On the State Complexity of the Shuffle of Regular Languages

Janusz Brzozowski, Galina Jirásková, Bo Liu, Aayush Rajasekaran, Marek Szykula

On the State Complexity of the Shuffle of Regular Languages*

Janusz Brzozowski[1], Galina Jirásková[2], Bo Liu[1**], Aayush Rajasekaran[1], and Marek Szykuła[3]

[1] David R. Cheriton School of Computer Science, University of Waterloo, Waterloo, ON, Canada N2L 3G1
{{brzozo, b23liu, arajasek}@uwaterloo.ca}
[2] Mathematical Institute, Slovak Academy of Sciences, Grešákova 6, 040 01 Košice, Slovakia
jiraskov@saske.sk
[3] Institute of Computer Science, University of Wrocław, Joliot-Curie 15, PL-50-383 Wrocław, Poland
{msz@cs.uni.wroc.pl}

Abstract. We investigate the shuffle operation on regular languages represented by complete deterministic finite automata. We prove that $f(m,n) = 2^{mn-1} + 2^{(m-1)(n-1)}(2^{m-1}-1)(2^{n-1}-1)$ is an upper bound on the state complexity of the shuffle of two regular languages having state complexities m and n, respectively. We also state partial results about the tightness of this bound. We show that there exist witness languages meeting the bound if $2 \leqslant m \leqslant 5$ and $n \geqslant 2$, and also if $m = n = 6$. Moreover, we prove that in the subset automaton of the NFA accepting the shuffle, all 2^{mn} states can be distinguishable, and an alphabet of size three suffices for that. It follows that the bound can be met if all $f(m,n)$ states are reachable. We know that an alphabet of size at least mn is required provided that $m, n \geqslant 2$. The question of reachability, and hence also of the tightness of the bound $f(m,n)$ in general, remains open.

Keywords: regular language, shuffle, state complexity, upper bound

1 An Upper Bound for the Shuffle Operation

The *state complexity of a regular language* L [6] is the number of states in a complete minimal deterministic finite automaton (DFA) recognizing the language; it will be denoted by $\kappa(L)$. The *state complexity of an operation* on regular languages is the maximal state complexity of the result of the operation expressed as a function of the state complexities of the operands.

* This work was supported by the Natural Sciences and Engineering Research Council of Canada under grant No. OGP0000871, by VEGA grant 2/0084/15, and by the National Science Centre, Poland under project number 2014/15/B/ST6/00615.

** Present address: Google Inc., 1600 Amphitheatre Parkway, Mountain View, CA 94043, USA.

Let Σ be a finite non-empty alphabet. The *shuffle* $u \sqcup\!\sqcup v$ of words $u, v \in \Sigma^*$ is defined as follows:

$$u \sqcup\!\sqcup v = \{u_1 v_1 \cdots u_k v_k \mid u = u_1 \cdots u_k, v = v_1 \cdots v_k, u_1, \ldots, u_k, v_1, \ldots, v_k \in \Sigma^*\}.$$

The shuffle of two languages K and L over Σ is defined by

$$K \sqcup\!\sqcup L = \bigcup_{u \in K, v \in L} u \sqcup\!\sqcup v.$$

Note that the shuffle operation is commutative on both words and languages.

The state complexity of the shuffle operation was first studied by Câmpeanu, Salomaa, and Yu [2], but they considered only bounds for incomplete deterministic automata. In particular, they proved that $2^{mn} - 1$ is a tight upper bound for that case. Since we can convert an incomplete deterministic automaton into complete one by adding the empty state, it follows that $2^{(m-1)(n-1)} - 1$ is a lower bound for the case of complete deterministic automata. Here we show that this lower bound can be improved, and we derive an upper bound for two regular languages represented by complete deterministic automata, but the question whether this bound is tight remains open.

A *nondeterministic finite automaton* (NFA) is a quintuple $\mathcal{A} = (Q, \Sigma, \delta, s, F)$, where Q is a finite non-empty set of states, Σ is a finite alphabet of input symbols, $\delta\colon Q \times \Sigma \to 2^Q$ is the transition function which is extended to the domain $2^Q \times \Sigma^*$ in the natural way, $s \in Q$ is the initial state, and $F \subseteq Q$ is the set of final states. The language accepted by NFA $\mathcal{A}$ is the set of words $L(\mathcal{A}) = \{w \in \Sigma^* \mid \delta(s, w) \cap F \neq \emptyset\}$.

An NFA $\mathcal{A}$ is *deterministic and complete* (DFA) if $|\delta(q, a)| = 1$ for each q in Q and each a in Σ. In such a case, we write $\delta(q, a) = q'$ instead of $\delta(q, a) = \{q'\}$. A DFA is *minimal* (with respect to the number of states) if all its states are reachable, and no two distinct states are equivalent.

Every NFA $\mathcal{A} = (Q, \Sigma, \delta, s, F)$ can be converted to an equivalent DFA $\mathcal{A}' = (2^Q, \Sigma, \delta, \{s\}, F')$, where $F' = \{R \in 2^Q \mid R \cap F \neq \emptyset\}$. The DFA $\mathcal{A}'$ is called the *subset automaton* of NFA $\mathcal{A}$. The subset automaton may not be minimal since some of its states may be unreachable or equivalent to other states.

Let K and L be regular languages over an alphabet Σ recognized by deterministic finite automata $\mathcal{K} = (Q_K, \Sigma, \delta_K, q_K, F_K)$ and $\mathcal{L} = (Q_L, \Sigma, \delta_L, q_L, F_L)$, respectively. Then $K \sqcup\!\sqcup L$ is accepted by the nondeterministic finite automaton

$$\mathcal{N} = (Q_K \times Q_L, \Sigma, \delta, (q_K, q_L), F_K \times F_L),$$

where

$$\delta((p, q), a) = \{(\delta_K(p, a), q), (p, \delta_L(q, a))\}.$$

Let $\mathcal{D} = (2^{Q_K \times Q_L}, \Sigma, \delta', \{(q_K, q_L)\}, F')$ be the subset automaton of $\mathcal{N}$. If $|Q_K| = m$ and $|Q_L| = n$, then NFA $\mathcal{N}$ has mn states. It follows that DFA $\mathcal{D}$ has at most 2^{mn} reachable and pairwise distinguishable states. However, this upper bound cannot be met, as we will show.

In the sequel, we assume that $Q_K = \{1, 2, \ldots, m\}$, $q_K = 1$, $Q_L = \{1, 2, \ldots, n\}$, and $q_L = 1$. We say that a state (p, q) of NFA $\mathcal{N}$ is in row i if $p = i$, and it is in column j if $q = j$.

Proposition 1. *Let $a \in \Sigma$. Let S be a state of $\mathcal{D}$. Let $\pi_{\mathrm{col}}(S) = \{p \mid (p, q) \in S \text{ for some } q\}$, and $\pi_{\mathrm{row}}(S) = \{p \mid (p, q) \in S \text{ for some } p\}$. Then $\pi_x(S) \subseteq \pi_x(S \cdot a)$ for $x \in \{\mathrm{col}, \mathrm{row}\}$.*

Proof. Let $p \in \pi_{\mathrm{col}}(S)$; then we have $(p, q) \in S$ for some q. Since $\delta((p, q), a) = \{(\delta_K(p, a), q), (p, \delta_L(q, a)\}$, we have $(p, \delta_L(q, a)) \in \delta(S, a)$, so $p \in \pi_{\mathrm{col}}(\delta(S, a))$. By symmetry, the same claim holds for π_{row}. □

We claim that in the subset automaton $\mathcal{D}$, every reachable subset S of $Q_K \times Q_L$ must contain a state in column 1 and a state in row 1, that is, it must satisfy the following condition.

Condition (C): There exist states $(s, 1)$ and $(1, t)$ in S for some $s \in Q_K$ and $t \in Q_L$.

Lemma 2. *Every reachable subset S of subset automaton $\mathcal{D}$ satisfies Condition (C).*

Proof. The initial subset of $\mathcal{D}$ is $\{(1, 1)\}$, and it satisfies Condition (C). By Proposition 1, for every $a \in \Sigma$ we get that $1 \in \pi_{\mathrm{col}}(\delta(S, a))$ and $1 \in \pi_{\mathrm{row}}(\delta(S, a))$, so $\delta(S, a)$ satisfies Condition (C). By induction, all reachable subsets satisfy Condition (C). □

Theorem 3 (Shuffle: Upper Bound). *Let $\kappa(K) = m$ and $\kappa(L) = n$. Then the state complexity of the shuffle of K and L is at most*

$$f(m, n) = 2^{mn-1} + 2^{(m-1)(n-1)}(2^{m-1} - 1)(2^{n-1} - 1). \tag{1}$$

Proof. By Lemma 2, every reachable subset of $\mathcal{D}$ must contain a state in row 1 and a state in column 1. There are 2^{mn-1} subsets containing state $(1, 1)$, and $2^{(m-1)(n-1)}(2^{m-1} - 1)(2^{n-1} - 1)$ subsets not containing $(1, 1)$ but containing $(s, 1)$ for some $s \in \{2, 3, \ldots, m\}$ and $(1, t)$ for some $t \in \{2, 3, \ldots, n\}$. This gives $f(m, n)$. □

Let K and L be two regular languages over Σ. If $\kappa(K) = \kappa(L) = 1$, then each of K, L, and $K \sqcup\!\sqcup L$ is either $\emptyset$ or Σ^*, and $\kappa(K \sqcup\!\sqcup L) = 1$; hence the bound $f(1, 1) = 1$ is tight.

Now suppose that $\kappa(K) = 1$; here we have two possible choices for K, the empty language or Σ^*. The first choice leads to $\kappa(K \sqcup\!\sqcup L) = 1$. Hence only the second choice is of interest, where the language $K \sqcup\!\sqcup L = \Sigma^* \sqcup\!\sqcup L$ is the all-sided ideal [1] generated by L. If $\kappa(L) = 2$, the upper bound $f(1, 2) = 2$ is met by the unary language $L = aa^*$. Hence assume that $\kappa(K) = 1$ and $\kappa(L) \geqslant 3$. The next observation shows that in such a case, the tight bound is less than $f(1, n) = 2^{n-1}$.

Proposition 4 (Okhotin [4]). *If $\kappa(L) \geqslant 3$, then the state complexity of $\Sigma^* \sqcup\!\sqcup L$ is at most $2^{n-2}+1$, and this bound can be reached only if $|\Sigma| \geqslant n-2$.*

Okhotin showed that the language $L = (a_1\Sigma^* a_1 \cup \cdots \cup a_{n-2}\Sigma^* a_{n-2})\Sigma^*$, where $\Sigma = \{a_1, \ldots, a_{n-2}\}$, meets this bound [4]. This takes care of the case $\kappa(K) = 1$ and, by symmetry, of the case $\kappa(L) = 1$.

In what follows we assume that $m \geqslant 2$ and $n \geqslant 2$. First, let us show that the upper bound $f(m,n)$ cannot be met by regular languages defined over a fixed alphabet.

Proposition 5. *Let K and L be regular languages over Σ with $\kappa(K) = m$ and $\kappa(L) = n$, where $m, n \geqslant 2$. If $\kappa(K \sqcup\!\sqcup L) = f(m,n)$, then $|\Sigma| \geqslant mn-1$.*

Proof. For $s = 2, 3, \ldots, m$ and $t = 2, 3, \ldots, n$ denote

$$\begin{aligned} A_s &= \{(1,1),(s,1)\}, \\ B_t &= \{(1,1),(1,t)\}, \\ C_{st} &= \{(s,1),(1,t)\}. \end{aligned}$$

If all the subsets satisfying Condition (C) are reachable, then, in particular, all the subsets A_s, B_t, and C_{st} must be reachable. Let us show that all these subsets must be reached from some subsets containing state $(1,1)$ by distinct symbols.

Suppose that a set A_s is reached from a reachable set S with $S \neq A_s$ by a symbol a, that is, we have $A_s = \delta(S,a)$ and $S \neq A_s$. The set A_s contains only states in column 1 and rows 1 or s. By Proposition 1, the set S may only contain states in column 1 and in rows 1 or s, that is, we have $S \subseteq \{(1,1),(s,1)\}$. Since $S \neq A_s$, we must have $S = \{(1,1)\}$.

By symmetry, each B_t can only be reached from $\{(1,1)\}$.

Suppose that a set C_{st} is reached from a reachable set S with $S \neq C_{st}$ by a symbol a. By Proposition 1, we must have $S \subseteq \{(1,1),(s,1),(1,t),(s,t)\}$. Let us show that $(1,1) \in S$. Suppose for a contradiction that $(1,1) \notin S$. Then, since S is reachable, it must contain a state in column 1 and a state in row 1, that is, we must have $\{(s,1),(1,t)\} \subseteq S$. But then $(s,t) \in S$ since $S \neq C_{st}$. However, then $\delta_K(s,a) = 1$ and $\delta_L(t,a) = 1$ which implies that $(1,1) \in \delta((s,1),a)$, and so $(1,1) \in C_{st}$. This is a contradiction. Therefore C_{st} is reached from a set containing $(1,1)$.

Thus each A_s is reached from $\{(1,1)\}$ by a symbol a_s, each B_t is reached from $\{(1,1)\}$ by a symbol b_t, each C_{st} is reached from a set containing $(1,1)$ by a symbol c_{st}, and we must have

$$\begin{aligned} \delta_K(1,a_s) &= s \text{ and } \delta_L(1,a_s) = 1, \\ \delta_K(1,b_t) &= 1 \text{ and } \delta_L(1,b_t) = t, \\ \delta_K(1,c_{st}) &= s \text{ and } \delta_L(1,c_{st}) = t. \end{aligned}$$

It follows that all the symbols a_s, b_t, and c_{st} must be pairwise distinct. Therefore we have $|\Sigma| \geqslant m-1+n-1+(m-1)(n-1) = mn-1$. □

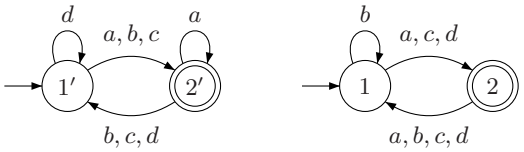

Fig. 1. Witness DFAs $\mathcal{K}$ and $\mathcal{L}$ for shuffle with $|Q_K| = 2$, $|Q_L| = 2$.

Unfortunately, this lower bound on the size of the alphabet is not tight, as is demonstrated by the following example:

Example 6. If t is a transformation of the set $\{1, 2, \ldots, n\}$ and $q \in \{1, 2, \ldots, n\}$, let qt be the image of q under t. Transformation t can now be denoted by $[1t, 2t, \ldots, nt]$.

(1) If $m = n = 2$, we have $f(2, 2) = 10$. Let $\Sigma = \{a, b, c, d\}$, and let the DFAs $\mathcal{K}$ and $\mathcal{L}$ be as shown in Fig. 1, and let K and L be their languages. Then $\kappa(K \sqcup\!\sqcup L) = 10$. We have used GAP [3] to show that the bound cannot be reached with a smaller alphabet, and that the DFAs of Fig. 1 are unique up to isomorphism.

(2) For $m = 2$ and $n = 3$, the minimal size of the alphabet of a witness pair is 6. We have verified this by a dedicated algorithm enumerating all pairs of non-isomorphic DFAs with 2 and 3 states. In contrast to the previous case, over a minimal alphabet there are more than 60 non-isomorphic DFAs of L – even if we do not distinguish them by sets of final states – that meet the bound with some K. One of the witness pairs is described below.

Let $\Sigma = \{a, b, c, d, e, f\}$. Let $\mathcal{K} = (\{1, 2\}, \Sigma, \delta_K, 1, \{2\})$, and let $a = [1, 2]$, $b = c = [2, 1]$, $d = [1, 1]$, $e = [2, 2]$, and $f = [2, 1]$. Let $\mathcal{L} = (\{1, 2, 3\}, \Sigma, \delta_L, 1, \{1\})$, and let $a = [2, 2, 3]$, $b = [2, 1, 3]$, $c = [1, 1, 1]$, $d = e = [3, 1, 2]$, $f = [3, 1, 1]$. Then $\kappa(K \sqcup\!\sqcup L) = 44 = f(2, 3)$.

The bound $mn - 1$ on the size of the alphabet is not tight for $m = n = 2$, where an alphabet of size four is required. For any $m, n \geqslant 2$ the subsets of $\{1, 2\} \times \{1, 2\}$ satisfying (C) must be also reachable, and to reach them we can use only transformations mapping 1 to either 1 or 2. There are only three such transformations counted in Proposition 5; thus we need one more letter.

2 Partial Results about Tightness

To prove that the upper bound $f(m, n)$ of Equation (1) is tight, we must exhibit two languages K and L with state complexities m and n, respectively, such that $\kappa(K \sqcup\!\sqcup L) = f(m, n)$. As usual, we use DFAs to represent the languages: Let $\mathcal{K}$ and $\mathcal{L}$ be minimal *complete* DFAs for K and L. We first construct the NFA $\mathcal{N}$ as defined in Section 1, and we consider the subset automaton $\mathcal{D}$ of NFA $\mathcal{N}$. We must then show that $\mathcal{D}$ has $f(m, n)$ states reachable from the initial state

Table 1. Computational verification of reachability of the bound. The fields with ✓* follow from the proofs of Subsection 2.1.

$m\backslash n$	2	3	4	5	6	7	$\geqslant 8$
2	✓	✓	✓	✓	✓	✓	✓*
3		✓	✓	✓	✓	✓	✓*
4			✓	✓	✓	✓	✓*
5				✓	✓	✓	✓*
6					✓	?	?
7						?	?
$\geqslant 8$							?

$\{(1,1)\}$, and that these states are pairwise distinguishable. We were unable to prove this for all m and n, but we have some partial results about reachability in Subsection 2.1, and we deal with distinguishability in Subsection 2.2.

2.1 Reachability

We performed computations verifying reachability of the upper bound for small values of m and n. These results are summarized in Table 1.

The computation in the hardest case with $m = n = 6$ took about 48 days on a computer with AMD Opteron(tm) Processor 6380 (2500 MHz) and 64 GB of RAM. Moreover, we verified that in all these cases, every subset of size at least 3 is directly reachable from some smaller subset. We also verified that for reachability in case of $m = n = 3$ an alphabet of size 12 is sufficient, and in case of $m = n = 4$ an alphabet of size 50 is sufficient. Using these results, we are going to prove reachability for all m, n with $2 \leqslant m \leqslant 5$ and $n \geqslant 2$.

Without loss of generality, the set of states of any n-state DFA is denoted by $Q_n = \{1, 2, \ldots, n\}$. Let $\mathcal{T}_n$ be the monoid of all transformations of the set Q_n. Let $p, q \in Q_n$ and $P \subseteq Q_n$. Let $\mathbf{1}$ denote the identity transformation. Let $(p \to q)$ denote the transformation that maps state p to state q and acts as the identity on all the other states. Let (p, q) denote the transformation that transposes p and q.

Here we deal only with reachability, so final states do not matter. We assume that the sets of final states are empty in this subsection.

Let $\Sigma_{m,n} = \{a_{s,t} \mid s \in \mathcal{T}_m \text{ and } t \in \mathcal{T}_n\}$ be an alphabet consisting of $m^m n^n$ symbols. If an input a induces transformations s in $\mathcal{T}_m$ and t in $\mathcal{T}_n$, this will be indicated by $a\colon s; t$.

Define DFAs $\mathcal{K}_{m,n} = (Q_m, \Sigma_{m,n}, \delta_m, 1, \emptyset)$ and $\mathcal{L}_{m,n} = (Q_n, \Sigma_{m,n}, \delta_n, 1, \emptyset)$, where $\delta_m(p, a_{s,t}) = ps$ if $p \in Q_m$ and $\delta_n(q, a_{s,t}) = qt$ if $q \in Q_n$. Let $\mathcal{N}_{m,n}$ be the NFA for the shuffle of languages recognized by DFAs $\mathcal{K}_{m,n}$ and $\mathcal{L}_{m,n}$ as described in Section 1, and let $\mathcal{D}_{m,n}$ be the subset automaton of $\mathcal{N}_{m,n}$. The NFA $\mathcal{N}_{m,n}$ has alphabet $\Sigma_{m,n}$, and so has an input letter for every pair of transformations in $\mathcal{T}_m \times \mathcal{T}_n$. Therefore the addition of another input letter to the DFAs $\mathcal{K}_{m,n}$ and

$\mathcal{L}_{m,n}$ cannot add any new set of states of $\mathcal{N}_{m,n}$ that would be reachable from $\{(1,1)\}$ in $\mathcal{D}_{m,n}$.

Let $m' \leqslant m$ and $n' \leqslant n$. Then DFA $\mathcal{K}_{m',n'} = (Q_{m'}, \Sigma_{m',n'}, \delta_{m'}, 1, \emptyset)$ (respectively, the DFA $\mathcal{L}_{m',n'} = (Q_{n'}, \Sigma_{m',n'}, \delta_{n'}, 1, \emptyset)$) is a sub-DFA of $\mathcal{K}_{m,n}$ (respectively, of $\mathcal{L}_{m,n}$), in the sense that $Q_{m'} \subseteq Q_m$, $\Sigma_{m',n'} \subseteq \Sigma_{m,n}$, and $\delta_{m'} \subseteq \delta_m$. As well, NFA $\mathcal{N}_{m',n'}$ is a sub-NFA of $\mathcal{N}_{m,n}$. Note that $\mathcal{D}_{m,n}$ is extremal for the shuffle: every language $K \sqcup\!\sqcup L$, where K and L are languages with state complexities m and n respectively, is recognized by some sub-DFA of $\mathcal{D}(m,n)$ after possibly renaming some letters.

For the next lemma it is convenient to consider a subset S of states (p,q) of $\mathcal{N}_{m,n}$ as an $m \times n$ matrix, where the entry in row p and column q is (p,q) if $(p,q) \in S$, and it is empty otherwise. We first introduce the following notions.

Definition 7. *Let $i, i' \in Q_m$, $i \neq i'$, and $j, j' \in Q_n$, $j \neq j'$.*
(a) A row i' contains *row i, if $(i,j) \in S$ implies $(i',j) \in S$ for all $j \in Q_n$.*
(b) A column j' contains *column j if $(i,j) \in S$ implies $(i,j') \in S$ for all $i \in Q_m$.*
(c) A subset of $Q_m \times Q_n$ is valid *if it satisfies Condition (C) from Lemma 2, that is, if it contains a state in row 1 and a state in column 1.*

Lemma 8. *Let S be a valid subset of $Q_m \times Q_n$ with the property that there are distinct i, i' or j, j' such that either row i' contains row i or column j' contains column j. Assume that every valid subset S' of $Q_{m'} \times Q_{n'}$, where $m' < m$, or $n' < n$, or $|S'| < |S|$, is reachable in DFA $\mathcal{D}_{m',n'}$. Then S is reachable in $\mathcal{D}_{m,n}$.*

Proof. If S contains an empty row or column, then without loss of generality we can renumber the n states of $\mathcal{L}_{m,n}$ in such a way that column n is the empty column in S. By the inductive assumption we know that S is reachable in $\mathcal{D}_{m,n-1}$ by some word w. Since $\mathcal{N}_{m,n-1}$ is a sub-NFA of $\mathcal{N}_{m,n}$, S is reachable in $\mathcal{D}_{m,n}$ as well by the same word. Suppose that S has neither an empty row nor an empty column. By symmetry, it is sufficient to consider the case with distinct i and i' such that row i' contains row i. Let $S' = S \backslash \{(i',j) \mid (i,j) \in S \text{ for } j \in \{1,\ldots,n\}\}$. Since $|S'| < |S|$, the set S' is reachable by assumption. To obtain S, we apply the letter that induces the transformation $i \to i'; \mathbf{1}$. □

Lemma 9. *Let S be a valid subset of $Q_m \times Q_n$ such that there is a column or a row with exactly one element. Assume that every valid subset S' of $Q_{m'} \times Q_{n'}$, where $m' < m$, or $n' < n$, or $|S'| < |S|$, is reachable in $\mathcal{D}_{m',n'}$. Then S is reachable in $\mathcal{D}_{m,n}$.*

Proof. Recall that we can assume $m \geqslant 2$ and $n \geqslant 2$. We may assume that there is neither an empty row nor an empty column in S; otherwise S is reachable by Lemma 8. It is sufficient to consider the case involving a column, since the case involving a row follows by symmetric arguments. Let (p,q) be the only element in column q. If there are more elements in row p, then column q is contained in another column and by Lemma 8, the set S is reachable.

Let S' be the subset of $Q_{m-1} \times Q_{n-1}$ obtained by removing row p and column q, and renumbering the states to $Q_{m-1} \times Q_{n-1}$ in the way such that

$i \in Q_m$ becomes $i-1$ if $i > p$ and otherwise remains the same, and $j \in Q_n$ becomes $j-1$ if $j > q$ and otherwise remains the same. We have that S' is a valid subset, and by the inductive assumption it is reachable in $\mathcal{D}_{m-1,n-1}$ by some word u'; let u be the word corresponding to u' in the original numbering of the states. We consider four cases.

Case $p \neq 1$ and $q \neq 1$: State $\{(1,1),(p,q)\}$ is reachable in $\mathcal{D}_{m,n}$ by word a^2, where $a\colon (1,p);(1,q)$. Then S is reachable by a^2u.

Case $p = 1$ and $q \neq 1$: State $\{(2,1),(1,q)\}$ is reachable in $\mathcal{D}_{m,n}$ by word a^2, where $a\colon (1,2);(1,q)$. Then state $(2,1)$ corresponds to state $(1,1)$ after the renumbering, and S is reachable by a^2u.

Case $p \neq 1$ and $q = 1$: This is symmetrical to the previous case.

Case $p = 1$ and $q = 1$: State $\{(1,1),(2,2)\}$ is reachable in $\mathcal{D}_{m,n}$ by word a^2, where $a\colon (1,2);(1,2)$. Then state $(2,2)$ corresponds to state $(1,1)$ after the renumbering, and S is reachable by a^2u. □

Theorem 10. *If for some h every valid subset can be reached in $\mathcal{D}_{h,\binom{h}{\lfloor h/2\rfloor}}$ then for every $m \leqslant h$ and every n, every valid subset can be reached in $\mathcal{D}_{m,n}$.*

Proof. This follows by induction on m, n, and $|S|$.

For $m = 1$ this follows by induction on n: if $n = 1$ then $\mathcal{D}_{1,1}$ consists of a single valid subset $\{(1,1)\}$, and if $n > 1$, then we apply Lemma 8. For $m \leqslant h$ and $n \leqslant \binom{h}{\lfloor h/2\rfloor}$ this holds by assumption, since $\mathcal{N}_{m,n}$ is a sub-NFA of $\mathcal{N}_{h,\binom{h}{\lfloor h/2\rfloor}}$. If $|S| = 1$, then $\{(1,1)\}$ is the only valid subset, and it is reachable since it is the initial subset of $\mathcal{D}_{m,n}$.

Let S be a valid subset of $Q_m \times Q_n$, where $m \leqslant h$ and $n > \binom{h}{\lfloor h/2\rfloor}$, and assume that every valid subset S' of $Q_{m'} \times Q_{n'}$ is reachable if $m' < m$, or $n' < n$, or $|S'| < |S|$. By Sperner's theorem [5], the maximal number of subsets of an m-element set such that none of them contains any other subset is $\binom{m}{\lfloor m/2\rfloor}$. This is not larger than $\binom{h}{\lfloor h/2\rfloor}$; hence, there exist some columns j, j' with $j \neq j'$ such that the j-th column is contained in j'-th column. By Lemma 8, the subset S is reachable. □

Corollary 11. *Let $1 \leqslant m \leqslant 4$ and $n \geqslant 1$. Then every valid subset can be reached in $\mathcal{D}_{m,n}$.*

Proof. Since we have verified the reachability of all valid subsets for $m = 4$ and $n = 6 = \binom{4}{2}$, Theorem 10 applies with $h = 4$. □

To strengthen this result and show reachability for $m \leqslant 5$, we need to introduce another concept with permutations. Let φ be any permutation of m rows. We split subsets of Q_m (subsets of rows) into equivalence classes under φ. For $U \subseteq Q_m$, $[U]_\varphi = \{V \subseteq Q_m \mid V = \varphi^i(U) \text{ for some } i \geqslant 0\}$ denotes the equivalence class of U. See Tables 2, 3, 4 for examples of subsets whose columns U are partitioned into equivalence classes under some φ.

For a subset S of $Q_m \times Q_n$, by $\mathrm{col}(S,i)$ we denote the subset of Q_m contained in the i-th column. Then $\mathrm{cols}(S) = \bigcup_{1\leqslant i\leqslant n} \mathrm{col}(S,i)$ is the set of the subsets in the columns of S.

The following lemma assures reachability (under an inductive assumption) of a special kind of subsets whose columns form only full and empty equivalence classes under some permutation φ.

Lemma 12. *Let φ be a permutation of m rows. Let S be a valid subset of $Q_m \times Q_n$ such that $[U]_\varphi \subseteq \mathrm{cols}(S)$ for every $U \in \mathrm{cols}(S)$, and there is a column $V \in \mathrm{cols}(S)$ such that $|[V]_\varphi| \geqslant 2$. Assume that every valid subset S' of $Q_{m'} \times Q_{n'}$, where $m' < m$, or $n' < n$, or $|S'| < |S|$, is reachable in $\mathcal{D}_{m',n'}$. Then S is reachable in $\mathcal{D}_{m,n}$.*

Proof. We can assume that no two columns contain the same subset of rows, no column is empty, and the first row contains at least two elements; otherwise S is reachable by Lemma 8 or by Lemma 9.

Let $S_j = \mathrm{col}(S, j)$ be the j-th column of a valid subset S. Thus we have $S = \{(i, j) \mid 1 \leqslant j \leqslant n \text{ and } i \in S_j\}$. Since $|[V]_\varphi| \geqslant 2$, we can always choose V so that $\varphi^{-1}(V)$ is in a k-th column S_k with $k \neq 1$. Let S' be the set obtained from S by omitting the states in the k-th column and by taking the pre-image of S_j under φ in any other column, that is,

$$S' = \{(i, j) \mid 1 \leqslant j \leqslant n, j \neq k, \text{ and } i \in \varphi^{-1}(S_j)\}.$$

Since $k \neq 1$ and the first row of S contains at least two elements, the set S' is valid. Since V is non-empty, we have $|S'| < |S|$. Let ψ be a permutation that maps a column j to the column containing $\varphi^{-1}(S_j)$, that is, we have $S_{\psi(j)} = \varphi^{-1}(S_j)$. Let t be the transformation given by $a_{\varphi,\psi}$. Let us show that $S't = S$.

Let $(i, j) \in S'$. Then $i \in \varphi^{-1}(S_j)$, so $\varphi(i) \in S_j$, and we have $(i, j)t = \{(\varphi(i), j), (i, \psi(j))\} \subseteq S$. Hence $S't \subseteq S$.

Now let $(i, j) \in S$. First let $j \neq k$. Then $i \in S_j$, so $\varphi^{-1}(i) \in \varphi^{-1}(S_j)$. Therefore $(\varphi^{-1}(i), j) \in S'$. Since $(i, j) \in (\varphi^{-1}(i), j)t$, we have $(i, j) \in S't$. Now let $j = k$. Then $i \in \varphi^{-1}(V)$ and $S_{\psi^{-1}(k)} = V$. Thus $(i, \psi^{-1}(k)) \in S'$, and we have $(i, k) \in (i, \psi^{-1}(k))t$. Hence $S \subseteq S't$. Our proof is complete. □

Table 2. A subset and the equivalence classes of columns under $\varphi = [2, 3, 1, 4, 5]$.

	1	2	3	4	5	6	7	8	9
1	∘		∘	∘			∘		
2	∘	∘			∘			∘	
3		∘	∘			∘			∘
4				∘	∘	∘			
5							∘	∘	∘
eq	*A*	*A*	*A*	*B*	*B*	*B*	*C*	*C*	*C*

Table 3. A subset and the equivalence classes of columns under $\varphi = [1, 2, 3, 5, 4]$.

	1	2	3	4	5	6	7	8
1	∘	∘			∘			
2	∘		∘			∘		
3	∘			∘			∘	
4		∘	∘	∘				∘
5					∘	∘	∘	∘
eq	A	B	C	D	B	C	D	E

Table 4. A subset and the equivalence classes of columns under $\varphi = [2, 3, 4, 1, 5]$.

	1	2	3	4	5	6	7	8
1		∘	∘	∘	∘			
2	∘		∘	∘		∘		
3	∘	∘		∘			∘	
4	∘	∘	∘					∘
5					∘	∘	∘	∘
eq	A	A	A	A	B	B	B	B

Corollary 13. *Let $1 \leqslant m \leqslant 5$ and $n \geqslant 1$. Then every valid subset can be reached in $\mathcal{D}_{m,n}$.*

Proof. The proof follows by analysis of valid subsets $S \subseteq Q_5 \times Q_n$, with the aid of Corollary 11, Lemma 8, Lemma 12, and the results from Table 1.

Suppose that there is a valid subset $S \subseteq Q_5 \times Q_n$ that is not reachable; let S be chosen so that n is the smallest number and S is a smallest non-reachable subset of $Q_5 \times Q_n$.

By Corollary 11 and the choice of n, every valid subset $S' \subset Q_{m'} \times Q_{n'}$, where $m' < 5$, or $n' < n$, or $|S'| < |S|$, is reachable. Hence, S has no column containing another column; otherwise, we can apply Lemma 8. Since we have verified the reachability of all valid subsets for $m = 5$ and $n \leqslant 7$ (Table 1), we must have $n \geqslant 8$ and so S has at least 8 distinct columns. Obviously there is neither an empty nor a full column. If there is a column U with $|U| = 1$ or $|U| = 4$, then by Sperner's theorem if $n > \binom{4}{2} = 6$, then S has a column containing another column; hence S can have only columns U with $|U| = 3$ or $|U| = 2$.

Let C_3 be the number of 3-element columns ($|U| = 3$), and C_2 be the number of 2-element columns ($|U| = 2$). We are searching for possible subsets S that do not have a column containing another column, and with $C_3 + C_2 \geqslant 8$. We consider the following six cases.

(1) Let $C_3 = 0$. If $C_2 = 10$, which implies that S contains all possible 2-element subsets, then under $\varphi = [2, 3, 4, 5, 1]$ we have two full and non-trivial equivalence classes. Hence S is reachable from a smaller subset by Lemma 12. If $C_2 = 9$, then without loss of generality let the missing 2-element subset be $\{4, 5\}$; see Table 2. Under $\varphi = [2, 3, 1, 4, 5]$ we have three full and non-trivial equivalence classes, and S is reachable by Lemma 12. Finally, if $C_2 = 8$, then we have two subcases. If the two missing 2-element subsets have a common element, then without loss of generality let them be $\{2, 3\}$ and $\{4, 5\}$. Under $\varphi = [1, 4, 5, 2, 3]$ we have four full and non-trivial equivalence classes, and S is reachable by Lemma 12. If they have a common element, then without loss of generality let them be $\{3, 4\}$ and $\{4, 5\}$. Under $\varphi = [1, 2, 5, 4, 3]$ we have six full equivalence classes and two of them are non-trivial. Thus S is reachable by Lemma 12.

(2) Let $C_3 = 1$. The only possible subset, up to permutation of columns and rows, is shown in Table 3. It has all columns with two elements that are not contained in the 3-element column. By Lemma 12 with $\varphi = [1, 2, 3, 5, 4]$, it is reachable.

(3) Let $C_3 = 2$. A simple analysis reveals that if the 3-element columns have only one common element, then C_2 is at most 4. If they have two common elements, then C_2 is at most 5. Thus in this case, we have $C_2 + C_3 \leqslant 7$.

(4) Let $C_3 = 3$. Here C_2 is at most 4.

(5) Let $C_3 = 4$. The only possible subset, up to permutation of columns and rows, is shown in Table 4. By Lemma 12 with $\varphi = [2, 3, 4, 1, 5]$, it is reachable.

(6) Let $C_3 \geqslant 5$. These cases are symmetrical to those with $C_3 \leqslant 3$; it is sufficient to consider the complement of S.

Since these cover all the possibilities for set S, this set is reachable. □

2.2 Proof of Distinguishability

The aim of this section is to show that there are regular languages defined over a three-letter alphabet such that the subset automaton of the NFA for their shuffle does not have equivalent states.

To this aim let $\mathcal{A} = (Q, \Sigma, \delta, s, F)$ be an NFA. We say that a state q in Q is *uniquely distinguishable* if there is a word w in Σ^* which is accepted by $\mathcal{A}$ from and only from the state q, that is, if there is a word w such that $\delta(p, w) \in F$ if and only if $p = q$. First, let us prove the following two observations.

Proposition 14. *If each state of an NFA $\mathcal{A}$ is uniquely distinguishable, then the subset automaton of $\mathcal{A}$ does not have equivalent states.*

Proof. Let S and T be two distinct subsets in 2^Q. Then, without loss of generality, there is a state q in Q with $q \in S \setminus T$. Since q is uniquely distinguishable, there is a word w which is accepted by $\mathcal{A}$ from and only from q. Therefore, the subset automaton of $\mathcal{A}$ accepts w from S and it rejects w from T. Hence w distinguishes S and T. □

Proposition 15. *Let a state q of an NFA $\mathcal{A} = (Q, \Sigma, \delta, s, F)$ be uniquely distinguishable. Assume that there is a symbol a in Σ and* exactly one *state p in Q that goes to q on a, that is, (p, a, q) is a unique in-transition on a going to q. Then the state p is uniquely distinguishable as well.*

Proof. Let w be a word which is accepted by $\mathcal{A}$ from and only from q. The word aw is accepted from p since $q \in \delta(p, a)$ and w is accepted from q. Let $r \neq p$. Then $q \notin \delta(r, a)$ since (p, a, q) is a unique in-transition on a going to q. It follows that the word w is not accepted from any state in $\delta(r, a)$. Thus $\mathcal{A}$ rejects aw from r, so p is uniquely distinguishable. □

Now we can prove the following result.

Theorem 16. *Let $m, n \geqslant 2$. There exist ternary languages K and L with $\kappa(K) = m$ and $\kappa(L) = n$ such that the subset automaton of the NFA accepting $K \sqcup\!\sqcup L$ does not have equivalent states.*

Proof. Let m and n be arbitrary but fixed integers with $m, n \geqslant 2$. Let K be accepted by the DFA $\mathcal{K} = (\{1, 2, \ldots, m\}, \{a, b, c\}, \delta_K, 1, \{m\})$, where for each i in $\{1, 2, \ldots, m\}$,

$\delta_K(i, a) = i + 1$ if $i \leqslant m - 1$ and $\delta_K(m, a) = 1$;
$\delta_K(i, b) = 1$;
$\delta_K(1, c) = 2$ and $\delta_K(i, c) = 1$ if $i \geqslant 2$.

Let L be accepted by the DFA $\mathcal{L} = (\{1, 2, \ldots, n\}, \{a, b, c\}, \delta_L, 1, \{n\})$, where for each j in $\{1, 2, \ldots, n\}$,

$\delta_L(j, a) = 1$;
$\delta_L(j, b) = j + 1$ if $j \leqslant n - 1$ and $\delta_L(n, b) = 1$;
$\delta_L(j, c) = n$.

The DFAs $\mathcal{K}$ and $\mathcal{L}$ are shown in Fig. 2.

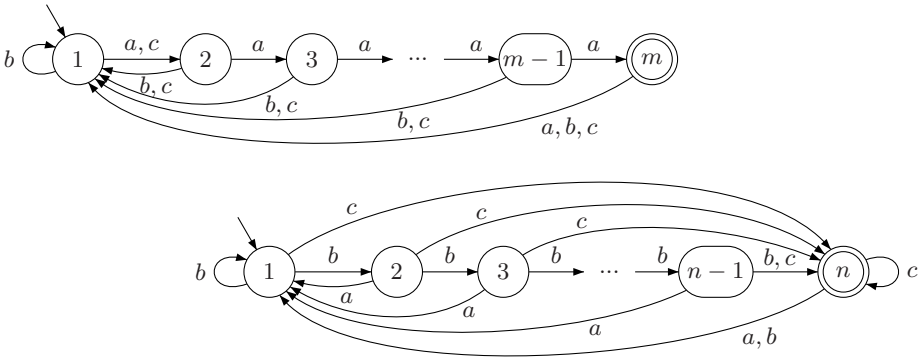

Fig. 2. The DFAs $\mathcal{K}$ and $\mathcal{L}$.

Construct the NFA $\mathcal{N}$ for $K \sqcup\!\sqcup L$ as described in Section 1 on page 2. The transitions on a, b, c in $\mathcal{N}$ for $m = 4$ and $n = 5$ are shown in Fig. 3. Notice that each state (i, j) with $2 \leqslant i \leqslant m$ and $2 \leqslant j \leqslant n$ has a unique in-transition on symbol a and this transition goes from state $(i-1, j)$; see the dashed transitions in Fig. 3 (top-left). Next, each state (m, j) with $2 \leqslant j \leqslant n$ has a unique in-transition on b which goes from $(m, j-1)$, and each state $(i, 2)$ with $2 \leqslant i \leqslant m$ has a unique in-transition on b going from $(i, 1)$; see the dashed transitions in Fig. 3 (top-right). Finally, the state $(2, 1)$ has a unique in-transition on c going from $(1, 1)$; see the dashed transition in Fig. 3 (bottom).

The empty word is accepted by $\mathcal{N}$ from and only from the state (m, n) since this is a unique accepting state of $\mathcal{N}$. Thus (m, n) is uniquely distinguishable. Next, consider the subgraph of unique in-transitions in $\mathcal{N}$. Fig. 4 shows this subgraph in the case of $m = 4$ and $n = 5$. Notice that from each state of $\mathcal{N}$, the state (m, n) is reachable in this subgraph. By Proposition 15, used repeatedly, we get that each state of $\mathcal{N}$ is uniquely distinguishable. Hence by Proposition 14, the subset automaton of $\mathcal{N}$ does not have equivalent states. □

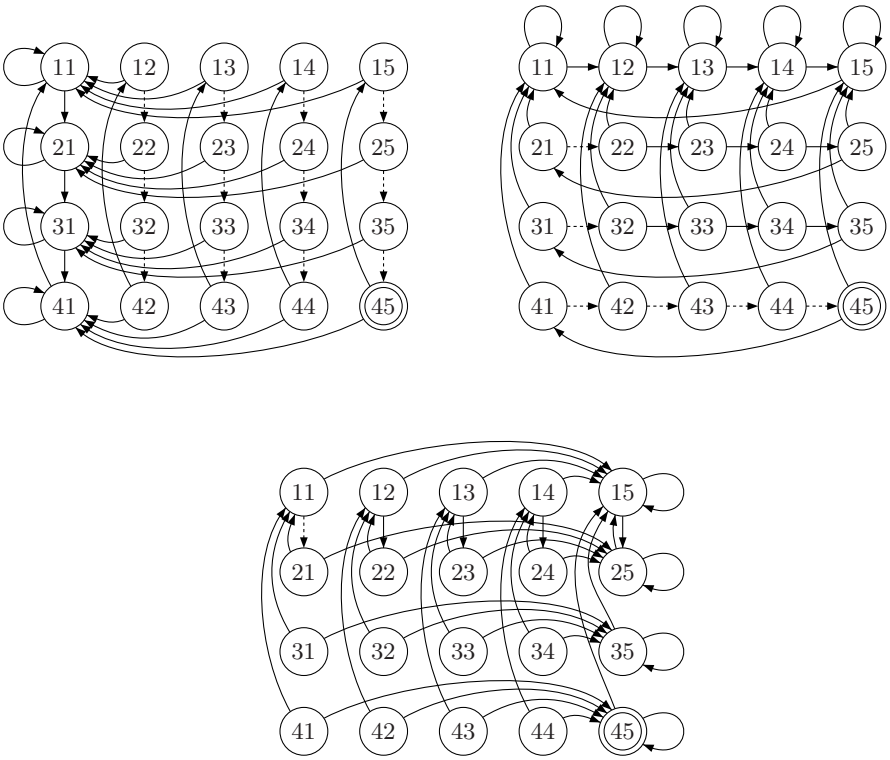

Fig. 3. NFA $\mathcal{N}$ for $m = 4$ and $n = 5$; the transitions on a (top-left), b (top-right), and c (bottom).

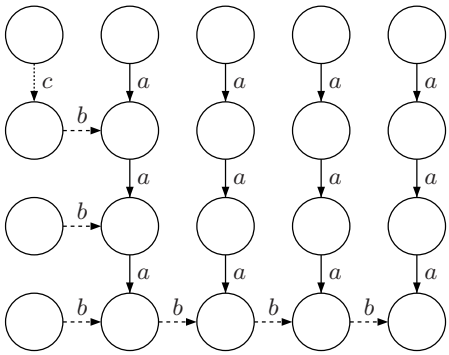

Fig. 4. The subgraph of unique in-transitions in NFA $\mathcal{N}$; $m = 4$ and $n = 5$.

3 Conclusions

We have examined the state complexity of the shuffle operation on two regular languages of state complexities m and n, respectively, and found an upper bound for it. We know that this bound can be reached for any m with $1 \leqslant m \leqslant 5$ and any $n \geqslant 1$, and also for $m = n = 6$. For the remaining values of m and n, however, the problem remains open. Since there exist two languages K and L for which all pairs of states in the subset automaton of the NFA accepting the shuffle $K \sqcup\!\sqcup L$ are distinguishable, the main difficulty consists of proving that all valid states in the subset automaton can be reached for the witness languages.

Acknowledgments

We would like to thank an anonymous referee for proposing the notions of a uniquely distinguishable state and of a subgraph of unique in-transitions which allow us to simplify the proof of distinguishability. We are also grateful for his comments and suggestions that helped us improve the presentation of the paper.

References

1. Brzozowski, J., Jirásková, G., Li, B.: Quotient complexity of ideal languages. Theoret. Comput. Sci. 470, 36–52 (2013)
2. Câmpeanu, C., Salomaa, K., Yu, S.: Tight lower bound for the state complexity of shuffle of regular languages. J. Autom. Lang. Comb. 7(3), 303–310 (2002)
3. The GAP Group: GAP – Groups, Algorithms, and Programming, Version 4.8.3 (2016), `http://www.gap-system.org`
4. Okhotin, A.: On the state complexity of scattered substrings and superstrings. Fund. Inform. 99(3), 325–338 (2010)

5. Sperner, E.: Ein Satz über Untermengen einer endlichen Menge. Math. Z. 27, 544–548 (1928)
6. Yu, S.: State complexity of regular languages. J. Autom. Lang. Comb. 6, 221–234 (2001)