Journal of Automata, Languages and Combinatorics **u** (v) w, x–y

ON THE SHUFFLE AUTOMATON SIZE FOR WORDS[1]

Franziska Biegler, Mark Daley
Department of Computer Science, University of Western Ontario,
London, ON N6A 5B7, Canada
e-mail: {fbiegler, daley}@csd.uwo.ca

and

Ian McQuillan
Department of Computer Science, University of Saskatchewan,
Saskatoon, SK S7N 5A9, Canada
e-mail: mcquillan@cs.usask.ca

ABSTRACT

We investigate the state size of DFAs accepting the shuffle of two words. We provide an infinite family of words u and v, such that the minimal DFA for $u \sqcup v$ requires an exponential number of states as a function of their lengths. We also show some conditions for the words u and v which ensure a quadratic upper bound on the state size of $u \sqcup v$. Moreover, switching only two letters within one of u or v is enough to trigger the change from quadratic to exponential.

Keywords: shuffle, words, finite languages, finite automata, state complexity

1. Introduction

Since its introduction, the shuffle operation has been aggressively studied as a model of nondeterministic interleaving in both purely theoretical and practical contexts. Perhaps due to the intrinsic nondeterminism of the operation, many problems concerning shuffle remain unsolved; e.g., shuffle decomposition for regular languages (though it is decidable [3] for commutative regular languages or locally testable languages while for context-free languages it is undecidable [3]).

We follow here the recent trend of attacking the special case of the shuffle of two words, inspired by attempts to solve the decomposition problem. It has been shown in [1] that shuffle decomposition on individual words is unique as long as there are two letters used within the words. In [2], the result from [1] was extended to show that if

[1]Research supported, in part, by the Natural Sciences and Engineering Research Council of Canada. Published in Journal of Automata, Languages and Combinatorics, 15, 53–70.
http://dx.doi.org/10.25596/jalc-2010-053

two words u and v both contain at least two letters, then the shuffle decomposition is the unique decomposition over arbitrary sets and not just words.

In this paper we ask a different type of question: what is the minimal state size for a DFA accepting the shuffle of two given words? For the more general case of languages, it has been shown in [4] that the shuffle of two DFAs can yield an exponential minimal DFA ($\Omega(2^{nm})$, where n, m were the sizes of the two DFAs). We show here that DFAs accepting the shuffle of two words also require an exponential number of states in general; however, for words obeying certain conditions, a DFA may be constructed with, at most, quadratically many states.

A striking reminder of the complexity of the shuffle is operation is illustrated by showing that two words which may be accepted by a quadratically-bounded shuffle DFA can only be accepted by an exponentially large DFA when only two letters in one word are exchanged.

2. Preliminaries

Let $\mathbb{N}$ be the set of non-negative integers. An alphabet Σ is a finite, non-empty set of letters. The set of all words over Σ is denoted by Σ^*, and this set contains the empty word, λ. The set of all non-empty words over Σ is denoted by Σ^+.

Let Σ be an alphabet and let $u, v \in \Sigma^*$. If $u = a_1^{\alpha_1} a_2^{\alpha_2} \cdots a_n^{\alpha_n}$ with $a_1, \ldots a_n \in \Sigma$, $\alpha_1, \ldots, \alpha_n \in \mathbb{N}$ and $a_i \neq a_{i+1}$, for $1 \leq i < n$, then the *skeleton* of u is defined as $\chi(u) = a_1 a_2 \cdots a_n$. The different occurrences of the same letter a in the skeleton of u are called the *a-sections* of u. Furthermore, for $a \in \Sigma$, $|u|_a$ denotes the number of a's in u. A word u over Σ is called non-repeating if $|u|_a \leq 1$ for all $a \in \Sigma$. Let $u, v \in \Sigma^*$. The *shuffle* of u and v is defined as $u \sqcup v = \{u_1 v_1 \cdots u_n v_n \mid u = u_1 \cdots u_n, v = v_1 \cdots v_n, u_i \in \Sigma^*, v_i \in \Sigma^*, 1 \leq i \leq n\}$. We say u is a *suffix* of v, written $u \leq_{\mathrm{s}} v$, if $v = xu$, for some $x \in \Sigma^*$.

A *trajectory* for two words u and v is a word $t \in \{0,1\}^*$, such that $|t|_0 = |u|$ and $|t|_1 = |v|$. Then the shuffle of u and v on t is denoted by $u \sqcup_t v$ and is the unique string in $u \sqcup v$, where a letter from u is used whenever t has a 0 at the respective position, and a letter from v is used whenever t has a 1. For details regarding shuffle on trajectories, consult [6].

We assume the reader to be familiar with nondeterministic and deterministic finite automata. See [5, 8] for an introduction and more details on finite automata. For each NFA we can effectively construct an equivalent DFA by using the so-called subset construction [5]. For an NFA with n states, the DFA constructed this way can have up to 2^n states. There exists a unique minimal DFA (up to isomorphism) for each regular language. States p and q of a DFA are *distinguishable* if there exists x such that $\delta(p, x)$ is a final state, but $\delta(q, x)$ is not, or vice versa. Moreover, if every state of a DFA is accessible and every pair of states are distinguishable, then the DFA is minimal [5]. For both NFAs and DFAs we use *size* synonymously with state size, and, thus, we define $|A| = |Q|$.

3. Shuffle NFAs for Words

In this section we discuss basic properties of shuffle NFAs for two words.

Definition 1 *Let Σ be an alphabet and let $u = u_1 \cdots u_m, v = v_1 \cdots v_n \in \Sigma^+$, where $u_i, v_j \in \Sigma$ for all $1 \leq i \leq m$ and $1 \leq j \leq n$. We say A is the* naive shuffle NFA *for u and v if $A = (Q, \Sigma, \delta, q_0, F)$ where $Q = \{0, \ldots, m\} \times \{0, \ldots, n\}$, $q_0 = (m, n)$, $F = \{(0, 0)\}$ and*

- *for $1 \leq k \leq m$, $0 \leq l \leq n$, we have $(k-1, l) \in \delta((k, l), u_{(m-k+1)})$; and*
- *for $0 \leq k \leq m$, $1 \leq l \leq n$, we have $(k, l-1) \in \delta((k, l), v_{(n-l+1)})$.*

For all i and j with $1 \leq i \leq m$ and $1 \leq j \leq n$ we denote by $\overline{u}_i$ and $\overline{v}_j$ the suffixes of length i and j or the words u and v, respectively. We furthermore define $L_A(i, j) = \overline{u}_i \sqcup\!\sqcup \overline{v}_j$, which is accepted by the automaton $A' = (Q, \Sigma, \delta, (i, j), F)$.

Note that the automaton as defined above is not complete. It is clear from Definition 1 that the naive shuffle NFA for u and v does, in fact, accept $u \sqcup\!\sqcup v$.

Definition 2 *Let A be the naive shuffle NFA for two words u and v over some alphabet Σ. The vertical layers and horizontal layers (shortly, v-layers and h-layers) are numbered $0, 1, \ldots, |u| + |v|$ and $|u|, |u| - 1, \ldots 1, 0, -1, \ldots, -|v|$, respectively. The* vertical layer *(*horizontal *respectively) k, contains all states (i, j) with $i + j = k$ (contains all states (i, j) with $k = i - j$).*

The vertical layer tells us how many letters we have read thus far, while the horizontal layer tells us the difference between the numbers of letters we have read from u and v. Note that the initial state $(|u|, |v|)$ is in horizontal layer $|u| - |v|$ if $|u| \geq |v|$, and in horizontal layer $|v| - |u|$ if $|v| \geq |u|$.

We now define what we mean by nondeterministic areas of a naive shuffle NFA.

Definition 3 *Let Σ be an alphabet and let A be the naive shuffle NFA for some words $u, v \in \Sigma^+$. Let $a \in \Sigma$ and $i_1, i_2, j_1, j_2 \in \mathbb{N}$. Then $R = (a, (i_1, j_1), (i_2, j_2))$ is a nondeterministic area of A if $|u| \geq i_1 \geq i_2 \geq 0$, $|v| \geq j_1 \geq j_2 \geq 0$ and*

1. *all states (i, j) with $i_1 \geq i > i_2$, $j_1 \geq j > j_2$ are nondeterministic on a,*
2. *if they exist, $(i_1 + 1, j_1)$ and $(i_1, j_1 + 1)$ are deterministic on a, and*
3. *$\delta((i_2, j_2), a)$ is undefined.*

The set of all nondeterministic areas of A is denoted by $\mathrm{Area}(A)$*, and we define the entrance and exit states of R and the states in $R = (a, (i_1, j_1), (i_2, j_2))$ as*

$$\begin{aligned} \mathrm{ent}(R) &= \{(i_1, j) \mid j_1 \geq j \geq j_2\} \cup \{(i, j_1) \mid i_1 \geq i \geq i_2\}; \\ \mathrm{ex}(R) &= \{(i_2, j) \mid j_1 \geq j \geq j_2\} \cup \{(i, j_2) \mid i_1 \geq i \geq i_2\}; \\ \mathrm{states}(R) &= \{(i, j) \mid i_1 \geq i > i_2, j_1 \geq j > j_2\}. \end{aligned}$$

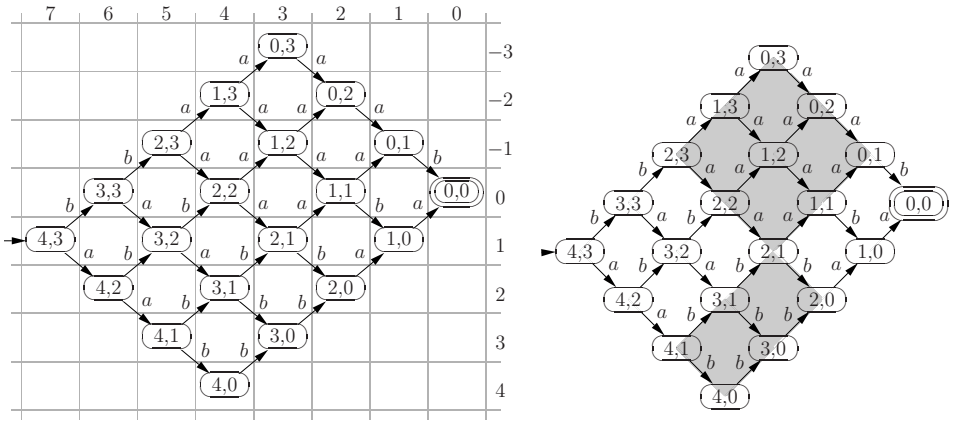

Figure 1: Naive shuffle NFA for $u = bbaa$ and $v = aab$

Example 1 Let $u = bbaa$, $v = aab$. Then the naive shuffle NFA A for u and v has $\text{Area}(A) = \{(a, (2, 3), (0, 1)), (b, (4, 1), (2, 0))\}$. A is depicted twice in Figure 1, first with the different horizontal and vertical layers labelled and then with the nondeterministic areas shown in grey.

We know from [7] that given an NFA accepting a finite language over a k letter alphabet with q states, a minimal DFA accepting the same language has at most $\mathcal{O}(k^{\frac{q}{\log_2(k)+1}})$ states in the worst case. Thus for a binary alphabet, $\mathcal{O}(2^{\sqrt{q}})$ states are both necessary and sufficient in the worst case.

In the case of naive shuffle NFAs, it is immediately obvious that during a subset-construction only state labels from the same vertical layer can appear within the same state of the DFA. If $|u| = m$ and $|v| = n$ with $0 \leq n \leq m$, then for each number i between 1 and n there are two vertical layers with i states, and there are $(m - n + 1)$ vertical layers with $(n + 1)$ states. If we assume that for each v-layer, all subsets of states except the empty set are possible (it is sufficient to add the empty set once) then this gives us an upper bound of

$$2\sum_{i=1}^{n}(2^i - 1) + (m - n + 1)(2^{n+1} - 1) + 1 = 2^{n+1}(m - n + 3) - m - n - 4 \quad (1)$$

for the number of states in the equivalent DFA. Recall that the NFA has $(m+1)(n+1)$ states, so the bound in (1) is better than the bound $\mathcal{O}(k^{\frac{(m+1)(n+1)}{\log_2(k)+1}})$ (where k is the size of the alphabet) from [7] for arbitrary finite languages, even when $k = 2$.

When u and v are over disjoint alphabets then the naive shuffle NFA for u and v is also the minimal DFA for $u \sqcup v$. This can be seen as every pair of states that are not distinguishable would have to be in the same vertical layer, however, every two states in the same layer have some different path to the final state. Thus, all pairs of states are distinguishable. So, in the worst case there is a lower bound of $(|u| + 1) \cdot (|v| + 1)$ on the size of the shuffle DFA for u and v.

We can also see that the bound (1) is not tight, as only labels of states of the NFA which have identical Parikh vectors can appear together as the label of a state in the DFA. Thus the bound (1) would be reached only if $u, v \in \{a\}^*$ for some $a \in \Sigma$. But then the minimal DFA for $u \sqcup v$ would only have $|u| + |v| + 1$ states, a contradiction.

Definition 4 *Let u and v be words over some finite alphabet Σ and let A be the naive shuffle NFA for u and v. A walk through A is a sequence of states $s_0, s_1, \ldots, s_{|u|+|v|}$, where $s_0 = (|u|, |v|)$, $s_{|u|+|v|} = (0, 0)$, and for all i with $0 \leq i < |u| + |v|$, we have $s_{i+1} \in \delta(s_i, a)$ for some $a \in \Sigma$. We say that a given vertical or horizontal layer is visited x-times during a given walk if exactly x states from that layer appear in the walk.*

Note that there exists a bijective mapping between the walks through a naive shuffle NFA and the set of possible trajectories for the shuffle of u and v.

Lemma 1 *Let u, v be words over some alphabet Σ and let A be the naive shuffle NFA for u and v. Then during each walk through A, every vertical layer has to be visited exactly once, while each horizontal layer may be visited once, multiple times or not at all. However, if $|u| \geq |v|$ then each of the horizontal layers $0, 1, \ldots |u| - |v|$ has to be visited at least once, and similarly if $|v| \geq |u|$.*

4. Shuffle DFAs for Periodic Words

In this section we focus on a special case of the shuffle of two words, namely the shuffle of two words that are periods of a common underlying word. Thus $u = w_1 w^k$ and $v = w_2 w^l$, where $w \in \Sigma^+$, $w \notin a^+$ for any $a \in \Sigma$, $k, l \geq 0$ and both w_1 and w_2 are suffixes of w. At first glance one could conjecture that these words lead to an exponential blow-up in the state size when converting the naive shuffle NFA to a DFA, because they induce long common factors. However we will show that this is not the case when the underlying word w contains at most one section per letter in Σ. We first show two subset-relations between different periodic shuffles over the same underlying word. These subset-relations are then used to construct the DFA in a more efficient manner.

Lemma 2 *Let Σ be a finite alphabet and let $w = a_1 \cdots a_n$ for some $n \geq 2$, such that $\mathrm{alph}(w) \geq 2$. Let $u = w_1 w^k$, $v = w_2 w^l$, $u' = w_1 w^{k'}$, $v' = w_2 w^{l'}$ where $0 \leq l < k' < k$, $0 \leq l < l' < k$, $k + l = k' + l'$ and each of w_1, w_2 is either empty or a proper suffix of w. Then $u \sqcup v \subsetneq u' \sqcup v'$.*

Proof. Let A be the naive shuffle NFA for u and v. Let t be a trajectory for u and v. We construct a trajectory t' for u' and v', such that $u \sqcup_t v = u' \sqcup_{t'} v'$. As mentioned above, the trajectory t correponds to a walk $\overline{w}$ through A.

As discussed in Lemma 1, the horizontal layers $0, \ldots, |u| - |v|$ have to be visited at least once during any walk through A. Let p be maximal such that

$$p \leq |u| - |v| \text{ and } p \bmod n = 0.$$

Thus $p \geq |u| - |v| - n$, which implies, as $|u| - |v| \geq n$, that layer p has to be visited at least once during any walk through A. Let $p' = p - n(l' - l)$ (see Figure 2). Then

$$p' \geq |u| - |v| - n - l'n + ln = kn - ln + |w_1| - |w_2| - n - l'n + ln > kn - l'n - 2n \geq -n.$$

Thus $p' > -n$, but as $p' \bmod n = 0$, this implies that $p' \geq 0$ and, thus, p' is also visited at least once during any walk through A.

We let (i, j) be the first occurrence of a state in h-layer p in $\overline{x}$ and we let (i', j') be the first occurrence of a state in h-layer p' in $\overline{x}$.

Then $i \bmod n = j \bmod n$ and $i' \bmod n = j' \bmod n$, which means that when in states (i, j) and (i', j') we are at the same point in the underlying period w for both words u and v. Let $t = t_1 t_2 t_3$ where t_1 is the part of t before visiting (i, j), t_2 is the part of t after visiting (i, j) but before visiting (i', j') and t_3 is the part of t after visiting (i', j'). Then $|t_2|_1 = |t_2|_0 + n(l' - l)$. Now let $t' = t_1 \bar{t}_2 t_3$, where $\bar{t}_2$ is obtained from t_2 by switching all 0's for 1's and vice versa. Then $u' \sqcup\!\sqcup_{t'} v' = u \sqcup\!\sqcup_t v$.

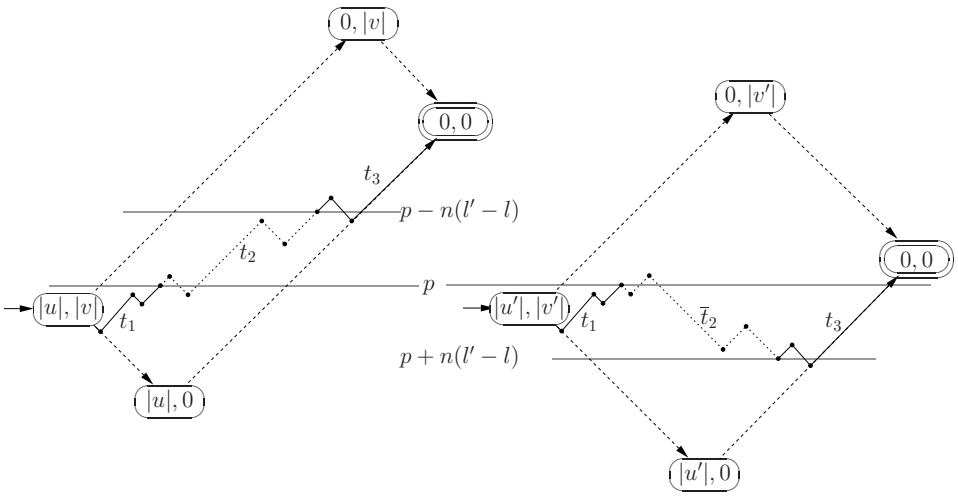

Figure 2: Transformation of a trajectory by switching all 0's and 1's in t_2.

Thus $u \sqcup\!\sqcup v \subseteq u' \sqcup\!\sqcup v'$ and by [1], we know that $u \sqcup\!\sqcup v \neq u' \sqcup\!\sqcup v'$, which implies $u \sqcup\!\sqcup v \subsetneq u' \sqcup\!\sqcup v'$. □

Example 2 Let $u = abc(babc)^5$, $v = bc(babc)^2$, $u' = abc(babc)^4$ and $v' = bc(babc)^3$. Then $u \sqcup\!\sqcup v \subsetneq u' \sqcup\!\sqcup v'$ by Lemma 2, and an example of a word $z \in u' \sqcup\!\sqcup v' \setminus u \sqcup\!\sqcup v$ is $z = ab^2c^2(b^2a^2b^2c^2)^3babc$.

The next result is similar to the previous one, but now only the suffixes of w at the beginning of the words are swapped and the number of repetitions of w do not change.

Lemma 3 *Let Σ be a finite alphabet and let $w = a_1 \cdots a_n$ for some $n \geq 2$, such that* $\mathrm{alph}(w) \geq 2$. *Let $u = w_1 w^k$, $v = w_2 w^l$, $u' = w_2 w^k$, $v' = w_1 w^l$ where $0 \leq l < k$ and $w_2 <_s w_1 \leq_s w$. Then $u \sqcup v \subsetneq u' \sqcup v'$.*

Proof. The proof is similar to the proof of the previous lemma. If $w_1 = w$, then $u = w^{k+1}$, $v = w_2 w^l$, $u' = w_2 w^k$ and $v' = w^{l+1}$ and the claim follows by Lemma 2 as $k+1 > k > l$ and $k+1 > l+1 > l$.

Assume that $|w_1| < |w| = n$. Let A be the naive shuffle NFA for u and v. We show again that for any trajectory t for u and v there exists a trajectory t' for u' and v', such that $u \sqcup_t v = u' \sqcup_{t'} v'$.

Let t be a trajectory for u and v. We define $p = |w_1| - |w_2|$, which implies that $1 \leq p < n$. Again t corresponds to a walk $\overline{x}$ through A. Let (i, j) be the first state in $\overline{x}$ that is in the horizontal layer $|u|-|v|-p$. As $|u|-|v| = |w_1|-|w_2|+n(k-l) \geq p+n$, this layer has to be visited at least once while reading $u \sqcup_t v$.

Let $t = t_1 t_2$, where t_1 is the part of t before visiting (i, j) and t_2 is the part of t after visiting (i, j). Now let $t' = \bar{t}_1 t_2$, where $\bar{t}_1$ is obtained from t_1 by switching all 0's for 1's and vice versa. Then $u' \sqcup_{t'} v' = u \sqcup_t v$.

Thus $u \sqcup v \subseteq u' \sqcup v'$ and by [1], we know that $u \sqcup v \neq u' \sqcup v'$, which implies $u \sqcup v \subsetneq u' \sqcup v'$. □

Example 3 Let $u = abc(babc)^3$, $v = bc(babc)^2$, $u' = bc(babc)^3$ and $v' = abc(babc)^2$. Then $u \sqcup v \subsetneq u' \sqcup v'$ by Lemma 3, and an example of a word $z \in u' \sqcup v' \setminus u \sqcup v$ is $z = bcba^2b^2c^2(b^2a^2b^2c^2)^2$.

We can use Lemma 2 and Lemma 3 to show a subset-relation between the languages defined by certain states of the naive shuffle NFA for two words that are periodic over the same underlying word. This result will be useful in the next subsection to show that the minimal DFA for the shuffle of periodic words over certain underlying words is smaller than the naive NFA for these words.

Lemma 4 *Let $u = w_1 w^k$ and $v = w_2 w^l$, where $w = a_1 \cdots a_n$ for some $n \geq 1$ such that $a_1, \ldots a_n \in \Sigma$ and w_1 and w_2 are suffixes of w. Let A be the naive shuffle NFA for u and v and let i, j, i', j' be natural numbers such that*

1. $1 \leq i \leq |u|$, $1 \leq i' \leq |u|$, $1 \leq j \leq |v|$, $1 \leq j' \leq |v|$;
2. $i + j = i' + j'$;
3. $\{i \bmod n, j \bmod n\} = \{i' \bmod n, j' \bmod n\}$; *and*
4. $|i - j| \geq |i' - j'|$.

Then $L_A(i, j) \subseteq L_A(i', j')$, and $L_A(i, j) = L_A(i', j')$ if and only if $\{i, j\} = \{i', j'\}$.

Proof. Obviously $\{i, j\} = \{i', j'\}$ implies $L_A(i, j) = L_A(i', j')$, so we only have to show that Conditions 1 through 3 and $|i-j| > |i'-j'|$ imply that $L_A(i, j) \subsetneq L_A(i', j')$.

By Condition 1 there exist suffixes $\overline{u}_i$ and $\overline{u}_{i'}$ of u and suffixes $\overline{v}_j$ and $\overline{v}_{j'}$ of v, such that $L(i, j) = \overline{u}_i \sqcup \overline{v}_j$ and $L(i', j') = \overline{u}_{i'} \sqcup \overline{v}_{j'}$. Condition 3 implies that there exist

suffixes $\overline{w}_1, \overline{w}_2$ of w, such that

$$\{\overline{u}_i, \overline{v}_j\} = \{\overline{w}_1 w^p, \overline{w}_2 w^q\} \text{ and } \{\overline{u}_{i'}, \overline{v}_{j'}\} = \{\overline{w}_1 w^{p'}, \overline{w}_2 w^{q'}\}$$

for some $p, q, p', q' \geq 0$. Furthermore, Condition 2 implies that the words in $L(i,j)$ and $L(i',j')$ all have the same length, which implies $p+q = p'+q'$. Now we get two cases, depending on whether i and j are from the same iteration of w as i' and j' or not.

If $\{i \text{ div } n, j \text{ div } n\} = \{i' \text{ div } n, j' \text{ div } n\}$, then $\{p,q\} = \{p',q'\}$. Then $|i-j| > |i'-j'|$ implies that $p = q'$ and $q = p'$ and either both $|\overline{w}_1| > |\overline{w}_2|$ and $p > q$, or both $|\overline{w}_1| < |\overline{w}_2|$ and $p < q$. We assume the former without loss of generality and obtain $L(i,j) \subsetneq L(i',j')$ by Lemma 3.

If $\{i \text{ div } n, j \text{ div } n\} \neq \{i' \text{ div } n, j' \text{ div } n\}$ then $\{i,j\} \neq \{i',j'\}$ follows immediately. Thus, by Condition 4, we have $|i-j| > |i'-j'|$, which implies without loss of generality that $q < q' < p$ and $q < p' < p$ (the case where $p < q' < q$ and $p < p' < q$ is symmetric). But this implies that $L(i,j) \subsetneq L(i',j')$ by Lemma 2.

Thus, $L_A(i,j) \subseteq L_A(i',j')$ and $L_A(i,j) = L_A(i',j')$ if and only if $\{i,j\} = \{i',j'\}$. □

4.1. Underlying Non-Repeating Words

We now show that the shuffle of periodic words over a non-repeating w yields deterministic finite automata that have at most a quadratic number of states.

The next two theorems show that the minimal DFA accepting $u \sqcup v$ for two words u and v that are periodic over the same non-repeating word has less than $(|u|+1)\cdot(|v|+1)$ states. The first theorem deals with the case where $u \neq v$ and the second theorem deals with the case where $u = v$.

Theorem 5 *Let $u = w_1 w^k$ and $v = w_2 w^l$, where $k > l \geq 0$ and $w = a_1 \cdots a_n$ for some $n \geq 2$ such that $a_i = a_j$ implies $i = j$ whenever $1 \leq i \leq n$ and $1 \leq j \leq n$ and w_1, w_2 are non-empty suffixes of w. Then the minimal DFA for $u \sqcup v$ has*

$$(|u|+1)\cdot(|v|+1) - \frac{1}{2}(|v|)\cdot(|v|+1) - \frac{1}{2}m\cdot(m+1)$$

states, where $m \leq |v|$ is maximal such that $(|u|-m) \bmod n = 0$.

Proof. We first construct a naive shuffle NFA $A = (Q, \Sigma, \delta, s_0, F)$ for u and v. Obviously $|Q| = (|u|+1)\cdot(|v|+1)$ by Definition 1. In the following we perform several transformations with the automaton A, so that in the end A has the properties that are mentioned in the theorem statement.

Removing $\frac{1}{2}|v|\cdot(|v|+1) + \frac{1}{2}m\cdot(m+1)$ states: We look at the horizontal layer 0, which contains the final state $(0,0)$ as well as the states $(|v|,|v|),\ldots,(1,1)$. All the states in this layer except the final state are nondeterministic, so for all i with $0 < i \leq |v|$ there exists $a \in \Sigma$, such that $\delta((i,i),a) = \{(i-1,i),(i,i-1)\}$. By Lemma 4 we know that $L(i,i-1) = L(i-1,i)$. Thus, we can modify the transition function δ to $\delta((i,i),a) = (i,i-1)$ without changing the accepted language. When we have

done this modification of δ for all nondeterministic states in the horizontal layer 0, the states in the horizontal layers $-1, \ldots, -|v|$ are unreachable and can be removed from Q. It is easy to see that the number of states removed in this way is

$$\sum_{i=1}^{|v|} i = \frac{1}{2}|v| \cdot (|v|+1).$$

We now look at the horizontal layer $|u| - m$, which contains the states $(|u|, m)$, $(|u|-1, m-1), \ldots, (|u|-m, 0)$. As $m \leq |v|$ is maximal, such that $(|u|-m) \bmod n = 0$, the horizontal layers $(|u| - |v|)$ (which contains the initial state), $(|u| - (|v| - 1)), \ldots, (|u| - (m+1))$ do not contain any nondeterministic states.

Furthermore, we know that all states in the horizontal layer $|u| - m$ except for the state $(|u| - m, 0)$ are nondeterministic. Thus, if we let (i, j) be one of the nondeterministic states in the horizontal layer $|u| - m$, then $(i, j) = (|u| - p, m - p)$ for some $0 \leq p < m$ and there exists $a \in \Sigma$, such that $\delta((i, j), a) = \{(i-1, j), (i, j-1)\}$. This implies that $\{(i-1) \bmod n, j \bmod n\} = \{i \bmod n, (j-1) \bmod n\}$, as the outgoing transitions of both states $(i-1, j)$ and $(i, j-1)$ carry the same labels and w is non-repeating. Also it is obvious that $(i-1) + j = i + (j-1)$ and $1 \leq i \leq |u|$, $1 \leq i - 1 \leq |u|$, $1 \leq j \leq |v|$, $1 \leq j - 1 \leq |v|$. Furthermore as $|u| > |v|$, we have

$$|(i-1) - j| = ||u| - m - 1| < ||u| - m + 1| = |i - (j-1)|.$$

Therefore by Lemma 4 we have $L(i-1, j) \subsetneq L(i, j-1)$, which implies that we can modify the transition function δ of A to $\delta((i, j), a) = (i, j-1)$ without changing the accepted language.

Once we have done that for all the states in the horizontal layer $|u| - m$ all the states in horizontal layers $|u| - m + 1, \ldots, |u|$ are no longer reachable and can be removed. It is again easy to see that the number of states removed in this way is

$$\sum_{i=1}^{m} i = \frac{1}{2}m \cdot (m+1).$$

We now have $|Q| = (|u|+1) \cdot (|v|+1) - \frac{1}{2}(|v|) \cdot (|v|+1) - \frac{1}{2}m \cdot (m+1)$, as claimed in the theorem statement, however our automaton A could still be nondeterministic. **Removing remaining nondeterminism:** The only horizontal layers left in A' are $|u| - m, \ldots, 0$. Furthermore we have already removed all nondeterminism from the horizontal layers $|u| - m$ and 0. Also note that all states $(i, j) \in Q$ now have $i \geq j$ and the only states with $i = j$ are those in the horizontal layer 0. Thus, all remaining nondeterminism must occur in the horizontal layers $|u| - m - 1, \ldots, 1$. However, a state (i, j) is nondeterministic precisely when $i \bmod n = j \bmod n$, which is only possible for states in the horizontal layers $|u| - m - pn$ where $1 \leq p < k - l$. Let (i, j) be such a state. As (i, j) has precisely two outgoing transitions, this implies that there exists a letter $a \in \Sigma$, such that $\delta((i, j), a) = \{(i-1, j), (i, j-1)\}$. As no letter appears more than once in w, we know that $\{(i-1) \bmod n, j \bmod n\} = \{i \bmod n, (j-1) \bmod n\}$. Thus, as $i > j$ implies that $|(i-1) - j| < |i - (j-1)|$, which implies, by Lemma

4, $L(i, j-1) \subsetneq L(i-1, j)$. We can, thus, redefine $\delta((i,j),a) = (i-1,j)$ without changing $L(A)$.

Showing minimality: Now A is deterministic, but we still have to show that A is accessible, co-accessible and minimal.

First note that all states of the naive NFA were accessible and co-accessible and that from a state (i,j) in the naive NFA all states (i',j') with $i' \leq i$ and $j' \leq j$ could be reached.

First assume that the initial state $(|u|,|v|)$ was deterministic in the naive NFA. Then all the states in the horizontal layers $|u|-|v|, |u|-|v|+1, \ldots, |u|-m-1$ were already deterministic in the naive NFA, and as the initial state is contained in the horizontal layer $|u|-|v|$, all states in the horizontal layers $|u|-|v|, |u|-|v|+1, \ldots, |u|-m-1$ are accessible. Furthermore, as all states in the horizontal layer $|u|-m-1$ were deterministic in the naive NFA, each state in horizontal layer $|u|-m$ is accessible from some state in horizontal layer $|u|-m-1$ and, thus, all states in horizontal layer $|u|-m$ are accessible.

Now assume that the initial state was nondeterministic. In this case there exists only one letter a such that $\delta((|u|,|v|),a)$ is defined and we have, by the above construction, $\delta((|u|,|v|),a) = (|u|-1,|v|)$. Then, as $|w| \geq 2$, we know that all the states in the horizontal layer $|u|-|v|-1$ were already deterministic in the naive NFA. Thus we can access all states in the horizontal layers $|u|-|v|$ and $|u|-|v|+1$.

Let (i,j) be a state in horizontal layers r, where $|u|-m \leq r \leq 1$. Then there exists a letter $a \in \Sigma$, such that $\delta((i,j),a) = (i-1,j)$ and, thus, by induction, all remaining states in the automaton are accessible.

All states in horizontal layers 0 and 1 are co-accessible, as for each i with $1 \leq i \leq |v|$ there exists a letter $a_i \in \Sigma$, such that we have $\delta((i,i),a_i) = (i,i-1)$ and for each j with $1 \leq j \leq |v|$ there exists a letter $b_j \in \Sigma$, such that $\delta((j,j-1),b_j) = (j-1,j-1)$. Then by an inductive argument similar to the one used to show that A is accessible, we can show that all remaining states are co-accessible.

To show that A is minimal, we first observe that state equivalence is only possible between states in the same vertical layer. We use induction on the vertical layers to show that no vertical layer has any equivalent states. Vertical layers 0 and 1 both only contain one state, namely $(0,0)$ and $(1,0)$, respectively. Now assume that in the vertical layer q, for $1 \leq q < |u|+|v|$ there are no equivalent states.

Then, in order to have two equivalent states (i',j') and (i'',j'') in vertical layer $q+1$, it is necessary (but not sufficient) that there exists a letter $a \in \Sigma$ and a state (i,j) in the vertical layer q, such that $\delta((i',j'),a) = \delta((i'',j''),a) = (i,j)$. This implies that $\{(i',j'),(i'',j'')\} = \{(i,j+1),(i+1,j)\}$. We have $i \geq j+1$, as this is true for all states in Q. If $i > j+1$, then, as shown in Figure 3, there has to exist a letter $b \in \Sigma$ such that $\delta((i,j+1),b) = (i-1,j+1)$ as in the construction above we never removed such transitions. But this implies that $(i,j+1)$ and $(i+1,j)$ cannot be equivalent as no transition can exist between states $(i+1,j)$ and $(i-1,j+1)$. If $i = j+1$, then this implies that the i-th and the $i+1$-st letter of u have to be equal, a contradiction as $|w| \geq 2$ and w is non-repeating. Thus, the automaton is minimal. □

Example 4 Let $u = bc(abc)^2$, $v = abc$. Then the naive NFA for $u \sqcup v$ is shown

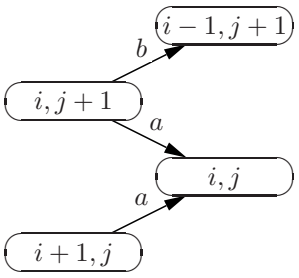

Figure 3: If $i > j+1$, then $(i, j+1)$ and $(i+1, j)$ cannot be equivalent.

on the left side of Figure 4. According to the proof of Lemma 5, we can remove all the shaded states and transitions and we can furthermore also remove the dashed non-shaded transitions. This then leaves the minimal DFA for $u \sqcup\!\sqcup v$, as shown on the right side of Figure 4.

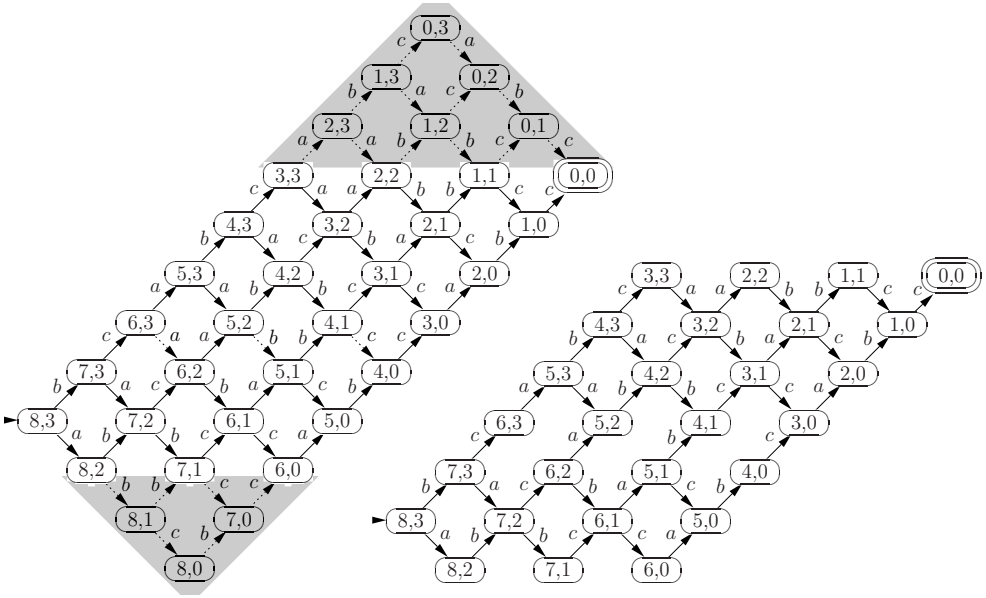

Figure 4: Naive shuffle NFA and minimal shuffle DFA for $u = bc(abc)^2$ and $v = abc$.

If $k = l$, then there are fewer than $|w|$ horizontal layers between the initial and final state and the proof of Theorem 6 has to be adapted.

Theorem 6 *Let $u = w_1 w^k$ and $v = w_2 w^k$ where $k \geq 0$ and $w = a_1 \cdots a_n$ for some $n \geq 2$ such that $a_i = a_j$ implies $i = j$ whenever $1 \leq i \leq n$ and $1 \leq j \leq n$ and w_1, w_2 are non-empty suffixes of w, such that $|w_1| \geq |w_2|$. Then the number of states in the minimal DFA for $u \sqcup v$ is*

$$(|u|+1)\cdot(|v|+1) - \frac{1}{2}|v|\cdot(|v|+1) - \frac{1}{2}(|u|-|w|)(|u|-|w|+1).$$

Proof. As in the previous proof, we construct a naive NFA $A = (Q, \Sigma, \delta, s_0, F)$ for $u \sqcup v$. Obviously $|Q| = (|u|+1)(|v|+1)$ by Definition 1. In the following we perform several transformations with the automaton A, so that in the end A has the properties that are mentioned in the theorem statement.

We first show that we can remove $\frac{1}{2}|v|\cdot(|v|+1)$ states without changing the accepted language.

We look at the horizontal layer 0, which contains the states $(|v|,|v|),(|v|-1,|v|-1),\ldots,(1,1)$ and the final state $(0,0)$. We know that all states in this horizontal layer except for the state $(0,0)$ are nondeterministic. Let (i,i) be one of these states, then there exists $a \in \Sigma$, such that $\delta((i,i),a) = \{(i-1,i),(i,i-1)\}$. By Lemma 4 we have $L(i-1,i) = L(i,i-1)$, which implies that we can modify the transition function δ of A to $\delta((i,i),a) = (i,i-1)$ without changing the accepted language.

Once we have done that for all the states in the horizontal layer 0 all the states in horizontal layers $-1,\ldots,-|v|$ are no longer reachable and can be removed. It is easy to see that the number of states removed in this way is

$$\sum_{i=1}^{|v|} i = \frac{1}{2}|v|\cdot(|v|+1).$$

We now look at the horizontal layer $|w|$, which contains the states $(|u|,|u|-|w|),(|u|-1,|u|-|w|-1),\ldots,(|w|,0)$. All the states in this layer except for the state $(|w|,0)$ are nondeterministic, so for all i with $0 < i \leq |u|$ there exists $a \in \Sigma$, such that $\delta((i,i-|w|),a) = \{(i,i-|w|-1),(i-1,i-|w|)\}$. This implies that $\{i \bmod n,(i-|w|-1) \bmod n\} = \{(i-1) \bmod n,(i-|w|) \bmod n\}$, as the outgoing transitions of both states $(i,i-|w|-1)$ and $(i-1,i-|w|)$ carry the same labels and w is non-repeating. Also $|i-(i-|w|-1)| > |(i-1)-(i-|w|)|$ and, therefore, we can apply Lemma 4 to show that $L(i,i-|w|-1) \subsetneq L(i-1,i-|w|)$, which implies that we can modify the transition function δ of A to $\delta((i,i-|w|),a) = (i-1,i-|w|)$ without changing the accepted language. When we have done this modification of δ for all nondeterministic states in the horizontal layer $|w|$, the states in the horizontal layers $|w|+1,\ldots,|u|$ are unreachable and can be removed from Q. It is easy to see that the number of states removed in this way is

$$\frac{1}{2}(|u|-|w|)\cdot(|u|-|w|+1).$$

We now have $|Q| = (|u|+1)(|v|+1) - \frac{1}{2}|v|\cdot(|v|+1) - \frac{1}{2}(|u|-|w|)(|u|-|w|+1)$, and, as there are only $|w|$ horizontal layers left and w is non-repeating, there are no more nondeterministic states in A.

Now all we are left to show is that A is accessible, co-accessible and minimal. This can be done in the same way as in the proof of Theorem 5. □

From the proofs of Theorem 5 and 6 and it is immediate that we can construct the minimal shuffle DFA for periodic words over a non-repeating underlying word directly without first constructing the NFA.

Corollary 7 *Let $u = w_1 w^k$ and $v = w_2 w^l$, where $k \geq l \geq 0$ and $w = a_1 \cdots a_n$ for some $n \geq 2$ such that $a_i = a_j$ implies $i = j$ whenever $1 \leq i \leq n$ and $1 \leq j \leq n$ and w_1 is a proper suffix of w or empty. Then we can effectively construct the minimal DFA A for $u \sqcup\!\sqcup v$, as mentioned in Theorems 5 and 6 in time $\mathcal{O}(|u| \cdot |v|)$.*

4.2. Periodic Words with one Section per Letter

We now generalize Theorem 5 to underlying words the skeletons of which are non-repeating. That is, we still consider only words $u = w_1 w^k$ and $v = w_2 w^l$, where $k \geq l \geq 0$ and w_1 and w_2 are proper (possibly empty) suffixes of w. However, w no longer has to be non-repeating, but we now have $w = a_1^{p_1} \cdots a_n^{p_n}$ for some $n \geq 2$ and positive integers $p_1, \ldots, p_n$ and where $a_1 \cdots a_n$ is non-repeating.

Theorem 8 *Let Σ be a finite alphabet and let $w \in \Sigma^+$, such that $|w| = n \geq 2$ and for all $a \in \Sigma$, we have $|\chi(w)|_a \leq 1$. Let $u = w_1 w^k$ and $v = w_2 w^l$ where w_1, w_2 are suffixes of w and $k, l \geq 0$. Then there exists a DFA A with $L(A) = u \sqcup\!\sqcup v$ and $|A| \in \mathcal{O}(|u| \cdot |v|)$.*

Proof. Let $A' = (Q', \Sigma, \delta', Q_0', F')$ be the naive shuffle NFA for u and v. Obviously $|A'| = (|u| + 1)(|v| + 1)$. We show that for each nondeterministic area $R \in \mathrm{Area}(A')$, we can determinize R in such a way, by using Lemma 4, that no state in the DFA contains more than one label from $\mathrm{ex}(R)$, and no more than $\mathcal{O}(|\mathrm{states}(R) \cup \mathrm{ex}(R)|)$ contain labels from $\mathrm{states}(R) \cup \mathrm{ex}(R)$.

Let $R = (a, (i_1, j_1), (i_2, j_2)) \in \mathrm{Area}(A')$ and let $(i, j) \in \mathrm{ent}(R)$. When determinizing R by using a subset construction it is easy to see that if both states $(i', j') \in Q'$ and $(i'', j'') \in Q'$ can be reached from (i, j) by reading k a's for some $k \in \mathbb{N}$, then also all states $(\bar{i}, \bar{j})$ with $\bar{i} + \bar{j} = i' + j'$ and either both $i' \leq \bar{i} \leq i''$ and $j' \geq \bar{j} \geq j''$ or both $i' \geq \bar{i} \geq i''$ and $j' \leq \bar{j} \leq j''$ can be reached from (i, j) by reading k a's. Furthermore if some state (i', j') can be reached from (i, j) by reading k a's, then also some state $(i''', j''') \in \mathrm{ent}(R) \cup \mathrm{ex}(R)$ can be reached from (i, j) by reading k a's. This implies that at most $2|\mathrm{states}(R) \cup \mathrm{ex}(R)|$ states can result from a subset construction on R, assuming that we are starting with states that contain only individual entrance state labels.

It is also obvious that each state q obtained by performing a subset construction on R contains at most 2 exit state labels (as there are only two exit states of R per vertical layer). If there is at most one exit state of R in q, then q does not induce any states with multiple labels outside of the states in $\mathrm{states}(R) \cup \mathrm{ex}(R)$ and we are done. If q contains distinct exit states (i', j') and (i'', j'') then there exists an $n \in \mathbb{N}$, with $1 \leq n \leq (i_1 - i_2)$ such that $(i', j') = (i_2 - 1, j_2 + n)$ and $(i'', j'') =$

$(i_2 + n, j_2 - 1)$ (or vice versa). But then, as $i_2 \bmod n = j_2 \bmod n$, we know that $\{i' \bmod n, j' \bmod n\} = \{i'' \bmod n, j'' \bmod n\}$. Furthermore we know that $i' + j' = i'' + j''$ and either $|i' - j'| \geq |i'' - j''|$ or $|i' - j'| < |i'' - j''|$. Thus by Lemma 4 we have either $L_A(i', j') \subseteq L_A(i'', j'')$ (if $|i' - j'| \geq |i'' - j''|$) or $L_{A'}(i'', j'') \subset L_{A'}(i', j')$ (if $|i'' - j''| > |i' - j'|$) and, hence, we can remove one of (i', j') and (i'', j'') from q without changing the accepted language.

Thus, the nondeterministic areas do not induce any states with multiple layers outside of the nondeterministic areas, which implies that $|A| \in \mathcal{O}(|u| \cdot |v|)$. □

5. Exponential Shuffle Automata

Theorem 9 *Let Σ be an alphabet of size at least 2. Then there exist words $u, v \in \Sigma^+$, $|u| = |v|$, such that the size of the minimal DFA accepting $u \sqcup v$, is $\Omega(\sqrt[8]{2}^{|u|})$.*

Note that, in the proof below, the numbering of the layers is different from the numbering used thus far.

Proof. For $n > 1$, let

$$u_n = (aabb)^n aabbaabb(aabb)^n aaaaa, v_n = (aabb)^n aababbabb(aabb)^n bbbbb,$$
$$X_n = a(aabb)^n aaa(bbbbaaaa + bbbabaaa)^{n+1} bbbb(aabb)^n aaaaabbbbb.$$

Let $A_n = (Q, \Sigma, q_0, F, \delta)$ be the naive shuffle NFA for u_n and v_n. We have A_2 pictured in Figure 5.

Let $m = |v_n| = |u_n| = 8n + 13$, and there are $2(8n + 13) + 1 = 16n + 27$ vertical layers. For each layer i, let Q_i be the set of states in that layer. Let $q_{i,j}$ be the jth state (along the diagonal) in the ith layer. There are i states in the ith layer for $i \leq 8n + 14$ and $(8n + 14) - (i - (8n + 14)) = 16n + 28 - i$ for $8n + 14 < i$. For each w which is a prefix of some word in $u_n \sqcup v_n$, let Q_w be the set of states $\delta(q_0, w)$. We will only consider input words in X_n. In Figure 5, we have the set of states Q_w, with each state denoted by bullet points.

We show by induction that for each i, $1 \leq i \leq n, Q_{a(aabb)^i} = \{q_{4i+2,j}, q_{4i+2,j+3} \mid j = 2 + 4l, 0 \leq l < i\}$. This is the "duplication stage", consisting of the states in the shaded top left corner of Figure 5.

One can see from Figure 5 that after reading *aaabb*, we are in either state $q_{6,2}$ or $q_{6,5}$ and thus $Q_{a(aabb)} = \{q_{6,2}, q_{6,5}\}$ which is equal to $\{q_{4+2,j}, q_{4+2,j+3} \mid j = 2\}$ and thus the base case holds.

Assume this holds for some $i < n$. For $0 \leq l < i$, $q_{4i+2,2+4l}$ must have read $2 + 4l - 1 = 4l + 1$ letters from u_n and $4(i - l)$ from v. Similarly, $q_{4i+2,5+4l}$ must have read $4l + 4$ letters from u_n and $4(i - l) - 3$ letters from v. Thus, from $q_{4i+2,2+4l}$, the next three letters to be read from u are *abb* and the next four from v_n are *aabb*, and from $q_{4i+2,5+4l}$, the next four from u_n are *aabb* whilst the next three from v_n are *abb*. Then consider $Q_{a(aabb)^{i+1}}$. From $q_{4i+2,2+4l}$, after reading *aabb*, we could be either in state $q_{4i+2+4,2+4l} = q_{4(i+1)+2,2+4l}$ or $q_{4i+2+4,5+4l} = q_{4(i+1)+2,5+4l}$. From $q_{4i+2,5+4l}$ after reading *aabb*, we could be either in state $q_{4(i+1)+2,6+4l} = q_{4(i+1)+2,2+4(l+1)}$ or

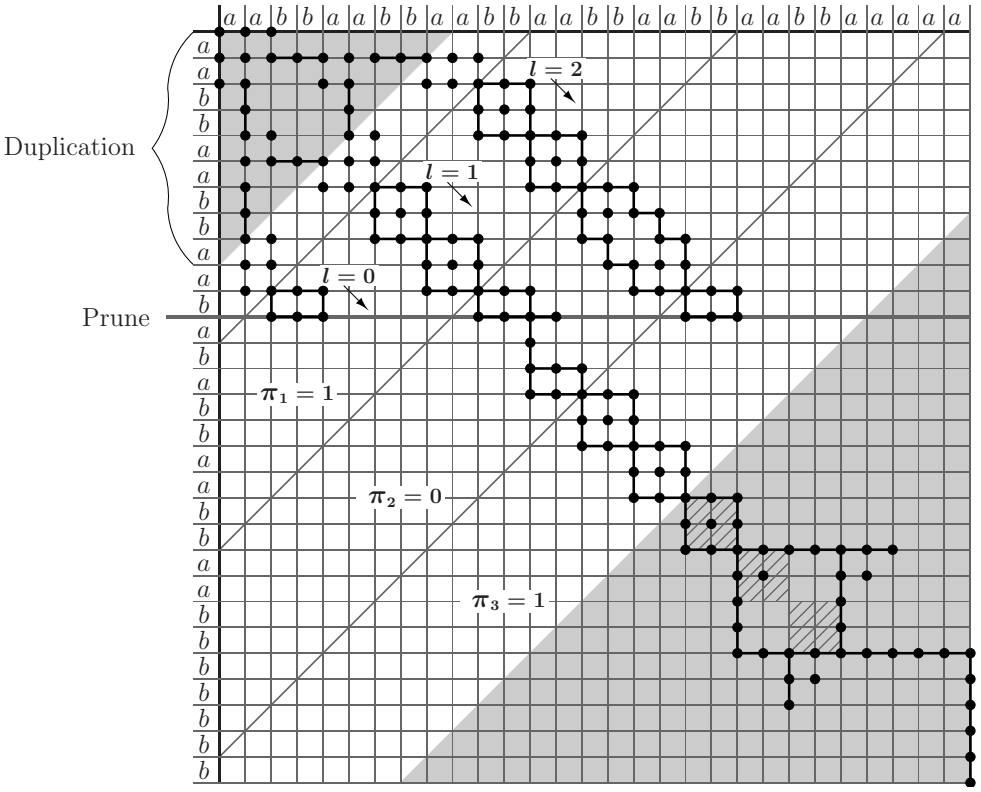

Figure 5: The diagram is the naive NFA A_2, with the top left corner as the initial state, the bottom right corner being the final state, and the lines of the grid being transitions on the letter labelling the axis, with u_2 along the horizontal and v_2 along the vertical axis. The input to A_2 is $a(aabb)^2aaa(bbbbaaaa)(bbbabaaa)(bbbbaaaa)bbbb(aabb)^2aaaaabbbbb$, with active states marked with bullet points.

$q_{4(i+1)+2,5+4l+4} = q_{4(i+1)+2,5+4(l+1)}$. Thus, for each l, we have $Q_{a(aabb)^{i+1}} =$

$$\{q_{4(i+1)+2,2+4l}, q_{4(i+1)+2,4l+5}, q_{4(i+1)+2,2+4(l+1)}, q_{4(i+1)+2,5+4(l+1)} \mid 0 \le l < i\}$$
$$= \{q_{4(i+1)+2,2+4l}, q_{4(i+1)+2,5+4l} \mid 0 \le l < i+1\}.$$

and thus the induction holds.

Thus, after reading $a(aabb)^n$, we are in one of the states in $Q_{a(aabb)^n} = \{q_{4n+2,j}, q_{4n+2,j+3} \mid j = 2+4l, 0 \le l < n\}$. This occurs at the bottom diagonal of the "duplication" section in Figure 5. Then $Q_{a(aabb)^n aaa} = \{q_{4n+5,j} \mid j = 3+4l, 0 \le l \le n\}$ which is of size $n+1$. The next set of input letters is in $(bbbbaaaa + bbbabaaa)^{n+1}$. This is the so called "filtering stage", marked in white in Figure 5. Intuitively, each element of $Q_{a(aabb)^n aaa}$, as determined by l, will continue roughly along a diagonal

(we get a diagonal for l being $0, 1, 2$ in the figure) until each reaches $baba$ along v marked by the "prune" line of the figure. If the input is then $bbbbaaaa$, this diagonal gets "cut off", while all other states in the vertical layer are able to continue along its diagonal. However, if the input is $bbbabaaa$, then every diagonal in the vertical layer is able to continue. Since each diagonal reaches the "prune" line at a different time, we can selectively keep or remove each diagonal one at a time.

More formally, assume that $x_1 \cdots x_{n+1}$ is the input, $x_i \in (bbbbaaaa + bbbabaaa)$. Let $\pi_i = 0$ if $x_i = bbbabaaa$, and $\pi_i = 1$ if $x_i = bbbbaaaa$. The sections of A_2 when reading x_1, x_2, x_3 are separated by lines in Figure 5 where $\pi_1 = 1, \pi_2 = 0, \pi_3 = 1$. We can then show by induction that for each i, $1 \leq i \leq n+1$,

$$Q_{a(aabb)^n aaax_1 \cdots x_i} = \{q_{4n+5+8i,j} \mid j = 3 + 4l + 4i, 0 \leq l \leq n, (l < i \Rightarrow \pi_l = 1)\}.$$

For the base case with $i = 1$, we can see that if $l > 0$, then the next four letters to be read from both u and v at state $q_{4n+5,3+4l}$ are $bbaa$. Thus, if $\pi_1 = 1$ then $q_{4n+5+8,j} \in Q_{a(aabb)^n x_1}$, where $j = 3 + 4l + 4$, (this is shown in the figure where l is 1 or 2) but $\delta(q_{4n+5+8,3+4}, bbbb)$ is undefined since the next two letters to be read from v are ba and bba from u and thus $bbbb$ is too many b's. However, if $\pi_1 = 0$, then $q_{4n+5+8,j} \in Q_{a(aabb)^i aaax_1}, j = 3 + 4l + 4, l > 0$, but also $\delta(q_{4n+5+8,3+4}, x_1)$ is defined since bbb can read two letters from u, one from v, then aba from v and aa from u (this is similar to the pattern where the $l = 1$ diagonal passes the "prune" line as $\pi_2 = 0$). Hence,

$$Q_{a(aabb)^n aaax_1} = \{q_{4n+5+8,j} \mid j = 3 + 4l + 4, 0 \leq l \leq n, (l = 0 \Rightarrow \pi_1 = 0)\}$$

and the base case holds.

Assume by way of induction that $i < n + 1$ and

$$Q_{a(aabb)^n aaax_1 \cdots x_i} = \{q_{4n+5+8i,j} \mid j = 3 + 4l + 4i, 0 \leq l \leq n, (l < i \Rightarrow \pi_l = 0)\}.$$

Assume first that $l > i$. Then the next four letters to be read from u and v are both $bbaa$. Then if either $\pi_l = 1$ or $\pi_l = 0$, $\delta(q_{4n+5+8i,3+4l+4i}, x_{i+1}) = \{q_{4n+5+8(i+1),3+4l+4(i+1)}\}$ (this occurs when either $i = 1, l = 2$ or $i = 2, l = 3$ in the figure). Assume $l < i$. Then $q_{4n+5+8i,3+4l+4i} \in Q_{a(aabb)^n aaax_1 \cdots x_i}$ if and only if $\pi_l = 0$ by the inductive hypothesis. Assume that $\pi_l = 0$. Then the next four letters to be read from both u and v are $bbaa$ (for example, when $l = 2, i = 3$ in the figure) and after reading x_{i+1}, and either $\pi_{i+1} = 0$ or $\pi_{i+1} = 1$, then $\delta(q_{4n+5+8i,3+4l+4i}, x_{i+1}) = \{q_{4n+5+8(i+1),3+4l+4(i+1)}\}$. Lastly assume $l = i$. Then the next four letters to be read from u are $bbaa$ and from v are $baba$. Then, if the next four input letters are b's ($\pi_{i+1} = 1$), then $\delta(q_{4n+5+8i,3+4l+4i}, bbbb)$ is undefined (as when $l = 0$ or $l = 2$ in the figure). Otherwise, if $\pi_{i+1} = 0$ (when $l = 1$ in the figure), then $\delta(q_{4n+5+8i,3+4l+4i}, x_{i+1}) = \{q_{4n+5+8(i+1),3+4l+4(i+1)}\}$ and the induction holds.

Hence, $Q_{a(aabb)^n aaax_1 \cdots x_{n+1}} = \{q_{4n+5+8(n+1),j} \mid j = 3 + 4l + 4(n+1), \pi_l = 0\}$. No matter the contents of this set, which depends on $x_1, \cdots, x_{n+1}$, every state can reach a final state on $bbbb(aabb)^n aaaaabbbbb$ since the rest of u is of the form $bb(aabb)^* aaaaa$ and the rest of v is of the form $bb(aabb)^* bbbbb$. Therefore, if we use the subset

construction [5] on A_n, there is only one set of states we can be in after reading each prefix of $a(aabb)^n aaa$. As we read each prefix w of $x_1 \cdots x_{n+1}, w = x_1 \cdots x_i y, |y| < 8, x_j \in (bbbbaaaa + bbbabaaa), j \leq i$, then $q_{4n+5+8i,3+4l+4i} \in Q_{a(aabb)^n x_1 \cdots x_i}$ if and only if $l \geq i$ or $\pi_l = 0$. There are 2^i such subsets. And indeed, if $|y| \geq 4$, then $\delta(q_{4n+5+8i,3+4l+4i}, y)$ is undefined if and only if $\pi_{i+1} = 1$. Hence, after reading each prefix of length 1 to $|x_1 \cdots x_{n+1}|$, there are

$$3 + 8 \cdot 2^1 + 8 \cdot 2^2 + \cdots + 8 \cdot 2^n + 5 \cdot 2^{n+1} = 3 + 5 \cdot 2^{n+1} + 8(2^1 + \cdots + 2^n)$$
$$= 3 + 5 \cdot 2^{n+1} + 8(2^{n+1} - 2) = 13(2^{n+1}) - 13 = 13(2^{n+1} - 1)$$

sets of states created in the subset construction. Thus, when reading every prefix of $a(aabb)^n aaax_1 \cdots x_{n+1}$, $4(n+1) + 13(2^{n+1} - 1)$ sets of states are created and thus the subset construction requires at least this many states, and the remaining input is of length $4(n+1) + 10$, the automaton from the subset construction has at least $8(n+1) + 13(2^{n+1} - 1) + 10$ states.

We can now show that each of the states created in the subset construction from each prefix of $a(aabb)^n aaax_1 \cdots x_{n+1}$ are distinguishable from each other. First, if two states were not distinguishable from each other, then they must be created from the same vertical layer as otherwise, there would be different lengths to reach the end of u_n and v_n. As there was only one state created in the subset construction automaton for each prefix of $a(aabb)^n aaa$, each such state is distinguishable from all other such states. Let Y_1, Y_2 be two different states created in the subset construction from reading some prefix of $a(aabb)^n aaax_1 \cdots x_{n+1}$. Assume without loss of generality that there exists $q \in Y_1 \setminus Y_2$. Then, there exists some path from state q to a state p on the remaining input of $a(aabb)^n aaax_1 \cdots x_{n+1} bbbb$. Then p cannot be reached on this input from any state in Y_2 (as is the path where $l = 1$ until it hits the bottom grey section and then reads $bbbb$). Moreover, there exists some word

$$w = (aaaabbbb)^+(aaaaa)(aabb)^*(bbbbb) + (aaaabbbb)^+(bbbbb)(aabb)^*(aaaaa)$$

such that $\delta(p, w) \in F$ (with state sets marked by the dashed lines in Figure 5, followed by $aaaaabbbbb$. In general, the path always marked by the states can move diagonally until it hits either the end of u or v with $aaaaa$ or $bbbbb$, then finishing one word followed by the other) and no word in Y_2 can reach any final state with this word. Hence, Y_1 and Y_2 are distinguishable. There is at least one state in each of the remaining vertical layers. From [5], this implies the minimal automaton accepting $u_n \sqcup v_n$ has at least $8(n+1)+13(2^{n+1}-1)+10$ states. As $m = |u_n| = |v_n| = 8n+13$, this implies $n = \frac{m-13}{8}$. Hence, the automaton has

$$8(\frac{m-13}{8} + 1) + 13(2^{\frac{m-13}{8}+1} - 1) + 10$$
$$= m - 13 + 8 + 13(2 \cdot 2^{\frac{m-13}{8}} - 1) + 10$$
$$= m - 8 + \frac{26}{\sqrt[8]{2}^{13}} \cdot \sqrt[8]{2}^m$$
$$\geq m - 8 + 8.42 \cdot \sqrt[8]{2}^m \geq m - 8 + 8.42 \cdot 1.09^m.$$

Hence, we get $\Omega(\sqrt[8]{2}^m)$ where $|u_n| = |v_n| = m$. □

Theorem 9 is especially interesting in light of Theorem 8, which showed that the minimal DFA for the shuffle of $u = (aabb)^{2n+2}$ and $v = (aabb)^{2n+2}$ is in $\mathcal{O}(n^2)$. It is easy to see that adding 5 a'a and 5 b's to the ends of these words does not change this bound. The words used in the proof of Theorem 9 differ from these u and v only by switching two letters in one of the words, and yet this subtle change is enough to cause an exponential blow-up in size.

References

[1] J. Berstel, L. Boasson, Shuffle Factorization is Unique. *Theoretical Computer Science* **273** (2002), 47–67.

[2] F. Biegler, M. Daley, M. Holzer, I. McQuillan, On the uniqueness of shuffle on words and finite languages. *Theoretical Computer Science* (2009). Accepted March 2009, 14 pages.

[3] C. Câmpeanu, K. Salomaa, S. Vágvölgyi, Shuffle Quotient and Decompositions. In: W. Kuich, G. Rozenberg, A. Salomaa (eds.), *Proceedings DLT 5*. Number 2295 in LNCS, Springer, Wien, Austria, 2001, 186–196.

[4] C. Câmpeanu, K. Salomaa, S. Yu, Tight Lower Bound for the State Complexity of Shuffle of Regular Languages. *Journal of Automata, Languages, and Combinatorics* **7** (2002), 303–310.

[5] J. Hopcroft, J. Ullman, *Introduction to Automata Theory, Languages, and Computation*. Addison-Wesley, Reading, MA, 1979.

[6] A. Mateescu, G. Rozenberg, A. Salomaa, Shuffle on trajectories: Syntactic constraints. *Theoretical Computer Science* **197** (1998) 1–2, 1–56.

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

[8] S. Yu, Regular Languages. In: G. Rozenberg, A. Salomaa (eds.), *Handbook of Formal Languages*. 1, Springer, Berlin Heidelberg, 1997, 41–110.