TUCS

Galina Jirásková | Alexander Okhotin

On the state complexity of star of union and star of intersection

Turku Centre *for* Computer Science

TUCS Technical Report
No 825, August 2007

On the state complexity of star of union and star of intersection

Galina Jirásková
Mathematical Institute, Slovak Academy of Sciences,
Grešákova 6, 040 01 Košice, Slovakia
jiraskov@saske.sk
Alexander Okhotin
Academy of Finland, *and*
Department of Mathematics, University of Turku, *and*
Turku Centre for Computer Science
Turku FIN–20014, Finland
alexander.okhotin@utu.fi

TUCS Technical Report
No 825, August 2007

Abstract

The state complexity of the star of union of an m-state DFA language and an n-state DFA language is proved to be $2^{m+n-1} - 2^{m-1} - 2^{n-1} + 1$ for every alphabet of at least two letters. The state complexity of the star of intersection is established as $\frac{3}{4} \cdot 2^{mn}$ for every alphabet of six or more letters. This improves the recent results of A. Salomaa, K. Salomaa and Yu ("State complexity of combined operations", *Theoret. Comput. Sci.*, 2007, to appear).

Keywords: descriptional complexity, finite automata, state complexity, combined operations

TUCS Laboratory
Discrete Mathematics for Information Technology

1 Introduction

One of the main subjects in the descriptional complexity of regular languages is the state complexity of operations on deterministic finite automata (DFAs). The state complexity of basic operations, such as union, intersection, concatenation and star, is known from Maslov [7], who used DFAs with a partially defined transition function. Similar results for complete DFAs were given by Yu, Zhuang and K. Salomaa [9]. These results were improved by Jirásková [4] by using smaller alphabets for witness languages in the lower bound arguments.

The study of the state complexity of further operations preserving regularity has led to a number of interesting results. Already Maslov [7] has found and investigated several operations with a nontrivial state complexity. Of the recent work, let us mention the state complexity of proportional removals shown to be $O(ne^{\sqrt{n \log n}})$ by Domaratzki [2], the $2^{mn} - 1$ state complexity of shuffle determined by Câmpeanu, K. Salomaa and Yu [1], and the $2^{n^2 + n \log n - O(n)}$ state complexity of the cyclic shift obtained by the authors [5].

A recent direction of research on the state complexity, initiated by A. Salomaa, K. Salomaa and Yu [8], concerns with combinations of basic operations regarded as separate operations. A. Salomaa, K. Salomaa and Yu [8] investigated the state complexity of the star of union of two languages given by DFAs, as well as of the star of intersection of two languages. Their work was followed by Gao, K. Salomaa and Yu [3], who similarly studied the star of concatenation and the star of reversal, and by a paper by Liu, Martín-Vide, A. Salomaa and Yu [6] dealing with the reversal of union and the reversal of concatenation.

This paper aims to refine two results by A. Salomaa, K. Salomaa and Yu [8]. For star of union, the known tight lower bound of $2^{m+n-1} - 2^{m-1} - 2^{n-1} + 1$ states is improved by using witness languages over the smallest possible alphabet $\{a, b\}$, cf. a 3-letter alphabet used by A. Salomaa, K. Salomaa and Yu [8]. For star of intersection we establish the first tight lower bound: the state complexity of this operation is proved to be exactly $\frac{3}{4} \cdot 2^{mn}$, which improves over the asymptotic estimation of $2^{O(mn)}$ due to A. Salomaa, K. Salomaa and Yu [8]. Our lower bound construction uses a 6-letter alphabet.

2 Basic definitions

A *deterministic finite automaton* (DFA) is a quintuple $(Q, \Sigma, \delta, q_0, F)$, in which Q is a finite set of states, Σ is an input alphabet, $\delta : Q \times \Sigma \to Q$ is the transition function, $q_0 \in Q$ is the initial state, and $F \subseteq Q$ is the set of accepting states. We consider only complete DFAs, that is, the transition function is total.

Nondeterministic finite automata (NFA) of the most general kind are defined as quintuples $(Q, \Sigma, \delta, Q_0, F)$ with a set of initial states $Q_0 \subseteq Q$ and with a nondeterministic transition function $\delta : Q \times (\Sigma \cup \{\varepsilon\}) \to 2^Q$. Any NFA can be converted to an equivalent DFA with the set of states 2^Q; this transformation is known as the *subset construction*.

The *state complexity* of a regular language L, denoted $sc(L)$, is the least number of states in any DFA accepting L.

Consider a k-ary operation on languages $f : (2^{\Sigma^*})^k \to 2^{\Sigma^*}$ that preserves regularity in the sense that for all regular $L_1, \ldots, L_k$ the language $f(L_1, \ldots, L_k)$ is regular as well. Define the state complexity function of f as $sc_f : \mathbb{N}^k \to \mathbb{N}$, so that $sc_f(n_1, \ldots, n_k)$ equals the greatest value of $sc(f(L_1, \ldots, L_k))$ over all vectors of languages $(L_1, \ldots, L_k)$ with $sc(L_i) = n_i$ for all i.

3 Star of union

Star of union is a binary operation on languages defined as $(K \cup L)^*$, where $K, L \subseteq \Sigma^*$ are its arguments. The state complexity of this operation over a k-letter alphabet is $f_k(m,n) = \max_{sc(K)=m, sc(L)=n} sc\big((K \cup L)^*\big)$.

A straightforward upper bound for this function is 2^{m+n+1}, which is given by the subset construction applied an $(m+n+1)$-state NFA for $(K \cup L)^*$. A. Salomaa, K. Salomaa and Yu [8] did a further analysis of this subset construction, showing that at most $2^{m+n-1} - 2^{m-1} - 2^{n-1} + 1$ states are reachable. At the same time, they established a matching lower bound over the alphabet $\{a, b, c\}$ by proving that $f_3(m,n) \geqslant 2^{m+n-1} - 2^{m-1} - 2^{n-1} + 1$ for all $m, n \geqslant 3$. For $m = 2$ or $n = 2$, A. Salomaa, K. Salomaa and Yu [8] have established the same precise lower bound using a different set of witness automata over the alphabet $\{a, b, c, d\}$. This settled the state complexity of this operation for alphabet of 4 letters and more (3 in the most interesting cases), while the state complexity over $\{a, b\}$ remained open.

The following stronger theorem, which uses witness automata of a very simple form, fills this gap.

Theorem 1. *For all integers $m \geqslant 2$ and $n \geqslant 2$, there exist binary DFAs A and B of m and n states, respectively, such that the state complexity of the language $(L(A) \cup L(B))^*$ is $2^{m+n-1} - 2^{m-1} - 2^{n-1} + 1$.*

Proof. Fix $m \geqslant 2$ and $n \geqslant 2$ and let $\Sigma = \{a, b\}$.

Define an m-state DFA $A = (Q, \Sigma, q_0, \delta_A, \{q_0\})$, where $Q = \{q_0, \ldots, q_{m-1}\}$ and for each i in $\{0, 1, \ldots, m-1\}$,

$$\delta_A(q_i, a) = \begin{cases} q_{i+1}, & \text{if } i < m-1, \\ q_0, & \text{if } i = m-1, \end{cases}$$

$$\delta_A(q_i, b) = \begin{cases} q_{i+1}, & \text{if } i < m-1, \\ q_1, & \text{if } i = m-1. \end{cases}$$

Define an n-state DFA $B = (R, \Sigma, r_0, \delta_A, \{r_0\})$, where $R = \{r_0, \ldots, r_{n-1}\}$ and for each j in $\{0, 1, \ldots, n-1\}$,

$$\delta_B(r_j, a) = \begin{cases} r_1, & \text{if } j = 0, \\ r_j, & \text{if } j > 0, \end{cases}$$

$$\delta_B(r_j, b) = \begin{cases} r_{j+1}, & \text{if } j < n-1, \\ r_0, & \text{if } j = n-1. \end{cases}$$

Automata A and B are shown in Figure 1.

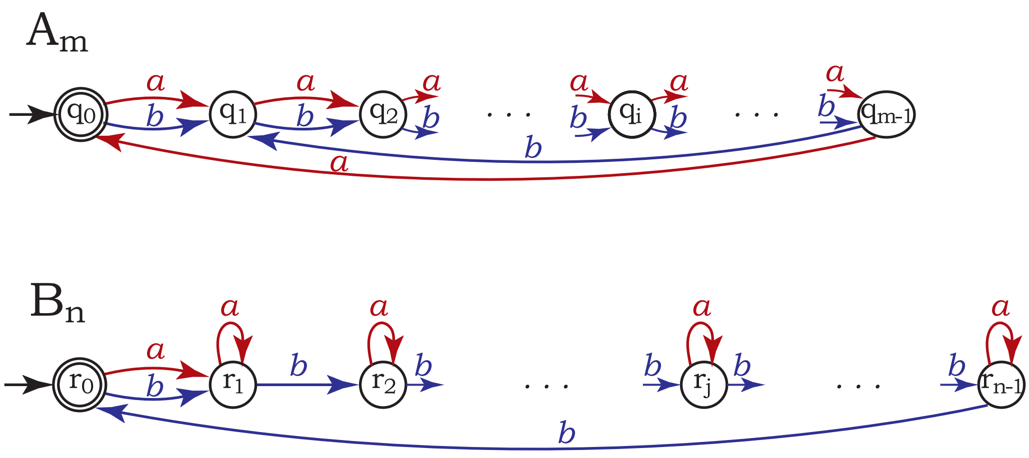

Figure 1: Witness DFAs A_m and B_n for the star of union.

Construct an NFA $C = (Q \cup R, \Sigma, \delta_C, \{q_0, r_0\}, \{q_0, r_0\})$ from the DFAs A and B by adding a transition on a from state q_{m-1} to state r_0 and a transition on b from state r_{n-1} to state q_0, as shown in Figure 2. The NFA C recognizes the language $(L(A) \cup L(B))^*$.

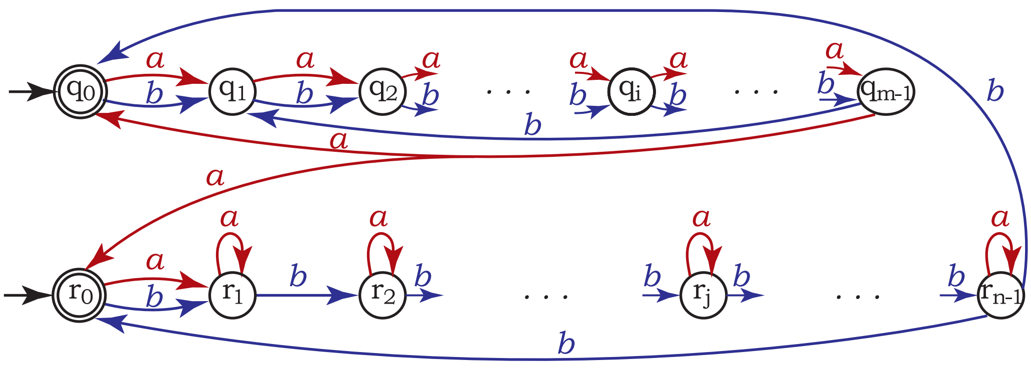

Figure 2: The nondeterministic finite automaton C.

Let $C' = (2^{Q \cup R}, \Sigma, \delta, \{q_0, r_0\}, F)$ be the DFA obtained from the NFA C by the subset construction. We are going to prove that the DFA C' has $2^{m+n-1} - 2^{m-1} - 2^{n-1} + 1$ reachable states that are pairwise inequivalent.

Let $\mathcal{R}$ be the following system of sets:

$$\mathcal{R} = \{S \cup T \mid S \subseteq \{q_1, \dots, q_{m-1}\}, T \subseteq \{r_1, \dots, r_{n-1}\}, S \neq \varnothing, T \neq \varnothing\} \cup$$

$$\cup \{\{q_0, r_0\} \cup R \mid R \subseteq \{q_1, \dots, q_{m-1}\} \cup \{r_1, \dots, r_{n-1}\}\}.$$

Notice that each set in $\mathcal{R}$ either contains both q_0 and r_0 or neither of them. The system $\mathcal{R}$ has

$$(2^{m-1} - 1)(2^{n-1} - 1) + 2^{m+n-2} = 2^{m+n-1} - 2^{m-1} - 2^{n-1} + 1$$

sets, and we will show that each of them is a reachable state of the DFA C'. The proof is by induction on the size of sets.

Basis: Let us prove that each set in $\mathcal{R}$ of size two is reachable. It is claimed that

$$\{q_i, r_j\} = \delta\big(\{q_0, r_0\}, (ba^{m-2})^{j-1} ba^{i-1}\big) \tag{1}$$

for all $i = 1, 2, \dots, m-1$ and $j = 1, 2, \dots, n-1$. First consider that both q_0 and q_{m-1} go to q_{m-1} by ba^{m-2}, and, more generally, both q_0 and q_{m-1} go to q_t by ba^{t-1} for any $t \in \{1, \dots, m-1\}$. This means that $\delta\big(\{q_0\}, (ba^{m-2})^{j-1} ba^{i-1}\big) = \{q_i\}$. As for the second component, each state r_k, with $0 \leqslant k < n-1$, goes to r_{k+1} by any string in ba^*, and therefore $\delta\big(\{r_0\}, (ba^{m-2})^{j-1} ba^{i-1}\big) = \{r_j\}$, which completes the proof of reachability of two-element subsets (1).

Induction step. Now, assume that $2 \leqslant t \leqslant m+n-1$ and that each set in the system $\mathcal{R}$ of size t is reachable. Let

$$\{q_{i_1}, q_{i_2}, \dots, q_{i_k}\} \cup \{r_{j_1}, r_{j_2}, \dots, r_{j_\ell}\},$$

where $0 \leqslant i_1 < i_2 < \dots < i_k \leqslant m-1$ and $0 \leqslant j_1 < j_2 < \dots < j_\ell \leqslant n-1$ be a set in $\mathcal{R}$ of size $t+1$, i.e., $k + \ell = t+1$. We will consider four cases:

(i) Let $i_1 = j_1 = 0$, and $i_2 > 1$ (or $k = 1$). Then

$$\begin{aligned}\{q_0, q_{i_2}, \dots, q_{i_k}\} \cup \{r_0, r_{j_2}, \dots, r_{j_\ell}\} = \\ = \delta(\{q_{m-1}, q_{i_2-1}, \dots, q_{i_k-1}\} \cup \{r_{j_2}, \dots, r_{j_\ell}\}, a),\end{aligned}$$

where the latter set of size t is in $\mathcal{R}$ (as it contains neither q_0 nor r_0), and hence is reachable by induction.

Figure 3(i) illustrates this transition from a set of size t (represented by the upper diagram) to the set of size $t+1$ (lower diagram) by symbol a. Grey circles refer to the states $q_{i_2}, \dots, q_{i_k}$ and $r_{j_2}, \dots, r_{j_\ell}$ and to the corresponding states in the upper diagram. Crossed out squares refer to states known not to be in the corresponding sets.

The subsequent cases are similarly illustrated in the rest of Figure 3.

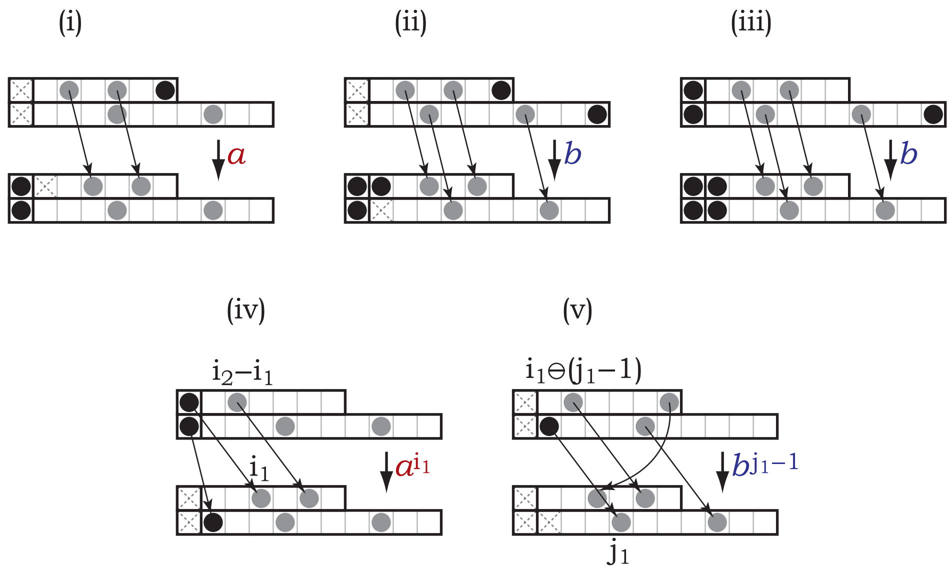

Figure 3: Five cases in the proof of reachability.

(ii) Let $i_1 = j_1 = 0$, $i_2 = 1$, and $j_2 > 1$ (or $\ell = 1$). Then

$$\{q_0, q_1, q_{i_3}, \dots, q_{i_k}\} \cup \{r_0, r_{j_2}, \dots, r_{j_\ell}\} = \\ = \delta\big(\{q_{m-1}, q_{i_3-1}, \dots, q_{i_k-1}\} \cup \{r_{j_2-1}, \dots, r_{j_\ell-1}, r_{n-1}\}, b\big).$$

The latter set of size t is in $\mathcal{R}$, and by the induction hypothesis it is reachable.

(iii) Let $i_1 = j_1 = 0$, $i_2 = 1$, and $j_2 = 1$. Then

$$\{q_0, q_1, q_{i_3}, \dots, q_{i_k}\} \cup \{r_0, r_1, r_{j_3}, \dots, r_{j_\ell}\} = \\ = \delta\big(\{q_0, q_{i_3-1}, \dots, q_{i_k-1}\} \cup \{r_0, r_{j_3-1}, \dots, r_{j_\ell-1}, r_{n-1}\}, b\big),$$

where the latter set of size t containing both q_0 and r_0 is reachable by induction.

(iv) Let $i_1 \geqslant 1$ and $j_1 = 1$. Then

$$\{q_{i_1}, q_{i_2}, \dots, q_{i_k}\} \cup \{r_1, r_{j_2}, \dots, r_{j_\ell}\} = \\ = \delta\big(\{q_0, q_{i_2-i_1}, \dots, q_{i_k-i_1}\} \cup \{r_0, r_{j_2}, \dots, r_{j_\ell}\}, a^{i_1}\big),$$

where the latter set of size $t+1$ containing both q_0 and r_0 is reachable as shown in cases *(i)-(iii)*.

(v) Let $i_1 \geqslant 1$ and $j_1 > 1$. Our construction uses subtraction modulo $m-1$, and it is convenient to assume that subtraction of identical numbers equals $m-1$. Denote this modified operation by $\ominus$, and let $i \ominus j$ be the unique number in $\{1, \dots, m-1\}$ equal to $i - j$ modulo $m-1$.

Then

$$\{q_{i_1}, q_{i_2}, \dots, q_{i_k}\} \cup \{r_{j_1}, r_{j_2}, \dots, r_{j_\ell}\} = \\ = \delta\big(\{q_{i_1 \ominus (j_1-1)}, \dots, q_{i_k \ominus (j_1-1)}\} \cup \{r_1, r_{j_2-j_1+1}, \dots, r_{j_\ell-j_1+1}\}, b^{j_1-1}\big),$$

where the latter set is considered in case *(iv)*.

This completes the proof of reachability of states in $\mathcal{R}$. It remains to prove that every two states are inequivalent. Let us show that for every state s of the NFA C, there is a string $w(s)$ that is accepted by C starting in this state s, but is not accepted by C starting in any other state. Then, the inequivalence of states of the DFA C' follows immediately since two distinct subsets of $Q \cup R$ must differ in some state s of the NFA C and so the string $w(s)$ distinguishes them.

Let s be a state of the NFA C. Let $w(s) = a^{m-i}$ if $s = q_i$ for some i in $\{0, 1, \dots, m-1\}$, and let $w(s) = b^{n-j}$ if $s = r_j$ for some j in $\{0, 1, \dots, n-1\}$. Then for every state s, the string $w(s)$ is accepted by the NFA C starting in state s since states q_i and p_j go to the accepting states q_0, r_0 by strings a^{m-i} and b^{n-j}, respectively.

On the other hand, by the string a^{m-i}, each state in R goes to a state in $\{r_1, r_2, \dots, r_{n-1}\}$, and if $k \neq i$, then state q_k goes to state $q_{(k+m-i) \bmod m}$ and also to r_1 if $k > i$. Next, by the string b^{n-j}, each state in Q goes to a state in $\{q_1, q_2, \dots, q_{m-1}\}$, and if $\ell \neq j$, then state p_ℓ goes to state $p_{k+n-j \pmod n}$ and also to a state in $\{q_1, q_2, \dots, q_{m-1}\}$ if $\ell > j$. This means that the string a^{m-1} is accepted by the NFA C only from state q_i and the string b^{n-j} is accepted only from state p_j.

Thus it was shown that all states in $\mathcal{R}$ are reachable and pairwise inequivalent, and hence the state complexity of $(L(A) \cup L(B))^*$ is $|\mathcal{R}|$, which establishes the theorem. □

The above theorem applies to $m, n \geqslant 2$, and it remains to consider the case of $m = 1$. There exist only two 1-state DFAs, one recognizing $\varnothing$ and the other recognizing Σ^*. In the former case star of union degenerates to star of the second argument, an operation of state complexity $\frac{3}{4} \cdot 2^n$ for $n \geqslant 2$, as shown by Maslov [7] and by Yu, Zhuang and K. Salomaa [9] using binary witness languages. Star of union with Σ^* always equals Σ^* and thus has state complexity of 1. In the case of $m = n = 1$ the state complexity is 2, reached by $(\varnothing \cup \varnothing)^* = \{\varepsilon\}$.

Thus we have established the following result:

Corollary 1. *For every alphabet Σ, such that $|\Sigma| \geqslant 2$, the state complexity* *of the star of union over Σ is:*

$$f(m,n) = \begin{cases} 2^{m+n-1} - 2^{m-1} - 2^{n-1} + 1, & \text{if} \quad m,n \geqslant 2, \\ \frac{3}{4} \cdot 2^m, & \text{if} \quad m \geqslant 2, n = 1, \\ \frac{3}{4} \cdot 2^n, & \text{if} \quad m = 1, n \geqslant 2, \\ 2, & \text{if} \quad m = n = 1. \end{cases}$$

4 Star of intersection

The state complexity of the star of intersection can be upper-estimated by combining the known state complexities of intersection and star, which gives an upper bound of $\frac{3}{4} \cdot 2^{mn}$.

Having stated this upper bound, A. Salomaa, K. Salomaa and Yu [8] have established two relatively close lower bounds: $2^{m(n-2)}$ over a 5-letter alphabet and $2^{m(n-2)} + 2^{n(m-2)} - 2^{mn-2(m+n+1)}$ over an 8-letter alphabet. This gives an asymptotic expression of $2^{mn-O(m+n)}$ for the state complexity function, but leaves the exact state complexity open.

We determine the state complexity of star of intersection precisely by showing that the straightforward upper bound $\frac{3}{4} \cdot 2^{mn}$ is in fact tight, using witness languages over a 6-letter alphabet.

Theorem 2. *For all integers $m \geqslant 2$ and $n \geqslant 2$, there exist DFAs A and B of m and n states, respectively, defined over a six-letter input alphabet and such that the state complexity of the language $(L(A) \cap L(B))^*$ is $\frac{3}{4} \cdot 2^{mn}$.*

Proof. Let $\Sigma = \{a, b, c, d, e, f\}$ and fix arbitrary $m, n \geqslant 2$.

Define an m-state DFA $A = A_m = (Q_m, \Sigma, \delta_A, 0, \{m-1\})$, where $Q_m = \{0, 1, \ldots, m-1\}$, and an n-state DFA $B = B_n = (Q_n, \Sigma, \delta_B, 0, \{n-1\})$ with $Q_n = \{0, 1, \ldots, n-1\}$, where the transition functions δ_A and δ_B are defined as follows: for each $i \in Q_m$ and $j \in Q_n$,

$\delta_A(i,a) = (i+1) \mod m,$ $\qquad$ $\delta_B(j,a) = (j+1) \mod n,$

$\delta_A(i,b) = i,$ $\qquad$ $\delta_B(j,b) = (j+1) \mod n,$

$$\delta_A(i,c) = \begin{cases} 0, & \text{if} \quad i = 0, \\ i+1, & \text{if} \quad 1 \leqslant i < m-1, \\ 1, & \text{if} \quad i = m-1, \end{cases} \qquad \delta_B(j,c) = j,$$

$$\delta_A(i,d) = i, \qquad \delta_B(j,d) = \begin{cases} 0, & \text{if} \quad j = 0, \\ j+1, & \text{if} \quad 1 \leqslant j < n-1, \\ 1, & \text{if} \quad j = n-1, \end{cases}$$

$$\delta_A(i,e) = \begin{cases} 0, & \text{if} \quad i = 1, \\ i, & \text{otherwise,} \end{cases} \qquad \delta_B(j,e) = j,$$

$$\delta_A(i,f) = i, \qquad \delta_B(j,f) = \begin{cases} 0, & \text{if} \quad j = 1, \\ j, & \text{otherwise.} \end{cases}$$

The following table gives a succinct semiformal explanation of these transitions. A graphic representation of these DFAs is given in Figure 4.

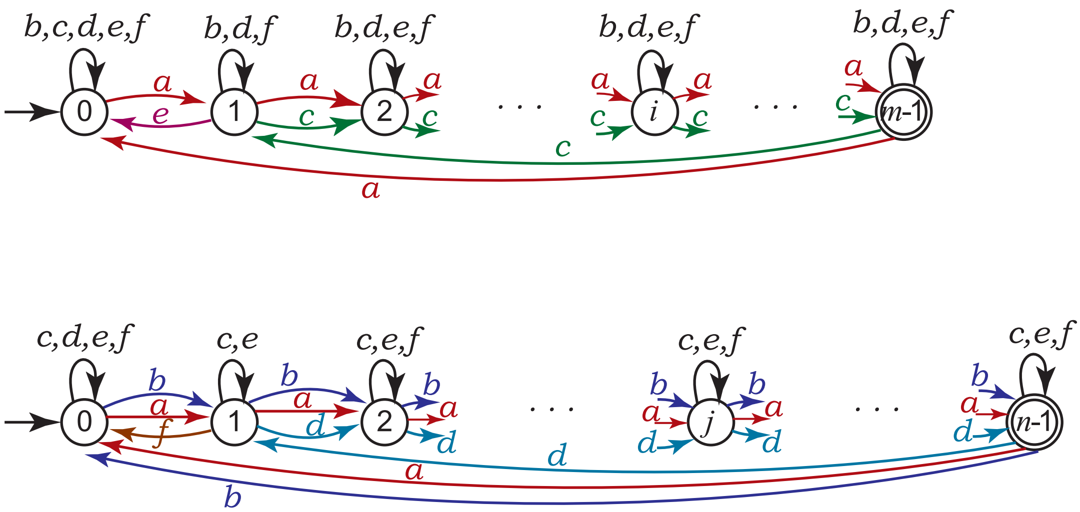

Figure 4: Witness DFAs A_m and B_n for the star of intersection.

	A_m	B_n
a	circle$(0..m-1)$	circle$(0..n-1)$
b	i to i	circle$(0..n-1)$
c	0 to 0, circle$(1..m-1)$	j to j
d	i to i	0 to 0, circle$(1..n-1)$
e	1 to 0, $i \neq 1$ to i	j to j
f	i to i	1 to 0, $j \neq 1$ to j

First construct the standard mn-state DFA for $L(A_m) \cap L(B_n)$, which is obtained as a direct product of A_m and B_n:

$$C = (Q_m \times Q_n, \Sigma, \delta_C, (0,0), \{(m-1, n-1)\}),$$

where $\delta_C((i,j), X) = (\delta_A(i, X), \delta_B(j, X))$ for each (i,j) in $Q_m \times Q_n$ and each X in Σ. The transitions of C by each symbol are given in Figure 5.

Next, construct an $(mn+1)$-state NFA D from the DFA C by adding a new initial and accepting state q_0 that goes to $(0,0)$ by ε, and add another epsilon transition from $(m-1, n-1)$ to $(0,0)$. The NFA D accepts the language $(L(A_m) \cap L(B_n))^*$.

Now, let D' be the DFA obtained from the NFA D by the subset construction. We will show that the DFA D' has $2^{mn-1} + 2^{mn-2}$ reachable states that are pairwise inequivalent.

Let $\mathcal{R}$ be the following system of sets:

$$\mathcal{R} = \{q_0\} \cup \big\{ S \subseteq Q_m \times Q_n \mid S \neq \varnothing \text{ and if } (m-1, n-1) \in S \text{ then } (0,0) \in S \big\}.$$

The system $\mathcal{R}$ consists of 2^{mn-2} sets containing both $(0,0)$ and $(m-1, n-1)$, $2^{mn-1} - 1$ nonempty sets containing neither $(0,0)$ nor $(m-1, n-1)$, and the state q_0, that is, there are $2^{mn-1} + 2^{mn-2}$ sets in $\mathcal{R}$. We will show that each

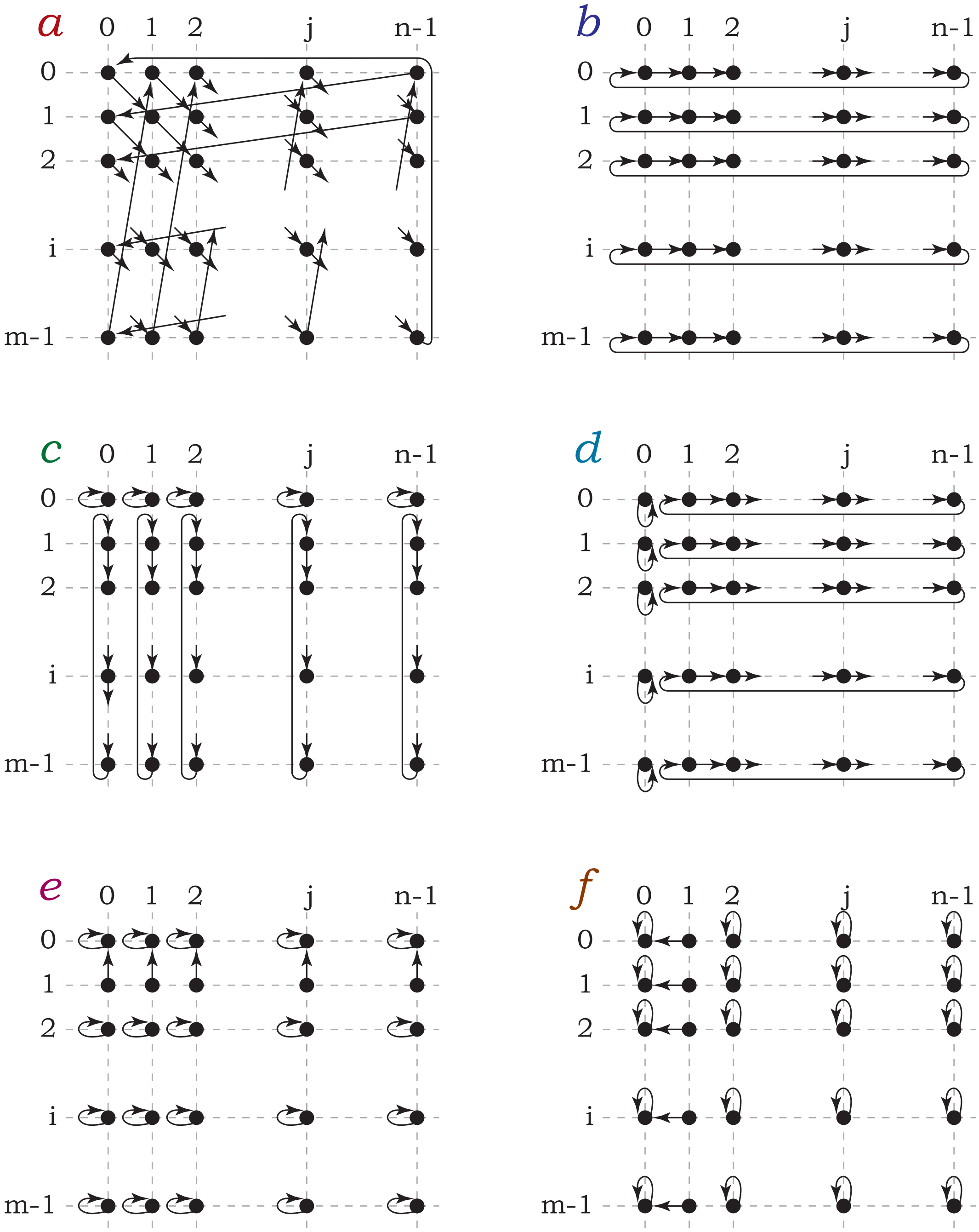

Figure 5: Transitions of DFA C (direct product of A_m and B_n).

of them is a reachable state of the DFA D'. The proof is by induction on the size of sets.

Basis. The subset $\{q_0\}$ is trivially reachable. Each one-element set $\big\{(i,j)\big\}$ in $\mathcal{R}$ is reachable, since q_0 goes to $(0,0)$ by ε and $(0,0)$ goes to (i,j) by $b^{(j-i) \bmod n}a^i$ for all (i,j) in $Q_m \times Q_n \setminus \big\{(m-1,n-1)\big\}$.

Induction step. Let $2 \leqslant k \leqslant mn$ and assume that all sets in $\mathcal{R}$ of size $k-1$ are reachable. Let S be a set in $\mathcal{R}$ of size k. Consider the following three possible cases:

(a) S contains both states $(0,0)$ and $(m-1,n-1)$;

(b) S contains state $(0,0)$, but not state $(m-1,n-1)$;

(c) S contains neither state $(0,0)$ nor state $(m-1,n-1)$.

In each case, we show that the set S is reachable.

(a) Let $S = \{(0,0),(m-1,n-1),(i_3,j_3),\dots,(i_k,j_k)\}$ be a set of size k containing both states $(0,0)$ and $(m-1,n-1)$. Let

$$S' = \big\{(m-2,n-2),(i_3-1,j_3-1),\dots,(i_k-1,j_k-1)\big\},$$

where subtraction is modulo m in first components and modulo n in second components of all pairs. The set S' does not contain state $(m-1,n-1)$ and is therefore in $\mathcal{R}$. Then, since S' is of size $k-1$, it is reachable by the induction hypothesis. And since S' goes to S by a, the set S is also reachable.

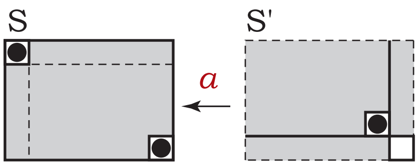

Figure 6: Proof of Theorem 2, reachability, case (a).

(b) Let $S = \{(0,0),(i_2,j_2),\dots,(i_k,j_k)\}$ be a set of size k that contains state $(0,0)$ but not state $(m-1,n-1)$. Let us write S as

$$S = \big\{(0,0)\big\} \cup U \cup V \cup W,$$

where

$$\begin{aligned} U &= \big\{(0,j_2),(0,j_3),\dots,(0,j_{r-1})\big\} \quad (1 \leqslant j_2 < j_3 < \cdots < j_{r-1} \leqslant n-1) \\ V &= \big\{(i_r,0),(i_{r+1},0),\dots,(i_{s-1},0)\big\} \quad (1 \leqslant i_r < \cdots < i_{s-1} \leqslant m-1) \\ W &= \big\{(i_s,j_s),\dots,(i_k,j_k)\big\} \quad (1 \leqslant i_\ell \leqslant m-1;\ 1 \leqslant j_\ell \leqslant n-1), \end{aligned}$$

that is, the set U contains the states in S from the top row, the set V contains states from the leftmost column, and the set W contains the other states of S.

There are four subcases to consider:

(i) W is nonempty. Let (i_s, j_s) be a state in S that is in W.

Define subsets U', V', W' as follows, using the notation $\ominus$ for subtraction modulo $\ell - 1$ with the result defined to be in $\{1, \ldots, \ell - 1\}$ (as in the proof of Theorem 1):

$$\begin{aligned} U' &= \big\{(0, j_2 \ominus j_s), (0, j_3 \ominus j_s), \ldots, (0, j_{r-1} \ominus j_s)\big\}, \\ V' &= \big\{(i_r \ominus i_s, 0), (i_{r+1} \ominus i_s, 0), \ldots, (i_{s-1} \ominus i_s, 0)\big\}, \\ W' &= \big\{(i_{s+1} \ominus i_s, j_{s+1} \ominus j_s), \ldots, (i_k \ominus i_s, j_k \ominus j_s)\big\}. \end{aligned}$$

In each pair, the subtraction in its first component is modulo $m - 1$, while the subtraction in second components is modulo $n - 1$. Let

$$S' = \big\{(0, 0), (m - 1, n - 1)\big\} \cup U' \cup V' \cup W'.$$

Then, S' is a set of size k that contains $(0, 0)$ and $(m - 1, n - 1)$, and such sets have been shown to be reachable in case (a).

Consider the state $\delta(S', c^{i_s} d^{j_s})$, which equals

$$\delta\big(\{(0, 0), (m-1, n-1)\}, c^{i_s} d^{j_s}\big) \cup \delta(U', c^{i_s} d^{j_s}) \cup \delta(V', c^{i_s} d^{j_s}) \cup \delta(W', c^{i_s} d^{j_s}).$$

Let us compute each of the four parts of this expression:

- From state $\big\{(0, 0), (m - 1, n - 1)\big\}$, the automaton goes to state $\big\{(0, 0), (i_s, n - 1)\big\}$ upon reading c^{i_s}, and then to $\big\{(0, 0), (i_s, j_s)\big\}$ by d^{j_s}.
- Having started in $U' = \{(0, j_t \ominus j_s) \mid 2 \leqslant t < r\}$, the automaton remains in U' upon reading c^{i_s} and then proceeds by d^{j_s} either to $\{(0, j_t) \mid 2 \leqslant t < r\} = U$ or to $U \cup \{(0, 0)\}$.
- Similarly, from V' the automaton goes by $c^{i_s} d^{j_s}$ either to V or $V \cup \{(0, 0)\}$.
- Finally, the automaton goes from $W' = \{(i_t \ominus i_s, j_t \ominus j_s) \mid s < t \leqslant k\}$ first to $\{(i_t, j_t \ominus j_s) \mid s < t \leqslant k\}$ or to $\{(i_t, j_t \ominus j_s) \mid s < t \leqslant k\} \cup \{(0, 0)\}$ by c^{i_s}, and then either to $\{(i_t, j_t) \mid s < t \leqslant k\} = W$ or to $W \cup \{(0, 0)\}$ by d^{j_s}.

The union of the above sets is $\big\{0, 0\big\} \cup U \cup V \cup W = S$. We have thus shown that the set S' goes to S by $c^{i_s} d^{j_s}$, and so the set S is reachable.

(ii) W and V are empty, that is, $S = \big\{(0, 0), (0, j_2), \ldots, (0, j_k)\big\}$. Let

$$S' = \big\{(0, 0), (1, j_2), \ldots, (1, j_k)\big\}.$$

The set S' is reachable as in case (b-i) and it goes to S by e. Thus, the set S is reachable.

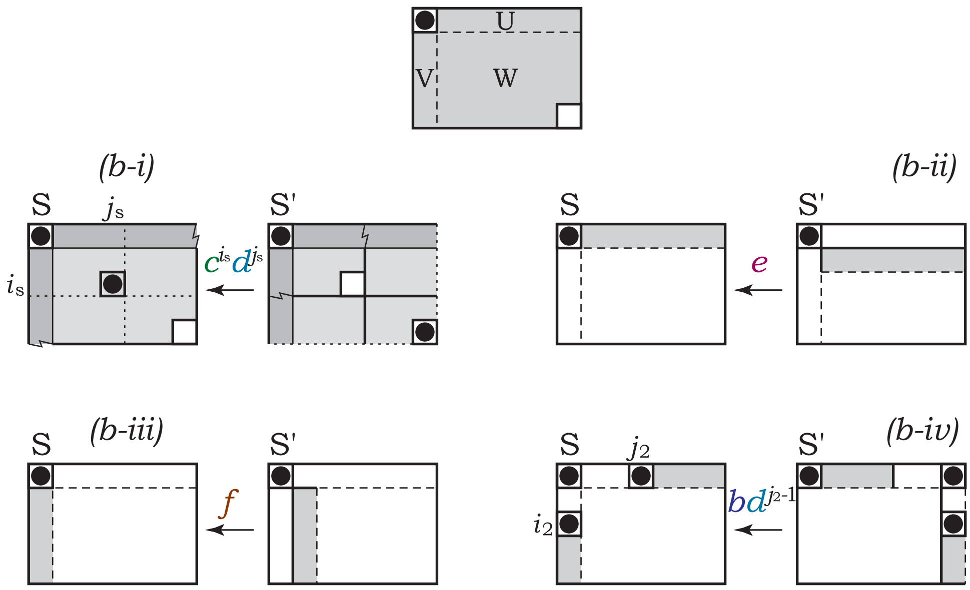

Figure 7: Proof of Theorem 2, reachability, case (b).

(iii) W and U are empty, that is, $S = \{(0,0),(i_2,0),\ldots,(i_k,0)\}$. The set S can be reached by f from the set $\{(0,0),(i_2,1),\ldots,(i_k,1)\}$, which is reachable as in case (b-i).

(iv) W is empty and U and V are not, that is,

$$S = \{(0,0)\} \cup \{(0,j_2),\ldots(0,j_{r-1})\} \cup \{(i_r,0),\ldots,(i_k,0)\},$$

where $1 \leqslant j_2 < \cdots < j_{r-1} \leqslant n-1$ and $1 \leqslant i_r < \cdots < i_k \leqslant m-1$. Let

$$S' = \{(0,n-1),(0,0),(0,j_3-j_2),\ldots,(0,j_{r-1}-j_2),(i_r,n-1),\ldots,(i_k,n-1)\},$$

where the subtraction is modulo n. The set S' is reachable as in case (b-i), and it goes to S by bd^{j_2-1}. This proves the reachability of S and concludes case (b).

(c) Let $S = \{(i_1,j_1),(i_2,j_2),\ldots,(i_k,j_k)\}$, where the pairs are sorted lexicographically as $(0,0) < (i_1,j_1) < (i_2,j_2) < \cdots < (i_k,j_k) < (m-1,n-1)$, be a set of size k that contains neither $(0,0)$ nor $(m-1,n-1)$. Consider three subcases:

(1) Let $i_1 \geqslant 1$, that is, the set S contains no states from the top row. Take

$$S' = \{(i-i_1, j-j_1) \mid (i,j) \in S\},$$

where subtraction in the second component is modulo n, while subtraction in the first component always produces a nonnegative number. The set S' is a subset of size k that contains $(0,0)$, and so it is reachable as in case (b). Consider the computation of D' by the string $b^{j_1-i_1 \bmod n}a^{i_1}$ starting from S'. Since $i_k - i_1 \leqslant m-2$, the set S' contains no states from the bottom row, and therefore

$$\begin{aligned}\delta(S', b^{j_1-i_1 \bmod n}) = \big\{(i-i_1, j-j_1+j_1-i_1) \mid (i,j) \in S\big\} = \\ = \big\{(i-i_1, j-i_1) \mid (i,j) \in S\big\},\end{aligned}$$

where subtraction in the second component is again modulo n.

Denote the latter set by S'', and let us see that it goes to S by a^{i_1}. In the case of $i_1 = 1$, since the set S does not contain state $(m-1, n-1)$, S'' does not contain state $(m-2, n-2)$, and so it goes to S by a. If $i_1 \geqslant 2$, then we have $i_k - i_1 < i_k - i_1 + 1 < \cdots < i_k - 2 < m-2$, which implies that S'' goes to $\big\{(i-1, j-1) \mid (i,j) \in S\big\}$ by a^{i_1-1}, since none of the intermediate subsets contains state $(m-2, n-2)$. From $\big\{(i-1, j-1) \mid (i,j) \in S\big\}$ it proceeds to S by a.

Thus we have shown that the set S can be reached from the set S' by the string $b^{j_1-i_1 \bmod n}a^{i_1}$.

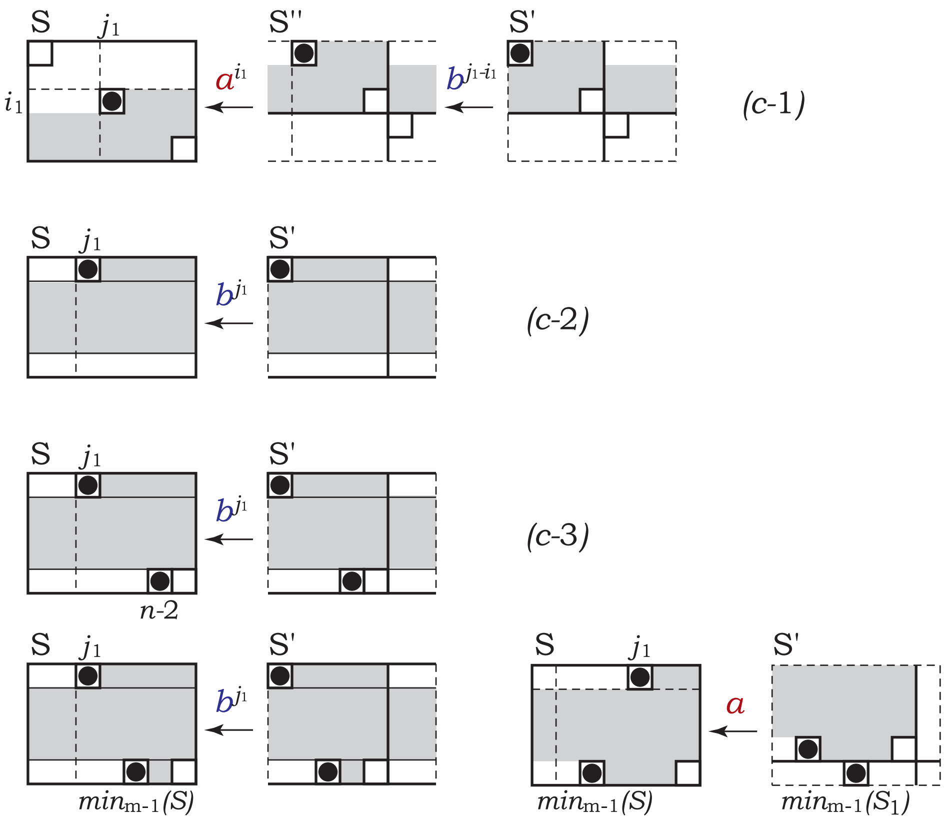

Figure 8: Proof of Theorem 2, reachability, case (c).

(2) Let $i_1 = 0$ (and so $j_1 > 0$) and $i_k < m-1$, that is, the set S contains

some states from the top row but no states from the bottom row. Take

$$S' = \{(i, j - j_1) \mid (i, j) \in S\}$$

where subtraction is modulo n. The set S' is a subset of size k that contains state $(0, 0)$, and so is reachable as in case (b). Since the bottom row of S' is empty, the set S' goes to S by b^{j_1}, which proves case (c-2).

(3) Let $i_1 = 0$ and $i_k = m - 1$, that is, the set S contains some states from the top row and some states from the bottom row (but does not contain $(0, 0)$ and $(m - 1, n - 1)$). Consider the leftmost state in the bottom row and denote its second component by

$$\min_{m-1}(S) = \min\{j \mid (m - 1, j) \in S\}.$$

The reachability of such states is proved by an induction on $\min_{m-1}(S)$.

Basis: $\min_{m-1}(S) = n - 2$, which is the greatest possible value of $\min_{m-1}(S)$, because $(m - 1, n - 1) \notin S$. Then $S' = \{(i, j - j_1) \mid (i, j) \in S\}$ is a subset of size k that contains state $(0, 0)$, and so it is reachable as in case (b). The bottom row of S' contains one state, $(m - 1, n - 2 - j_1)$. This means that the set S' goes to S by b^{j_1}, and thus S is reachable.

Induction step. Let $\min_{m-1}(S) < n - 2$ and assume that every subset S_0 of size k with nonempty top and bottom rows and with $\min_{m-1}(S_0) > \min_{m-1}(S)$ is reachable. Let us prove that the set S is also reachable.

If $j_1 \leqslant \min_{m-1}(S) + 1$, consider the set $S' = \{(i, j - j_1) \mid (i, j) \in S\}$ of size k, which contains $(0, 0)$, and hence is reachable as in case (b). Since $(m - 1, j') \notin S$ for all $j' \leqslant j_1 - 2$, it follows that the bottom row of S' does not contain states $(m - 1, n + j' - j_1)$ for all such j'. That is, $(m - 1, j'') \notin S'$ for all $n - j_1 \leqslant j'' \leqslant n - 2$, while state $(m - 1, n - 1 - j_1)$ is not in S' because $(m - 1, n - 1) \notin S$. The absence of these states is sufficient to ensure that S' goes to S by b^{j_1}.

Assume $j_1 > \min_{m-1}(S) + 1$ and consider the subset

$$S' = \{(i - 1, j - 1) \mid (i, j) \in S\},$$

where subtraction in the first (second) component is modulo m (modulo n, respectively). This set of size k does not contain states $(m - 2, n - 2)$ and $(m - 1, n - 1)$, since S does not contain states $(0, 0)$ and $(m - 1, n - 1)$. Therefore, the set S_1 goes to S by a. If the set S' contains state $(0, 0)$, then it is reachable as in case (b) and we are done. If S' contains no state from the top row, then, as in case (c-1), it can be reached from a subset of size k containing state $(0, 0)$. Otherwise, the set S' contains some states from the top row (but not $(0,0)$), and since the bottom row in S' is obtained from the top row in S, we have

$$\min_{m-1}(S') = j_1 - 1 > \min_{m-1}(S).$$

Then the induction hypothesis is applicable to S', and hence it is reachable. Since S' goes to S by a, the state S is reachable as well.

This concludes the proof of reachability of all subsets in $\mathcal{R}$.

It remains to demonstrate that all subsets in $\mathcal{R}$ are pairwise inequivalent. The initial state q_0 is an accepting state and cannot be equivalent to any state of the DFA D' that does not contain $(m-1, n-1)$. However, the string b^n is accepted by the NFA D from $(m-1, n-1)$ and is not accepted from q_0. To prove that no two different subsets of $Q_m \times Q_n$ are equivalent it is sufficient to show that for all (i,j) in $Q_m \times Q_n$, the string

$$w_{ij} = b^{n-j-m+i \pmod n} a^{m-i-1}$$

is accepted by the NFA D only from state (i,j).

Indeed, the string w_{ij} is accepted by the NFA D from state (i,j), since this state goes to the accepting state $(m-1, n-1)$ by w_{ij}. On the other hand, the length of this string is at most $m+n-2$. So if the computation of the DFA from another state (i', j') on this string passes through $(m-1, n-1)$, then the newly added state (0,0) has no chance to reach $(m-1, n-1)$ and lead to acceptance. And the direct path from (i', j') leads to the state $(i'+m-i-1, j'+n-j-1)$, which is accepting if and only if $i' = i$ and $j' = j$.

Altogether it has been established that the constructed DFA for $(L(A_m) \cap L(B_n))^*$ contains $\frac{3}{4} \cdot 2^{mn}$ reachable and pairwise inequivalent subsets, and therefore every DFA for this language must contain at least as many states. Together with the matching upper bound on the state complexity of star of intersection, this establishes it precisely. □

Corollary 2. *For every alphabet Σ, such that $|\Sigma| \geqslant 6$, the state complexity of the star of intersection over Σ is:*

$$f(m,n) = \begin{cases} \frac{3}{4} \cdot 2^{mn}, & \text{if} \quad m,n \geqslant 1, \ m+n > 2, \\ 2, & \text{if} \quad m = n = 1. \end{cases}$$

5 Calculations

Our lower bound for star of union uses the smallest possible alphabet $\{a, b\}$.

On the other hand, for star of intersection we had to use an alphabet of as many as six letters to establish the precise lower bound $\frac{3}{4} \cdot 2^{mn}$. It is natural to ask how many symbols are actually needed to reach this upper bound.

Exhaustive computations of star of intersection for all small DFAs over the alphabets $\{a,b\}$ and $\{a,b,c\}$ show that the upper bound cannot be reached. The computed values of the state complexity function for small m,n are given in Table 1(left, centre), and in each case they are less than $\frac{3}{4}\cdot 2^{mn}$.

$\{a,b\}$

	2	3	4
2	7		
3	41	304	
4	165		

$\{a,b,c\}$

	2	3	4
2	11		
3	46	375	
4			

$\{a,b,c,d\}$

	2	3	4
2	12		
3	48	384	
4		3072	
5		24576	786432

Table 1: State complexity of star of intersection for small alphabets and for small m,n.

On the other hand, our computations determined some pairs of automata over the alphabet $\{a,b,c,d\}$ that reach the $\frac{3}{4}\cdot 2^{mn}$ upper bound. These reached values are shown in Table 1(right), and following are some witness pairs of automata (quite many such pairs were found):

(

	a	b	c	d
0	0	0	0	1
1	1	0	1	0

$\cap$

	a	b	c	d
0	0	0	1	1
1	0	1	0	0

)* : 12 states

(

	a	b	c	d
0	0	0	0	1
1	0	1	1	0

$\cap$

	a	b	c	d
0	0	0	1	1
1	1	2	0	2
2	2	0	2	0

)* : 48 states

(

	a	b	c	d
0	0	0	0	1
1	0	1	2	0
2	1	2	1	2

$\cap$

	a	b	c	d
0	0	1	1	2
1	1	0	2	0
2	2	0	0	1

)* : 384 states

(

	a	b	c	d
0	0	0	0	1
1	0	1	2	2
2	0	2	3	0
3	1	3	1	3

$\cap$

	a	b	c	d
0	1	0	2	1
1	2	0	0	0
2	0	1	1	2

)* : 3072 states

$$\left(\begin{array}{|c|cccc|} \hline & a & b & c & d \\ \hline 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 2 & 2 \\ 2 & 0 & 2 & 3 & 3 \\ 3 & 0 & 3 & 4 & 0 \\ \mathbf{4} & 1 & 4 & 1 & 4 \\ \hline \end{array} \cap \begin{array}{|c|cccc|} \hline & a & b & c & d \\ \hline 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 2 & 0 \\ \mathbf{2} & 2 & 1 & 1 & 2 \\ \hline \end{array}\right)^* \quad : \quad 24576 \text{ states}$$

$$\left(\begin{array}{|c|cccc|} \hline & a & b & c & d \\ \hline 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 2 & 2 \\ 2 & 0 & 2 & 3 & 3 \\ 3 & 0 & 3 & 4 & 0 \\ \mathbf{4} & 1 & 4 & 1 & 4 \\ \hline \end{array} \cap \begin{array}{|c|cccc|} \hline & a & b & c & d \\ \hline 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 2 & 2 \\ 2 & 2 & 0 & 3 & 0 \\ \mathbf{3} & 3 & 1 & 1 & 3 \\ \hline \end{array}\right)^* \quad : \quad 786432 \text{ states}$$

In all cases 0 is the initial state and the accepting states are given in bold. We could not obtain any general form of such automata.

Acknowledgements

Research of the first author supported by VEGA grant 2/6089/26. Research of the second author supported by the Academy of Finland under grant 118540.

References

[1] 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.

[2] M. Domaratzki, "State complexity and proportional removals", *Journal of Automata, Languages and Combinatorics*, 7 (2002), 455–468.

[3] Y. Gao, K. Salomaa, S. Yu, "State complexity of star of catenation and reversal", *DCFS 2006* (Las Cruces, USA), 153–164.

[4] G. Jirásková, "State complexity of some operations on binary regular languages", *Theoretical Computer Science*, 330 (2005), 287–298.

[5] G. Jirásková, A. Okhotin, "State complexity of cyclic shift", *Proceedings of DCFS 2005* (Como, Italy, June 30–July 2, 2005), 182–193.

[6] G. Liu, C. Martín-Vide, A. Salomaa, S. Yu, "State complexity of basic operations combined with reversal", *LATA 2007* (Tarragona, Spain).

[7] A. N. Maslov, “Estimates of the number of states of finite automata”, *Soviet Mathematics Doklady*, 11 (1970), 1373–1375.

[8] A. Salomaa, K. Salomaa, S. Yu, “State complexity of combined operations”, *Theoretical Computer Science*, 2007, to appear, doi:10.1016/j.tcs.2007.04.015

[9] S. Yu, Q. Zhuang, K. Salomaa, “The state complexity of some basic operations on regular languages”, *Theoretical Computer Science*, 125 (1994), 315–328.

TURKU

CENTRE *for*

COMPUTER

SCIENCE

Lemminkäisenkatu 14 A, 20520 Turku, Finland | www.tucs.fi

University of Turku

- Department of Information Technology
- Department of Mathematical Sciences

Åbo Akademi University

- Department of Computer Science
- Institute for Advanced Management Systems Research

Turku School of Economics and Business Administration

- Institute of Information Systems Sciences

ISBN 978–952–12–1914–6
ISSN 1239-1891