Asymptotic enumeration of Minimal Automata

Frédérique Bassino, Julien David, Andrea Sportiello

Asymptotic enumeration of Minimal Automata *

Frédérique Bassino[1], Julien David[1], and Andrea Sportiello[2]

1 **LIPN, Université Paris 13, and CNRS, UMR 7030**
99, av. J.-B. Clément, 93430 Villetaneuse, France
Frederique.Bassino@lipn.univ-paris13.fr Julien.David@lipn.univ-paris13.fr

2 **Università degli Studi di Milano, Dip. di Fisica, and INFN**
Via G. Celoria 16, 20133 Milano, Italy
Andrea.Sportiello@mi.infn.it

Abstract

We determine the asymptotic proportion of minimal automata, within n-state accessible deterministic complete automata over a k-letter alphabet, with the uniform distribution over the possible transition structures, and a binomial distribution over terminal states, with arbitrary parameter b. It turns out that a fraction $\sim 1 - C(k,b)\, n^{-k+2}$ of automata are minimal, with $C(k,b)$ a function, explicitly determined, involving the solution of a transcendental equation.

1998 ACM Subject Classification F.2 Analysis of algorithms and problem complexity

Keywords and phrases minimal automata, regular languages, enumeration of random structures

Digital Object Identifier 10.4230/LIPIcs.STACS.2012.88

1 Introduction

To any regular language, one can associate in a unique way its minimal automaton, i.e. the only accessible complete deterministic automaton recognizing the language, with a minimal number of states. Therefore the *space complexity* of a regular language can be seen as the number of states of its minimal automaton. The worst-case complexity of algorithms dealing with finite automata is usually known [27]. But the average-case analysis of algorithms requires weighted sums on the set of possible realizations, and in particular the enumeration of the objects that are handled [10]. Therefore a precise enumeration is often required for the algorithmic study of regular languages.

The enumeration of finite automata according to various criteria (with or without initial state [18], non-isomorphic [13], up to permutation of the labels of the edges [13], with a strongly connected underlying graph [21, 18, 26, 19], acyclic [22],...) has been investigated since the fifties.

In [18] Korshunov determines the asymptotic estimate of the number of accessible complete and deterministic n-state automata over a finite alphabet. His derivation, and even the formulation of the result, are quite complicated. In [4] a reformulation of Korshunov's result leads to an estimate of the number of such automata involving the Stirling number of the second kind. On the other side, in [20] a different simplification of the involved expressions is achieved, by highlighting the role of the Lagrange Inversion Formula in the analysis.

A natural question is to ask what is the fraction of minimal automata, among accessible complete and deterministic automata of a given size n and alphabet cardinality k. Nicaud

* This work was completed with the support of the ANR project MAGNUM number 2010-BLAN-0204.

29th Symposium on Theoretical Aspects of Computer Science (STACS'12).
Editors: Christoph Dürr, Thomas Wilke; pp. 88–99

Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

[25] shows that, asymptotically, half of the complete deterministic accessible automata over a unary alphabet are minimal, thus solving the question for $k = 1$.

In this paper we solve this question for a generic integer $k \geq 2$ (see Theorem 1 later on). At a slightly higher level of generality, we give a precise estimation of the asymptotic proportion of minimal automata, within n-state accessible deterministic complete automata over a k-letter alphabet, for the uniform distribution over the possible transition structures, and a binomial distribution over terminal states, with arbitrary parameter $0 < b < 1$ (the uniform case corresponding to $b = \frac{1}{2}$). Our theoretical results are in agreement with the experimental ones.

The paper is organized as follows. In Section 2 we recall some basic notions of automata theory, and we set a list of notations that will be used in the remainder of the paper. Then, we state our main theorem, and give a short and simple heuristic argument. In Section 3 we give a detailed description of the proof structure, and its subdivision into separate lemmas. In Section 4 we prove in detail the most difficult lemmas, and give indications for those that are provable through standard methods. Finally, in Section 5 we discuss some of the implications of our result.

2 Statement of the result

For a given set E, $|E|$ denotes the cardinality of E. The symbol $[n]$ denotes the canonical n-element set $\{1, 2, \ldots, n\}$. Let $\mathcal{E}$ be a Boolean condition, the Iverson bracket $[\![\mathcal{E}]\!]$ is equal to 1 if $\mathcal{E} =$ true and 0 otherwise. We use $\mathbb{E}(X)$ to denote the expectation of the quantifier X, and $\mathbb{P}(\mathcal{E}) = \mathbb{E}([\![\mathcal{E}]\!])$ for the probability of the event $\mathcal{E}$. For $\{\mathcal{E}_i\}$ a collection of events, we define a shorthand for the first moment

$$\boldsymbol{m}(\{\mathcal{E}_i\}) := \sum_i \mathbb{P}(\mathcal{E}_i) = \mathbb{E}\Big(\sum_i [\![\mathcal{E}_i]\!]\Big)\,. \tag{1}$$

If $p(c)$ is the probability that exactly c events occur, we have $\boldsymbol{m}(\{\mathcal{E}_i\}) = \sum_c c\,p(c) \geq \sum_{c\geq 1} p(c) = 1 - p(0)$, i.e. $p(0) \geq 1 - \boldsymbol{m}(\{\mathcal{E}_i\})$. This elementary inequality, known as *first-moment bound*, is used repeatedly in the following. It is in fact a special case of a more general relation (named *Markov's inequality*), stating that, for x a random variable, $\mathbb{P}(|x| \geq a) \leq \mathbb{E}(|x|)/a$ (the specialisation is $x \in \mathbb{N}$ and $a = 1$).

A *finite deterministic automaton* A is a quintuple $A = (\Sigma, Q, \delta, q_0, \mathcal{T})$ where Q is a finite set of *states*, Σ is a finite set of *letters* called *alphabet*, the *transition function* δ is a mapping from $Q \times \Sigma$ to Q, $q_0 \in Q$ is the *initial state* and $\mathcal{T} \subseteq Q$ is the set of *terminal* (or *final*) states. With abuse of notations, we identify $\mathcal{T}(i) \equiv [\![i \in \mathcal{T}]\!]$.

An automaton is *complete* when its transition function is total. The transition function can be extended by morphism to all words of Σ^*: $\delta(p, \varepsilon) = p$ for any $p \in Q$ and for any $u, v \in \Sigma^*$, $\delta(p, (uv)) = \delta(\delta(p, u), v)$. A word $u \in \Sigma^*$ is *recognized* by an automaton when $\delta(q_0, u) \in \mathcal{T}$. The *language* recognized by an automaton is the set of words that it recognizes. An automaton is *accessible* when for any state $p \in Q$, there exists a word $u \in \Sigma^*$ such that $\delta(q_0, u) = p$.

We say that two states p, q are *Myhill-Nerode-equivalent* (or just *equivalent*), and write $p \sim q$, if, for all finite words u, $\mathcal{T}(\delta(p, u)) = \mathcal{T}(\delta(q, u))$ [24]. This property is easily seen to be an equivalence relation. An automaton is said to be *minimal* if all the equivalence classes are atomic, i.e. $p \not\sim q$ for all $p \neq q$. In other words, the minimal automaton A' recognizing the same language as A has set of states Q' corresponding to the set of equivalence classes of A. For a general reference on automata see e.g. [14].

At the aim of enumeration, the actual labeling of states in Q and letters in Σ is immaterial, and we can canonically assume that $Q = [n]$, $\Sigma = [k]$, and $q_0 = 1$. In this case, when there is no ambiguity on the values of n and k, we will associate an automaton A to a pair $(\mathcal{D}, \mathcal{T})$ of a transition structure and a set of terminal states. The set of complete deterministic accessible automata with n states over a k-letter alphabet is denoted $\mathcal{A}_{n,k}$.

We will determine statistical averages of quantities associated to automata $A \in \mathcal{A}_{n,k}$. This requires the definition of a measure $\mu(A)$ over $\mathcal{A}_{n,k}$. The simplest and more natural case is just the uniform measure. We generalise this measure by introducing a continuous parameter. For S a finite set, the *multi-dimensional Bernoulli distribution* of parameter b over subsets $S' \subseteq S$ is defined as $\mu_b(S') = b^{|S'|}(1-b)^{|S|-|S'|}$. The distribution associated to the quantifier $|S'|$ is thus the binomial distribution. We will consider the family of measures $\mu_b^{(n,k)}(A) = \mu_{\text{unif}}^{(n,k)}(\mathcal{D})\mu_b^{(n)}(\mathcal{T})$, with $\mu_{\text{unif}}^{(n,k)}(\mathcal{D})$ the uniform measure over the transition structures of appropriate size, and $\mu_b^{(n)}(\mathcal{T})$ the Bernoulli measure of parameter b over $Q \equiv [n]$. The uniform measure over all accessible deterministic complete automata is recovered setting $b = \frac{1}{2}$. Superscripts will be omitted when clear.

The result we aim to prove in this paper is

▶ **Theorem 1.** *In the set $\mathcal{A}_{n,k}$, with the uniform measure, the asymptotic fraction of minimal automata is*

$$\exp\left(-\tfrac{1}{2}c_k n^{-k+2}\right), \tag{2}$$

with

$$c_k = \tfrac{1}{2}\,{\omega_k}^k; \qquad -k\,\omega_k = \ln(1-\omega_k)\,. \tag{3}$$

More generally, for any $0 < b < 1$, with measure $\mu_b^{(n,k)}(A)$, the asymptotic fraction is

$$\exp\left(-(1-2b(1-b))c_k n^{-k+2}\right). \tag{4}$$

We singled out the constant ω_k, instead of only c_k, because the former appears repeatedly, in the evaluation of several statistical properties of random automata. Solving (3), it can be written in terms of (a branch of) the Lambert W-function, as $\omega_k = 1 + \frac{1}{k}W(-ke^{-k})$, however the implicit definition (3) is of more practical use. See Table 1 for a numerical table of values.

k	2	3	4	5	6
ω_k	0.796812	0.940480	0.980173	0.993023	0.997484
c_k	0.317455	0.415928	0.461509	0.482799	0.492498

Table 1 The constants involved in the statement of Theorem 1, for the first values of k.

The result above, specialised to $b = 1/2$, provides as a corollary the asymptotic number of minimal automata, when combined with the known asymptotics for $|\mathcal{A}_{n,k}|$ [18, 4, 20]

$$|\mathcal{A}_{n,k}| = \omega_k \left\{ {kn+1 \atop n} \right\} 2^n\big(1 + \mathcal{O}(n^{-1})\big)\,, \tag{5}$$

where $\left\{ {n \atop m} \right\}$ denotes a Stirling number of second type, i.e., the number of ways of partitioning n elements into m non-empty blocks, and ω_k is defined as in (3) (the asymptotics of Stirling numbers in this regime is then extracted through a saddle-point analysis [11]).

When it is understood that $|\Sigma| = k$, a transition function δ is identified with a k-tuple of maps (or, for short, a *k-map*) $\delta_\alpha : Q \to Q$, as $\delta_\alpha(p) \equiv \Delta(p, \alpha)$ (in this case, to avoid confusion, we use Δ for the k-tuple of $\{\delta_\alpha\}_{1\le\alpha\le k}$). And, clearly, a k-map is identified with the corresponding vertex-labeled, edge-coloured digraph over n vertices, with uniform out-degree k, such that, for each vertex $i \in [n]$ and each colour $\alpha \in [k]$, there exists exactly one edge of colour α outgoing from i. The terminology of graph theory will occasionally be used in the following.

We use the word *motif* for an unlabeled oriented graph M, when it is intended as denoting the class of (edge-coloured) subgraphs of a k-map that are isomorphic to M. The core of our proof is in the analysis of the probability of occurrence of certain motifs, that we now introduce.

▶ **Definition 2.** A *M-motif* M of a transition structure $\mathcal{D}$ is a pair of states $i \neq j$, and an ordered k-tuple of states $\{\ell_\alpha\}_{1\le\alpha\le k}$, such that $\delta_\alpha(i) = \delta_\alpha(j) = \ell_\alpha$ (see Figure 1, left). Repetitions among ℓ_α's are allowed.

A *three-state M-motif* $M^{(3)}$ of a transition structure $\mathcal{D}$ is the analogue of a M-motif, with three distinct states i, j and h, such that $\delta_\alpha(i) = \delta_\alpha(j) = \delta_\alpha(h) = \ell_\alpha$ for all $1 \le \alpha \le k$ (see Figure 1, right).

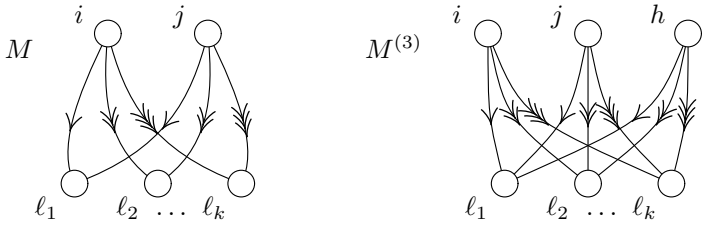

Figure 1 Left: a M-motif. Right: a three-state M-motif. The colouring of the edges is represented through the multiplicity of arrows. The examples are for $k = 3$.

The reason for studying M-motifs is in the two following easy remarks:

▶ Remark. If the transition structure of an automaton A contains a M-motif, with states i, j and $\{\ell_\alpha\}_{1\le\alpha\le k}$, and $\mathcal{T}(i) = \mathcal{T}(j)$, then $i \sim j$ and A is not minimal.

▶ Remark. Consider a transition structure $\mathcal{D}$ containing no three-state M-motifs, and r M-motifs with states $\left\{i^a, j^a, \{\ell^a_\alpha\}_{1\le\alpha\le k}\right\}_{1\le a\le r}$. Averaging over the possible sets of terminal states with the measure $\mu_b(\mathcal{T})$, the probability that $\mathcal{T}(i^a) = \mathcal{T}(j^a)$ for some $1 \le a \le r$ is $1 - (2b(1-b))^r$.

Our theorem results as a consequence of a number of statistical facts, on the structure of random automata, which are easy to believe although hard to prove. Thus, there is a short, non-rigorous path leading to the theorem, that we now explain.

1. A fraction $1-o(1)$ of non-minimal automata contain two Myhill-Nerode-equivalent states $i \sim j$ that are the incoming states of a M-motif.
2. Random transition structures locally "look like" random k-maps – this despite the highly non-local, and non-trivial, accessibility condition – the only remarkable difference being in the distribution of the incoming degrees r of the states, $p_r = 0$ if $r = 0$, and $\frac{1}{\omega_k}\mathrm{Poiss}_{k\omega_k}(r)$ if $r \ge 1$.

3. With this in mind, it is easy to calculate that the average number of M-motifs with equivalent incoming states is $(1-2b(1-b))\binom{n}{2}n^{-k}\left[\frac{\mathbb{E}(r(r-1)p_r)}{k^2}\right]^k$, at leading order in n, that is, $\frac{1}{2}(1-2b(1-b))\,{\omega_k}^k\,n^{-k+2}$.
4. Random transition structures also show weak correlations between distant parts, and M-motifs are 'small', thus, with high probability, pairs of M-motifs are non-overlapping. This suggests that the distribution of the number of M-motifs is a Poissonian, with the average calculated above (as if the corresponding events were decorrelated). As a corollary, we get the probability that there are no M-motifs. By the first claim, on the dominant role of M-motifs, this allows to conclude.

3 Structure of the proof

As it often happens, what seems the easiest way to get convinced of a claim is not necessarily the easiest path to produce a rigorous proof. Our proof strategy will be in fact very different from the sequence of claims collected above. As it is quite composite, in this section we will outline the subdivision of the proof into lemmas, and postpone the proofs to Section 4.

Call P_{rare} the probability, w.r.t. $\mu_b(\mathcal{D},\mathcal{T})$ above, that the transition structure contains no M-motif, and still the automaton is non-minimal. Call P_{confl} the probability that the transition structure contains some three-state M-motifs. Call $P(r)$ the probability that the transition structure contains no three-state M-motif, and exactly r M-motifs. Thus $P_{\text{confl}}+\sum_{r\geq 0}P(r)=1$.

The fraction of pairs $(\mathcal{D},\mathcal{T})$ of transition structures $\mathcal{D}$ with no three-state M-motif, and sets of terminal states $\mathcal{T}$ taken with the Bernoulli measure of parameter b such that $\mathcal{T}(i^a)=\mathcal{T}(j^a)$ for some M-motif, are $\sum_{r\geq 1}P(r)\,(1-(2b(1-b))^r)$. As a consequence, w.r.t. the measure $\mu_b(A)$ above, the probability that an automaton is non-minimal is

$$\mathbb{P}(A \text{ is non-minimal}) = \sum_{r\geq 1} P(r)\,(1-(2b(1-b))^r) + \mathcal{O}(P_{\text{rare}}) + \mathcal{O}(P_{\text{confl}})\,. \tag{6}$$

(To be precise, above we are neglecting summands $P^{(r)}_{\text{rare}}$, for $r\geq 1$, describing the probability of having no three-state M-motif, and exactly r M-motifs, all of which satisfying $\mathcal{T}(i^a)\neq\mathcal{T}(j^a)$, and still there exist pairs of equivalent states. From the treatment of the following section, it would be easily seen that these terms are negligible.)

If one can prove that $P_{\text{rare}}, P_{\text{confl}} = o\big(1-P(0)\big)$, then

$$\mathbb{P}(A \text{ is non-minimal}) = \sum_{r\geq 1} P(r)\big(1-(2b(1-b))^r + o(1)\big)\,. \tag{7}$$

In particular, if we can prove that $P(r)=\text{Poiss}_\rho(r)(1+o(1))$, with $\rho=\sum_r rP(r)$, it would follow that

$$\mathbb{P}(A \text{ is non-minimal}) = \big(1-e^{-\rho(1-2b(1-b))}\big)(1+o(1))\,. \tag{8}$$

This corresponds to the statement of Theorem 1, with $\rho=c_k n^{-k+2}$.

Note that our error term is not only small w.r.t. 1: as important for probabilities, it is small also w.r.t. $\min(p,1-p)$, with p the probability of our event of interest. As, for an alphabet with k letters, $p\sim n^{-k+2}$ has a non-trivial scaling with size when $k>2$, this difference is relevant.

So we see that Theorem 1 is implied by

▶ Proposition 1. The statements in the following list do hold

1. $P(r) = \mathrm{Poiss}_\rho(r)(1+o(1))$, for some ρ;
2. $\rho = c_k n^{-k+2}(1+o(1))$;
3. $P_{\mathrm{confl}} = o(n^{-k+2})$;
4. $P_{\mathrm{rare}} = o(n^{-k+2})$.

A collection of related, more explicit probabilistic statements is the following

▶ Proposition 2. The average number of occurrences of M-motifs M and three-state M-motifs $M^{(3)}$ in uniform random transition structures are respectively given by

$$\boldsymbol{m}(M) = \frac{1}{2} n^{-k+2} {\omega_k}^k \left(1+o(1)\right); \tag{9}$$

$$\boldsymbol{m}(M^{(3)}) = \frac{1}{6} n^{-2k+3} {\omega_k}^{2k} \left(1+o(1)\right). \tag{10}$$

Given that there are no three-state M-motifs, the average number of r-tuples $(M_1, \dots, M_r)$ of distinct M-motifs is given by

$$\frac{1}{r!}\boldsymbol{m}\big((M_1,\dots,M_r)\big) = \frac{1}{r!}\left(\frac{1}{2} n^{-k+2} {\omega_k}^k \left(1+o(1)\right)\right)^r. \tag{11}$$

The proof of this proposition is postponed to the end of Section 4.

Equation (9) proves $\rho = c_k n^{-k+2}\left(1+o(1)\right)$, that is, Part 2 of Proposition 1. Using the first-moment bound, equation (10) proves $P_{\mathrm{confl}} = \mathcal{O}(n^{-k+1})$ as required for Part 3 of Proposition 1.

The result in equation (11) concerning higher moments of M-motifs implies the proof of convergence of $P(r)$ to a Poissonian distribution, Part 1 of Proposition 1. The idea behind this claim is the fact that the occurrence of a M-motif with given states $\{i, j\}$ (and any k-tuple $\{\ell_\alpha\}$) is a 'rare' event, as it has a probability $\sim n^{-k}$, and, as the motifs are 'small' subgraphs, involving $\mathcal{O}(1)$ vertices, and parts of the transition structure $\mathcal{D}$ far away from each other (in the sense of distance on the graph) are weakly correlated, we expect the "Poisson Paradigm" to apply in this case, as discussed, for example, in Alon and Spencer [1, ch. 8]. A rigorous proof of this phenomenon can be achieved using the strategy called *Brun's sieve* (see e.g. [1, sec. 8.3]). The verification of the hypotheses discussed in the mentioned reference is exactly the statement of equation (11).

Thus, assuming Proposition 2, there is a single missing item in our 'checklist', namely, Part 4 of Proposition 1. We need to determine that $P_{\mathrm{rare}} = o(n^{-k+2})$. The idea behind this is that, in absence of M-motifs, there is a high probability that, for all pairs of states (i, j), certain isomorphic subgraphs of the two breadth-first search trees started from i and j visit a large number of states which are all distinct. For $i \sim j$, we need in particular that, for all the pairs of homologous states in these subgraphs, they are either both or none terminal states. The fact that they are all distinct implies that the probability for this to occur is a product of factors $1 - 2b(1-b)$, one for each such pair, thus we can concentrate on the estimation of the size of these subgraphs.

For the bounds that we need, it would suffice to produce isomorphic subgraphs of size of order $\frac{\ln n}{-\ln(1-2b(1-b))}$, but it will turn out that the largest possible subgraph is provably of size at least of order $n^{\frac{1}{4(k+1)}}$, and in fact conjecturally $\mathcal{O}(n)$.

Note that we need only an upper bound on P_{rare} (and no lower bound), and we have some freedom in producing bounds, as, at a heuristic level, we expect $P_{\mathrm{rare}} = \mathcal{O}(n^{-k+1}) \ll o(n^{-k+2})$. Our proof strategy will exploit this fact, and the following property of accessible transition functions (see [7]): given a random k-map $\Delta = \{\delta_\alpha(i)\}_{1\le i\le n, 1\le\alpha\le k}$, the number of states accessible from state 1 is a random variable $m = m(n,k)$, with average $\Theta(n)$ and

probability around the modal value[1] of order $n^{-\frac{1}{2}}$. Remarkably, if the accessible part has size m, then the induced transition structure is sampled uniformly among all transition structures of size m.

This has a direct simple consequence: if the average number of occurrences of a family of events on a random k-map is $\boldsymbol{m}(\{\mathcal{E}_i\})_{k\text{-maps}} = \mathcal{O}(n^{-\gamma})$, then the same average over random accessible transition functions of fixed size is bounded as $\boldsymbol{m}(\{\mathcal{E}_i\})_{\text{acc.}} \leq \mathcal{O}(n^{-\gamma+\frac{1}{2}})$. Actually, this bound is very generous and, if needed (but this is not our case), the extra exponent $\frac{1}{2}$ could be replaced by any $\epsilon > 0$ with some extra effort.

Thus, instead of proving that $P_{\text{rare}} = o(n^{-k+2})$, we will define the quantity P'_{rare}, exactly as P_{rare} but on random k-maps over n states. Note that the definitions of P_{rare} and P'_{rare} are based on two notions: not containing certain motifs, and not presenting pairs of Myhill-Nerode-equivalent states, and that both this notions are not confined to accessible automata, but are well-defined also for maps which are not accessible. Then we will prove that

▶ Proposition 3. $P'_{\text{rare}} = o(n^{-k+\frac{3}{2}})$.

In summary, as this proposition implies Part 4 of Proposition 1, Proposition 2 implies Parts 1 to 3 of Proposition 1, and Proposition 1 implies our main Theorem 1, we need to provide proofs of Propositions 2 and 3. This task is fulfilled in the following sections.

4 Proofs of the lemmas

Proof of Proposition 3. In a k-map, we say that a state i is a *sink state* if $\delta_\alpha(i) = i$ for all α. We say that two states $\{i, j\}$ form a *sink pair* if the set

$$N_{ij} = \{i, j, \delta_1(i), \delta_1(j), \cdots, \delta_k(i), \delta_k(j)\}$$

has cardinality $k+1$ or smaller. As easily seen through the first-moment bound, the probability of having any sink state or sink pair in a random k-map is at most of order n^{-k+1} (precisely, the overall constant is bounded by $1 + \frac{(k+1)^{2k}}{2(k-1)!}$). So, to prove that $P'_{\text{rare}} = o(n^{-k+\frac{3}{2}})$, it is enough to prove the same statement conditioned on the property that the k-map does not contain any sink state or pair.

We say that two states $\{i, j\}$ form a *quasi-sink pair* if the set N_{ij} has cardinality $k+2$. The average number of quasi-sink pairs in a random k-map is of order n^{-k+2}, thus this case must be analysed at our level of accuracy.

There exist three families of quasi-sink pairs: one family corresponds to pairs producing a M-motif; another family, that we call of *type-1*, corresponds to pairs for which there exists a value α such that $\{i, j, \delta_\alpha(i), \delta_\alpha(j)\}$ are all distinct; and a further family, that we call of *type-2*, corresponds to pairs for which any letter α is such that $\delta_\alpha(i)$ or $\delta_\alpha(j)$ is not repeated within N_{ij} (say that the first case occurs for h letters, and the second one for the remaining $k-h$ ones).

In evaluating P'_{rare}, we have excluded the M-motif case, and we are left only with type-1 and type-2 quasi-sinks. Furthermore, we have excluded sink states, so in type-2 quasi-sinks we must have both h and $k-h$ non-zero.

For a type-1 quasi-sink $\{i, j\}$, define the *pair following* $\{i, j\}$ as the pair $\{i', j'\}$ such that $i' = \delta_\alpha(i)$, $j' = \delta_\alpha(j)$, for α the first lexicographic letter such that $\{i, j, \delta_\alpha(i), \delta_\alpha(j)\}$ are all distinct. For a type-2 quasi-sink $\{i, j\}$ define the *pair following* $\{i, j\}$ as the pair $\{i', j'\}$ with $i' = \delta_1(i)$, $j' = \delta_1(j)$. Again, by first-moment estimate, the probability that there exists

[1] I.e., the most probable value.

a quasi-sink pair $\{i,j\}$, such that also the pair following it is a quasi-sink, is bounded by $\mathcal{O}(n^{-k+1})$ (for which we need that $h(k-h)>0$ in a type-2 quasi-sink), and we can further condition our k-map not to contain such motifs. If $\{i,j\}$ is a quasi-sink pair, a necessary condition for $i\sim j$ is that also $i'\sim j'$. Thus, we can bound P'_{rare} by the probability that there exist no non–quasi-sink pairs in the k-map. This is the formulation of the problem that we ultimately address.

Consider a non–quasi-sink pair $\{i,j\}$, and construct the lexicographic breadth-first tree exploration, simultaneously on the two states i and j, neglecting those branches in which, in one or both of the two trees, there is a state already visited by the exploration (call *leaves* these nodes).

Call $(v_1,v_2,\dots)$ the ordered sequence of steps in the breadth-first search, at which a leaf node is visited. For fixed integers v and h, we want to determine the probability of the event $v_h\le v$, conditioned to the event that the list has at least h items. By standard estimate of factorials, and crucially making use of the exclusion of sink and quasi-sink motifs, it can be proved for this quantity

$$\mathbb{P}(v_h\le v)\le\frac{1}{h!}\left(\frac{v(v+1)}{n-2v}\right)^h. \tag{12}$$

Set $h=k+1$. By definition, in a non–quasi-sink pair, we certainly have at least $k+1$ entries v_j. If $v=\mathcal{O}(n^\gamma)$ for some $0<\gamma<1$, we have that for each non–quasi-sink pair $\{i,j\}$

$$\mathbb{P}(v_{k+1}^{(ij)}\le v)\le\mathcal{O}(n^{-(k+1)(1-2\gamma)}). \tag{13}$$

The number of non–quasi-sink pairs is bounded by $\binom{n}{2}$, thus by first-moment bound

$$\mathbb{P}(v_{k+1}^{(ij)}\le v\text{ for all }\{i,j\})\le\mathcal{O}(n^{-k+1+2\gamma(k+1)}). \tag{14}$$

For $\gamma<\frac{1}{4(k+1)}$ we thus get $\mathbb{P}(v_{k+1}^{(ij)}\le v\text{ for all }\{i,j\})\le o(n^{-k+\frac{3}{2}})$ as needed. Thus, we know that, with probability larger than $1-o(n^{-k+\frac{3}{2}})$, all the non–quasi-sink pairs in our k-map have $v_{k+1}\gtrsim n^\gamma$, for any $\gamma<\frac{1}{4(k+1)}$. This means that, if we truncate the breadth-first search tree exploration to a depth of order $\gamma\frac{\ln n}{\ln k}$, we have at most k leaves in the tree. Thus, for all the trees, we have at least order n^γ internal nodes, i.e. pairs of states (i',j') for which it is required $\mathcal{T}(i')=\mathcal{T}(j')$ for having $i\sim j$.

But, as all these states appear not repeated in the exploration, the probability that $i\sim j$ is bounded by an exponential of the form $(1-2b(1-b))^{n^\gamma}$, which decreases faster than any power law. The overall factor $\binom{n}{2}$ from the first-moment bound is irrelevant, and we are able to conclude that $P'_{\text{rare}}=o(n^{-k+\frac{3}{2}})$, as needed. Note that this proof works not only for finite values of b in the open interval $]0,1[$, as required for our purposes, but even up to $b\gg n^{-\gamma}$. ◀

Before passing to the proof of Proposition 2, we need to recall the relation between accessible deterministic complete automata and combinatorial objects known as *k-Dyck tableaux* [4], and determine a collection of statistical properties of these tableaux.

Given the integers M and n, a *tableau* T in the set $\mathcal{T}[M\times n]$ is a map from $[M]$ to $[n]$ such that:

1. every value $y\in[n]$ has at least one preimage;
2. calling $x_T(y)$ the smallest preimage, we have $x_T(1)<x_T(2)<\ldots<x_T(n)$.

The tableau T may be represented graphically, on a $M\times n$ grid, by marking the M pairs $\{(x,T(x))\}_{1\le x\le M}$. Then the conditions above translate as follows. There is exactly one

marked entry per column. Mark in red the pairs $(x_T(y), y)$, and in black the remaining ones: there is exactly one red entry per row, which is at the left of all black entries in the same row (if any), and the polygonal line connecting the red entries in sequence is monotonically increasing. We call the collections of positions of red and black marks respectively the *backbone* B_T and *wiring part* W_T of the tableau T. It is easily seen that the number of tableaux in $\mathcal{T}[M \times n]$ is given by the Stirling number of second type $\left\{ {M \atop n} \right\}$. The asymptotic evaluation of $\left\{ {M \atop n} \right\}$, for n large and $M/n = \mathcal{O}(1)$, can be done through the general methods of analytic combinatorics (see e.g. [10], and in particular [11] for this specific problem). A result of this calculation that we shall need is the following

▶ Proposition 4. If $M, M' = \kappa n + \mathcal{O}(1)$, with $\kappa > 1$, calling ω the only solution of the equation $-\kappa\omega = \ln(1-\omega)$ in $[0,1]$,

$$\left\{ {M \atop n} \right\} = \left\{ {M' \atop n} \right\} \left(\frac{n}{\omega}\right)^{M-M'} \left(1 + o(1)\right). \tag{15}$$

For a fixed integer k, when $M = N(n,k) = kn+1$, we have a special subfamily of tableaux in $\mathcal{T}[N \times n]$. A tableau is *k-Dyck* if $x_T(\ell) \leq k(\ell-1)+1$, i.e. if the backbone cells lie above the line of slope $1/k$ containing the origin of the grid. A small example of k-Dyck tableau is shown in Figure 2.

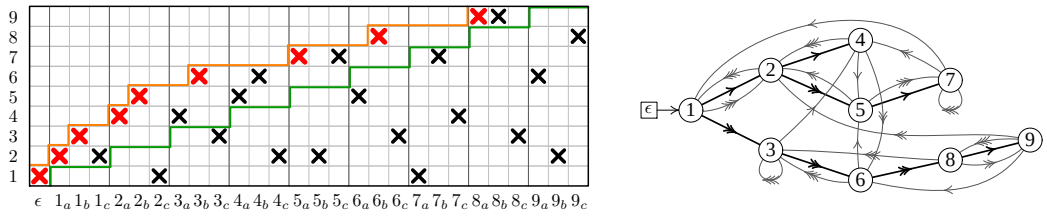

Figure 2 Left: a tableau with $n = 9$ and $k = 3$. The backbone part is in red. This tableau is valid because the red entries are monotonic (as shown by the orange profile), and k-Dyck because they are all on the left of the green staircase line. Right: the associated k-map. Backbone edges, corresponding to the breadth-first search tree, are in black, and wiring edges are in gray.

There exists a canonical bijection between k-Dyck tableaux and transition structures $\mathcal{D}$ of accessible deterministic complete automata. It suffices to associate the indices $(1, 2, \ldots, n)$ of the states to the rows of the tableaux, and the indices $(\epsilon, 1_1, \ldots, 1_k, \cdots, n_1, \ldots, n_k)$ of the oriented edges of $\mathcal{D}$ to the columns. Then, for $x = i_\alpha$, the entry (x, y) is marked in T if and only if $\delta_\alpha(i) = y$, and it is part of the backbone if and only if it is part of the breadth-first search tree on $\mathcal{D}$ started at the initial state.

Given a function $\hat{f}(y) : [n] \to [M]$, consider the restriction of the set $\mathcal{T}[M \times n]$ to tableaux T in which the backbone function $x_T(y)$ is dominated by $\hat{f}$, i.e., such that $x_T(y) \leq \hat{f}(y)$ for all $1 \leq y \leq n$. Call $\mathcal{T}[M \times n; \hat{f}]$ this set. Our k-Dyck tableaux correspond to the special case $\mathcal{T}[N \times n; \hat{f}^{\varnothing}]$, with $\hat{f}^{\varnothing}(y) := N - k(n - y + 1)$. A required technical lemma, that we state without proof, is the following

▶ Proposition 5. Take an integer n, $N = \mathcal{O}(n)$, $B = \mathcal{O}(1)$, and $\ell \gg \sqrt{n}$. Let $M = N - B$, and take a function $\hat{f}$ such that $\hat{f}(y) = \hat{f}^{\varnothing}(y)$ for all $y \leq \ell$, $\hat{f}(y) = \hat{f}^{\varnothing}(y) - B$ for all

$y \geq n-\ell$, and $\hat{f}^{\varnothing}(y) - B \leq \hat{f}(y) \leq \hat{f}^{\varnothing}(y)$ for all y. Then

$$\frac{\left|cT[M\times n;\hat{f}]\right|}{\left|cT[M\times n]\right|} - \frac{\left|cT[N\times n;\hat{f}^{\varnothing}]\right|}{\left|cT[N\times n]\right|} = o(1)\,. \tag{16}$$

With these tools at hand, we are now ready to prove Proposition 2.

Proof of Proposition 2 (p.93). Given three distinct states i, j, h, with $i<j<h$, call $\mathcal{M}_{ijh}(T)$ the event that in the tableau T there is a three-state motif on states $\{i,j,h\}$ and $\{\ell_\alpha\}$, for some ℓ_α's. Similarly, given $2r$ distinct states $\{(i_a,j_a)\}_{1\leq a\leq r}$, with $i_a<j_a$ and $j_a<j_{a+1}$, call $\mathcal{M}_{(i_1,j_1;\ldots;i_r,j_r)}(T)$ the event that in the tableau T there is a r-tuple of M-motifs, such that the a-th motif has states i_a, j_a, and $\{\ell^a_\alpha\}$, for some ℓ^a_α's. Proposition 2 consists in evaluating the two quantities

$$\sum_{i<j<h} \mathbb{E}[\![\mathcal{M}_{ijh}]\!]_{\mathcal{T}[N\times n;\hat{f}^{\varnothing}]}\,; \qquad \sum_{(i_1,j_1;\ldots;i_r,j_r)} \mathbb{E}[\![\mathcal{M}_{(i_1,j_1;\ldots;i_r,j_r)}]\!]_{\mathcal{T}[N\times n;\hat{f}^{\varnothing}]}\,. \tag{17}$$

We now make a crucial remark: given a backbone structure B, the average over all possible completions of the indicator variables $[\![\mathcal{M}_{ijh}]\!]$ (respectively $[\![\mathcal{M}_{(i_1,j_1;\ldots;i_r,j_r)}]\!]$) is zero if any column of index in the set $C=\{k(j-1)+1+\alpha, k(h-1)+1+\alpha\}_{1\leq\alpha\leq k}$ has a red mark (respectively, in the set $C=\{k(j_a-1)+1+\alpha\}_{1\leq a\leq r;1\leq\alpha\leq k}$), otherwise, it is $\prod_{i\in C} y_i^{-1}$, where y_i is the height of the backbone profile at column i. As a consequence, backbone structures contributing to the quantities in (17), weighted with the factor $\mu(\boldsymbol{c})\prod_{i\in C} y_i^{-1}$, correspond to generic backbone structures, weighted with the factor $\mu(\boldsymbol{c})$, over $(N-kr)\times n$ tableaux. The correspondence is done by just erasing the columns in C. The function $\hat{f}$ is modified accordingly. Define

$$\hat{f}^{i_1,\ldots,i_r}(y) = \hat{f}^{\varnothing}(y) - k\sum_{a=1}^{r}[\![y\geq j_a]\!]\,. \tag{18}$$

Then, the precise statement of the remark above is

$$\mathbb{E}[\![\mathcal{M}_{ijh}]\!]_{\mathcal{T}[N\times n;\hat{f}^{\varnothing}]} = \frac{\left|\mathcal{T}[(N-2k)\times n;\hat{f}^{j,h}]\right|}{\left|\mathcal{T}[N\times n;\hat{f}^{\varnothing}]\right|}\,; \tag{19}$$

$$\mathbb{E}[\![\mathcal{M}_{(i_1,j_1;\ldots;i_r,j_r)}]\!]_{\mathcal{T}[N\times n;\hat{f}^{\varnothing}]} = \frac{\left|\mathcal{T}[(N-kr)\times n;\hat{f}^{j_1,\ldots,j_r}]\right|}{\left|\mathcal{T}[N\times n;\hat{f}^{\varnothing}]\right|}\,. \tag{20}$$

Thus, the right-hand side of (19) is just the special case $r=2$ of (20). Of course we have

$$\frac{\left|\mathcal{T}[(N-kr)\times n;\hat{f}^{j_1,\ldots,j_r}]\right|}{\left|\mathcal{T}[N\times n;\hat{f}^{\varnothing}]\right|} = \frac{\frac{\left|\mathcal{T}[(N-kr)\times n;\hat{f}^{j_1,\ldots,j_r}]\right|}{\left|\mathcal{T}[(N-kr)\times n]\right|}}{\frac{\left|\mathcal{T}[N\times n;\hat{f}^{\varnothing}]\right|}{\left|\mathcal{T}[N\times n]\right|}}\,\frac{\left|\mathcal{T}[(N-kr)\times n]\right|}{\left|\mathcal{T}[N\times n]\right|}\,. \tag{21}$$

We can apply Proposition 4 to the rightmost ratio. Then, if the j_a's are within the range for application of Proposition 5, we can also simplify the leftmost ratio, to get

$$\mathbb{E}[\![\mathcal{M}_{ijh}]\!]_{\mathcal{T}[N\times n;\hat{f}^{\varnothing}]} \simeq \left(\frac{\omega_k}{n}\right)^{2k}\,; \tag{22}$$

$$\mathbb{E}[\![\mathcal{M}_{(i_1,j_1;\ldots;i_r,j_r)}]\!]_{\mathcal{T}[N\times n;\hat{f}^{\varnothing}]} \simeq \left(\frac{\omega_k}{n}\right)^{kr}\,. \tag{23}$$

As in Proposition 5 we just asked for $\ell \gg \sqrt{n}$, which is compatible with $\ell \ll n$, the fraction of $2r$-tuples $(i_1, j_1; \ldots; i_r, j_r)$ such that some j_a's are out of range is subleading, and, using the reasonings at the beginning of Section 3, the corresponding contribution can be included in P_{confl}.

Then, the straightforward calculation of the number of triplets (i, j, h), and $2r$-tuplets $\{(i_a, j_a)\}_{1 \leq a \leq r}$, at leading order in n, allows to conclude. ◀

5 Algorithmic consequences

The results obtained in this paper open new possibilities for the study in average of the properties of regular languages, and of the average-case complexity of algorithms applied to minimal automata. In this section we mention just a few among these consequences.

▶ **Corollary 3.** *Minimal automata with n states over a k-letter alphabet can be randomly generated with $\mathcal{O}(n^{3/2})$ average complexity, using Boltzmann samplers.*

The random generator for complete deterministic accessible automata given in [4] is based on a Boltzmann sampler [9], its average complexity is $\mathcal{O}(n^{3/2})$. As from Theorem 1 there is a constant proportion of minimal automata amongst accessible ones, the rejection method can be efficiently applied to randomly generate a minimal automaton. Note that such a generator (described in [4]) has already been implemented in REGAL,[2] a C++-library for the random generation of automata [2], though there were no theoretical result on the efficiency of this algorithm at that time.

▶ **Corollary 4.** *For the uniform distribution on complete deterministic accessible automata, the average complexity of Moore's state minimization algorithm is $\Theta(n \log\log n)$.*

Proof. The average complexity of Moore's state minimization algorithm for the uniform distribution on n-state deterministic automata over a finite alphabet is $\mathcal{O}(n \log\log n)$ [8]. The upper bound for accessible automata is then obtained studying the size of the accessible part of a k-random map [7, 18]. Moreover from [3] the lower bound of Moore's algorithm applied on minimal automata with n states is $\Omega(n \log\log n)$. Using Theorem 1, this is also a lower bound for complete deterministic accessible automata. ◀

▶ **Corollary 5.** *For the uniform distribution on complete deterministic accessible automata, there exists a family of implementations of Hopcroft's state minimization algorithm whose average complexity is $\mathcal{O}(n \log\log n)$.*

From [8] a family of implementations of Hopcroft's state minimization algorithm are always faster than Moore's algorithm. The result follows from Corollary 4. In [5] the lower bound on the algorithm is proved to be $\mathcal{O}(n \log n)$ for any implementation. Though it is still unknown whether there exists an implementation whose average complexity is $\Theta(n)$.

References

1 N. Alon and J. Spencer. *The Probabilistic Method.* 2nd ed., John Wiley, 2000.
2 F. Bassino, J. David and C. Nicaud. REGAL: A library to randomly and exhaustively generate automata. In J. Holub and J. Zdárek eds, *12th Int. Conference Implementation and Application of Automata (CIAA 2007)*, LNCS 4783, 303–305. Springer, 2007.

[2] Available at `http://regal.univ-mlv.fr/`

3 F. Bassino, J. David and C. Nicaud. Average-case analysis of Moore's state minimization algorithm. *Algorithmica,* to appear.

4 F. Bassino and C. Nicaud. Enumeration and random generation of accessible automata. *Theor. Comput. Sci.*, **381** 86–104, 2007.

5 J. Berstel, L. Boasson and O. Carton. Continuant polynomials and worst-case behavior of Hopcroft's minimization algorithm. *Theor. Comput. Sci.*, **410** 2811–2822, 2009.

6 J.R. Buchi. Weak second-order arithmetic and finite automata. *Math. Logic Quart.*, **6** 66–92, 1960.

7 A. Carayol and C. Nicaud. Distribution of the number of accessible states in a random deterministic automaton. In *29th International Symposium on Theoretical Aspects of Computer Science (STACS 2012)*, Paris, March 2012.

8 J. David. Average complexity of Moore's and Hopcroft's algorithms. *Theor. Comput. Sci.* to appear.

9 P. Duchon, P. Flajolet, G. Louchard and G. Schaeffer. Boltzmann Samplers for the Random Generation of Combinatorial Structures. In *Combinatorics, Probability, and Computing,* Special issue on Analysis of Algorithms **13** 577–625, 2004.

10 P. Flajolet and R. Sedgewick. *Analytic Combinatorics.* Cambridge Univ. Press, 2009.

11 I.J. Good, An Asymptotic Formula for the Differences of the Powers at Zero. *Ann. Math. Stat.* **32** 249–256, 1961.

12 F. Harary. Unsolved problems in the enumeration of graphs. *Publ. Math. Inst. Hungar. Acad. Sci.*, **5** 63–95, 1960.

13 M.A. Harrison. A census of finite automata, *Canad. Journ. of Math.*, **17** 100–113, 1965.

14 J.E. Hopcroft and J.D. Ullman. *Introduction to Automata Theory, Languages and Computation.* Addison-Wesley, 1979.

15 J.E. Hopcroft. An $n \log n$ algorithm for minimizing states in a finite automaton. Technical report, Stanford CA, USA, 1971.

16 R. Iranpour and P. Chacon. *Basic Stochastic Processes: The Mark Kac Lectures.* Macmillan Publ. Co., 1988.

17 S. Kleene. Representation of Events in Nerve Nets and Finite Automata. In C. Shannon and J. Mccarthy eds., *Automata Studies*, 3–42. Princeton University Press, 1956.

18 A.D. Korshunov. Enumeration of finite automata. *Problemy Kibernetiki*, **34** 5–82, 1978. In Russian.

19 A.D. Korshunov. On the number of non-isomorphic strongly connected finite automata. *Journal of Information Processing and Cybernetics*, **9** 459–462, 1986.

20 E. Lebensztayn. On the asymptotic enumeration of accessible automata. *Discr. Math. Theor. Comp. Science* **12** 75–80, 2010

21 V.A. Liskovets. Enumeration of non-isomorphic strongly connected automata, *Vesci Akad. Navuk BSSR, Ser. Fiz.-Mat. Navuk,* **3** 26–30, 1971. In Russian.

22 V.A. Liskovets. Exact enumeration of acyclic automata. In *15-th International Conference on Formal Power Series and Algebraic Combinatorics (FPSAC'03)*, 2003.

23 E.F. Moore. Gedanken experiments on sequential machines. In *Automata Studies,* Princeton Univ., 129–153, 1956.

24 A. Nerode. Linear automaton transformations. *Proc. of the American Math. Society*, **9** 541–544, 1958.

25 C. Nicaud. Average state complexity of operations on unary automata. In *24th Int. Symposium on Mathematical Foundations of Computer Science (MFCS 1999)*, 231–240, 1999.

26 R. Robinson, Counting strongly connected finite automata, In *Graph theory with Applications to Algorithms and Computer Science*, Y. Alavi et al. eds., Wiley, 671–685, 1985.

27 S. Yu, Q. Zhuang and K. Salomaa, The state complexities of some basic operations on regular languages, *Theoret. Comput. Sci.*, **125** 315–328, 1994.