by letters in $\{0, 1\}$. Each time the random DFA selects a state $p = \delta(q, w)$, if p is already in the tree, the edge that would create a directed cycle is not drawn. We refer to the resulting tree as the tree *growing* from q. Its size is the accessibility spectrum of q, denoted by $S(q)$.

Let $C > 0$ be a constant to be determined later. A state's accessibility spectrum is said to be *small* if it is below $C\log_2 n$. As in the proof of Theorem 4, the probability of a given state having a small accessibility spectrum is $O(1/n^2)$. A similar argument shows that the joint probability of any two states q, q' having small accessibility spectra is $\tilde{O}(1/n^4)$. Indeed, consider the event of $S(q')$ being small, conditioned on $S(q)$ being such. Draw the states $\delta(q', 0), \delta(q', 1), \delta(q', 00), \ldots$ similarly to the proof of Theorem 4. The event in question is contained in the event whereby, in the course of the first $C\log_2 n$ steps of the process of "closing" the open edges, we encounter at least twice either a state visited already or a state belonging to the tree growing from q. The probability of the latter event is clearly $O(\log^4 n/n^2)$. Hence,

$$\mathbb{P}\big(S(q), S(q') \text{ are both small}\big) = O\left(\frac{\log^4 n}{n^4}\right).$$

Carrying this line of reasoning over to triples, we find that the probability of any three states having small accessibility spectra is $\tilde{O}(1/n^6)$ — and therefore,

$$\begin{aligned}\mathbb{P}(\text{there are 3 distinct states with small accessibility spectra}) &\in \tilde{O}\left(\binom{n}{3}\cdot\frac{1}{n^6}\right)\\ &= \tilde{O}\left(\frac{1}{n^3}\right) \subset O\left(\frac{1}{n^2}\right).\end{aligned}$$

In view of the discussion above, we may assume (after removing up to 2 states) that all states have large accessibility spectra. Consider two states q, q'. Let T be a tree of size $m = C\log_2 n$ growing from q (this will typically be a subtree of a larger tree of size $\alpha n(1 + o(1))$). The nodes of T are $\delta(q, w_1), \delta(q, w_2), \ldots, \delta(q, w_m)$ for certain words $w_1, w_2, \ldots, w_m \in \{0, 1\}^*$.

We claim that if $\delta(q, w_i) \neq \delta(q', w_i)$ for $i = 1, 2, \ldots, m$, then the probability of q, q' being equivalent is at most $\frac{1}{2^{\lfloor m/2\rfloor}} \leq \frac{1}{n^C}$. In fact, consider the graph $G = (V, E)$, where $V = \big\{\delta(q, w_i), \delta(q', w_i) : 1 \leq i \leq m\big\}$ and $E = \big\{(\delta(q, w_i), \delta(q', w_i)) : 1 \leq i \leq m\big\}$. Note that the states $\delta(q', w_i)$ are not necessarily all distinct, nor are they necessarily different from the states in T. Thus, the size m_1 of V may be anywhere between m and $2m$. The choice of E ensures that no vertex in G is isolated. Hence the number s of connected components of G is at most $\lfloor m/2\rfloor$. The requirement that q and q' be equivalent implies that, in each connected component of G, either all of the states are accepting or all are rejecting. The probability for the states in any single connected component of size C to satisfy this condition is $\frac{1}{2^{C-1}}$. Hence, the probability that q and q' are equivalent is indeed at most $\frac{1}{2^{m-s}} \leq \frac{1}{2^{\lfloor m/2\rfloor}}$.

The probability that both $\delta(q, 0) = \delta(q', 0)$ and $\delta(q, 1) = \delta(q', 1)$ is $1/n^2$. States satisfying these equalities are *coalescing*. Clearly, coalescence is an equivalence relation on Q. An equivalence class of size r of coalescing states contains $\lfloor r/2\rfloor$ mutually disjoint pairs of coalescing states. When passing to the minimal equivalent automaton, these r states are merged either into a single state (if all are accepting or all are rejecting) or into two states (otherwise). Thus, the number of states collapsing due to coalescence is in any case at most twice the maximal number of mutually disjoint coalescing pairs.

Clearly, the probability that d specific pairwise disjoint pairs will coalesce is $1/n^{2d}$. Now the probability that there will be d disjoint coalescing pairs is at most

$$\frac{1}{n^{2d}}\binom{n}{2d}\cdot(2d-1)\cdot(2d-3)\cdot\ldots\cdot 1 \leq \frac{1}{n^{2d}}\cdot\frac{n^{2d}}{(2d)!}\cdot\frac{(2d)!}{2^d d!} = \frac{1}{2^d d!}.$$

Choosing $d = 3\log n/\log\log n$ and applying Stirling's formula, we see that the probability of there being d disjoint coalescing pairs is $O(1/n^2)$.[6]

Coalescing states q, q' have the property that there exists only one word w (namely, the null word $w = \varepsilon$) for which $\delta(q, w) \neq \delta(q', w)$. In general, we grow the trees from q and q' in parallel, opening the arrows leading to the states $\delta(q, w)$ and $\delta(q', w)$ for words w of length 1, then for words of length 2, and so forth. Each branch is stopped when it either leads to a state already visited in that tree, or coincides with the parallel state (i.e., corresponding to the same w) in the second tree. If both trees are bounded, there is a significant (i.e., bounded away from zero) probability that q and q' are equivalent. However, the main term in the probability of both trees being small is due to q and q' coalescing. Hence, this phenomenon contributes a term of $O(1/n^2) + O(1/n^3) = O(1/n^2)$ to the probability of equivalence. □

Proof of Theorem 1. Follows immediately from Theorems 4 and 6 since the former estimates the number of states remaining after `REMOVE-UNREACHABLE` and the latter bounds the number of states lost after `COLLAPSE-EQUIVALENT`. □

Acknowledgements

We thank Borja de Balle Pigem for bringing [9] to our attention, and the referee for carefully reading the manuscript and catching several errors. A. K. was supported in part by the Israel Science Foundation (grants No. 1141/12 and 755/15) and a Yahoo Faculty award.

References

[1] Dana Angluin, David Eisenstat, Leonid (Aryeh) Kontorovich, Lev Reyzin, Lower bounds on learning random structures with statistical queries, in: ALT, 2010, pp. 194–208.
[2] Borja Balle, Ergodicity of random walks on random DFA, arXiv:1311.6830, 2013.
[3] Frédérique Bassino, Julien David, Andrea Sportiello, Asymptotic enumeration of minimal automata, in: STACS, 2012, pp. 88–99.
[4] Frédérique Bassino, Cyril Nicaud, Enumeration and random generation of accessible automata, Theoret. Comput. Sci. 381 (1–3) (2007) 86–104.
[5] Jean-Marc Champarnaud, Thomas Paranthoën, Random generation of DFAs, Theoret. Comput. Sci. 330 (2) (2005) 221–235.
[6] Robert M. Corless, David J. Jeffrey, Donald E. Knuth, A sequence of series for the Lambert W function, in: Proceedings of the 1997 International Symposium on Symbolic and Algebraic Computation, Kihei, HI, ACM, New York, 1997, pp. 197–204 (electronic).
[7] Michael Domaratzki, Derek Kisman, Jeffrey Shallit, On the number of distinct languages accepted by finite automata with n states, J. Autom. Lang. Comb. 7 (4) (2002) 469–486.
[8] Devdatt Dubhashi, Desh Ranjan, Balls and bins: a study in negative dependence, Random Structures Algorithms 13 (2) (September 1998) 99–124.
[9] Aleksandr Aleksandrovich Grusho, Limit distributions of certain characteristics of random automaton graphs, Math. Notes Acad. Sci. USSR 14 (1) (1973) 633–637.
[10] John Hopcroft, An n log n algorithm for minimizing states in a finite automaton, in: Theory of Machines and Computations, Proc. Internat. Sympos., Technion, Haifa, 1971, Academic Press, New York, 1971, pp. 189–196.
[11] John E. Hopcroft, Rajeev Motwani, Jeffrey D. Ullman, Introduction to Automata Theory, Languages, and Computation, Addison–Wesley, 2003.
[12] Leonid (Aryeh) Kontorovich, Boaz Nadler, Universal kernel-based learning with applications to regular languages, J. Mach. Learn. Res. 10 (2009) 997–1031.
[13] Boris Pittel, On distributions related to transitive closures of random finite mappings, Ann. Probab. 11 (2) (1983) 428–441.
[14] Boris Pittel, On spreading a rumor, SIAM J. Appl. Math. 47 (1) (1987) 213–223.
[15] Boris A. Trakhtenbrot, Janis M. Barzdin', Finite Automata: Behavior and Synthesis, Fundamental Studies in Computer Science, vol. 1, North-Holland, Amsterdam, 1973.
[16] Sheng Yu, Regular Languages, in: Handbook of Formal Languages, vol. 1: Word, Language, Grammar, Springer-Verlag New York, Inc., New York, NY, USA, 1997, pp. 41–105.

[6] It is probably easy to show that the number of such pairs is asymptotically Poi($\frac{1}{2}$)-distributed, but for our purposes, this level of precision is not needed.