# On a Structural Property in the State Complexity of Projected Regular Languages[☆]

Galina Jirásková[a,1,*], Tomáš Masopust[b,2]

[a] *Mathematical Institute, Slovak Academy of Sciences*
*Grešákova 6, 040 01 Košice, Slovak Republic*
[b] *Institute of Mathematics, Academy of Sciences of the Czech Republic*
*Žižkova 22, 616 62 Brno, Czech Republic*

## Abstract

A transition is unobservable if it is labeled by a symbol removed by a projection. The present paper investigates a new structural property of incomplete deterministic finite automata – a number of states incident with an unobservable transition – and its effect on the state complexity of projected regular languages. We show that the known upper bound can be met only by automata with one unobservable transition (up to unobservable multi-transitions). We improve this upper bound by taking into consideration the structural property of minimal incomplete automata, and prove the tightness of new upper bounds. Special attention is focused on the case of finite languages. The paper also presents and discusses several fundamental problems which are still open.

*Keywords:* Projections, state complexity, descriptional complexity.

## 1. Introduction

A typical automaton model of a real-world system usually consists of a huge number of states. Therefore, the simplification of the system plays an important role in many fields of computer science and engineering, such as compositional verification, fault diagnoses, or supervisory control [2, 4, 21, 22, 23, 24, 33, 37]. Projections, also called natural projections because they can be seen as natural transformations of category theory, are one of the forms of abstraction methods that are used for such a simplification. Given a regular language $L$ and a projection $P$, it is well-known that the minimal deterministic finite automaton (dfa) accepting the language $P(L)$ can be of exponential size in comparison with the dfa accepting the language $L$. However, from the practical point of view, only those projections which ensure that the automaton for the projected language is significantly smaller than the automaton of the original language are of interest. In this paper, we summarize the known results on this topic, improve the known upper bounds of the projected regular languages, and formulate several open problems.

Wong in [36] proved that the upper bound on the state complexity of projections of regular languages is $3 \cdot 2^{n-2} - 1$. However, Wong did not consider the structure of automata in his result. This is of interest because, as we show in this paper, this upper bound can be met only by automata with one *unobservable transition*, that is, with one transition which is labeled by a symbol removed by the projection. In that result and in what follows, we disregard unobservable multi-transitions, thus, several unobservable transitions connecting the same two states in the same direction are considered as only one unobservable transition.

[☆] The previous version of this work has been presented at the International Workshop on Descriptional Complexity of Formal Systems (DCFS 2011).

[*] Corresponding author.

*Email addresses:* jiraskov@saske.sk (Galina Jirásková), masopust@math.cas.cz (Tomáš Masopust)

[1] Research supported by the Slovak Research and Development Agency under contract APVV-0035-10 "Algorithms, Automata, and Discrete Data Structures".

[2] The research leading to these results received funding from the European Community's Seventh Framework Programme under grant agreement no. INFSO-ICT-224498, from the Academy of Sciences of the Czech Republic, Institutional Research Plan no. AV0Z10190503, and from the GAČR grant no. P202/11/P028.

In this paper, we improve the known upper bound by considering the structure of the automata. Specifically, we study the state complexity with respect to the number of states incident with unobservable transitions. This parameter turns out to be more convenient for this study than the number of unobservable transitions. We show that, given a projection and a minimal incomplete dfa with $n$ states, $m$ of which are incident with the unobservable transitions, the minimal incomplete dfa accepting the projected language has no more than $2^{n-1} + 2^{n-m} - 1$ states. This bound can be met if the number of unobservable transitions is $m - 1$. However, any additional unobservable transition may introduce a new unreachable subset, which means that the bound is not tight if there are more than $m - 1$ unobservable transitions. Therefore, we also discuss the case the automaton has at least $m$ unobservable transitions, and show that in this case the tight upper bound can be improved to $3 \cdot 2^{n-3} + 2^{n-m} - 1$.

The paper also discusses the case of projected finite languages, and shows that the upper bounds on the number of states correspond to the upper bounds on the nfa to dfa conversion [32].

For several operations, $op(\cdot)$, such as the determinization of nfa's, it has been shown that for all integers $n$ and $\alpha$ with $f(n) \leq \alpha \leq g(n)$, where $f(n)$ and $g(n)$ are the tight lower and upper bounds for the operation $op(\cdot)$, there exists a regular language $L$ represented by a minimal dfa of size $n$ such that the minimal dfa for $op(L)$ is of size $\alpha$. A number $\alpha$ for which no such language exists is called *magic* for $n$ with respect to $op(\cdot)$. For instance, there are no magic numbers for the determinization of nfa's with the input alphabet of cardinality at least three, where $f(n) = n$ and $g(n) = 2^n$. During the last few years, this topic has widely been discussed in the literature. The reader is referred to [9, 11, 13, 14, 15, 16, 18, 35] for more information on this topic. We solve the magic number problem for projections by showing that all the values in the range from 1 to $2^{n-1} + 2^{n-2} - 1$ can be produced as the state complexity of projected regular languages.

We conclude the paper with several remarks on sub-regular languages and a short overview of fundamental open problems concerning projected regular languages.

## 2. Preliminaries and Definitions

It is assumed that the reader is familiar with automata theory and regular languages. For all unexplained notions, we refer the reader to monograph [34].

For an alphabet $\Sigma$, denote by $\Sigma^*$ the set of all finite strings over the alphabet $\Sigma$ including the empty string $\varepsilon$. A *language* over $\Sigma$ is any subset of $\Sigma^*$. A language $L$ is *finite* if $L$ is a finite set; otherwise, $L$ is an *infinite* language.

Let $\Sigma_o \subseteq \Sigma$ be two alphabets. A homomorphism $P : \Sigma^* \to \Sigma_o^*$ is called the *(natural) projection* if it is defined so that $P(a) = \varepsilon$ for all $a \in \Sigma \setminus \Sigma_o$, and $P(a) = a$ for all $a \in \Sigma_o$.

An *(incomplete) deterministic finite automaton* (dfa) is a quintuple $A = (Q, \Sigma, \delta, s, F)$, where $Q$ is a finite set of *states*, $\Sigma$ is an *input alphabet*, $\delta : Q \times \Sigma \to Q$ is a *(partial) transition function*, $s \in Q$ is the *initial state*, and $F \subseteq Q$ is the set of *final states*. In the usual way, transition function $\delta$ can be extended to the domain $Q \times \Sigma^*$ by induction. The language *accepted* by $A$ is defined as the set $L(A) = \{w \in \Sigma^* \mid \delta(s, w) \in F\}$. A transition $\delta(p, a) = q$ is said to be *unobservable* with respect to a projection $P$ if $a \in \Sigma \setminus \Sigma_o$, that is, if $P(a) = \varepsilon$.

A *nondeterministic finite automaton* (nfa) is a quintuple $M = (Q, \Sigma, \delta, S, F)$, where $Q$, $\Sigma$, $F$ are as in a dfa, $S \subseteq Q$ is a set of initial states, and $\delta : Q \times (\Sigma \cup \{\varepsilon\}) \to 2^Q$ is a transition function which can be generalized to the function $\hat{\delta} : 2^Q \times \Sigma^* \to 2^Q$. The language accepted by $M$ is defined as the set $L(M) = \{w \in \Sigma^* \mid \delta(S, w) \cap F \neq \emptyset\}$. The *subset automaton* $M' = (2^Q, \Sigma, \hat{\delta}_\Sigma, S, F')$, where $\hat{\delta}_\Sigma : 2^Q \times \Sigma \to 2^Q$ is a restriction of $\hat{\delta}$, and $F' = \{R \subseteq Q \mid R \cap F \neq \emptyset\}$, is a dfa equivalent with $M$, that is, $L(M) = L(M')$.

For a regular language $L$, we denote by $\|L\|$ the smallest number of states in any incomplete dfa accepting the language $L$. In comparison with complete dfa's, each incomplete dfa $A$ represents two languages: A *marked language*, which is the language accepted by $A$ as defined above, and a *generated language*, which is the language accepted by the dfa obtained from $A$ by setting every state of $A$ to be final. For complete dfa's, the latter language is equal to $\Sigma^*$. This pair of languages associated with each incomplete dfa is of interest in the theory of discrete-event systems, cf. [4, 37].

## 3. Motivation

To motivate the investigation discussed in this paper, let us consider an example of a small model of a real system. The dfa in Figure 1 describes the behavior of this simple system, which is a paint factory, located at the Faculty of

Mechanical Engineering of the Eindhoven University of Technology. This machine produces cups of coloured fluids. It has vessels to store and mix fluids, a switched network of pipes and pumps to drive the fluids, and a turntable where the cups are eventually filled. The operations include the pumping of the fluids between vessels, from a vessel to the turn table, and the cleaning operations of the mixing vessel and of the pipes. The interested reader will find the details in [3].

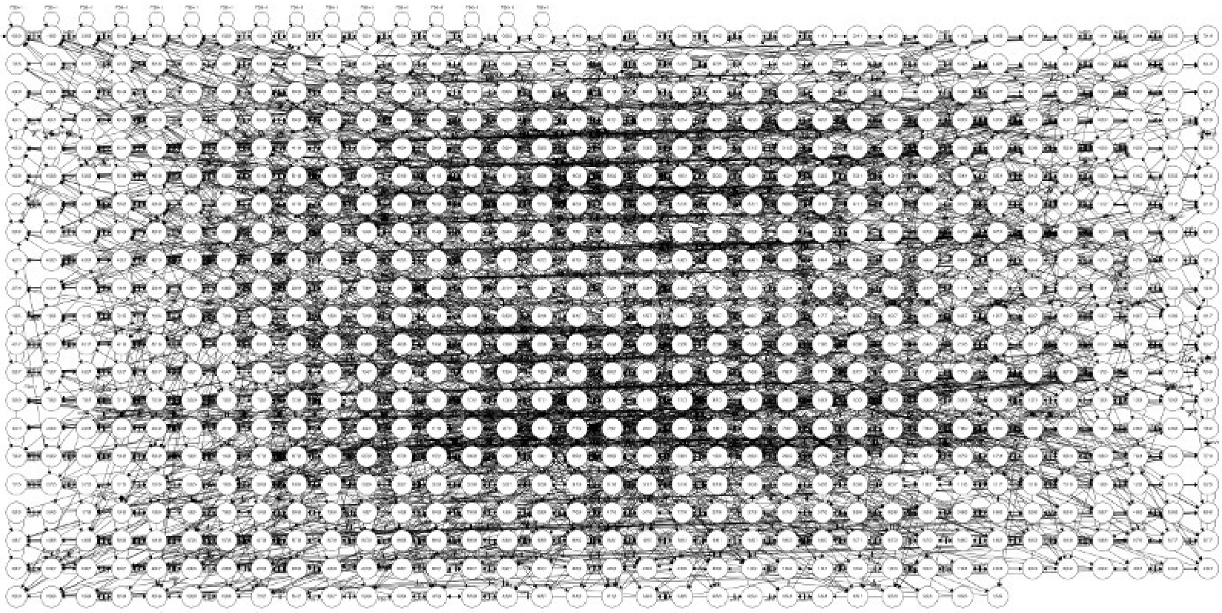


Figure 1: An example of a simple system $G$: 729 states, 4400 transitions, 19 events.

The dfa of this system consists of 729 states, 4400 transitions, and the cardinality of the alphabet is 19. It is obvious that for a human reader, it is impossible to understand the behavior and to verify any properties. Fortunately, to verify a property, it is not always necessary to have the complete model, but it is sufficient to keep only an important abstracted part. Assume that the property we need to check concerns only seven of those 19 symbols of the alphabet. We can then use the projection to keep only these symbols and to remove all the other symbols. The minimal dfa for the projected language of our example is depicted in Figure 2. The dfa has only 27 states and 62 transitions, and it is quite readable even for a human reader. However, the computation of the projection takes (on a current PC) about 23 minutes, which means that to produce one state of the resulting automaton takes one minute, on average. For bigger systems this becomes infeasible. Furthermore, the projection used in this example satisfies the so-called observer property, see [36], which ensures that the minimal automaton for the projected language has no more states than the minimal automaton for the input language, cf. also [29]. Moreover, Wong [36] proposed a polynomial-time algorithm running in time $O(n^7 m^2)$, where $n$ is the number of states and $m$ is the cardinality of the co-domain of the projection satisfying the observer property. This implies that the best known algorithm for this special case needs on the order of 5361530467444105601241 steps to produce the dfa with 27 states of Figure 2.

Based on this example, let us summarize the questions we are able to answer, and the questions which are open and of interest. Given a language $L$ with $\|L\| = n$ and a projection $P$. In time $O(1)$, we can immediately get an answer to the question "What is the state complexity of $P(L)$?" The answer is "$1 \leq \|P(L)\| \leq 2^{n-1} + 2^{n-2} - 1$", that is, "$\|P(L)\|$ is a number between 1 and $2^{n-1} + 2^{n-2} - 1$." In what follows, we improve this result by considering the structural property of the automaton representation of the given language. The required time complexity to verify our structural property is linear with respect to the size of the minimal dfa for $L$. In addition, the observer property [36] can be verified in time $O(n^3)$, so in this time we get that if $L$ satisfies the observer property, then $1 \leq \|P(L)\| \leq \|L\|$. Note that the languages satisfying the observer property (with respect to a given projection $P$) and the class of finite languages projected onto unary finite languages are the only known language classes for which it is true that $\|P(L)\| \leq \|L\|$.

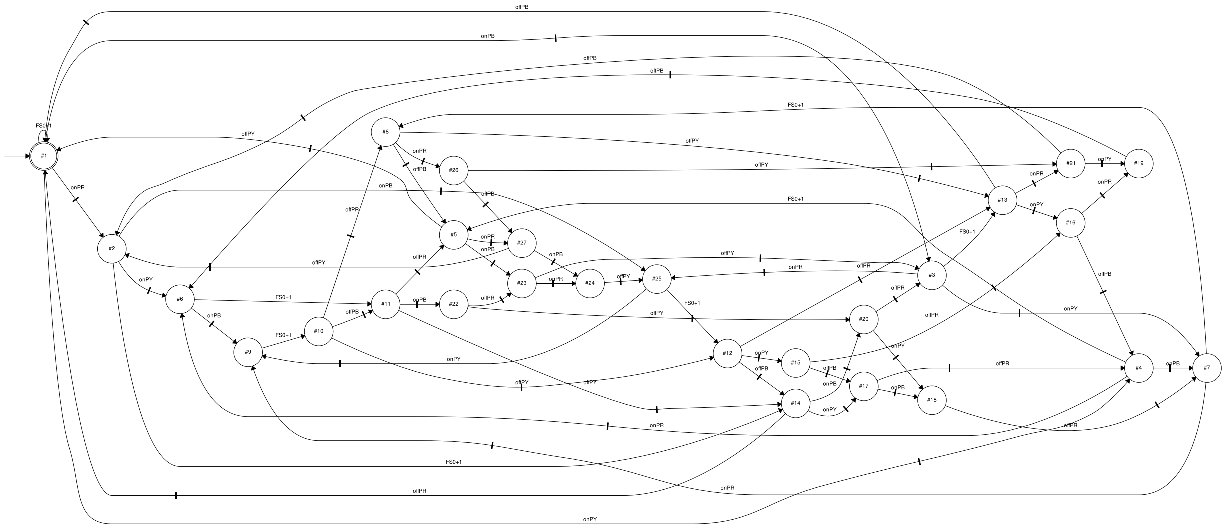


Figure 2: Projection of $G$: 27 states, 62 transitions, 7 events.

On the other hand, what we expect for practical applications is a bit different. The fundamental question is "What is $\|P(L)\|$?" Of course, this can be computed, but the known algorithm is exponential and, therefore, infeasible in general. Are there feasible algorithms to answer this question? Or at least to produce an answer of the form $\|P(L)\| = O(x)$, for some $x \in \{n, n\log n, n^2, \ldots\}$?

## 4. DFAs as Graphs

In this section, we concentrate our attention on the number of states potentially reachable in the subset automaton constructed from a given dfa after the application of a projection. For simplification, we consider the important parts of the automata as graphs.

A *directed graph* is a pair $G = (V, E)$, where $V$ is a finite set of nodes, and $E \subseteq V \times V$ is a set of edges. An edge $(u, v) \in E$ is called a *loop* if $u = v$. Let $v \in V$ be a node, then we define *in-degree* and *out-degree* of $v$ as the sizes of sets $\{u \in V \mid (u, v) \in E\}$ and $\{w \in V \mid (v, w) \in E\}$, respectively. A node with in-degree 0 and out-degree 1, or with in-degree 1 and out-degree 0 is called a *leaf*. This definition requires that the node is incident to an edge. Thus, a node incident to no edge is not considered to be a leaf.

A *path* in $G$ is a sequence of nodes $v_0, v_1, \ldots, v_k$, for some $k \geq 1$, such that $v_i \neq v_j$ if $i \neq j$, and $(v_i, v_{i+1})$ is an edge in $E$, for $i = 0, 1, \ldots, k-1$. A *non-oriented path* is a sequence $v_0, v_1, \ldots, v_k$, for some $k \geq 1$, such that $v_i \neq v_j$ if $i \neq j$, and either $(v_i, v_{i+1})$ or $(v_{i+1}, v_i)$ is an edge in $E$, for $i = 0, 1, \ldots, k-1$. Graph $G$ is *connected* if for all nodes $u, v$ in $V$, there is a non-oriented path from $u$ to $v$. For a node $v$ in $V$, let $G \setminus \{v\}$ denote the graph constructed from $G$ by removing node $v$ and all edges incident to $v$.

A subset $X$ of $V$ is said to be *bad* in graph $G = (V, E)$ if there exists an edge $(u, v)$ in $E$ such that $u \in X$ and $v \notin X$. A set is said to be *good* if it is not bad; thus, each good subset of $V$ is closed under outgoing transitions. Let $b(G)$ denote the number of bad subsets in $G$, and $g(G)$ the number of good subsets in $G$. Our first lemma studies the number of bad subsets in a graph.

**Lemma 1.** *Let $m, n \geq 2$ and $G = (V, E)$ be a directed graph without loops with $n$ nodes. Let $U = \{u, v \in V \mid (u, v) \in E\}$ and assume that $U$ is of size $m$. Then $b(G) \geq (2^{m-1} - 1)\, 2^{n-m}$.*

*Proof.* Let $G$ and $U$ be as assumed in the theorem, and consider a special case where the edges involved in nodes of $U$ go only from $m - 1$ different nodes to the last $m$-th node, see Figure 3. This means that there exists a node $v$ in $V$ such that for each node $u$ in $U \setminus \{v\}$, the edge $(u, v)$ is in $E$, while for each node $z$ in $V$, the edge $(z, u)$ is not in $E$. Then there are $2^{m-1} - 1$ nonempty subsets of $U$ which do not contain node $v$, and so are bad. This gives $b(G) \geq (2^{m-1} - 1)\, 2^{n-m}$.

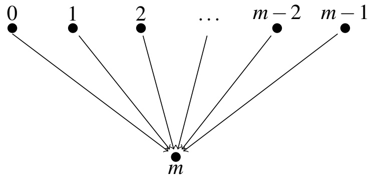


Figure 3: The worst case of Lemma 1.

Now we show the theorem to be true in general, and not just under the assumption that the edges in $U$ go only from $m-1$ different nodes to the last $m$-th node as was done in the paragraph above. The proof is by induction on $m$.

If $m = 2$, then $U$ involves either one or two edges. Note first that if $X$ is a bad subset in $G$, then $X$ is bad after addition of any number of edges to $G$. Thus, we can consider that there is only one edge because the other one cannot decrease the number of bad subsets. Then, if we have one edge, say $(a, b)$, we can have $a$ along with any combination of elements of $V \setminus \{a, b\}$ in a bad subset, and thus we have $b(G) \geq 2^{n-2} = (2^{2-1} - 1)\, 2^{n-2}$. Assume that the statement holds for all sets $U$ of size less than $m$, and consider the case $U$ is of size $m$. There are two possibilities. Either the number of edges is strictly less than $m$, or it is greater then or equal to $m$. In the former case, consider the number of edges and denote it by $t$, and in the latter case, consider the subset of edges of size $t$ forming the minimal spanning tree (forest). Thus $t < m$ and there is a leaf $v$ in $U$ such that $v$ is connected with a node $u$ in $U \setminus \{v\}$. Then, either (i) all nodes in $U \setminus \{v\}$ are incident with some of the $t$ edges, or (ii) node $u$ was connected only with $v$ and now it is not incident with any other node in $U \setminus \{v\}$.

In case (i), the set $U \setminus \{v\}$ is of size $m-1$, and by the induction hypothesis, there are at least $2^{m-2} - 1$ bad subsets of $U \setminus \{v\}$. If $(v, u) \in E$, then for each subset $A$ of $U \setminus \{v\}$ that is bad in $U \setminus \{v\}$, the sets $A$ and $A \cup \{v\}$ are bad in $U$, and $\{v\}$ is a new bad set. This gives $b(G) \geq (2^{m-2} - 1 + 2^{m-2} - 1 + 1)\, 2^{n-m}$. Similarly, if $(u, v) \in E$, then for each subset $A$ of $U \setminus \{v\}$ that is bad in $U \setminus \{v\}$, the sets $A$, $A \cup \{v\}$ are bad in $U$, and the set $U \setminus \{v\}$ is a new bad set.

In case (ii), the set $U \setminus \{u, v\}$ is of size $m-2$, and so, there are at least $2^{m-3} - 1$ bad subsets of $U \setminus \{u, v\}$. We now have $m \geq 4$. The sets $\emptyset$ and $U \setminus \{u, v\}$ are not bad. Thus $\{v\}$ or $\{u\}$, and $U \setminus \{u\}$ or $U \setminus \{v\}$, depending on the direction of the edge connecting $u$ and $v$, are two new bad subsets. Moreover, all bad subsets of $U \setminus \{u, v\}$ are also bad in $U$. If there is at least one more proper non-empty good subset $B$ of $U \setminus \{u, v\}$, then $B \cup \{u\}$ or $B \cup \{v\}$ is the third new bad subset of $U$. Summarized, this gives $b(G) \geq (2^2\,(2^{m-3} - 1) + 3)\, 2^{n-m} = (2^{m-1} - 1)\, 2^{n-m}$. If there are only two good subsets of $U \setminus \{u, v\}$, namely $\emptyset$ and $U \setminus \{u, v\}$, then the number of bad subsets of $U \setminus \{u, v\}$ is $2^{m-2} - 2$, which, since $m \geq 4$, gives $b(G) \geq 2^2(2^{m-2} - 2)\, 2^{n-m} = (2^{m-1} - 1 + 2^{m-1} - 7)\, 2^{n-m} \geq (2^{m-1} - 1)\, 2^{n-m}$. □

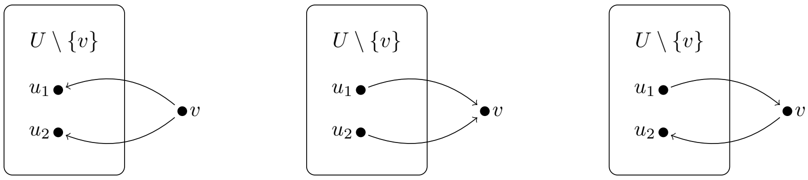
Consider the statement of Lemma 1. Then the number of all the subsets of $V \setminus U$ is $2^{n-m}$ while the number of bad subsets of $U$ is at least $2^{m-1} - 1$. Moreover, there is a graph $G = (V, E)$ with $U$ of size $|E| - 1$, for which the equality holds. However, if $m \leq |E|$, each additional transition can introduce a new bad subset. This problem is discussed in the following result that gives a lower bound that is strictly greater than $(2^{m-1} - 1)2^{n-m}$.

**Lemma 2.** *Let $m, n \geq 2$ and $G = (V, E)$ be a directed graph without loops with n nodes. Let $U = \{u, v \in V \mid (u, v) \in E\}$ and assume that $|U| = m \leq |E|$. Then $b(G) \geq (5 \cdot 2^{m-3} - 1)\, 2^{n-m}$.*

*Proof.* The proof is by induction on $m$. If $m = 2$, then the graph consists of two nodes connected by two edges. This gives two bad subsets of $U$, which results in $b(G) = 2 \cdot 2^{n-m} \geq 3/2 \cdot 2^{n-m}$. Assume that the statement holds for all sets $U$ of cardinality less then $m$, and consider the case $U$ is of cardinality $m$. Recall that $m \leq |E|$. Consider a subset of $m$ edges forming a minimal spanning tree (forest). Then there is a leaf $v$ in $U$. If $|U \setminus \{v\}| \leq |E(G \setminus \{v\})|$ then by the induction hypothesis, the set $U \setminus \{v\}$ has at least $5 \cdot 2^{m-4} - 1$ bad subsets. Otherwise, by Lemma 1, the set $U \setminus \{v\}$ has at least $2^{m-2} - 1$ bad subsets by examining the subgraph of $G$ with $U \setminus \{v\}$ as vertices.

In the former case, if $(v, u) \in E$, then for each bad subset $A$ of $U \setminus \{v\}$, the set $A \cup \{v\}$ is a new bad subset of $U$ and, in addition, $\{v\}$ is a new bad subset of $U$. If $(u, v) \in E$, then for each bad subset $A$ of $U \setminus \{v\}$, the set $A \cup \{v\}$ is a new bad subset of $U$ and, in addition, the set $U \setminus \{v\}$ is a new bad subset of set $U$. Thus $b(G) \geq (5 \cdot 2^{m-4} - 1 + 5 \cdot 2^{m-4})\, 2^{n-m} = (5 \cdot 2^{m-3} - 1)\, 2^{n-m}$.

In the latter case, notice that there are at least two edges connecting $v$ and $U \setminus \{v\}$ in $G$. We have three possibilities illustrated in Figure 4:



(a) Case 1. (b) Case 2. (c) Case 3.

Figure 4: The three possibilities in the proof of Lemma 2.

1. Node $v$ is connected with $U \setminus \{v\}$ by edges $(v, u_1)$ and $(v, u_2)$ with $u_1 \neq u_2$. Then the sets $A \cup \{v\}$, $A \cup \{v, u_1\}$, and $A \cup \{v, u_2\}$ are bad in $U$ for every subset $A$ of $U \setminus \{v, u_1, u_2\}$. Thus we have at least $3 \cdot 2^{m-3}$ new bad subsets in $U$.
2. Node $v$ is connected with $U \setminus \{v\}$ by edges $(u_1, v)$ and $(u_2, v)$. Then for each subset $A$ of $U \setminus \{u_1, u_2, v\}$, if $A \cup \{u_1\}$ is bad in $U \setminus \{v\}$, then $A \cup \{v, u_1\}$ is bad in $U$, otherwise $A \cup \{u_1\}$ is bad in $U$; if $A \cup \{u_2\}$ is bad in $U \setminus \{v\}$, then $A \cup \{v, u_2\}$ is bad in $U$, otherwise $A \cup \{u_2\}$ is bad in $U$; if $A \cup \{u_1, u_2\}$ is bad in $U \setminus \{v\}$, then $A \cup \{u_1, u_2, v\}$ is bad in $U$, otherwise $A \cup \{u_1, u_2\}$ is bad in $U$. Summarized, there are $3 \cdot 2^{m-3}$ new bad subsets in $U$.
3. Node $v$ is connected with $U \setminus \{v\}$ by edges $(u_1, v)$ and $(v, u_2)$. Then the sets $A \cup \{v\}$ and $A \cup \{u_1, v\}$ are bad in $U$ for each subset $A$ of $U \setminus \{u_1, u_2, v\}$. In addition, if $A \cup \{u_1, u_2\}$ is bad in $U \setminus \{v\}$, then the set $A \cup \{u_1, u_2, v\}$ is a new bad subset of $U$. Otherwise, the set $A \cup \{u_1, u_2\}$ is a new bad subset of $U$. Thus there are at least $3 \cdot 2^{m-3}$ new bad subsets of $U$.

Summarized, this gives $b(G) \geq (2^{m-2} - 1 + 3 \cdot 2^{m-3})\, 2^{n-m} = (5 \cdot 2^{m-3} - 1)\, 2^{n-m}$. □

## 5. State Complexity of Projected Regular Languages

Recall that it is shown in [36] that the worst-case tight upper bound on projected regular languages is $2^{n-1}+2^{n-2}-1$, where $n$ is the number of states of the minimal incomplete dfa recognizing the given language.

**Theorem 3** ([36])**.** *Let $n \geq 2$ and $L$ be a regular language over $\Sigma$ with $\|L\| = n$. Let $\Sigma_o \subseteq \Sigma$ and $P$ be the projection of $\Sigma^*$ onto $\Sigma_o^*$. The tight upper bound on the size of the minimal incomplete dfa for the projected language $P(L)$ is $3 \cdot 2^{n-2} - 1$.*

In what follows, we improve the upper bound by taking into account the structure of nonloop unobservable transitions. More specifically, we consider the number of states that are incident with nonloop unobservable transitions. Note that it follows from the following results that the previous bound is reachable only by dfa's with one unobservable transition, disregarding the unobservable multi-transitions.

**Theorem 4.** *Let $m, n \geq 2$, $\Sigma_o \subseteq \Sigma$, and $P$ be the projection of $\Sigma^*$ onto $\Sigma_o^*$. Let $L$ be a regular language over alphabet $\Sigma$ with $\|L\| = n$, and $(Q, \Sigma, \delta, s, F)$ be the minimal incomplete dfa recognizing language $L$, in which*

$$|\{p, q \in Q \mid p \neq q \text{ and } q \in \delta(p, \Sigma \setminus \Sigma_o)\}| = m.$$

*Then $\|P(L)\| \leq 2^{n-1} + 2^{n-m} - 1$.*

*Proof.* Consider the minimal incomplete dfa $(Q, \Sigma, \delta, s, F)$ accepting $L$, and construct a directed graph $G = (Q, E)$ without loops so that $E$ contains an edge $(p, q)$ in $Q \times Q$ if and only if $p \neq q$ and there is a transition $\delta(p, a) = q$ for some unobservable symbol $a$ in $\Sigma \setminus \Sigma_o$. Construct an nfa for language $P(L)$ from dfa $A$ by replacing all the unobservable transitions with $\varepsilon$-transitions. Observe that each subset of $Q$ that contains $p$, but not $q$, is not reachable in the corresponding subset automaton because every string leading the nfa to state $p$ also leads the automaton to state $q$. This means that no subset of $Q$ that is bad in graph $G$ is reachable. By Lemma 1, for the number $g(G)$ of good subsets (that is, subsets closed under outgoing transitions) we have $g(G) = 2^n - b(G) \leq 2^n - (2^{m-1} - 1)\, 2^{n-m} = 2^{n-1} + 2^{n-m}$. Good subsets of $Q$ in graph $G$ correspond to potentially reachable states in the subset automaton. This number is decreased by one because the empty set (the dead state) is potentially reachable but it is not present in the minimal incomplete dfa. □

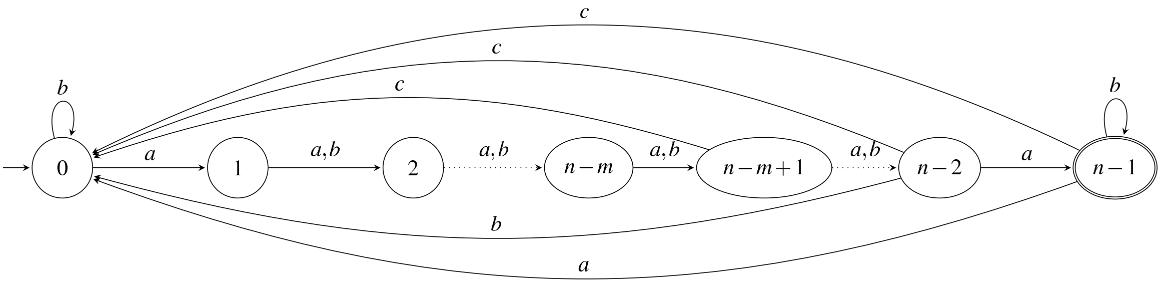


Figure 5: The minimal incomplete dfa for a language $L$ with $\|P(L)\| = 2^{n-1} + 2^{n-m} - 1$. $\Sigma = \{a, b, c\}$ and $\Sigma_o = \{a, b\}$.

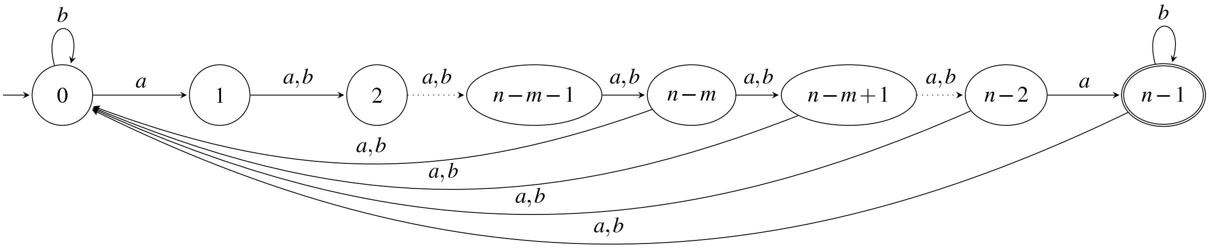


Figure 6: An nfa accepting the projection of the language from Figure 5.

Notice that Theorem 3 is a consequence of Theorem 4 since $\|P(L)\|$ is maximal if $m = 2$. The next result shows that the bound $2^{n-1} + 2^{n-m} - 1$ is tight. Moreover, the worst-case example in the following theorem is defined over a three-letter alphabet which also improves the result of Wong [36].

**Theorem 5.** *Let $m, n \geq 2$ and $P$ be the projection of $\{a, b, c\}^*$ onto $\{a, b\}^*$. There exists a regular language $L$ over $\{a, b, c\}$ with $\|L\| = n$, such that the minimal incomplete dfa accepting $L$ has $m - 1$ unobservable nonloop transitions connecting $m$ states, and $\|P(L)\| = 2^{n-1} + 2^{n-m} - 1$.*

*Proof.* Let $L$ be the language over $\{a, b, c\}$ accepted by the incomplete dfa shown in Figure 5. After applying the projection onto $\{a, b\}$ and removing $\varepsilon$-transitions, we get the $n$-state nfa shown in Figure 6. The nfa accepts the string $b^n$ only from state $n - 1$, and the string $a^i b^n$ only from state $n - 1 - i$ $(0 \leq i \leq n - 1)$. It follows that the states in the corresponding subset automaton are pairwise distinguishable. To prove the theorem, we only need to show that the subset automaton has $2^{n-1} + 2^{n-m} - 1$ reachable non-empty states.

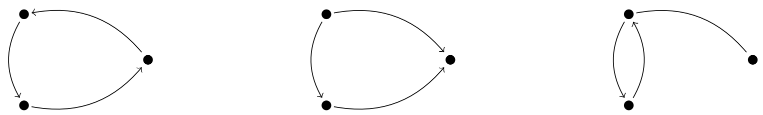
We first prove by induction that every subset of $\{0, 1, \ldots, n - 1\}$ containing state 0 is reachable. The initial state $\{0\}$ goes to state $\{n - m\}$ by $a^{n-m}$, then by a string in $b^*$ to states $\{0, i\}$ with $n - m + 1 \leq i \leq n - 2$. State $\{0, n - 2\}$ goes to state $\{0, 1, n - 1\}$ by $a$, and then by a string in $b^*$ to states $\{0, i, n - 1\}$ with $1 \leq i \leq n - 2$. State $\{0, n - 2, n - 1\}$ goes to $\{0, n - 1\}$ by $b$, and then to $\{0, 1\}$ by $a$. By a string in $b^*$, state $\{0, 1\}$ goes to states $\{0, i\}$ with $1 \leq i \leq n - m$. Thus each subset of size 2 containing state 0 is reachable. Now let $X = \{0, i_1, i_2, \ldots, i_t\}$ be a set of size $t + 1$, where $2 \leq t \leq n - 1$ and $1 \leq i_1 < i_2 < \cdots < i_t \leq n - 1$. Consider two cases:

(*i*) $i_t = n - 1$. Then $X$ is reached from $\{0, i_2 - i_1, \ldots, i_{t-1} - i_1, n - 2\}$ by $ab^{i_1 - 1}$, and the latter set of size $t$ is reachable by the induction hypothesis.

(*ii*) $i_t < n - 1$. Then $X$ is reached from $\{0, i_2 - i_1, \ldots, i_t - i_1, n - 1\}$ by $ab^{i_1 - 1}$, and the latter set of size $t + 1$ contains state $n - 1$, and is reachable by (*i*).

This proves reachability of all subsets containing state 0. Next, if $\{i_1, i_2, \ldots, i_t\}$ is a non-empty subset of the set $\{1, 2, \ldots, n - m\}$, then it is reached from the set $\{0, i_2 - i_1, i_3 - i_1, \ldots, i_t - i_1\}$ containing state 0 by $a^{i_1}$. This gives $2^{n-1} + 2^{n-m} - 1$ reachable non-empty states, and completes our proof. □

In the theorems above, the number of unobservable transitions is considered to be less than the size of the set $\{p, q \in Q \mid p \neq q \text{ and } q \in \delta(p, \Sigma \setminus \Sigma_o)\}$. However, an additional unobservable transition may introduce a new



$(a)$ Case 1. $(b)$ Case 2. $(c)$ Case 3.

Figure 7: The three possibilities of Example 1.

unreachable subset. The following example shows that if the size of this set is less than or equal to the number of unobservable nonloop transitions, then the upper bound is not tight. The precise upper bound for this case is open.

**Example 1.** Let $m, n \geq 2$. Consider a minimal incomplete dfa $(Q, \Sigma, \delta, s, F)$ of $n$ states. Let the incomplete automaton have at least $m$ unobservable transitions. Let $U = \{p, q \in Q \mid p \neq q \text{ and } q \in \delta(p, \Sigma \setminus \Sigma_o)\}$ and assume that $|U| = m$. Construct a directed graph $G = (Q, E)$ without loops so that the set $E$ contains an edge $(p, q)$ in $Q \times Q$ if and only if $p \neq q$ and there is a transition $\delta(p, a) = q$ for some unobservable symbol $a$ in $\Sigma \setminus \Sigma_o$.
In the case of $m = 2$, there must be a cycle of length two in $G$. In this case, however, we have $g(G) = 2^n - 2 \cdot 2^{n-2} = 2^{n-1}$. In the case of $m = 3$, there are three possibilities, see Figure 7:

1. if $U$ contains a cycle of length three, then there are at least 6 subsets that are bad for $U$ because all but the empty set and the whole set $U$ are bad;
2. if $U$ contains a cycle with one transition reversed, then there are at least 4 bad subsets of $U$;
3. if $U$ contains a cycle of length two and an edge to (or from) the third node, then there are at least 5 bad subsets of $U$.

In all three cases, we get $g(G) \leq 2^n - 4 \cdot 2^{n-3} = 2^{n-1}$. Since only non-empty good subsets for $G$ can be reached in the incomplete dfa for the projected language, we get the bound $2^{n-1} - 1$ on the size of this dfa in both cases. This is strictly less than $2^{n-1} + 2^{n-m} - 1$ given by Theorem 4.

Finally, the situation is significantly different for projections of regular languages with one-letter co-domains. Note that it is not hard to construct an incomplete dfa with $n + 1$ states such that its projection results in the Chrobak's unary automaton [5] with $n$ states meeting the upper bound from the following theorem. However, to do this, we need an alphabet of size linear with respect to $n$. It is an open problem whether the upper bound can be met using a fixed alphabet.

**Theorem 6.** *Let $a$ be a symbol in an alphabet $\Sigma$ and $P$ be the projection of strings in $\Sigma^*$ to strings in $a^*$. Let $L$ be a regular language over $\Sigma$ with $\|L\| = n$. Then $\|P(L)\| \leq e^{(1+o(1))\sqrt{n \ln n}}$.*

*Proof.* Replace all the transitions unobservable for projection $P$ in the minimal incomplete dfa recognizing language $L$ with $\varepsilon$-transitions to get an $n$-state unary nfa for language $P(L)$. This unary nfa can be simulated by a dfa with no more than $e^{(1+o(1))\sqrt{n \ln n}}$ states [5, 9, 26], and the upper bound follows. □

The following theorem discusses a special case that gives an idea how to treat the cases with more and more unobservable transitions.

**Theorem 7.** *Let $m, n \geq 2$ and $\Sigma_o \subseteq \Sigma$. Let $P$ be the projection of strings in $\Sigma^*$ to strings in $\Sigma_o^*$. Let $L$ be a regular language over alphabet $\Sigma$ with $\|L\| = n$, and $(Q, \Sigma, \delta, s, F)$ be the minimal incomplete dfa recognizing language $L$, in which $|\{p, q \in Q \mid p \neq q \text{ and } q \in \delta(p, \Sigma \setminus \Sigma_o)\}| = m$. If at least $m$ transitions in the dfa are unobservable for the projection, then $\|P(L)\| \leq 2^{n-2} + 2^{n-3} + 2^{n-m} - 1$.*

*Proof.* Consider the minimal incomplete dfa $(Q, \Sigma, \delta, s, F)$ for $L$, and construct a directed graph $G = (Q, E)$ without loops so that $E$ contains an arc $(p, q)$ if and only if $p \neq q$ and there is a transition $\delta(p, a) = q$ for some unobservable symbol $a$ in $\Sigma \setminus \Sigma_o$. Construct an nfa for language $P(L)$ from the dfa for $L$ by replacing all the unobservable transitions with $\varepsilon$-transitions. Then every subset that is reachable in the corresponding subset automaton must be good for $G$. By Lemma 2, we have $g(G) \leq 2^n - (5 \cdot 2^{m-3} - 1)\, 2^{n-m} = 2^{n-2} + 2^{n-3} + 2^{n-m}$. This number is decreased by one because of the empty set (the dead state). □

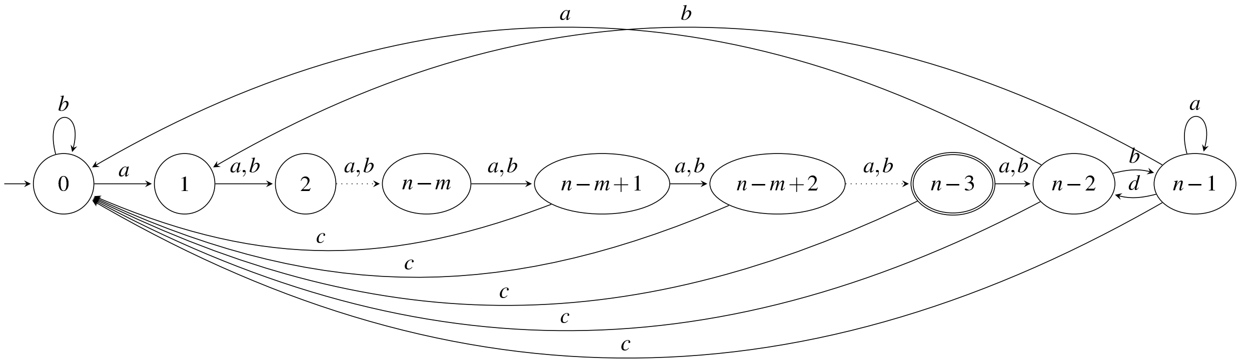


Figure 8: The minimal incomplete dfa for a language $L$ with $\|P(L)\| = 2^{n-2} + 2^{n-3} + 2^{n-m} - 1$. $\Sigma = \{a, b, c, d\}$ and $\Sigma_o = \{a, b\}$.

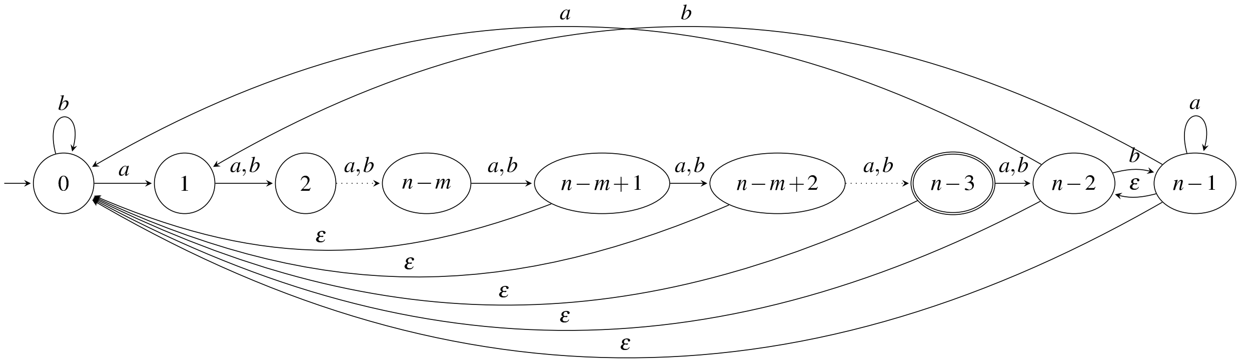


Figure 9: An $\varepsilon$-nfa for the projection of the language from Figure 8.

The next result proves the tightness of the bound $2^{n-2} + 2^{n-3} + 2^{n-m} - 1$ in the case of a four-letter domain alphabet and a two-letter co-domain alphabet. Let us remark that in the preliminary version of this paper [20], there is an error in defining the corresponding dfa. In fact, the automaton in Fig. 3 on page 206 of the DCFS paper is not deterministic since two transitions on symbol $d$ go from state 2. Therefore, to get a deterministic automaton, one more symbol must be used, so the domain alphabet should be of size 5. The following theorem fixes this error, and moreover, decreases the size of domain and co-domain alphabets.

**Theorem 8.** *Let $3 \le m \le n$ and $P$ be the projection of $\{a, b, c, d\}^*$ onto $\{a, b\}^*$. There exists a regular language $L$ over $\{a, b, c, d\}$ with $\|L\| =\ n$ such that the minimal incomplete dfa accepting $L$ has $m$ unobservable nonloop transitions on no more than $m$ states, and $\|P(L)\| = 2^{n-2} + 2^{n-3} + 2^{n-m} - 1$.*

*Proof.* Let $L$ be the language accepted by the dfa of Figure 8. Here by $a$, state $n - 1$ goes to itself, state $n - 2$ goes to state 0, and every other state $i$ goes to state $i + 1$. By $b$, state 0 goes to itself, state $n - 1$ goes to state 1, and every other state $i$ goes to state $i + 1$. By $c$, every state $i$ with $n - m + 1 \le i \le n - 1$ goes to state 0. By $d$, state $n - 1$ goes to state $n - 2$. State $n - 3$ is the sole accepting state. Construct the nfa for $P(L)$ by replacing transitions on $c, d$ by transitions on $\varepsilon$, as shown in Figure 9. Let us show that $2^{n-2} + 2^{n-3} + 2^{n-m} - 1$ states are reachable and pairwise distinguishable in the corresponding incomplete subset automaton.

Denote by $\mathcal{R}$ the following family of $2^{n-2} + 2^{n-3} + 2^{n-m} - 1$ subsets of $\{0, 1, \ldots, n - 1\}$:

$$\begin{aligned}\mathcal{R} =&\{X \subseteq \{0, 1, \ldots, n - 1\} \mid 0 \in X \text{ and } n - 1 \notin X\} \cup\\ &\{X \subseteq \{0, 1, \ldots, n - 1\} \mid \{0, n - 2, n - 1\} \subseteq X\} \cup\\ &\{X \subseteq \{1, 2, \ldots, n - m\} \mid X \neq \emptyset\},\end{aligned}$$

that is, family $\mathcal{R}$ consists of all the subsets containing state 0 but not containing state $n-1$, all the subsets containing states $0, n-2$, and $n-1$, and all the non-empty subsets of $\{1, 2, \ldots, n-m\}$. The proof of reachability of all the subsets in $\mathcal{R}$ is by induction on the size of subsets. All the subsets in $\mathcal{R}$ of size 1 and 2 are reachable since for all $i, j$ with $0 \le i < j \le n-2$, we have $\{0\} \xrightarrow{a} \{1\} \xrightarrow{a} \{2\} \xrightarrow{a} \cdots \xrightarrow{a} \{n-m\} \xrightarrow{b} \{0, n-m+1\} \xrightarrow{b} \{0, n-m+2\} \xrightarrow{b} \cdots \xrightarrow{b} \{0, n-2\} \xrightarrow{a} \{0, 1\} \xrightarrow{b^{j-1}} \{0, j\}$; and $\{0, j-i\} \xrightarrow{a^i} \{i, j\}$. Now let $3 \le k \le n$ and assume that all the subsets in $\mathcal{R}$ of size $k-1$ are reachable. Let $X = \{i_1, i_2, \ldots, i_k\}$, where $0 \le i_1 < i_2 < \cdots < i_k \le n-1$, be a set in $\mathcal{R}$ of size $k$. We have nine cases:

1. $i_1 = 0, i_2 = 1, n-m+1 \le i_k \le n-2$. Then $\{0, i_3-1, \ldots, i_k-1\}$ is a set in $\mathcal{R}$ of size $k-1$. This set is reachable by the induction hypothesis and goes to $X$ by $a$.
2. $i_1 = 0, i_2 \ge 2, i_k = n-m+1$. Then $\{i_2-1, \ldots, i_k-1\}$ is reachable by induction and goes to $X$ by $a$.
3. $i_1 = 0, i_2 \ge 2, i_k = n-m+2$. Then $\{0, i_2-1, \ldots, i_k-1\}$ is reachable as shown either in case (1) or in case (2), and it goes to $X$ by $b$. In such a way, by induction on $i_k$, we prove the reachability of all the subsets $X$ with $i_1 = 0, i_2 \ge 2$, and $i_k = n-m+1, n-m+2, \ldots, n-2$.
4. $i_1 = 0, i_2 = 1, i_k \le n-m$. Then $\{0, i_3-1, \ldots, i_k-1, n-2\}$ is reachable as shown in (1) or (3) and goes to $X$ by $a$.
5. $i_1 = 0, i_2 \ge 2, i_k \le n-m$. Then $\{0, 1, i_3-(i_2-1), \ldots, i_k-(i_2-1)\}$ is reachable as in (4) and goes to $X$ by $b^{i_2-1}$.
6. $i_1 \ge 1, i_k \le n-m$. Then $\{0, i_2-i_1, \ldots, i_k-i_1\}$ is reachable as shown in (4) or (5), and goes to $X$ by $a^{i_1}$.
7. $i_1 = 0, i_2 = n-2, i_3 = n-1$ if $k = 3$. Then $\{0, n-2\}$ goes to $X$ by $b$.
8. $i_1 = 0, i_2 = 1, i_{k-1} = n-2, i_k = n-1$ if $k \ge 4$. Then $\{0, i_3-1, \ldots, i_{k-2}-1, n-2, n-1\}$ is a set in $\mathcal{R}$ of size $k-1$ that is reachable by the induction hypothesis, and it goes to $X$ by $a$.
9. $i_1 = 0, i_2 \ge 2, i_{k-1} = n-2, i_k = n-1$ if $k \ge 4$. Then $\{0, i_2-1, \ldots, i_{k-2}-1, n-2\}$ is reachable by the induction hypothesis and goes to $X$ by $b$. This completes the proof of reachability of all the subsets in family $\mathcal{R}$.

To prove distinguishability, notice that the string $b^i$ with $0 \le i \le n-4$ is accepted by the nfa for $P(L)$ only from state $n-3-i$, and the string $b^{n-3}$ only from state $n-1$. Thus, if two subsets differ in a state $1, 2, \ldots, n-3$, or $n-1$, then they can be distinguished by the corresponding string in $b^*$. If $0 \in X$ and $0 \notin Y$, then $Y \subseteq \{1, 2, \ldots, n-m\}$, and therefore $a^{n-3}$ is accepted from $X$ but not from $Y$. If $n-2 \in X$ and $n-2 \notin Y$, then $n-1 \notin Y$, and therefore $a^{n-2}$ is accepted from $X$ but not from $Y$. This proves distinguishability. □

## 6. State Complexity of Projected Finite Languages

In this section, we consider the state complexity of projected finite languages. First, let us consider the case of projections with co-domains of size one.

**Proposition 9.** *Let $a$ be a symbol in an alphabet $\Sigma$ and let $P$ be the projection of $\Sigma^*$ onto $a^*$. If $L$ is a finite regular language over $\Sigma$, then $\|P(L)\| \le \|L\|$. The bound is tight for any alphabet.*

*Proof.* Consider the minimal complete dfa with $n$ states accepting language $L$. Since $L$ is finite, there must exist a string that leads the dfa to the dead state. Hence the minimal incomplete dfa accepting $L$ has $n-1$ states. After replacing all the unobservable transitions with $\varepsilon$-transitions and eliminating $\varepsilon$-transitions, the resulting nfa with $n-1$ states accepts the finite language $P(L)$. Therefore, this nfa can be simulated by an $n$-state complete dfa [32]. Again, some string must lead this complete dfa to the dead state, which implies that the minimal incomplete dfa accepting $P(L)$ has at most $n-1$ states. Thus $\|P(L)\| \le \|L\|$. The bound is met if we only have loops by unobservable letters. □

The following theorem deals with finite languages and binary co-domain alphabets.

**Theorem 10.** *Let $a$ and $b$ be symbols in an alphabet $\Sigma$ and $P$ be the projection of $\Sigma^*$ onto $\{a, b\}^*$. Let $L$ be a finite language over $\Sigma$ with $\|L\| = n$. Then*

$$\|P(L)\| \le \begin{cases} 2 \cdot 2^{\lfloor n/2 \rfloor} - 2 & \textit{if } n \textit{ is even,} \\ 3 \cdot 2^{\lfloor n/2 \rfloor} - 2 & \textit{if } n \textit{ is odd.} \end{cases}$$

*In addition, the bound is tight in the case of a ternary domain alphabet.*

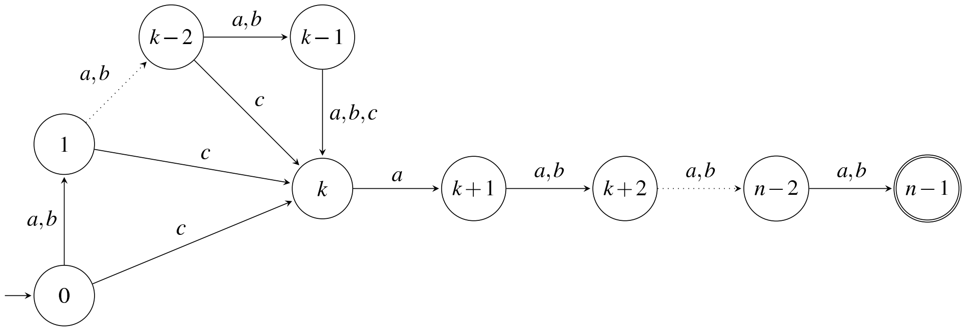


Figure 10: The minimal incomplete dfa over $\{a, b, c\}$ accepting a finite language meeting the upper bound on the projection onto $\{a, b\}^*$; $k = \lceil n/2\rceil - 1$.

*Proof.* We first prove the upper bound. Consider an incomplete dfa accepting language $L$, and construct an $n$-state nfa for $P(L)$ by replacing all the unobservable transitions with $\varepsilon$-transitions, and eliminating the $\varepsilon$-transitions. The $n$-state nfa for finite language $P(L)$ can be simulated by a complete dfa of $2^{n/2+1} - 1$ states if $n$ is even, or of $3 \cdot 2^{\lfloor n/2\rfloor} - 1$ states if $n$ is odd [32]. Since some string must lead this complete dfa to the dead state, this state is removed from the minimal incomplete dfa representation of $P(L)$.

For tightness, consider the ternary finite regular language recognized by the incomplete dfa shown in Figure 10, where $k = \lceil n/2\rceil - 1$. The application of the projection $P$ results in the language

$$P(L) = \bigcup_{i=0}^{\lceil n/2\rceil-1} (a + b)^i a(a + b)^{\lfloor n/2\rfloor-1}$$

that can be written as $P(L) = \{uav \in \{a, b\}^* \mid |uav| < n \text{ and } |v| = \lfloor n/2\rfloor - 1\}$. However, the minimal complete dfa accepting $P(L)$ has $2^{n/2+1} - 1$ states if $n$ is even, or $3 \cdot 2^{\lfloor n/2\rfloor} - 1$ states if $n$ is odd, as shown in [32]. Since $P(L)$ is finite, the minimal incomplete dfa for $P(L)$ has one less state than the complete dfa. Hence the bounds are tight. □

In the next theorem, we consider the case of projections of finite languages with co-domains of size $k$ with $k \geq 2$. In comparison with the previous result, where the sizes of the domain and co-domain differ by one, note that the size of the domain of the projection is required to be of linear size with respect to the number of states. It remains open if it can be limited by a constant.

**Theorem 11.** *Let $k, n \geq 2$. There exist alphabets $\Sigma$ and $\Sigma_o$ with $\Sigma_o \subseteq \Sigma$ and $|\Sigma_o| = k$, and a finite language $L$ over $\Sigma$ with $\|L\| = n$ such that*

$$\|P(L)\| = \frac{k^{\lfloor n/(\log k+1)\rfloor+1} - 1}{k - 1} - 1$$

*where $P$ is the projection of strings in $\Sigma^*$ onto strings in $\Sigma_o^*$. In addition, for any finite language $L'$ over $\Sigma$,*

$$\|P(L')\| \leq \frac{k^{\lceil (n/(\log k+1)\rceil+1} - 1}{k - 1} - 1.$$

*Proof.* The upper bound follows from [32, Theorem 5] in a similar way as shown in the proof of Theorem 10. To prove the lower bound, let $t = \lceil \log k\rceil$ and let $m = \lfloor n/(t+1)\rfloor$. Let $\Sigma_o = \{0, 1, \ldots, k-1\}$, let $\Sigma = \{a_1, a_2, \ldots, a_{n-m-1}\} \cup \Sigma_o$, and let $P$ be the projection of $\Sigma^*$ onto $\Sigma_o^*$.

Set $S_i = \{j \in \Sigma_o \mid j \bmod 2^i \geq 2^{i-1}\}$ for $i = 1, 2, \ldots, t$. Notice that a symbol $j$ is in $S_i$ if and only if the $i$-th digit from the end in the binary notation of $j$ is 1.

Now let $L'$ be the language over $\Sigma_o$ consisting of all strings of length $n - 1$ that have a symbol from $S_i$ in position $im$ from the end ($i = 1, 2, \ldots, t$). Language $L'$ is accepted by an $n$-state incomplete dfa $A'$ over $\Sigma_o$ with states $0, 1, \ldots, n - 1$, of which 0 is the initial state, and $n - 1$ is the sole final state.

Construct an incomplete dfa $A$ over $\Sigma$ from dfa $A'$ by adding an unobservable transition on $a_\ell$ from the initial state 0 to state $\ell$ for $\ell = 1, 2, \ldots, n - m - 1$. Let $L$ be the language over $\Sigma$ recognized by $A$. The projected language $P(L)$ consists of all suffixes of length at least $m$ of strings in $L'$. As shown in [31, 32], every complete dfa for $P(L)$ needs at least $(k^{\lfloor n/(\log k+1)\rfloor+1} - 1)/(k - 1)$ states. □

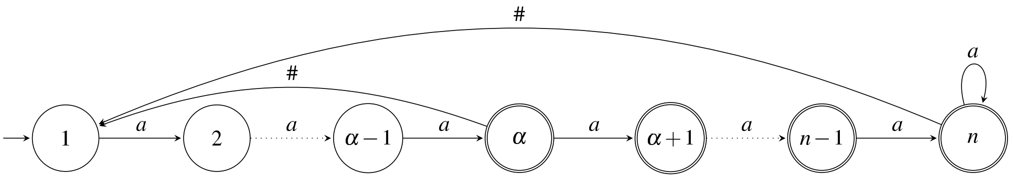


Figure 11: The minimal incomplete $n$-state dfa over $\{a, \#\}$ for a language $L$ with $\|P(L)\| = \alpha$; $1 \le \alpha \le n-2$.

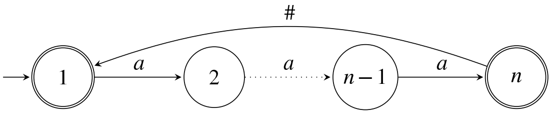


Figure 12: The minimal incomplete $n$-state dfa over $\{a, \#\}$ for a language $L$ with $\|P(L)\| = n-1$.

## 7. Magic Number Problem for Projections of Regular Languages

Here we consider the state complexity of projections not only in the worst case, but we rather ask what values may be produced as the state complexity of a projection of a regular language $L$ with $\|L\| = n$. The problem is known as Magic Number Problem in the literature [9, 11, 13, 14, 15, 16, 17, 18, 35] and values that possibly cannot be produced are called "magic numbers". This section proves that no magic numbers exist for projections of regular languages. We show that for every number $\alpha$ in the range from 1 to $2^{n-1} + 2^{n-2} - 1$, there exist a projection $P$ and a regular language $L$ with $\|L\| = n$ such that $\|P(L)\| = \alpha$. The result can be obtained using a similar result for star operation [17], however, paper [17] does not provide all the proofs. Moreover, the constructions can be simplified in the case of projections. Therefore, we provide all the proofs here. The three lemmata below deal with the following three cases:

(1) $1 \le \alpha \le n-1$;<br>
(2) $\alpha = n - k + 2^{k-1} + 2^{k-2} - 1$ for an integer $k$ with $2 \le k \le n$;<br>
(3) all the other values of $\alpha$ from $n$ to $2^{n-1} + 2^{n-2} - 1$.

**Lemma 12.** *Let $n \ge 2$, $1 \le \alpha \le n-1$, and $P$ be the projection of $\{a, \#\}^*$ onto $\{a\}^*$. There exists a regular language $L$ over $\{a, \#\}$ with $\|L\| = n$ such that $\|P(L)\| = \alpha$.*

*Proof.* If $1 \le \alpha \le n-2$, then take the minimal incomplete $n$-state dfa of Figure 11. The projected language is $\{a^i \mid i \ge \alpha - 1\}$, for which the minimal incomplete dfa has $\alpha$ states. If $\alpha = n-1$, then take the minimal incomplete dfa of Figure 12. The projected language is $(a^{n-1})^*$, for which the minimal incomplete dfa has $n-1$ states. □

**Lemma 13.** *Let $2 \le k \le n$ and $P$ be the projection of $\{a, b, c, \#\}^*$ onto $\{a, b, c\}^*$. There exists a regular language $L$ over $\{a, b, c, \#\}$ with $\|L\| = n$ such that $\|P(L)\| = n - k + 2^{k-1} + 2^{k-2} - 1$.*

*Proof.* Consider the language accepted by the minimal incomplete $n$-state dfa $B_{n,k}$ of Figure 13. Construct the $\varepsilon$-nfa for $P(L)$ by replacing the transition on # by the transition on the empty string. Let us show that $n - k + 2^{k-1} + 2^{k-2} - 1$ subsets are reachable and pairwise distinguishable in the incomplete subset automaton corresponding to this nfa. Every singleton set $\{i\}$ with $k \le i \le n-1$ is reached from the initial state $\{n-1\}$ by a string in $a^*$. We prove the reachability of all the subsets of $\{0, 1, \ldots, k-1\}$ containing state 0 by induction on the size of subsets. The set $\{0\}$ is reached from the singleton set $\{k\}$ by $a$. Every set $\{0, i_1, \ldots, i_t\}$ of size $t+1$, where $1 \le t \le k-1$ and $1 \le i_1 < i_2 < \cdots < i_t \le k-1$, is reached from the set $\{0, i_2 - i_1, \ldots, i_t - i_1\}$ of size $t$ by $ab^{i_1 - 1}$. Finally, every non-empty set $\{i_1, i_2, \ldots, i_t\}$ with $i_1 \ge 2$ is reached from the $\{0, i_1 - 1, i_2 - 1, \ldots, i_t - 1\}$ containing state 0 by $c$.

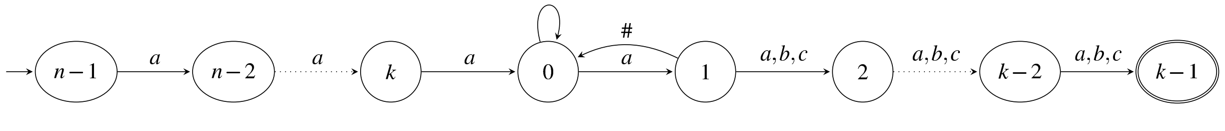


Figure 13: The minimal incomplete $n$-state dfa $B_{n,k}$ over $\{a, b, c, \#\}$ with $\|P(L(B_{n,k}))\| = n - k + 2^{k-1} + 2^{k-2} - 1$.

To prove distinguishability, notice that the string $a^i$ is accepted by the nfa for $P(L)$ only from state $k-1-i$ for $i = 0, 1, \ldots, k-2$, and the string $a^j b^k a^{k-1}$ is accepted by the nfa only from state $k-1+j$ for $j = 1, 2, \ldots, n-k$. Thus, for every state $q$, except for state 0, there exists a string $w_q$ that is accepted only from state $q$. Therefore, if two subsets differ in a state $q$ with $q \neq 0$, then the string $w_q$ distinguishes the two subsets. If $0 \in X$ and $0 \notin Y$, then $1 \notin Y$, and therefore the string $b^k a^{k-1}$ is accepted from $X$ and rejected from $Y$. This concludes our proof. □

**Lemma 14.** *Let $2 \le n \le \alpha \le 2^{n-1} + 2^{n-2} - 1$. There exists an alphabet $\Sigma$ of size $O(\alpha)$ with $\# \notin \Sigma$, a projection $P$ of $(\Sigma \cup \{\#\})^*$ onto $\Sigma^*$, and a regular languages $L$ over $\Sigma \cup \{\#\}$ with $\|L\| = n$ such that $\|P(L)\| = \alpha$.*

*Proof.* If $\alpha = n - k + 2^{k-1} + 2^{k-2} - 1$ for an integer $k$ with $2 \le k \le n$, then take the automaton $B_{n,k}$ from the previous lemma. Otherwise, number $\alpha$ is between two such values, that is,

$$n - k + 2^{k-1} + 2^{k-2} - 1 < \alpha < n - (k+1) + 2^k + 2^{k-1} - 1.$$

Then

$$\alpha = n - k + 2^{k-1} + 2^{k-2} - 1 + m,$$

where $m$ is an integer such that $1 \le m \le 2^{k-1} + 2^{k-2} - 2$. The idea is to start with automaton $B_{n,k}$, and define transitions on $m$ new symbols $d_1, d_2, \ldots, d_m$ so that every new symbol $d_i$ produces *exactly one new subset* of the form $\{k\} \cup S_i$ with $S_i \subseteq \{0, 1, \ldots, k-1\}$ in the subset automaton corresponding to the nfa for the new projected language. To guarantee that just one new subset is added by a new symbol, the new subsets $\{k\} \cup S_i$ for $i = 1, 2, \ldots, m$ will be produced according to their cardinality. To this aim, let

$$S_1, S_2, \ldots, S_\ell, \tag{1}$$

where $\ell = 2^{k-1} + 2^{k-2} - 2$, be all the non-empty subsets of $\{0, 1, \ldots, k-1\}$ containing state 0, or not containing states 0 and 1, except for the whole set $\{0, 1, \ldots, k-1\}$. The sets are ordered in such a way that $S_1 = \{0\}$, and if $|S_i| < |S_j|$ then $i < j$. Now let $S_1, S_2, \ldots, S_m$ be the first $m$ sets in sequence (1), and consider the input alphabet

$$\Sigma = \{a, b, c, \#, d_1, d_2, \ldots, d_m\}.$$

Construct minimal incomplete $n$-state dfa $C = C_{n,k,m}$ over $\Sigma$ from automaton $B_{n,k}$ of Figure 13 by adding transitions on symbols $d_1, d_2, \ldots, d_m$ in the following way. For $i = 1, 2, \ldots, m$, by $d_i$, every state in $S_i$ goes to itself, and every state in $\{0, 1, \ldots, k-1\} \setminus S_i$ goes to state $k$. In the incomplete subset automaton corresponding to projected language $P(L(C))$, the following subsets are reachable:

- the singleton sets $\{n-1\}, \{n-2\}, \ldots, \{k\}$; and all the non-empty subsets of $\{0, 1, \ldots, k-1\}$ containing state 0 or not containing states 0 and 1 since they are reached by strings over $\{a, b, c\}$ as shown in the above lemma;
- the subsets $\{k\} \cup S_i$ for $i = 1, 2, \ldots, m$ since the set $\{0, 1, \ldots, k-1\}$ goes to $\{k\} \cup S_i$ by $d_i$.

Denote the family of the above mentioned $n - k + 2^{k-1} + 2^{k-2} - 1 + m$ reachable subsets by $\mathcal{R}$, and let us show that no other subset is reachable. Since the initial state $\{n-1\}$ is in $\mathcal{R}$, it is enough to show that each subset in $\mathcal{R}$ goes to a subset in $\mathcal{R}$ or to the empty set by $a, b, c$, and by every $d_i$. This is straightforward for symbols $a, b, c$. By every $d_i$, the singleton sets $\{n-1\}, \ldots, \{k\}$ go to the empty set. Every subset $S$ of $\{0, 1, \ldots, k-1\}$ in $\mathcal{R}$ goes by $d_i$ to itself if $S \subseteq S_i$, and to $\{k\} \cup (S_i \cap S)$ otherwise. If $S_i \cap S$ is not $S_i$, then it either is empty, or is non-empty with smaller cardinality than $S_i$, and either contains state 0, or does not contain states 0 and 1. Therefore, if it is non-empty, then it precedes $S_i$ in our sequence (1), and the set $\{k\} \cup (S_i \cap S)$ is in family $\mathcal{R}$. Similarly, every set $\{k\} \cup S_j$ $(1 \le j \le m)$ goes by $d_i$ either to the empty set, or to $S_j$, or to $\{k\} \cup S_i$, or to $\{k\} \cup S_t$ with $t < i$.

Two distinct subsets in $\mathcal{R}$ can be distinguished by strings over $\{a, b, c\}$ in the same way as in the previous lemma, and our proof is complete. □

Putting the above three lemmata together, we get the following result showing that no magic numbers exist for projections. However, the alphabet used to prove this result grows exponentially with $n$.

**Theorem 15.** *Let $n \ge 2$ and $1 \le \alpha \le 2^{n-1} + 2^{n-2} - 1$. There exist an alphabet $\Sigma_n$ with $\# \notin \Sigma_n$, the size of which grows exponentially with n, a projection $P$ of strings in $(\Sigma_n \cup \{\#\})^*$ onto strings in $\Sigma_n^*$, and a regular language $L$ over $\Sigma_n \cup \{\#\}$ with $\|L\| = n$ such that $\|P(L)\| = \alpha$.* □

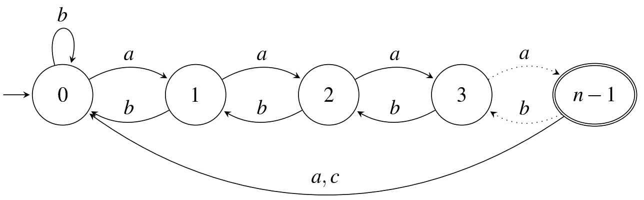


Figure 14: A dfa for a union-free language meeting the exponential upper bound on the number of states for projection.

## 8. Conclusions and Future Directions

The results for finite languages immediately imply the reachable exponential upper bounds for sub-regular languages such as definite languages, strictly locally testable languages, locally testable languages, generalized definite languages, ordered languages, star-free languages, and power separating languages, see [1] for the definitions and more information. In [19] we studied union-free languages. The next result shows that the upper bound for projections can also be met by union-free languages.

**Theorem 16.** *Let $n \geq 2$ and $P$ be the projection of $\{a, b, c\}^*$ onto $\{a, b\}^*$. There exists a union-free regular language $L$ over $\{a, b, c\}$ with $\|L\| = n$ such that $\|P(L)\| = 2^{n-1} + 2^{n-2} - 1$.*

*Proof.* Consider the language $L$ given by the dfa of Figure 14. Then, the application of projection $P$ from $\{a, b, c\}^*$ to $\{a, b\}^*$ results in an nfa of Theorem 5(8) of [19] without the initial state $q_0$. It is shown there that the nfa-to-dfa conversion results in a dfa with $2^{n-1} + 2^{n-2}$ states. As the initial state is missing, we get $2^{n-1} + 2^{n-2} - 1$ states. □

The dfa accepting a projected language is obtained from the dfa accepting an input language by replacing unobservable transitions with $\varepsilon$-transitions and by applying the subset construction to the resulting nfa. The minimal dfa for the projected language, however, may be of exponential size in comparison with the input automaton [12, 25, 27, 28]. This observation gives rise to a challenging open problem. How can we characterize classes of dfa's, for which the minimal dfa for the projections is of a linear (polynomial, logarithmic) size?

**Problem 1.** Let $P$ be a projection, and let $\mathbb{A}_P^f$ denote the class of all minimal dfa's such that $A \in \mathbb{A}_P^f$ if and only if the minimal dfa accepting $P(L(A))$ has no more than $f(n)$ states, where $f$ is a (recursive) upper bound state-space function. Given a projection $P$ and a function $f$, characterize the class $\mathbb{A}_P^f$.

It follows from the results of this paper that the class $\mathbb{A}_P^f$ does not include all minimal acyclic dfa's for any reasonable upper bound $f$ (such as linear or polynomial). Note that there exists a property called an *observer property* [36] ensuring that the minimal automaton for the projected language has no more states than the minimal automaton for the input language, see also [29]. This property is well known and widely used in supervisory control of hierarchical and distributed discrete-event systems, and, as mentioned in [30], also in compositional verification [8] and modular synthesis [6, 10]. If the projection does not satisfy the property, the co-domain of the projection can be extended so that it satisfies it. However, the computation of such a minimal extension is NP-hard. Nevertheless, there exists a polynomial-time algorithm that finds an acceptable extension [7]. A different approach with further references can be found in [30]. Although we know that the result is of polynomial size, the problem is how to compute it in polynomial time. Consider the determinization procedure of an nfa. This procedure can produce an exponential number of states where most of the states are equivalent. In [36], a polynomial-time algorithm running in $O(n^7m^2)$, where $n$ is the number of states and $m$ is the cardinality of the co-domain of the projection satisfying the observer property, has been proposed. However, the precise time complexity of this problem is open.

**Problem 2.** How to compute the minimal dfa accepting the projected language when the projection satisfies the observer property?

## Acknowledgement

We would like to thank Professor J. H. van Schuppen for his useful comments on the previous version of this work.

## References

[1] H. Bordihn, M. Holzer, M. Kutrib, Determination of finite automata accepting subregular languages, Theoret. Comput. Sci. 410 (2009) 3209–3222.

[2] O. Boutin, J. Komenda, T. Masopust, K. Schmidt, J.H. van Schuppen, Hierarchical control with partial observations: Sufficient conditions, in: Proc. of IEEE Conference on Decision and Control and European Control Conference (CDC-ECC 2011), Orlando, USA, pp. 1817–1822.

[3] O. Boutin, J.H. van Schuppen, On the Control of the Paint Factory Scale Model, Technical Report MAC-1103, CWI, Amsterdam, 2011. [Online]. Available: http://oai.cwi.nl/oai/asset/18598/18598D.pdf.

[4] C.G. Cassandras, S. Lafortune, Introduction to Discrete Event Systems, Springer, second edition, 2008.

[5] M. Chrobak, Finite automata and unary languages, Theoret. Comput. Sci. 47 (1986) 149–158. Errata: Theoret. Comput. Sci. 302 (2003) 497-498.

[6] L. Feng, W.M. Wonham, Computationally efficient supervisor design: Abstraction and modularity, in: Proc. of Workshop on Discrete Event Systems (WODES 2006), Ann Arbor, USA, pp. 3–8.

[7] L. Feng, W.M. Wonham, On the computation of natural observers in discrete-event systems, Discrete Event Dyn. Syst. 20 (2010) 63–102.

[8] H. Flordal, R. Malik, Compositional verification in supervisory control, SIAM J. Control Optim. 48 (2009) 1914–1938.

[9] V. Geffert, Magic numbers in the state hierarchy of finite automata, Inform. Comput. 205 (2007) 1652–1670.

[10] R.C. Hill, D.M. Tilbury, Modular supervisory control of discrete event systems with abstraction and incremental hierarchical construction, in: Proc. of Workshop on Discrete Event Systems (WODES 2006), Ann Arbor, USA, pp. 399–406.

[11] M. Holzer, S. Jakobi, M. Kutrib, The magic number problem for subregular language families, in: Proc. of Descriptional Complexity of Formal Systems (DCFS 2010), volume 31 of *EPTCS*, Saskatoon, Canada, pp. 110–119.

[12] M. Holzer, M. Kutrib, Descriptional complexity – an introductory survey, in: Scientific Applications of Language Methods, volume 2, Imperial College Press, 2010.

[13] K. Iwama, Y. Kambayashi, K. Takaki, Tight bounds on the number of states of DFAs that are equivalent to $n$-state NFAs, Theoret. Comput. Sci. 237 (2000) 485–494.

[14] K. Iwama, A. Matsuura, M. Paterson, A family of NFAs which need $2^n - \alpha$ deterministic states, Theoret. Comput. Sci. 301 (2003) 451–462.

[15] J. Jirásek, G. Jirásková, A. Szabari, Deterministic blow-ups of minimal nondeterministic finite automata over a fixed alphabet, Internat. J. Found. Comput. Sci. 19 (2008) 617–631.

[16] G. Jirásková, Note on minimal finite automata, in: Proc. of Mathematical Foundations of Computer Science (MFCS 2001), volume 2136 of *Lecture Notes in Comput. Sci.*, Springer, Mariánské Lázně, Czech Republic, 2001, pp. 421–431.

[17] G. Jirásková, On the state complexity of complements, stars, and reversals of regular languages, in: Proc. of Developments in Language Theory (DLT 2008), volume 5257 of *Lecture Notes in Comput. Sci.*, Springer, Kyoto, Japan, 2008, pp. 431–442.

[18] G. Jirásková, Magic numbers and ternary alphabet, Internat. J. Found. Comput. Sci. 22 (2011) 331–344.

[19] G. Jirásková, T. Masopust, Complexity in union-free regular languages, Internat. J. Found. Comput. Sci. 22 (2011) 1639–1653.

[20] G. Jirásková, T. Masopust, State complexity of projected languages, in: Proc. of Descriptional Complexity of Formal Systems (DCFS 2011), volume 6808 of *Lecture Notes in Comput. Sci.*, Springer, Gießen/Limburg, Germany, 2011, pp. 198–211.

[21] J. Komenda, T. Masopust, J.H. van Schuppen, Coordinated control of discrete event systems with nonprefix-closed languages, in: Proc. of World Congress of the International Federation of Automatic Control (IFAC WC 2011), volume 18, Milano, Italy, pp. 6982–6987.

[22] J. Komenda, T. Masopust, J.H. van Schuppen, Synthesis of controllable and normal sublanguages for discrete-event systems using a coordinator, Systems Control Lett. 60 (2011) 492–502.

[23] J. Komenda, T. Masopust, J.H. van Schuppen, Supervisory control synthesis of discrete-event systems using a coordination scheme, Automatica 48 (2012) 247–254.

[24] J. Komenda, J.H. van Schuppen, Coordination control of discrete event systems, in: Proc. of Workshop on Discrete Event Systems (WODES 2008), Göteborg, Sweden, pp. 9–15.

[25] O.B. Lupanov, Über den vergleich zweier typen endlicher quellen, Probl. Kybernetik 6 (1966) 328–335. Translation from Probl. Kibernetiki 9, 321-326 (1963).

[26] Y.I. Lyubich, Estimates for optimal determinization of nondeterministic autonomous automata, Sib. Matemat. Zhu. 5 (1964) 337–355. In Russian.

[27] A.R. Meyer, M.J. Fischer, Economy of description by automata, grammars, and formal systems, in: Proc. of Symposium on Switching and Automata Theory (SWAT 1971), East Lansing, USA, pp. 188–191.

[28] F.R. Moore, On the bounds for state-set size in the proofs of equivalence between deterministic, nondeterministic, and two-way finite automata, IEEE Trans. Comput. 20 (1971) 1211–1214.

[29] P.N. Pena, J.E.R. Cury, S. Lafortune, Polynomial-time verification of the observer property in abstractions, in: Proc. of American Control Conference (ACC 2008), Seattle, USA, pp. 465–470.

[30] P.N. Pena, J.E.R. Cury, R. Malik, S. Lafortune, Efficient computation of observer projections using OP-verifiers, in: Proc. of Workshop on Discrete Event Systems (WODES 2010), Berlin, Germany, pp. 416–421.

[31] K. Salomaa, NFA to DFA conversion for finite languages over a $k$-letter alphabet, 2011. Personal communication.

[32] K. Salomaa, S. Yu, NFA to DFA transformation for finite languages, in: Proc. of Workshop on Implementing Automata (WIA 1996), volume 1260 of *Lecture Notes in Comput. Sci.*, Springer, London, Canada, 1996, pp. 149–158.

[33] J.H. van Schuppen, O. Boutin, P.L. Kempker, J. Komenda, T. Masopust, N. Pambakian, A.C.M. Ran, Control of distributed systems: Tutorial and overview, Eur. J. Control 17 (2011) 579–602.

[34] M. Sipser, Introduction to the Theory of Computation, PWS Publishing Company, Boston, USA, 1997.

[35] A. Szabari, Descriptional Complexity of Regular Languages, Ph.D. thesis, Mathematical Institute, Slovak Academy of Sciences, Košice, Slovakia, 2010.

[36] K. Wong, On the complexity of projections of discrete-event systems, in: Proc. of Workshop on Discrete Event Systems (WODES 1998), Cagliari, Italy, pp. 201–206.

[37] W.M. Wonham, Supervisory control of discrete-event systems, 2011. Lecture Notes, Dept. of Electrical and Computer Engineering, Univ. of Toronto, Canada, [Online]. Available: http://www.control.utoronto.ca/cgi-bin/dldes.cgi.