

# Register Set Automata (Technical Report)

Sabína Gulčíková ✉

Faculty of Information Technology, Brno University of Technology, Czech Republic

Ondřej Lengál ✉ 

Faculty of Information Technology, Brno University of Technology, Czech Republic

## Abstract

We present *register set automata* (RsAs), a register automaton model over data words where registers can contain *sets* of data values and the following operations are supported: **adding values to registers**, **clearing registers**, and **testing (non-)membership**. We show that the emptiness problem for RsAs is decidable and complete for the  $\mathbf{F}_\omega$  class. Moreover, we show that a large class of *register automata* can be transformed into *deterministic* RsAs, which can serve as a basis for (i) fast matching of a family of regular expressions with back-references and (ii) language inclusion algorithm for a subclass of register automata. **RsAs are incomparable in expressive power to other popular automata models over data words, such as alternating register automata and pebble automata.**

**2012 ACM Subject Classification** Theory of computation → Automata over infinite objects; Theory of computation → Modal and temporal logics

**Keywords and phrases** register automata, register set automata, data words, deterministic automata

**Digital Object Identifier** 10.4230/LIPIcs.CVIT.2016.23

## 1 Introduction

Automata over infinite alphabets have found applications in many areas of computer science, for instance in modelling network protocols [5], reasoning about XML (XML schema definition via XSD, XML transformation via XSLT) [31], pattern matching with back-references [38], runtime verification [22], verification of MapReduce programs [8, 7], or even testing satisfiability in the SMT theory of strings [6].

A popular formal model for languages over infinite alphabets are the so-called register automata. (Nondeterministic) *register automata* (RAs, NRAs) are an extension of (non-deterministic) finite automata (FAs, NFAs) equipped with a finite set of registers, each of which can store a single value copied from the input tape (which contains symbols from an infinite data domain  $\mathbb{D}$ ). On transitions, the automata can then test (non-)equality of the current input symbol and values stored inside the registers, and save the input symbol into some of the registers. Compared to NFAs, the complexity of decision problems for NRAs is much higher, for example, language emptiness, which is solvable in linear time for NFAs, is PSPACE-complete for NRAs. Furthermore, universality (“Does an automaton accept all words?”), language inclusion, and equivalence, which is PSPACE-complete for NFAs, are, in general, undecidable for NRAs [32]. The models also significantly differ in their expressive power: NFAs can be easily determined but NRAs cannot. For instance, the language of strings containing two occurrences of the same data value can be accepted by an NRA but by no DRA.

In this paper, we study *register set automata* (RsAs), a formal model that extends RAs with the possibility to store *sets of data values* into registers.<sup>1</sup> Our study of the model is motivated by the fact that *deterministic* RsAs can capture a quite large fragment of NRA languages. The need for deterministic automata models comes (not only) from two applications: (i) regex matching with back-references and (ii) checking language inclusion or equivalence of NRAs, which are both described in more detail below.

<sup>1</sup>The RsA model is inspired by the model of *counting set automata* of Turoňová *et al.* [48], which allow compact deterministic representation of a subclass of NFAs extended with bounded counters.



**Matching Regexes with Back-references.** *Pattern matching using regular expressions with back-references* is performed ubiquitously, e.g., in validating user inputs on web pages, processing text using the `grep` and `sed` tools, transforming XML documents, or detecting network incidents [36, 3]. Consider, for instance, the (extended) regular expression (regex)

$$R_{\setminus 3 \setminus 2 \setminus 1} = /(.) .*; .*(.) .*; .*(.) .* \setminus 3 \setminus 2 \setminus 1 /,$$

where we use the wildcard "." to denote any symbol except the semicolon ";" (i.e., it stands for the character class "[^;]"), parentheses "(...)" to denote the so-called *capture groups*, and "\x" to denote a *back-reference* to the string captured by the x'th capture group. Intuitively, the regex matches input strings  $w$  that can be seen as a concatenation of three strings  $w = uvz$  such that  $u$  has the structure  $u = u_1 ; u_2 ; u_3$  with  $u_1, u_2, u_3 \in (\Sigma \setminus \{;\})^+$ ,  $v$  is a string of three characters  $a_3 a_2 a_1$  such that  $a_i \in u_i$  for  $i \in \{1, 2, 3\}$ , and  $z \in (\Sigma \setminus \{;\})^*$  (such a regex can describe, e.g., a simplified version of the match rule of some XML transformation). Trying to find a match of the regex in the randomly generated 42-character-long string

"ah;jk2367ash;la5akv451wkjb9f.dj5fqkbsfyrf"

using the state-of-the-art PCRE2 regex matcher available at regex101.com [34] takes 10,416 steps before reporting *no match*.<sup>2</sup> Ideally, the matcher should take only 42 steps, one step for every character in the string. This  $\frac{10,416}{42} = 248\times$  slowdown is caused by the so-called *catastrophic backtracking*—the PCRE2 matcher is based on backtracking and, since the regex is nondeterministic, the backtracking algorithm needs to try all possibilities of placing the three capture groups before concluding that there is no match.

If such a scenario with catastrophic backtracking happens, e.g., in a web application by providing a malicious user input tailored to take advantage of a feature of a regex that makes it perform poorly, this can have serious consequences: the web may become nonresponsive and unusable. In industry, this is called a *regular expression denial of service* (ReDoS) attack [1]. ReDoS is a real-world threat; it caused, for instance, the 2016 outage of StackOverflow [16] or rendered vulnerable websites that were using the Express.js framework [2].

A practical way how to avoid poor performance of regex matchers is to use a matching algorithm with no backtracking. Indeed, the majority of industrial-strength regex matchers avoid backtracking by using a matching algorithm based on *deterministic* FAs (DFAs), which guarantee constant per-symbol matching time<sup>3</sup>. Such algorithms are, however, not available when back-references are used, witnessed by the fact that neither of the two currently most advanced regex matchers in the industry, RE2 [20] and HyperScan [49], supports back-references, due to a missing efficient expressive deterministic automaton model [24, 21].

**Language Inclusion of Register Automata.** Testing RA inclusion is an essential problem in many of their applications, such as in their minimization, learning [5, 26, 19], checking for fixpoint in regular model checking [6], checking XML schema subsumption [47], verification of parameterized concurrent programs with shared memory [25], and is an essential component of the RA toolkit. Unfortunately, inclusion of RAs is in general an undecidable problem [32], which has forced researchers to either find ways to approximate the inclusion test—e.g., via several membership tests [26, 19] or by an abstraction refinement semi-algorithm using interpolation [25]—or restrict themselves to other models with decidable inclusion problem, such as deterministic RAs (whose expressive power is quite limited) [29] or session automata [5].

<sup>2</sup>Finding a match of the regex  $/(.) .*; .*(.) .*; .*(.) .* \setminus 3 \setminus 2 \setminus 1 /$  in the same text took 179,372 steps before *no match* was reported. The regex is more challenging—it does not use delimiters (semicolons in  $R_{\setminus 3 \setminus 2 \setminus 1}$ ).

<sup>3</sup>In practice, the matchers are based on Thompson's algorithm [46], which does not build the (possibly prohibitively large) DFA *a priori* but, instead, on the fly, while using cache to store already constructed parts of the DFA. This gives almost constant per-symbol matching time for usual inputs if the cache is utilized well, malicious inputs can, however, cause high cache miss ratio and performance degradation.

The standard approach to testing the language inclusion  $\mathcal{L}(\mathcal{A}) \subseteq \mathcal{L}(\mathcal{B})$  is to (i) determinise  $\mathcal{B}$  into  $\mathcal{B}^D$ , (ii) complement  $\mathcal{B}^D$  into  $\overline{\mathcal{B}^D}$ , and (iii) test emptiness of the intersection of  $\mathcal{A}$  and  $\overline{\mathcal{B}^D}$ . Here, the critical point is already the *determinisation* part: RAs cannot always be determinised (and it is undecidable whether a particular RA *can be determinised* [9, 10]).

The previous two applications show a strong need for expressive deterministic models of automata with registers. We address this need by introducing *register set automata* (RsAs), an extension of RAs where every register can hold a *set of values* instead of a single one. This model is quite powerful: it strictly generalizes RAs and is incomparable to (one-way) alternating RAs, another popular powerful generalization of RAs. The core property of RsAs is that their emptiness problem is decidable, although for a higher price than for RAs—the complexity blows from PSPACE-complete for RAs to  $\mathbf{F}_\omega$ -complete (i.e., Ackermannian) for RsAs. Furthermore, RsAs are practically relevant: many NRAs (e.g., the NRA for the regex  $R_{\setminus 3 \setminus 2 \setminus 1}$ ) can be algorithmically determinised into *deterministic* RsAs (DRsAs), which, together with the emptiness decidability, gives us an algorithm for testing language inclusion of a large class of RAs (the last missing component—complementing a DRsA—is done easily by swapping final and non-final states).

**Contribution.** We list the main contributions of the paper:

1. introduction of the model of register set automata and showing its closure properties,
2. proving  $\mathbf{F}_\omega$ -completeness of the emptiness problem for RsAs by showing interreducibility with the coverability problem for transfer Petri nets,
3. designing a (semi-)algorithm that can determinise an RA into a DRsA and showing that it is complete for the class of languages that can be obtained from RAs with one register and no disequality tests by Boolean operations (i.e., union, intersection, and complement),
4. placing the RsA model into the landscape of automata-with-registers models, and
5. discussing the power of several extensions of the model.

## 2 Preliminaries

We use  $\mathbb{N}$  to denote the set of natural numbers without 0,  $\mathbb{N}_0$  to denote  $\mathbb{N} \cup \{0\}$ , and  $[n]$  for  $n \in \mathbb{N}$  to denote the set  $\{1, \dots, n\}$  (we note that  $[0] = \emptyset$ ). We sometimes use “.” to denote an *ellipsis*, i.e., a value that can be ignored.

**Data Words.** Let us fix a finite nonempty *alphabet*  $\Sigma$  and an infinite *data domain*  $\mathbb{D}$ . A (finite) *data word* of *length*  $n$  is a function  $w: [n] \rightarrow (\Sigma \times \mathbb{D})$ ; we use  $|w| = n$  to denote its length and  $w_1, \dots, w_n$  to denote its symbols. The *empty word* of length 0 is denoted  $\epsilon$ . We use  $\Sigma[w]$  and  $\mathbb{D}[w]$  to denote the *projection* of  $w$  onto the respective domain (e.g., if  $w = \langle a, 1 \rangle \langle b, 2 \rangle \langle b, 3 \rangle$ , then  $\Sigma[w] = abc$  and  $\mathbb{D}[w] = 123$ ) and, given  $a \in \Sigma$ , we use  $a[w_i]$  as a shortcut for  $\Sigma[w_i] = a$ .

**Register Automata on Data Words.** A (nondeterministic one-way) *register automaton* (on data words), abbreviated as (N)RA, is a tuple  $\mathcal{A} = (Q, \mathbf{R}, \Delta, I, F)$  where  $Q$  is a finite set of *states*,  $\mathbf{R}$  is a finite set of *registers*,  $I \subseteq Q$  is a set of *initial states*,  $F \subseteq Q$  is a set of *final states*, and  $\Delta \subseteq Q \times \Sigma \times 2^{\mathbf{R}} \times 2^{\mathbf{R}} \times (\mathbf{R} \rightarrow \mathbf{R} \cup \{in, \perp\}) \times Q$  is a *transition relation* such that if  $t: (q, a, g^=, g^\neq, up, s) \in \Delta$ , then  $g^= \cap g^\neq = \emptyset$ . We use  $q \xrightarrow{a \mid g^=, g^\neq, up} s$  to denote  $t$  (and often drop from  $up$  mappings  $r \mapsto r$  for  $r \in \mathbf{R}$ , which we treat as implicit). The semantics of  $t$  is that  $\mathcal{A}$  can move from state  $q$  to state  $s$  if the  $\Sigma$ -symbol at the current position of the input word is  $a$  and the  $\mathbb{D}$ -value at the current position is equal to all registers from  $g^=$  and not equal to any register from  $g^\neq$ ; the content of the registers is updated so that  $r_i \leftarrow up(r_i)$  (i.e.,  $r_i$  can be assigned the value of some other register, the current  $\mathbb{D}$ -symbol, denoted by  $in$ , or it can be cleared by being assigned  $\perp$ ).

A *configuration* of  $\mathcal{A}$  is a pair  $c \in Q \times (\mathbf{R} \rightarrow \mathbb{D} \cup \{\perp\})$ , i.e., it consists of a state and an assignment of data values to registers. An *initial configuration* of  $\mathcal{A}$  is a pair  $c_{init} \in I \times \{\{r \mapsto \perp \mid r \in \mathbf{R}\}\}$ . Suppose  $c_1 = (q_1, f_1)$  and  $c_2 = (q_2, f_2)$  are two configurations of  $\mathcal{A}$ . We say that  $c_1$  can make a *step* to  $c_2$  over  $\langle a, d \rangle \in \Sigma \times \mathbb{D}$  using transition  $t: q - \langle a \mid g^-, g^+, up \rangle \rightarrow s \in \Delta$ , denoted as  $c_1 \vdash_t^{(a,d)} c_2$ , iff

1.  $d = f_1(r)$  for all  $r \in g^-$ ,
2.  $d \neq f_1(r)$  for all  $r \in g^+$ , and
3. for all  $r \in \mathbf{R}$ , we have  $f_2(r) = \begin{cases} f_1(r') & \text{if } up(r) = r' \in \mathbf{R}, \\ d & \text{if } up(r) = in, \text{ and} \\ \perp & \text{if } up(r) = \perp. \end{cases}$

A *run*  $\rho$  of  $\mathcal{A}$  over the word  $w = \langle a_1, d_1 \rangle \dots \langle a_n, d_n \rangle$  from a configuration  $c$  is a sequence of alternating configurations and transitions  $\rho = c_0 t_1 c_1 t_2 \dots t_n c_n$  such that  $\forall 1 \leq i \leq n: c_{i-1} \vdash_{t_i}^{(a_i, d_i)} c_i$  and  $c_0 = c$ . We say that  $\rho$  is accepting if  $c$  is an initial configuration,  $c_n = (s, f)$ , and  $s \in F$ . The language accepted by  $\mathcal{A}$ , denoted as  $\mathcal{L}(\mathcal{A})$ , is defined as  $\mathcal{L}(\mathcal{A}) = \{w \in (\Sigma \times \mathbb{D})^* \mid \mathcal{A} \text{ has an accepting run over } w\}$ .

We say that  $\mathcal{A}$  is a *deterministic RA* (DRA) if for all states  $q \in Q$  and all  $a \in \Sigma$ , it holds that for any two distinct transitions  $q - \langle a \mid g_1^-, g_1^+, up_1 \rangle \rightarrow s_1, q - \langle a \mid g_2^-, g_2^+, up_2 \rangle \rightarrow s_2 \in \Delta$  we have that  $g_1^- \cap g_2^+ \neq \emptyset$  or  $g_2^- \cap g_1^+ \neq \emptyset$ .  $\mathcal{A}$  is *complete* if for all states  $q \in Q$ , symbols  $a \in \Sigma$ , and  $g \subseteq \mathbf{R}$ , there is a transition  $q - \langle a \mid g^-, g^+, up \rangle \rightarrow s$  such that  $g^- \subseteq g$  and  $g \cap g^+ = \emptyset$ .

**Universal RAs.** A *universal RA* (URA)  $\mathcal{A}_U$  is defined exactly as an NRA with the exception of its language. The language of  $\mathcal{A}_U$  is the set  $\mathcal{L}(\mathcal{A}_U) = \{w \in (\Sigma \times \mathbb{D})^* \mid \text{every run of } \mathcal{A}_U \text{ on } w \text{ is accepting}\}$  (we emphasize that if a run cannot continue from some state over the current input symbol, it is not accepting).

There is indeed duality between NRAs and URAs, as stated by the following fact.

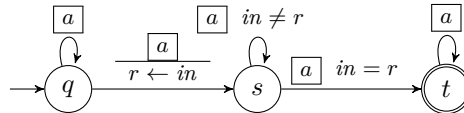
► **Fact 1.** For every NRA  $\mathcal{A}_N$ , there is a URA accepting the complement of  $\mathcal{L}(\mathcal{A}_N)$ . Conversely, for every URA  $\mathcal{A}_U$ , there is an NRA accepting the complement of  $\mathcal{L}(\mathcal{A}_U)$ .

**Proof.** For both parts, the complement automaton is obtained by (i) adding a rejecting *sink* state to the automaton, (ii) completing the transition relation (i.e., adding transitions for undefined combinations of symbols and guards to the sink state) of the input automaton, and (iii) swapping final and non-final states. ◀

► **Example 2.** Consider the language of words over  $\Sigma = \{a\}$  that contain two occurrences of some data value, i.e., the language

$$L_{\exists repeat} = \{w \mid \exists i, j: i \neq j \wedge \mathbb{D}[w_i] = \mathbb{D}[w_j]\}.$$

An NRA recognising this language is in the following figure:



Formally, it is an NRA  $\mathcal{A} = (\{q, s, t\}, \{r\}, \Delta, \{q\}, \{t\})$  with  $\Delta = \{q - \langle a \mid \emptyset, \emptyset, \{r \mapsto in\} \rangle \rightarrow s, q - \langle a \mid \emptyset, \emptyset, \emptyset \rangle \rightarrow q, s - \langle a \mid \emptyset, \{r\}, \emptyset \rangle \rightarrow s, s - \langle a \mid \{r\}, \emptyset, \emptyset \rangle \rightarrow t, t - \langle a \mid \emptyset, \emptyset, \emptyset \rangle \rightarrow t\}$  (actually, the guard  $in \neq r$  on the self-loop over  $s$  is redundant). We note that this language is not expressible by any DRA or URA. Intuitively,  $\mathcal{A}$  waits in  $q$  until it nondeterministically guesses the input data value that should be repeated, stores it into register  $r$ , and moves to state  $s$ . In state  $s$ , it is waiting to see the data value again, upon which it moves to the accepting state  $t$  and reads out the rest of the word.

On the other hand, the complement of the language, i.e., the language of words where no two positions have the same data value, formally,

$$L_{\neg \exists \text{repeat}} = \{w \mid \forall i, j: i \neq j \implies \mathbb{D}[w_i] \neq \mathbb{D}[w_j]\},$$

is not expressible by any DRA or NRA, but is expressible by a URA. The URA accepting  $L_{\neg \exists \text{repeat}}$  looks similar to the NRA above with the exception of final states, which are  $\{q, s\}$  (note that in URAs, in order to accept a word, all runs over the word need to accept, so in order to accept in this example, all runs of the URA need to avoid the state  $t$ ). ◀

**Alternating RAs.** Intuitively, *alternating RAs* (ARAs) allow to combine the accepting conditions of NRAs and URAs in a fine-grained way (states are marked *existential* or *universal* to denote whether, in order to accept from the state, *at least one* run or *all* runs respectively need to accept). Since in this paper we do not explicitly work with ARAs and only compare to ARAs in their expressive power, we avoid giving their quite involved formal definition and point the interested reader to the definition of one-way ARAs in the paper of Demri and Lazić [12].

**Alternating RAs with guess and spread.** ARAs can also be extended with the two following operations: *guess* and *spread*. The *guess* operation allows a thread to assign a nondeterministically chosen data value into some register (in each thread a different one). On the other hand, the *spread* operation allows to make a certain kind of universal quantification over the data values seen so far on the run of the automaton (even in different threads). ARAs extended with both these operations are denoted as  $\text{ARA}(\mathbf{g}, \mathbf{s})$ . As with ARAs, we do not directly work with this extension and only compare with its expressive power. The interested reader can find the formal definition in the paper of Figueira [17].

We use  $\text{NRA}^-$ ,  $\text{URA}^-$ , and  $\text{DRA}^-$  to denote the sub-classes of NRAs, URAs, and DRAs with *no disequality* guards, i.e., automata where for every transition  $q - (a \mid g^=, g^\neq, up) \rightarrow s$  it holds that  $g^\neq = \emptyset$ . Furthermore, for a class  $\mathcal{C}$  of automata with registers and  $n \in \mathbb{N}$ , we use  $\mathcal{C}_n$  to denote the sub-class of  $\mathcal{C}$  containing automata with at most  $n$  registers (e.g.,  $\text{DRA}_2$ ). We abuse notation and use  $\mathcal{C}$  to also denote the class of languages defined by  $\mathcal{C}$ .

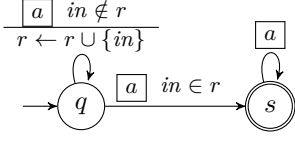
### 3 Register Set Automata

A (nondeterministic) *register set automaton* (on data words), abbreviated as (N)RsA is a tuple  $\mathcal{A}_S = (Q, \mathbf{R}, \Delta, I, F)$  where  $Q, \mathbf{R}, I, F$  are the same as for RAs and  $\Delta \subseteq Q \times \Sigma \times 2^{\mathbf{R}} \times 2^{\mathbf{R}} \times (\mathbf{R} \rightarrow 2^{\mathbf{R} \cup \{in\}}) \times Q$  such that if  $q - (a \mid g^=, g^\neq, up) \rightarrow s \in \Delta$ , then  $g^= \cap g^\neq = \emptyset$  (as with NRAs, we often do not write mappings  $r \mapsto \{r\}$  for  $r \in \mathbf{R}$  when defining  $up$ ). The semantics of a transition  $q - (a \mid g^=, g^\neq, up) \rightarrow s$  is that  $\mathcal{A}_S$  can move from state  $q$  to state  $s$  if the  $\Sigma$ -symbol at the current position of the input word is  $a$  and the  $\mathbb{D}$ -value at the current position is in all registers from  $g^=$  and in no register from  $g^\neq$ ; the content of the registers is updated so that  $r_i \leftarrow \bigcup \{x \mid x \in up(r_i)\}$  (i.e.,  $r_i$  can be assigned the union of values of several registers, possibly including the current  $\mathbb{D}$ -symbol denoted by  $in$ ).

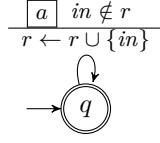
A *configuration* of  $\mathcal{A}_S$  is a pair  $c \in Q \times (\mathbf{R} \rightarrow 2^{\mathbb{D}})$ , i.e., it consists of a state and an assignment of sets of data values to registers. An *initial configuration* of  $\mathcal{A}_S$  is a pair  $c_{init} \in I \times \{\{r \mapsto \emptyset \mid r \in \mathbf{R}\}\}$ . Suppose  $c_1 = (q_1, f_1)$  and  $c_2 = (q_2, f_2)$  are two configurations of  $\mathcal{A}_S$ . We say that  $c_1$  can make a *step* to  $c_2$  over  $\langle a, d \rangle \in \Sigma \times \mathbb{D}$  using transition  $t: q - (a \mid g^=, g^\neq, up) \rightarrow s \in \Delta$ , denoted as  $c_1 \vdash_t^{(a,d)} c_2$ , iff

1.  $d \in f_1(r)$  for all  $r \in g^=$ ,
2.  $d \notin f_1(r)$  for all  $r \in g^\neq$ , and
3. for all  $r \in \mathbf{R}$ , we have  $f_2(r) = \bigcup \{f_1(r') \mid r' \in \mathbf{R}, r' \in up(r)\} \cup \begin{cases} \{d\} & \text{if } in \in up(r) \text{ and} \\ \emptyset & \text{otherwise.} \end{cases}$

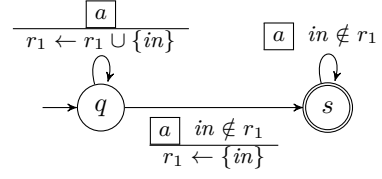
The definition of a run and language of  $\mathcal{A}_S$  is then the same as for NRAs.



■ **Figure 1** DRsA<sub>1</sub> for  $L_{\exists repeat}$



■ **Figure 2** DRsA<sub>1</sub> for  $L_{\neg \exists repeat}$



■ **Figure 3** RsA<sub>1</sub> for  $L_{\neg \forall repeat}$

We say that the RsA  $\mathcal{A}_S$  is *deterministic* (DRSA) if for all states  $q \in Q$  and all  $a \in \Sigma$ , it holds that for any two distinct transitions  $q \xrightarrow{a \mid g_1^\epsilon, g_1^\zeta, up_1} s_1, q \xrightarrow{a \mid g_2^\epsilon, g_2^\zeta, up_2} s_2 \in \Delta$  we have that  $g_1^\epsilon \cap g_2^\zeta \neq \emptyset$  or  $g_2^\epsilon \cap g_1^\zeta \neq \emptyset$ .

► **Example 3.** A DRsA accepting the language  $L_{\exists repeat}$  from Example 2 is in Figure 1. Formally, it is a DRsA<sub>1</sub>  $\mathcal{A} = (\{q, s\}, \{r\}, \Delta, \{q\}, \{s\})$  where  $\Delta = \{q \xrightarrow{a \mid \emptyset, \{r\}, \{r \mapsto \{r, in\}\}} q, q \xrightarrow{a \mid \{r\}, \emptyset, \emptyset} s, s \xrightarrow{a \mid \emptyset, \emptyset, \emptyset} s\}$ . Intuitively, the DRsA waits in  $q$  and accumulates the so far seen input data values in register  $r$  (we use  $r \leftarrow r \cup \{in\}$  to denote the update  $r \mapsto \{r, in\}$ ). Once the DRsA reads a value that is already in  $r$ , it moves to  $s$  and accepts. ◀

► **Example 4.** A DRsA<sub>1</sub> accepting the language  $L_{\neg \exists repeat}$  from Example 2 is in Figure 2. Intuitively, the automaton stays in state  $q$  and accumulates input data values in register  $r$ , making sure the input data value has not been seen previously. ◀

► **Example 5.** Consider the following language:

$$L_{\neg \forall repeat} = \{w \mid \exists i \forall j: i \neq j \implies \mathbb{D}[w_i] \neq \mathbb{D}[w_j]\}$$

Intuitively,  $L_{\neg \forall repeat}$  is the language of all words containing a data value with exactly one occurrence. This language is accepted, e.g., by the RsA<sub>1</sub> in Figure 3. The RsA stays in state  $q$ , collecting the seen values into its register, and at some point, when it sees a value not seen previously, it nondeterministically moves to  $s$ , remembering the value in its register. Then, at state  $s$ , the RsA just checks that it does not see the remembered value any more. ◀

## 4 Properties of Register Set Automata

In this section, we establish decidability of basic decision problems for RsAs and their closure properties. First, we claim that RsAs generalise NRAs.

► **Fact 6.** For every  $n \in \mathbb{N}$  and  $\text{NRA}_n$ , there exists an RsA<sub>n</sub> accepting the same language.

The next theorem shows the core property of RsAs: that their emptiness problem is decidable, however, for a much higher price than for NRAs, for which it is PSPACE-complete<sup>4</sup> [12]. For classifying the complexity of the problem, we use the hierarchy of fast-growing complexity classes of Schmitz [39], in particular the class  $\mathbf{F}_\omega$ , which, intuitively, corresponds to Ackermannian problems closed under primitive-recursive reductions.

► **Theorem 7.** The emptiness problem for RsA is decidable, in particular,  $\mathbf{F}_\omega$ -complete.

**Sketch of proof.** The proof is done by showing interreducibility of RsA emptiness with coverability in *transfer Petri nets* (TPNs) (often used for modelling the so-called *broadcast protocols*), which is a known  $\mathbf{F}_\omega$ -complete problem [42, 40, 41]. In the following, we briefly describe both directions of the reduction (see Appendix A.1 for details and examples).

<sup>4</sup>Note that for an alternative definition of NRAs considered in [27, 37], where no two registers can contain the same data value, the problem is NP-complete [37].



(RsA emptiness  $\leq$  TPN coverability) Intuitively, the conversion of an RsA  $\mathcal{A} = (Q, \mathbf{R}, \Delta, I, F)$  into a TPN  $\mathcal{N}_{\mathcal{A}}$  is done in the following way. The set of places of  $\mathcal{N}_{\mathcal{A}}$  will be as follows: (i) one place for each state of  $\mathcal{A}$ , (ii) two special places *init* and *fin*, and (iii) one place for every subset  $rgn \subseteq \mathbf{R}$ ; these places are used to represent all possible intersections of values held in the registers. For instance, if there are four tokens in the place representing  $r_1 \cap r_2$ , it means that there are exactly four different data values stored in both  $r_1$  and  $r_2$  and in no other register. Each transition  $t$  of  $\mathcal{A}$  is simulated by one or more TPN transitions between places representing its source and target states. The number of respective TPN transitions depends on how specific the guard is in the original automaton, since we need to distinguish every possible option of *in* being in some region  $rgn \in 2^{\mathbf{R}}$ . The transitions move the token between the places corresponding to  $t$ 's source and target states and, moreover, use the *broadcast* arcs to move tokens between the places representing regions, according to the manipulation of the set-registers in the update function of  $t$ . The special place *init* is used to have a single starting marking (it just nondeterministically chooses one state from  $I$ ) and the place *fin* is used as the target for the coverability test; all places corresponding to final states of  $\mathcal{A}$  can simply transition into it.

(TPN coverability  $\leq$  RsA emptiness) Given a TPN  $\mathcal{N}$ , the RsA  $\mathcal{A}_{\mathcal{N}}$  simulating it will have the following structure. There will be a state  $q_{main}$ , which will be active before and after the simulation of firing each transition of  $\mathcal{N}$ . Moreover, there will be one register for every place of  $\mathcal{N}$ ; individual tokens in the places will be simulated by unique data values from  $\mathbb{D}$  stored in the corresponding registers. For each transition of  $\mathcal{N}$ , there will be a *gadget*, doing a cycle on  $q_{main}$ , that represents the semantics of  $\mathcal{N}$ 's transition. Each such gadget is composed of several *protogadgets*, which simulate basic actions performed during the transition (adding a token to a place, removing a token, moving all tokens between places). Implementation of adding a token and moving tokens is relatively easy, the tricky part is removing a token, since RsAs do not support removing a data value from a register. We solve this by using a *lossy remove*: i.e., if *one* token is to be removed from a place, we simulate it by removing *at least one* token (but potentially more). This will not preserve *reachability*, but it is enough to preserve *coverability*. Moreover, there will also be an *initial* part setting the contents of the registers to reflect the initial marking of  $\mathcal{N}$  (terminating in  $q_{main}$ ) and a *final* part that checks the coverability by removing (again in a lossy way) tokens from places, terminating in a single final state. ◀

► **Remark 8.** Since RsAs generalise NRAs, their universality, equivalence, and language inclusion problems are all undecidable.

## 4.1 Closure Properties

The closure properties of RsAs are the same as for NRAs.

► **Theorem 9.** *The following closure properties hold for RsA:*

1. RsA is closed under union and intersection.
2. RsA is not closed under complement.

**Sketch of proof.** The proofs for closure under union and intersection are standard. For showing the non-closure under complement, consider the language  $L_{\neg \forall repeat}$  from Example 5, which can be accepted by RsA. We use a similar technique as in the proof of Proposition 3.2 in [17] and show that if there were an RsA accepting its complement, namely, the language

$$L_{\forall repeat} = \{w \mid \forall i \exists j: i \neq j \wedge \mathbb{D}[w_i] = \mathbb{D}[w_j]\},$$

RsAs could be used to decide the emptiness problem of a Minsky machine, which is a known undecidable problem. See the full proof in Appendix A.2 for more details. ◀

For RsAs with a limited number of registers, we lose the closure under intersection.

► **Theorem 10.** *For each  $n \in \mathbb{N}$ , the following closure properties hold for  $\text{RsA}_n$ :*

1.  $\text{RsA}_n$  is closed under union.
2.  $\text{RsA}_n$  is not closed under intersection and complement.

**Sketch of proof.** The proof of closure under union is standard. For proving non-closure of  $\text{RsA}_1$  under intersection, we consider the two following  $\text{RsA}_1$  languages

$$\mathcal{L}_1^A = \{w \mid \mathbb{D}[w_1] = \mathbb{D}[w_{|w|}]\} \quad \text{and} \quad \mathcal{L}_1^B = \{w \mid \mathbb{D}[w_2] = \mathbb{D}[w_{|w|-1}]\}$$

and show that their intersection cannot be accepted by  $\text{RsA}_1$ . This argument can be extended to  $\text{RsA}_n$  for  $n > 1$ . Non-closure under complement then follows from De Morgan's laws. ◀

► **Theorem 11.**  *$\text{DRsA}$  is closed under union, intersection, and complement.*

**Proof.** The proofs are standard (product construction and swapping (non)-final states). ◀

► **Theorem 12.** *For each  $n \in \mathbb{N}$ , the following closure properties hold for  $\text{DRsA}_n$ :*

1.  $\text{DRsA}_n$  is closed under complement.
2.  $\text{DRsA}_n$  is not closed under union and intersection.

**Proof.** Proof of closure under complement is standard. Non-closure under union and intersection is done in the same way as in the proof of Theorem 10. ◀

As with RAs, nondeterminism also allows bigger expressivity for RsAs.

► **Theorem 13.**  $\text{DRsA} \subsetneq \text{RsA}$

**Proof.** Let us consider the language  $L_{\neg\forall\text{repeat}}$  from the proof of Theorem 9, which is expressible using RsAs, and its complement  $L_{\forall\text{repeat}}$ , which is not expressible using RsAs. Since  $\text{DRsAs}$  are closed under complement (Theorem 11), if they could accept  $L_{\neg\forall\text{repeat}}$ , they could also accept  $L_{\forall\text{repeat}}$ , which is a contradiction. Therefore,  $L_{\neg\forall\text{repeat}} \notin \text{DRsA}$ . ◀

## 5 Determinising Register Automata

RAs have the following interesting property: a large class of NRA languages can be determinised into  $\text{DRsAs}$ . In this section, we give a determinisation semi-algorithm and specify properties of a class of NRAs for which it is complete.

Let  $\mathcal{A} = (Q, \mathbf{R}, \Delta, I, F)$  be an NRA. We use  $\mathbf{R}[q]$  for  $q \in Q$  to denote the set of registers  $r$  such that there exists a transition  $s \xrightarrow{(\cdot \mid g^-, g^+, up)} t \in \Delta$  with (i)  $up(r) \neq \perp$  and  $t = q$  or (ii)  $r \in g^- \cup g^+$  and  $s = q$ . Intuitively,  $\mathbf{R}[q]$  denotes the set of registers *active* in  $q$  and given a set of states  $S$ , we define  $\mathbf{R}[S] = \bigcup_{q \in S} \mathbf{R}[q]$ . We call  $\mathcal{A}$  *register-local* if for all  $r \in \mathbf{R}$  it holds that if  $r \in \mathbf{R}[q]$  and  $r \in \mathbf{R}[s]$  for some states  $q, s \in Q$ , then  $q = s$ . It is easy to see that every NRA can be transformed into the register-local form by creating a new copy of a register for every state that uses it, potentially increasing the number of registers to  $|Q| \cdot |\mathbf{R}|$ .

Furthermore, we call  $\mathcal{A}$  *single-valued* if there is no reachable configuration  $(q, f)$  such that  $f(r_1) = f(r_2)$  for a pair of distinct registers  $r_1, r_2 \in \mathbf{R}$ , i.e., there is at most one copy of each data value in  $\mathcal{A}$ . Again, any NRA can be converted into the single-valued form, however, the number of states can increase to  $B_{|\mathbf{R}|} \cdot |Q|$  where  $B_n$  is the  $n$ -th Bell number. Intuitively, the transformation is done by creating one copy of each state for every possible partition of  $\mathbf{R}$  (the partitions denote which registers hold the same value), and modifying the transition function correspondingly.

The determinisation (semi-)algorithm for a single-valued NRA  $\mathcal{A}$  is shown in Algorithm 1. On the high level, it is similar to the classical Rabin-Scott subset construction for determinising



■ **Algorithm 1** Determinisation of an NRA into a DRsA

---

**Input** : Single-valued NRA  $\mathcal{A} = (Q, \mathbf{R}, \Delta, I, F)$

**Output** : DRsA  $\mathcal{A}' = (\mathcal{Q}', \mathbf{R}, \Delta', I', F')$  with  $\mathcal{L}(\mathcal{A}') = \mathcal{L}(\mathcal{A})$  or  $\perp$

- 1  $\mathcal{Q}' \leftarrow \text{worklist} \leftarrow I' \leftarrow \{(I, c_0 = \{r \mapsto 0 \mid r \in \mathbf{R}\})\};$
- 2  $\Delta' \leftarrow \emptyset;$
- 3 **while**  $\text{worklist} \neq \emptyset$  **do**
- 4    $(S, c) \leftarrow \text{worklist.pop}();$
- 5   **foreach**  $a \in \Sigma, g \subseteq \mathbf{R}[S] \setminus \{r \in \mathbf{R} \mid c(r) = 0\}$  **do**
- 6      $T \leftarrow \{q - \boxed{a \mid g^-, g^\neq, \cdot} \rightarrow q' \in \Delta \mid q \in S, g^- \subseteq g, g^\neq \cap g = \emptyset\};$
- 7      $S' \leftarrow \{q' \mid \cdot - \boxed{\cdot \mid \cdot, \cdot, \cdot} \rightarrow q' \in T\};$
- 8     **if**  $\exists q - \boxed{\cdot \mid \cdot, g^\neq, \cdot} \rightarrow q' \in T, \exists r \in g^\neq : c(r) = \omega$  **then return**  $\perp$  ;
- 9     **foreach**  $r_i \in \mathbf{R}$  **do**
- 10        $\text{tmp} \leftarrow \emptyset;$
- 11       **foreach**  $\cdot - \boxed{\cdot \mid g^-, \cdot, \text{up}} \rightarrow \cdot \in T$  **do**
- 12           $\text{tmp} \leftarrow \text{tmp} \cup x$  where
- $$x = \begin{cases} \{y\} & \text{if } \text{up}(r_i) = y, y \in \mathbf{R} \cup \{\text{in}\}, c(y) \neq 0, \text{ and } y \notin g^-, \\ \{\text{in}\} & \text{if } \text{up}(r_i) = y \text{ and } y \in g^-, \text{ and} \\ \emptyset & \text{otherwise.} \end{cases}$$
- 13        $\text{op}_{r_i} \leftarrow \begin{cases} \text{tmp} \setminus \{\text{in}\} & \text{if } \text{tmp} \cap g \neq \emptyset \text{ and} \\ \text{tmp} & \text{otherwise.} \end{cases}$
- 14       **foreach**  $q' \in S'$  **do**
- 15           $P \leftarrow \text{op}_{r_1} \times \dots \times \text{op}_{r_n}$  for  $\{r_1, \dots, r_n\} = \mathbf{R}[q'];$
- 16          **foreach**  $(x_1, \dots, x_n) \in P$  **do**
- 17           **if**  $\nexists (\cdot - \boxed{\cdot \mid \cdot, \cdot, \text{up}} \rightarrow q') \in T$  s.t.  $\bigwedge_{1 \leq i \leq n} \text{up}(r_i) = x_i$  **then return**  $\perp$  ;
- 18        $\text{up}' \leftarrow \{r_i \mapsto \text{op}_{r_i} \mid r_i \in \mathbf{R}\};$
- 19        $c' \leftarrow \{r_i \mapsto \sum_{x \in \text{up}'(r_i)} \hat{c}(x, g) \mid r_i \in \mathbf{R}\};$
- 20       **if**  $(S', c') \notin \mathcal{Q}'$  **then**
- 21           $\text{worklist.push}((S', c'));$
- 22           $\mathcal{Q}' \leftarrow \mathcal{Q}' \cup \{(S', c')\};$
- 23        $\Delta' \leftarrow \Delta' \cup \{(S, c) - \boxed{a \mid g, \mathbf{R} \setminus g, \text{up}'} \rightarrow (S', c')\};$
- 24 **return**  $\mathcal{A}' = (\mathcal{Q}', \mathbf{R}, \Delta', I', \{(S, c) \in \mathcal{Q}' \mid S \cap F \neq \emptyset\});$

---

finite automata [33] with additional treatment of registers superimposed onto it.<sup>5</sup> During the construction, we track (i) all states of  $\mathcal{A}$  in which the runs of  $\mathcal{A}$  might be at a given point, represented by a set of states  $S \subseteq Q$  and (ii) the sizes of sets stored in each register, represented by a mapping  $c: \mathbf{R} \rightarrow \{0, 1, \omega\}$  ( $\omega$  denotes any number  $\geq 2$ )<sup>6</sup>; the macrostate is then a pair  $(S, c)$ . The initial state of the constructed DRsA is the macrostate  $(I, c_0)$  where  $c_0$  is a mapping assigning zero to each register (the run of a DRsA starts with all registers initialized to  $\emptyset$ ) (Line 1).

The main loop of the algorithm then constructs successors of reachable macrostates for

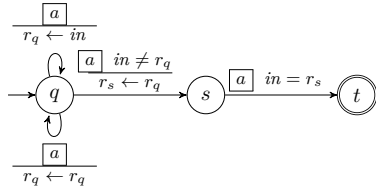
<sup>5</sup>The algorithm can be seen as a simplification of the algorithm for converting counting automata to deterministic counting-set automata from [48] (we do not need to deal with operations on the values stored in registers), but with additional features necessary to deal with NRA-specific issues.

<sup>6</sup>We keep track of the sizes to detect when our simulation of a disequality test  $\text{in} \neq r$  by the non-membership test  $\text{in} \notin r$  is imprecise due to  $r$  containing two or more elements.

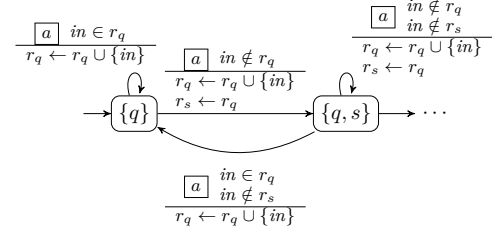
each  $a \in \Sigma$  and each  $g \subseteq \mathbf{R}$  on Line 5; each pair  $a, g$  corresponds to the so-called *minterm* (minterms denote combinations of guards whose semantics do not overlap [11]). For each minterm, we collect all transitions of  $\mathcal{A}$  compatible with this minterm (Line 6) and generate the successor set of states  $S'$  (Line 7). The  $\mathcal{A}'$  update function  $up'$  for register  $r$  is then set to collect into  $r$  all possible values that might be stored into  $r$  in  $\mathcal{A}$  on any run over the input word at the given position (Lines 9–18).

The algorithm needs to avoid the following possible issues:

1. Since the algorithm collects in the set-register  $r$  all possible values that could have been stored into the standard register  $r$  in  $\mathcal{A}$ , if the disequality tests in  $g^\neq$  were changed for non-membership tests in  $g^\notin$ , this could mean that  $\mathcal{A}'$  might not be able to simulate some transition of  $\mathcal{A}$  (the transition would not be enabled). Consider the following example:



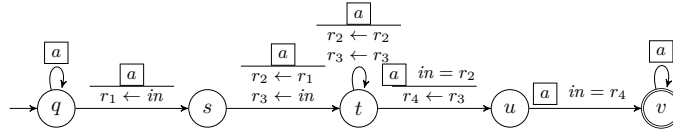
(a) An NRA  $\mathcal{A}$  with a disequality guard



(b) A part of the RsA obtained for  $\mathcal{A}$

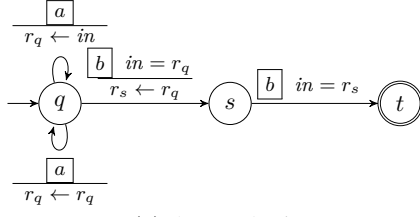
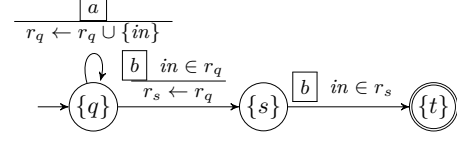
where (b) contains a part of the RsA obtained if Algorithm 1 did not use the  $c$ -component of macrostates. Notice that while  $\mathcal{A}$  accepts the word  $\langle a, 1 \rangle \langle a, 2 \rangle \langle a, 2 \rangle \langle a, 1 \rangle$ , the RsA obtained in this way does not. The reason for this is that after reading the third symbol (i.e.,  $\langle a, 2 \rangle$ ), the RsA goes to the macrostate  $\{q\}$ —it thinks it cannot be in  $s$  any more. This is the reason why we augment macrostates with the  $c$ -component. If we detect that a disequality test is performed on a register containing more than one element, we terminate the algorithm (Line 8). The tracking of sizes of sets stored in registers is done on Line 19 where  $\hat{c}(x, g)$  is defined as (i)  $c(x)$  if  $x \in \mathbf{R} \setminus g$ , (ii) 0 if  $x \in g$ , and (iii) 1 if  $x = in$ ; moreover, the sum is *saturated* to  $\omega$  for values  $\geq 2$ .

2. By collecting all possible values that can occur in registers, the algorithm is performing the so-called *Cartesian abstraction* (i.e., it is losing information about dependencies between components in tuples). This can lead to a scenario where, for some set-register assignment  $f'$  of  $\mathcal{A}'$ , we would have  $d_1 \in f'(r_1)$  and  $d_2 \in f'(r_2)$ , but there would be no corresponding configuration of  $\mathcal{A}$  with register assignment  $f$  such that  $d_1 = f(r_1)$  and  $d_2 = f(r_2)$ . Consider, e.g., an NRA for the language  $\{uvw v z \mid u, w, z \in (\Sigma \times \mathbb{D})^*, |v| = 2\}$ :



When the algorithm computes the successor of the macrostate  $(\{q, s, t\}, \{r_1:1, r_2:1, r_3:1\})$  over  $a \in \Sigma$  and the guard  $g = \emptyset$ , it would obtain the following update of registers:  $r_1 \leftarrow \{in\}$  (transition from  $q$  to  $s$ ),  $r_2 \leftarrow r_1 \cup r_2$  (transition from  $s$  to  $t$  and transition from  $t$  to  $t$ ), and  $r_3 \leftarrow r_3 \cup \{in\}$  (transition from  $s$  to  $t$  and transition from  $t$  to  $t$ ). This would simulate the update  $r_2 \leftarrow r_2, r_3 \leftarrow in$ , which is nowhere in the original NRA. The algorithm detects the possibility of such an overapproximation on Lines 14–17.

3. We need to avoid a situation when a set-register has collected all possible nondeterministic choices of a standard NRA register and then is tested twice with a different result. Consider the following example of an NRA  $\mathcal{A}$  and an RsA obtained from  $\mathcal{A}$  by Algorithm 1 without Line 13 (to save space, we collapse all macrostates with the same set of states into one):

(a) An NRA  $\mathcal{A}$ (b) An RsA overapproximating  $\mathcal{A}$ 's language

One can see that while the NRA cannot accept the word  $\langle a, 1 \rangle \langle a, 2 \rangle \langle b, 1 \rangle \langle b, 2 \rangle$ , the RsA accepts it. This happens because the RsA did not “collapse” the possible nondeterministic choices that are kept in the registers for the value of  $r_q$  after the first membership test (on the transition from  $\{q\}$  to  $\{s\}$ ) succeeded. We avoid this situation by the code on Line 13, which performs the collapse of the set of nondeterministic choices into a single value when it is positively tested. The update on the RsA transition from  $\{q\}$  to  $\{s\}$  constructed by the algorithm will then become  $r_s \leftarrow \{in\}$  and the result will be precise. One might also imagine similar scenario as the previous but with several registers copying a nondeterministically chosen value (e.g., when a data value is copied from  $r_1$  to  $r_2$  and, later,  $r_1$  is positively tested for equality, we need to guarantee that the value of  $r_2$  also collapses to the given data value). In order to avoid this, we require that the input NRA is *single-valued*, i.e., it never happens that a data value is in more than one register.

► **Theorem 14.** *When Algorithm 1 returns a DRsA  $\mathcal{A}'$ , then  $\mathcal{L}(\mathcal{A}) = \mathcal{L}(\mathcal{A}')$ .*

Naturally, we wish to syntactically characterise the class of NRAs for which Algorithm 1 is complete. We observe that when we start with an  $\text{NRA}_1^-$  and transform it into the single-valued register-local form, the algorithm always returns a DRsA.

► **Theorem 15.** (a) *For every  $\text{NRA}_1^-$ , there exists a DRsA accepting the same language.*  
 (b) *For every  $\text{URA}_1^-$ , there exists a DRsA accepting the same language.*

Let  $\mathcal{B}(\text{NRA}_1^-)$  be the class of languages that can be expressed using a Boolean combination of  $\text{NRA}_1^-$  languages, i.e., it is the closure of  $\text{NRA}_1^-$  languages under union, and intersection, and complement (it could also be denoted as  $\mathcal{B}(\text{URA}_1^-)$ ).

► **Example 16.** For instance, the language  $L_{\exists, \neg \exists \text{repeat}}$  composed as the concatenation of  $L_{\exists \text{repeat}}$  and  $L_{\neg \exists \text{repeat}}$  with a delimiter, formally

$$L_{\exists, \neg \exists \text{repeat}} = L_{\exists \text{repeat}} \cdot \{\langle b, d \rangle \mid d \in \mathbb{D}\} \cdot L_{\neg \exists \text{repeat}},$$

is in  $\mathcal{B}(\text{NRA}_1^-)$ , since it is the intersection of languages

$$L_{\exists \text{repeat}} \cdot \{\langle b, d \rangle \mid d \in \mathbb{D}\} \cdot \{\langle a, d \rangle \mid d \in \mathbb{D}\}^* \text{ and } \{\langle a, d \rangle \mid d \in \mathbb{D}\}^* \cdot \{\langle b, d \rangle \mid d \in \mathbb{D}\} \cdot L_{\neg \exists \text{repeat}},$$

but is expressible neither by an NRA nor by a URA (URAs cannot express the part *before* the delimiter and NRAs cannot express the part *after* the delimiter). ◀

From Theorem 15 we can conclude that  $\mathcal{B}(\text{NRA}_1^-)$  is captured by DRsA.

► **Corollary 17.** *For any language in  $\mathcal{B}(\text{NRA}_1^-)$ , there exists a DRsA accepting it.*

**Sketch of proof.** Follows from Theorems 11 and 15. ◀

► **Corollary 18.** *The inclusion problem between RsA and  $\mathcal{B}(\text{NRA}_1^-)$  is decidable.*

**Proof.** We just write  $\mathcal{L}(\mathcal{A}_1) \subseteq \mathcal{L}(\mathcal{A}_2)$  as  $\mathcal{L}(\mathcal{A}_1) \cap \overline{\mathcal{L}(\mathcal{A}_2)} = \emptyset$  and use Corollary 17 and Theorems 7 and 9. ◀

■ **Table 1** Distinguishing languages for a selection of register automata models. Grey cells denote that the result is implied from the class being a sub/super-class of another class where the result is established.  $\text{DRA}_1^{(=)}$  denotes both  $\text{DRA}_1$  and  $\text{DRA}_1^-$ , similarly for  $\text{URA}_1^{(=)}$ .  $\text{DRsA}_{(1)}$  denotes both  $\text{DRsA}$  and  $\text{DRsA}_1$ . None of the languages is accepted by  $\text{DRA}$ .

Language	$\text{NRA}_1^{(=)}$	$\text{URA}_1^{(=)}$	$\mathcal{B}(\text{NRA}_1^-)$	$\text{ARA}_1$	$\text{DRsA}_{(1)}$	$\text{RsA}_1$	$\text{ARA}_1(\mathbf{g}, \mathbf{s})$
$L_{\exists \text{repeat}}$	✓ Ex. 2	✗ Ex. 2	✓	✓	✓ Ex. 3	✓	✓
$L_{\neg \exists \text{repeat}}$	✗ Ex. 2	✓ Ex. 2	✓	✓	✓ Ex. 4	✓	✓
$L_{\exists, \neg \exists \text{repeat}}$	✗ Ex. 16	✗ Ex. 16	✓ Ex. 16	✓	✓	✓	✓
$L_{\forall \text{repeat}}$	✗	✗	✗	✗ [17]	✗ Thm. 13	✗ Thm. 9	✗ [17]
$L_{\neg \forall \text{repeat}}$	✗	✗	✗	✗ [17]	✗ Thm. 13	✓ Ex. 5	✓ [17]
$L_{\exists a \text{-no-} b}$	✗	✗	✗	✗	✓ Ex. 19	✓	✓ [17]
$L_{\neg \exists a \text{-no-} b}$	✗	✗	✗	✗	✓ Ex. 19	✓	✗ [17]
$L_{\forall a \exists b}$	✗ Ex. 20	✗ Ex. 20	✗	✓ [12]	✗	✗ Ex. 20	✓

## 6 Expressivity of Register Set Automata

In this section, we position RsAs in the landscape of automata over data words. For this, we use the languages introduced previously and some other languages defined below; these languages are then used to distinguish various register automata models (due to space constraints, this listing is by no means exhaustive).

► **Example 19.** First, we define the language over  $\Sigma = \{a, b\}$

$$L_{\exists a \text{-no-} b} = \{w \mid \exists i: a[w_i] \wedge \nexists j < i: b[w_j] \wedge \mathbb{D}[w_j] = \mathbb{D}[w_i]\}$$

from [17, Proof of Proposition 3.2] and its complement  $L_{\neg \exists a \text{-no-} b}$ . Intuitively,  $L_{\exists a \text{-no-} b}$  is the language of words  $w$  such that there exists an input element  $\langle a, d \rangle$  that is not preceded by an occurrence of a  $\langle b, d \rangle$  element. Neither  $L_{\exists a \text{-no-} b}$  nor  $L_{\neg \exists a \text{-no-} b}$  can be accepted by  $\text{ARA}_1$  while  $\text{ARA}_1(\mathbf{g}, \mathbf{s})$  accepts  $L_{\exists a \text{-no-} b}$  but cannot accept  $L_{\neg \exists a \text{-no-} b}$ . On the other hand, as shown in Figures 5 and 6,  $\text{DRsA}_1$  can accept both  $L_{\exists a \text{-no-} b}$  and  $L_{\neg \exists a \text{-no-} b}$ .

(It might seem suspicious that  $\text{DRsA}_1$  can express the language  $L_{\neg \exists a \text{-no-} b}$ , since according to [17, Proof of Proposition 3.2], if an  $\text{ARA}_1$   $\mathcal{A}$  could accept the language,  $\mathcal{A}$  could be used to decide language-emptiness of a Minsky machine. So how come that we can express  $L_{\neg \exists a \text{-no-} b}$  using  $\text{DRsA}_1$ , which have a decidable emptiness problem (cf. Theorem 7)? The reason is that in the construction of the automaton representing the accepting runs of a Minsky machine from [17], apart from  $L_{\neg \exists a \text{-no-} b}$ , we also need to be able to express the property “every counter increment is matched with its decrement”, which is not expressible by  $\text{RsA}$  (cf. the proof of Theorem 9).) ◀

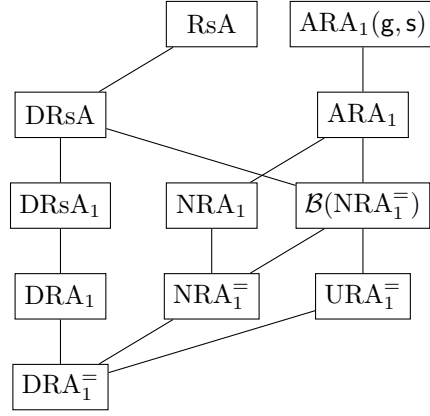
► **Example 20.** Moreover, let us define the following language from [12, Example 2.2]:

$$L_{\forall a \exists b} = \{w \mid \forall i: a[w_i] \implies ((\forall j \neq i: a[w_j] \implies \mathbb{D}[w_i] \neq \mathbb{D}[w_j]) \wedge (\exists k > i: b[w_k] \wedge \mathbb{D}[w_i] = \mathbb{D}[w_k]))\}$$

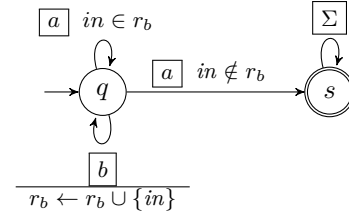
Intuitively,  $L_{\forall a \exists b}$  denotes the language of words where no two  $a$ -positions contain the same data value and every  $a$ -position is followed by a matching  $b$ -position.  $L_{\forall a \exists b}$  is recognizable by  $\text{ARA}_1$  [12, Example 2.6], but is not recognizable by any  $\text{URA}$ ,  $\text{NRA}$ , or  $\text{RsA}$ . ◀

The Hasse diagram comparing the expressive power of a selection of register automata models with decidable emptiness problem is in Figure 4 and languages that distinguish the various classes are in Table 1.

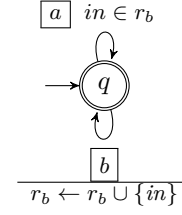
► **Remark 21.** RsAs are also incomparable to the class of *pebble automata* [32], since  $\text{DRsAs}$  generalize  $\text{DRAs}$ , they can accept a language not expressible by  $\text{PAs}$  (as shown by Tan in [44]). On the other hand, RsAs cannot express  $L_{\forall \text{repeat}}$ , which is expressible by  $\text{PAs}$ . ◀



■ **Figure 4** Hasse diagram comparing the expressive power of a selection of register automata models with decidable emptiness problem. All inclusions are strict. Languages distinguishing the different models can be found in Table 1.



■ **Figure 5** A DRsA<sub>1</sub> recognizing  $L_{\exists a-no-b}$



■ **Figure 6** A DRsA<sub>1</sub> recognizing  $L_{\neg \exists a-no-b}$

## 7 Extensions of Register Set Automata

We also consider several extensions of the RsA model. In this section, we only give intuitive definitions and statement of main results, see Appendix D for more details.

**RsAs with Register Emptiness Test.** The first extension  $\text{RsA}^{=\emptyset}$  that we consider allows to test on each transition *emptiness* of arbitrary registers, i.e., a transition  $q \xrightarrow{a \mid g^{\in}, g^{\notin}, g^{=\emptyset}, up} s$  is enabled if, in addition to the standard semantics, it also holds that all registers in  $g^{=\emptyset}$  are empty. It is easy to see that each such  $\text{RsA}^{=\emptyset}$  can be transformed into a standard RsA with  $2^{|\mathbf{R}|} \cdot |Q|$  states by keeping track of which registers are empty and which are not, in each state.

This extension could be generalized to an extension where on each transition, we could ask, for every register  $r \in \mathbf{R}$ , whether the number of elements in the register is equal to  $k$  for  $k < n$ , where  $n \in \mathbb{N}_0$  is an *a priori* fixed number. The translation to an RsA with  $(n+1)^{|\mathbf{R}|} \cdot |Q|$  states can be done in a similar way as for  $\text{RsA}^{=\emptyset}$ .

**RsAs with Register Equality Test.** The second extension  $\text{RsA}^{=r}$  enables to test *equality* of registers. In particular, a transition  $q \xrightarrow{a \mid g^{\in}, g^{\notin}, g^{=}, up} s$  is enabled if, in addition to the standard semantics, the equality of the content of registers specified by  $g^{=}$  also holds. Unfortunately, this model is too powerful: already its emptiness problem is undecidable. The undecidability can be proved by a reduction from the reachability problem of Petri nets with inhibitor arcs, which is undecidable.

**RsAs with Removal.** The next extension  $\text{RsA}^{rm}$  allows the update operation to also *remove* from registers the current data value on the input and test emptiness. Although it may seem as a simple extension of the RsA model, it is much more powerful: emptiness is, again, undecidable. In this case, undecidability of emptiness can be shown by a reduction from reachability in transfer Petri nets, which is undecidable [13].

## 8 Related Work

The literature on automata over infinite alphabets is rich, see, e.g., the excellent survey by Segoufin [43] and the paper by Neven *et al.* [32].

Register automata were introduced (under the name *finite memory automata*) by Kaminski and Francez in [27] and their basic properties further studied by Sakamoto and Ikeda in [37]. Demri and Lazić study in [12] (non-)deterministic, universal, and alternating one-way and two-way register automata, and their relation to the linear temporal logic with the *freeze* quantifier, which can store the current data value into a register. In particular, they show that  $LTL_1^\downarrow(X, U)$ , i.e., the fragment of the logic with one freeze register and the *next* ( $X$ ) and *until* ( $U$ ) temporal operators captures the class of languages accepted by one-way alternating register automata with one register ( $ARA_1$ ).

Figueira [17] introduces alternating register automata with one register and two extra operations: **guess** and **spread** ( $ARA_1(g, s)$ ). Intuitively, **guess** can guess an arbitrary data value and store it in the register (which can be used to encode an existential quantification over future data values) and **spread** orchestrates all threads whose register value matches the current data value on the input tape to make a transition into a new state (which can be used to encode a sort of universal quantification over past data values).

Symbolic register automata [14] extend the register automaton model with the ability to test predicates over the data values (which need to come from some decidable first-order theory) while preserving most of the theoretical properties of register automata.

Figueira *et al.* [18] noticed the tight connection between register and timed automata, allowing to transfer results between the models. Clemente *et al.* [9, 10] consider the problem of *determinisability* of register/timed automata, i.e., whether a nondeterministic register/timed automaton can be transformed into a deterministic register/timed automaton with the same language. In this paper, we use a more general model (RsAs) for the determinisation.

Bojańczyk *et al.* [4] introduce a model of *nominal automata*, register automata where not only equality, but also total and partial order tests between symbols are allowed. On the other hand, Chen *et al.* [8, 7] extend register automata with the capability to compute affine functions over rationals and show decidability of reachability and the commutativity problem for fragments of the model.

In our work, we were inspired by *counting-set automata* introduced by Turoňová *et al.* [48], which use *sets of counter values* to compactly represent configurations of *counting automata* [23] (a restricted version of counter automata [28] with a bound on the value of counters for compact representation of finite automata), to obtain a deterministic model for efficient matching of regular expressions with repetitions.

Pebble automata (PAs) [47] are a model for automata over infinite words based on a different principle than RAs: a PA can *mark* certain positions in the input word using *pebbles* and make decisions based on equality of data values under the automaton head and the pebbles. The class of PAs is quite robust—they are closed under Boolean operations, determinisable, and the two-way and one-way versions have the same expressive power. On the other hand, PAs are not really practical—language emptiness is already undecidable.

There have been several attempts of automata models for matching regexes with back-references, but they have generally failed expectations of the community. Memory automata of Schmid [38] provide a natural extension of finite automata with a register bank, each register able to store a word, but do not allow determinisation. Becchi and Crowley [3] encode back-references into an extended model of NFAs using backtracking. Namjoshi and Narlikar [30] extend Thompson’s algorithm [46] to a model similar to the models of [38, 3] in the obvious way by tracking the set of all possible configurations, which is inefficient (for each symbol, the number of operations proportional to the (potentially exponential) size of the set of configurations is required). Moreover, there are also some proprietary ad hoc industrial solutions that support fast matching of regexes with back-references [45].



## 9 Conclusion and Future Work

We have introduced register set automata, a class of automata over data words providing an underlying formal model for efficient pattern matching of a subclass of regexes with backreferences. There are many challenges that we wish to address in the future: (i) improvement of our determinisation algorithm to work on a larger class of input NRAs, (ii) explore other, more expressive, formal models that would allow deterministic automata models for a large class of regexes with backreferences (e.g., the hard regex in footnote<sup>2</sup> is beyond the power of DRsAs), (iii) develop efficient toolbox for working with RsAs occurring in practice.

---

### References

- 1 Adar Weidman. Regular expression denial of service — ReDoS. [https://owasp.org/www-community/attacks/Regular\\_expression\\_Denial\\_of\\_Service\\_-\\_ReDoS](https://owasp.org/www-community/attacks/Regular_expression_Denial_of_Service_-_ReDoS), 2021. [Online; accessed 19-December-2021].
- 2 Adam Baldwin. Regular expression denial of service affecting Express.js. <https://medium.com/node-security/regular-expression-denial-of-service-affecting-express-js-9c397c164c43>, 2016.
- 3 Michela Becchi and Patrick Crowley. Extending finite automata to efficiently match Perl-compatible regular expressions. In *Proceedings of the 2008 ACM CoNEXT Conference on - CONEXT '08*, pages 1–12, Madrid, Spain, 2008. ACM Press. URL: <http://portal.acm.org/citation.cfm?doid=1544012.1544037>, doi:10.1145/1544012.1544037.
- 4 Mikolaj Bojanczyk, Bartek Klin, and Slawomir Lasota. Automata with group actions. In *Proceedings of the 26th Annual IEEE Symposium on Logic in Computer Science, LICS 2011, June 21-24, 2011, Toronto, Ontario, Canada*, pages 355–364. IEEE Computer Society, 2011. doi:10.1109/LICS.2011.48.
- 5 Benedikt Bollig, Peter Habermehl, Martin Leucker, and Benjamin Monmege. A Fresh Approach to Learning Register Automata. In *Developments in Language Theory*, Lecture Notes in Computer Science, pages 118–130, Berlin, Heidelberg, 2013. Springer. doi:10.1007/978-3-642-38771-5\_12.
- 6 Yu-Fang Chen, Vojtěch Havlena, Ondřej Lengál, and Andrea Turrini. A Symbolic Algorithm for the Case-Split Rule in String Constraint Solving. In *Programming Languages and Systems*, volume 12470, pages 343–363, Cham, 2020. Springer International Publishing. Series Title: Lecture Notes in Computer Science. URL: [https://link.springer.com/10.1007/978-3-030-64437-6\\_18](https://link.springer.com/10.1007/978-3-030-64437-6_18), doi:10.1007/978-3-030-64437-6\_18.
- 7 Yu-Fang Chen, Ondřej Lengál, Tony Tan, and Zhilin Wu. Register automata with linear arithmetic. In *2017 32nd Annual ACM/IEEE Symposium on Logic in Computer Science (LICS)*, pages 1–12, June 2017. doi:10.1109/LICS.2017.8005111.
- 8 Yu-Fang Chen, Lei Song, and Zhilin Wu. The Commutativity Problem of the MapReduce Framework: A Transducer-Based Approach. In Swarat Chaudhuri and Azadeh Farzan, editors, *Computer Aided Verification*, Lecture Notes in Computer Science, pages 91–111, Cham, 2016. Springer International Publishing. doi:10.1007/978-3-319-41540-6\_6.
- 9 Lorenzo Clemente, Slawomir Lasota, and Radosław Piórkowski. Determinisability of One-Clock Timed Automata. In Igor Konnov and Laura Kovács, editors, *31st International Conference on Concurrency Theory (CONCUR 2020)*, volume 171 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 42:1–42:17, Dagstuhl, Germany, 2020. Schloss Dagstuhl–Leibniz-Zentrum für Informatik. URL: <https://drops.dagstuhl.de/opus/volltexte/2020/12854>, doi:10.4230/LIPIcs.CONCUR.2020.42.
- 10 Lorenzo Clemente, Slawomir Lasota, and Radosław Piórkowski. Determinisability of register and timed automata. *arXiv:2104.03690 [cs]*, April 2021. URL: <http://arxiv.org/abs/2104.03690>.

- 11 Loris D’Antoni and Margus Veanes. Minimization of symbolic automata. In Suresh Jagannathan and Peter Sewell, editors, *The 41st Annual ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages, POPL ’14, San Diego, CA, USA, January 20-21, 2014*, pages 541–554. ACM, 2014. doi:10.1145/2535838.2535849.
- 12 Stéphane Demri and Ranko Lazić. LTL with the freeze quantifier and register automata. *ACM Transactions on Computational Logic*, 10(3):1–30, April 2009. URL: <https://dl.acm.org/doi/10.1145/1507244.1507246>, doi:10.1145/1507244.1507246.
- 13 Catherine Dufourd, Alain Finkel, and Philippe Schnoebelen. Reset nets between decidability and undecidability. In *Automata, Languages and Programming, 25th International Colloquium, ICALP’98, Aalborg, Denmark, July 13-17, 1998, Proceedings*, volume 1443 of *LNCS*, pages 103–115. Springer, 1998. doi:10.1007/BFb0055044.
- 14 Loris D’Antoni, Tiago Ferreira, Matteo Sammartino, and Alexandra Silva. Symbolic Register Automata. In Isil Dillig and Serdar Tasiran, editors, *Computer Aided Verification*, Lecture Notes in Computer Science, pages 3–21, Cham, 2019. Springer International Publishing. doi:10.1007/978-3-030-25540-4\_1.
- 15 Javier Esparza, Alain Finkel, and Richard Mayr. On the verification of broadcast protocols. In *14th Annual IEEE Symposium on Logic in Computer Science, Trento, Italy, July 2-5, 1999*, pages 352–359. IEEE Computer Society, 1999. doi:10.1109/LICS.1999.782630.
- 16 Stack Exchange. Outage postmortem. <http://stackstatus.net/post/147710624694/outage-postmortem-july-20-2016>, 2016.
- 17 Diego Figueira. Alternating register automata on finite words and trees. *Logical Methods in Computer Science*, 8(1):22, March 2012. URL: <http://arxiv.org/abs/1202.3957>, doi:10.2168/LMCS-8(1:22)2012.
- 18 Diego Figueira, Piotr Hofman, and Sławomir Lasota. Relating timed and register automata. *Mathematical Structures in Computer Science*, 26(6):993–1021, September 2016. URL: <https://www.cambridge.org/core/journals/mathematical-structures-in-computer-science/article/abs/relating-timed-and-register-automata/7F8C6A8E36FA8550287D0E0F67C8EA0E>, doi:10.1017/S0960129514000322.
- 19 Bharat Garhewal, Frits Vaandrager, Falk Howar, Timo Schrijvers, Toon Lenaerts, and Rob Smits. Grey-Box Learning of Register Automata. In Brijesh Dongol and Elena Troubitsyna, editors, *Integrated Formal Methods*, Lecture Notes in Computer Science, pages 22–40, Cham, 2020. Springer International Publishing. doi:10.1007/978-3-030-63461-2\_2.
- 20 Google. RE2, 2022. URL: <https://github.com/google/re2>.
- 21 Google. RE2 issue tracker: Feature request #101: Add backreference support, 2022. URL: <https://github.com/google/re2/issues/101>.
- 22 Radu Grigore, Dino Distefano, Rasmus Lerchedahl Petersen, and Nikos Tzevelekos. Runtime Verification Based on Register Automata. In *Tools and Algorithms for the Construction and Analysis of Systems*, Lecture Notes in Computer Science, pages 260–276, Berlin, Heidelberg, 2013. Springer. doi:10.1007/978-3-642-36742-7\_19.
- 23 Lukáš Holík, Ondřej Lengál, Olli Saarikivi, Lenka Turoňová, Margus Veanes, and Tomáš Vojnar. Succinct Determinisation of Counting Automata via Sphere Construction. In Anthony Widjaja Lin, editor, *Programming Languages and Systems*, Lecture Notes in Computer Science, pages 468–489, Cham, 2019. Springer International Publishing. doi:10.1007/978-3-030-34175-6\_24.
- 24 Intel. Hyperscan manual, unsupported features, 2022. URL: <https://intel.github.io/hyperscan/dev-reference/compilation.html#unsupported-constructs>.
- 25 Radu Iosif and Xiao Xu. Alternating Automata Modulo First Order Theories. In *Computer Aided Verification*, volume 11562 of *Lecture Notes in Computer Science*, pages 43–63, Cham, 2019. Springer International Publishing. URL: [http://link.springer.com/10.1007/978-3-030-25543-5\\_3](http://link.springer.com/10.1007/978-3-030-25543-5_3), doi:10.1007/978-3-030-25543-5\_3.

- 26 Malte Isberner, Falk Howar, and Bernhard Steffen. Learning register automata: from languages to program structures. *Machine Learning*, 96(1):65–98, July 2014. doi:10.1007/s10994-013-5419-7.
- 27 Michael Kaminski and Nissim Francez. Finite-memory automata. *Theoretical Computer Science*, 134(2):329–363, November 1994. URL: <https://www.sciencedirect.com/science/article/pii/0304397594902429>, doi:10.1016/0304-3975(94)90242-9.
- 28 Marvin L. Minsky. Recursive Unsolvability of Post's Problem of "Tag" and other Topics in Theory of Turing Machines. *Annals of Mathematics*, 74(3):437–455, 1961. URL: <https://www.jstor.org/stable/1970290>, doi:10.2307/1970290.
- 29 Andrzej S. Murawski, Steven J. Ramsay, and Nikos Tzevelekos. Polynomial-Time Equivalence Testing for Deterministic Fresh-Register Automata. page 14 pages, 2018. URL: <http://drops.dagstuhl.de/opus/volltexte/2018/9654/>, doi:10.4230/LIPICS.MFCS.2018.72.
- 30 Kedar S. Namjoshi and Girija J. Narlikar. Robust and fast pattern matching for intrusion detection. In *INFOCOM'10*, pages 740–748. IEEE, 2010. doi:10.1109/INFCOM.2010.5462149.
- 31 Frank Neven. Automata, Logic, and XML. In Julian Bradfield, editor, *Computer Science Logic*, Lecture Notes in Computer Science, pages 2–26, Berlin, Heidelberg, 2002. Springer. doi:10.1007/3-540-45793-3\_2.
- 32 Frank Neven, Thomas Schwentick, and Victor Vianu. Finite state machines for strings over infinite alphabets. *ACM Trans. Comput. Log.*, 5(3):403–435, 2004. doi:10.1145/1013560.1013562.
- 33 Michael O. Rabin and Dana S. Scott. Finite automata and their decision problems. *IBM J. Res. Dev.*, 3(2):114–125, 1959. doi:10.1147/rd.32.0114.
- 34 Regex101.com. <https://regex101.com>, 2021. [Online; accessed 19-December-2021].
- 35 Klaus Reinhardt. Reachability in petri nets with inhibitor arcs. *Electronic Notes in Theoretical Computer Science*, 223:239–264, 2008.
- 36 M. Roesch et al. Snort: A network intrusion detection and prevention system,, 2022. URL: <http://www.snort.org>.
- 37 Hiroshi Sakamoto and Daisuke Ikeda. Intractability of decision problems for finite-memory automata. *Theor. Comput. Sci.*, 231(2):297–308, 2000. doi:10.1016/S0304-3975(99)00105-X.
- 38 Markus L Schmid. Characterising REGEX Languages by Regular Languages Equipped with Factor-Referencing. 249:1–17, August 2016. URL: <https://www.sciencedirect.com/science/article/pii/S0890540116000109>, doi:10.1016/j.ic.2016.02.003.
- 39 Sylvain Schmitz. Complexity Hierarchies Beyond Elementary. *ACM Transactions on Computation Theory*, 8(1):1–36, February 2016. URL: <http://arxiv.org/abs/1312.5686>, doi:10.1145/2858784.
- 40 Sylvain Schmitz. *Algorithmic Complexity of Well-Quasi-Orders*. Habilitation à diriger des recherches, École normale supérieure Paris-Saclay, November 2017. URL: <https://tel.archives-ouvertes.fr/tel-01663266>.
- 41 Sylvain Schmitz and Philippe Schnoebelen. Algorithmic Aspects of WQO Theory. August 2012. URL: <https://cel.archives-ouvertes.fr/cel-00727025>.
- 42 Sylvain Schmitz and Philippe Schnoebelen. The power of well-structured systems. In Pedro R. D'Argenio and Hernán C. Melgratti, editors, *CONCUR 2013, Buenos Aires, Argentina, August 27-30, 2013. Proceedings*, volume 8052 of *Lecture Notes in Computer Science*, pages 5–24. Springer, 2013. doi:10.1007/978-3-642-40184-8\_2.
- 43 Luc Segoufin. Automata and Logics for Words and Trees over an Infinite Alphabet. In *Computer Science Logic*, volume 4207, pages 41–57. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006. URL: [http://link.springer.com/10.1007/11874683\\_3](http://link.springer.com/10.1007/11874683_3), doi:10.1007/11874683\_3.
- 44 Tony Tan. Graph reachability and pebble automata over infinite alphabets. *ACM Trans. Comput. Log.*, 14(3):19:1–19:31, 2013. doi:10.1145/2499937.2499940.
- 45 Nvidia Mellanox team. Personal communication, 2021.
- 46 Ken Thompson. Programming Techniques: Regular expression search algorithm. *Communications of the ACM*, 11(6):419–422, 1968. doi:10.1145/363347.363387.

- 47 Milo Tova, Suciú Dan, and Vianu Victor. Typechecking for XML transformers. In *POD'00*, pages 11–22, Dallas, USA, 2000. ACM. URL: <https://dl.acm.org/doi/abs/10.1145/335168.335171>, doi:10.1145/335168.335171.
- 48 Lenka Turoňová, Lukáš Holík, Ondřej Lengál, Olli Saarikivi, Margus Veanes, and Tomáš Vojnar. Regex matching with counting-set automata. *Proceedings of the ACM on Programming Languages*, 4(OOPSLA):218:1–218:30, November 2020. doi:10.1145/3428286.
- 49 Xiang Wang, Yang Hong, Harry Chang, KyoungSoo Park, Geoff Langdale, Jiayu Hu, and Heqing Zhu. Hyperscan: A fast multi-pattern regex matcher for modern CPUs. In Jay R. Lorch and Minlan Yu, editors, *NSDI'19*, pages 631–648. USENIX Association, 2019. URL: <https://www.usenix.org/conference/nsdi19/presentation/wang-xiang>.

## A Proofs for Section 4

► **Fact 6.** For every  $n \in \mathbb{N}$  and  $\text{NRA}_n$ , there exists an  $\text{RsA}_n$  accepting the same language.

**Proof.** We transform every NRA transition  $q \xrightarrow{a \mid g^-, g^\neq, up} s$  into the RsA transition  $q \xrightarrow{a \mid g^\epsilon, g^\neq, up'} s$  such that  $g^\epsilon = g^-$ ,  $g^\neq = g^\neq$ , and for every register  $r_i$  and  $up(r_i) = x$ ,

$$up'(r_i) = \begin{cases} \{x\} & \text{for } x \in \mathbf{R} \cup \{in\} \\ \emptyset & \text{for } x = \perp. \end{cases}$$

Intuitively, every simple register of NRA will be represented by a set register of RsA that will always hold the value of either an empty or a singleton set. ◀

### A.1 Proof of Theorem 7

► **Theorem 7.** The emptiness problem for RsA is decidable, in particular,  $\mathbf{F}_\omega$ -complete.

The proof is done by showing interreducibility of RsA emptiness with coverability in *transfer Petri nets* (often used for modelling the so-called *broadcast protocols*), which is a known  $\mathbf{F}_\omega$ -complete problem [42, 40, 41]. We start with defining transfer Petri nets and then give the reductions.

#### A.1.1 Transfer Petri Nets

Intuitively, transfer Petri net is an extension of Petri nets where transitions can *transfer* all tokens from one place to another place at once. They are closely related to *broadcast protocols* [15].

Formally, a *transfer Petri net* (TPN) is a triple  $\mathcal{N} = (P, T, M_0)$ , s.t.  $P$  is a finite set of *places*,  $T$  is a finite set of *transitions*, and  $M_0: P \rightarrow \mathbb{N}$  is an *initial marking*. The set of transitions  $T$  is such that  $P \cap T = \emptyset$  and every transition  $t \in T$  is of the form  $t = \langle In, Out, Transfer \rangle$  where  $In, Out: P \rightarrow \mathbb{N}$  define  $t$ 's *input* and *output places* respectively and  $Transfer: P \rightarrow P$  is a (total) *transfer function*.

A *marking* of  $\mathcal{N}$  is a function  $M: P \rightarrow \mathbb{N}$  assigning a particular number of tokens to each state. Given a pair of markings  $M$  and  $M'$ , we use  $M \leq M'$  to denote that for all  $p \in P$  it holds that  $M(p) \leq M'(p)$ . Moreover, we use  $M + M'$  and  $M - M'$  (for  $M' \leq M$ ) to denote the pointwise addition and subtraction of markings and we use  $\mathbf{u}_p$  to denote the marking such that  $\mathbf{u}_p(p') = 1$  if  $p = p'$  and  $\mathbf{u}_p(p') = 0$  otherwise. The *identity* function (over an arbitrary set that is clear from the context) is denoted as  $\mathbf{id}$ . Given a marking  $M$ , a transition  $t = \langle In, Out, Transfer \rangle$  is *enabled* if  $In \leq M$ , i.e., there is a sufficient number of tokens in each of its input places. We use  $M[t]M'$  to denote that

1.  $t$  is enabled in  $M$  and
2.  $M'$  is the marking such that for every  $p \in P$  we have

$$M'(p) = Out(p) + \sum \{M_{aux}(p') \mid Transfer(p') = p\} \text{ where } M_{aux} = M - In. \quad (1)$$

That is, the successor marking  $M'$  is obtained by (i) removing  $In$  tokens from inputs of  $t$ , (ii) transferring tokens according to  $Transfer$ , and (iii) adding  $Out$  tokens to  $t$ 's outputs.

We say that a marking  $M$  is *reachable* if there is a (possibly empty) sequence  $t_1, t_2, \dots, t_n$  of transitions such that it holds that  $M_0[t_1]M_1[t_2] \dots [t_n]M$ , where  $M_0$  is the initial marking. A marking  $M$  is *coverable*, if there exists a reachable marking  $M'$ , such that  $M \leq M'$ .

The *Coverability* problem for TPNs asks, given a TPN  $\mathcal{N}$  and a marking  $M$ , whether  $M$  is coverable in  $\mathcal{N}$ .

► **Proposition 22** ([42]). *The Coverability problem for TPNs is  $\mathbf{F}_\omega$ -complete.*

Let us now prove the two directions of the proof of Theorem 7.

► **Lemma 23.** *The emptiness problem for RsA is in  $\mathbf{F}_\omega$ .*

**Proof.** The proof of the lemma is based on reducing the RsA emptiness problem to coverability in TPNs, which is  $\mathbf{F}_\omega$ -complete (Proposition 22). Intuitively, the reduction consists of creating a TPN with places representing both individual states of RsA and individual *regions* of the Venn diagram of  $\mathbf{R}$ . Transitions of RsA are represented by one or more transitions of TPN, distinguishing every possible option of *in* being in some region  $rgn \in 2^{\mathbf{R}}$ . The set of arcs leading to and from each transition is calculated in a way which preserves the semantics and position of values defined by the *guard* and *update* formulae. Finally, the marking to be covered requires one token to be present in the places representing the final states of the RsA.

Formally, let  $\mathcal{A} = (Q, \mathbf{R}, \Delta, I, F)$  be an RsA. In the following, we will construct a TPN  $\mathcal{N}_{\mathcal{A}} = (P, T, M_0)$  and a marking  $M_F$  such that  $\mathcal{L}(\mathcal{A}) \neq \emptyset$  iff  $M_F$  is coverable in  $\mathcal{N}_{\mathcal{A}}$ . We set the components of  $\mathcal{N}_{\mathcal{A}}$  as follows:

- $P = Q \uplus \{init, fin\} \uplus 2^{\mathbf{R}}$  where *init* and *fin* are two new places,
- $T = \{\langle \mathbf{u}_{init}, \mathbf{u}_{q_i}, \mathbf{id} \rangle \mid q_i \in I\} \cup \{\langle \mathbf{u}_{q_f}, \mathbf{u}_{fin}, \mathbf{id} \rangle \mid q_f \in F\} \cup T'$  with  $T'$  defined below, and
- $M_0 = \mathbf{u}_{init}$ .

Intuitively, the set of places contains the states of  $\mathcal{A}$  (there will always be at most one token in those places), two new places *init* and *fin* that are used for the initial nondeterministic choice of some initial state of  $\mathcal{A}$  and for a unique *final place* (whose coverability will be checked) respectively, and, finally, a new place for every *region* of the Venn diagram of  $\mathbf{R}$ , which will track the number of data values that two or more registers share (e.g., for  $\mathbf{R} = \{r_1, r_2, r_3\}$ , the subset  $\{r_1, r_3\}$  denotes the region  $r_1 \cap \overline{r_2} \cap r_3$ , i.e., the data values that are stored in  $r_1$  and  $r_3$  but are not stored in  $r_2$ . The region  $\overline{r_1} \cap \overline{r_2} \cap \overline{r_3}$  is denoted as  $\{\emptyset\}$ . Regions  $rgn_1, rgn_2$  are *distinct*, if  $rgn_1 \triangle rgn_2 \neq \emptyset$ , i.e.  $\exists r: (r \in rgn_1 \vee r \in rgn_2) \wedge r \notin rgn_1 \cap rgn_2$ . The set of TPN transitions  $T'$  is defined below.

We now proceed to the definition of  $T'$ . Let  $t = q - \boxed{a \mid g^\epsilon, g^\not\epsilon, up} \rightarrow s \in \Delta$  be a transition in  $\mathcal{A}$ . Then, we create a TPN transition for every possible option of *in* being in some region  $rgn_g \in 2^{\mathbf{R}}$  (e.g., for  $in \in r_1 \cap \overline{r_2} \cap r_3$  or  $in \in \overline{r_1} \cap \overline{r_2} \cap r_3$ ). For  $t$  and  $rgn_g$ , we define

$$\gamma(t, rgn_g) = \begin{cases} \{\langle In, Out, Transfer \rangle\} & \text{if } (g^\epsilon \subseteq rgn_g) \wedge (g^\not\epsilon \cap rgn_g = \emptyset) \text{ and} \\ \emptyset & \text{otherwise.} \end{cases} \quad (2)$$

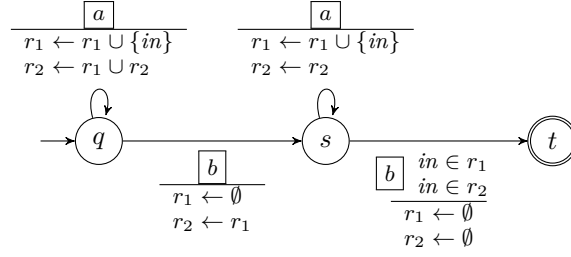
Then  $T' = \bigcup \{\gamma(t, rgn_g) \mid t \in \Delta, rgn_g \in 2^{\mathbf{R}}\}$ .

In the definition of  $\gamma(t, rgn_g)$  above, *In*, *Out*, and *Transfer* are defined as follows:

- $In = \mathbf{u}_{rgn_g} + \mathbf{u}_q$  and
- $Out = \mathbf{u}_{dst} + \mathbf{u}_s$  where  $dst = \{r_i \in \mathbf{R} \mid in \in up(r_i)\}$ .
- Before we give a formal definition of *Transfer*, let us start with an intuition given in the following example.

► **Example 24.** Let us consider the RsA in Figure 7 and its transition  $q - \boxed{a \mid \emptyset, \emptyset, up} \rightarrow q$  with  $up(r_1) = \{r_1, in\}$  and  $up(r_2) = \{r_1, r_2\}$ . We need to update the following four





■ **Figure 7** An example RsA

regions of the Venn diagram of  $r_1$  and  $r_2$ :  $r_1 \cap r_2$ ,  $r_1 \cap \overline{r_2}$ ,  $\overline{r_1} \cap r_2$ , and  $\overline{r_1} \cap \overline{r_2}$ . From the update function  $up$ , we see that the new values stored in  $r_1$  and  $r_2$  will be (we used primed versions of register names to denote their value after update)  $r'_1 = r_1$  (we do not consider  $\{in\}$  here because it has been discharged within *Out* in the previous step) and  $r'_2 = r_1 \cup r_2$ . The values of the regions will therefore be updated as follows:

$$\begin{aligned}
 r'_1 \cap r'_2 &= r_1 \cap (r_1 \cup r_2) & r'_1 \cap \overline{r'_2} &= r_1 \cap \overline{(r_1 \cup r_2)} & \overline{r'_1} \cap r'_2 &= \overline{r_1} \cap (r_1 \cup r_2) & \overline{r'_1} \cap \overline{r'_2} &= \overline{r_1} \cap \overline{(r_1 \cup r_2)} \\
 &= (r_1 \cap r_1) \cup (r_1 \cap r_2) & &= r_1 \cap \overline{r_1} \cap \overline{r_2} & &= (\overline{r_1} \cap r_1) \cup (\overline{r_1} \cap r_2) & &= \overline{r_1} \cap \overline{r_1} \cap \overline{r_2} \\
 &= r_1 \cup (r_1 \cap r_2) & &= \emptyset & &= \overline{r_1} \cap r_2 & &= \overline{r_1} \cap \overline{r_2} \\
 &= (r_1 \cap \overline{r_2}) \cup (r_1 \cap r_2) & & & & & & \\
 & & & & & & & (3)
 \end{aligned}$$

Note that in the last step of the calculation of  $r'_1 \cap r'_2$ , we used the fact that  $r_1 = (r_1 \cap r_2) \cup (r_1 \cap \overline{r_2})$  in order to obtain a union of regions. From the previous calculation, we see that *Transfer* should be set as follows:  $Transfer(\{r_1, r_2\}) = \{r_1, r_2\}$ ,  $Transfer(\{r_1\}) = \{r_1, r_2\}$ ,  $Transfer(\{r_2\}) = \{r_2\}$ , and  $Transfer(\{\emptyset\}) = \{\emptyset\}$ .  $\triangleleft$

Formally, *Transfer* is computed as follows. For every  $rgn_o \in 2^{\mathbf{R}}$ , let us compute the sets of sets of registers

$$pst_{\Pi}(rgn_o) = \{up(r_i) \cap \mathbf{R} \mid r_i \in rgn_o\} \quad \text{and} \quad ngt(rgn_o) = \bigcup \{up(r_i) \cap \mathbf{R} \mid r_i \notin rgn_o\} \quad (4)$$

(*pst* is for “positive” and *ngt* is for “negative”, which represent registers that occur positively and negatively, respectively, in the specification of the region of the Venn diagram  $rgn_o$ ). The intuition is that  $pst_{\Pi}(rgn_o)$  represents the update of  $rgn_o$  as the *product of sums* (intersection of unions), cf. the first line in Equation (3) in Example 24. Next, we convert the product of sums  $pst_{\Pi}(rgn_o)$  into a sum of products

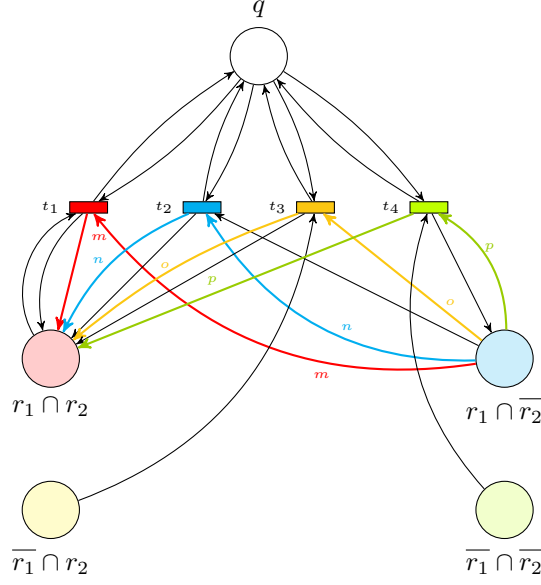
$$pst_{\Sigma}(rgn_o) = \coprod pst_{\Pi}(rgn_o) \quad (5)$$

where  $\coprod \{D_1, \dots, D_n\}$  is the *unordered Cartesian product* of sets  $D_1, \dots, D_n$ , i.e.,

$$\coprod \{D_1, \dots, D_n\} = \{\{d_1, \dots, d_n\} \mid (d_1, \dots, d_n) \in D_1 \times \dots \times D_n\}. \quad (6)$$

► **Example 25.** In the transition considered in Example 24, we obtain the following:

$$\begin{aligned}
 pst_{\Pi}(\{r_1, r_2\}) &= \{\{r_1\}, \{r_1, r_2\}\} & pst_{\Pi}(\{r_1\}) &= \{\{r_1\}\} & pst_{\Pi}(\{r_2\}) &= \{\{r_1, r_2\}\} & pst_{\Pi}(\{\emptyset\}) &= \{\{\emptyset\}\} \\
 pst_{\Sigma}(\{r_1, r_2\}) &= \{\{r_1\}, \{r_1, r_2\}\} & pst_{\Sigma}(\{r_1\}) &= \{\{r_1\}\} & pst_{\Sigma}(\{r_2\}) &= \{\{r_1\}, \{r_2\}\} & pst_{\Sigma}(\{\emptyset\}) &= \{\{\emptyset\}\} \\
 ngt(\{r_1, r_2\}) &= \emptyset & ngt(\{r_1\}) &= \{r_1, r_2\} & ngt(\{r_2\}) &= \{r_1\} & ngt(\{\emptyset\}) &= \{r_1, r_2\} \quad \triangleleft
 \end{aligned}$$



■ **Figure 8** TPN for the transition  $q - [a \mid \emptyset, \emptyset, \{r_1 \mapsto \{r_1, in\}, r_2 \mapsto \{r_1, r_2\}\}] \rightarrow q$  of the RsA from Figure 7 (corresponding colours represent the source position of *in* for the given transition).

Next, we modify  $pst_\Sigma$  into  $pst'_\Sigma$  by removing regions that are incompatible with  $ngt$  to obtain

$$pst'_\Sigma(rgn_o) = \{x \in pst_\Sigma(rgn_o) \mid x \cap ngt(rgn_o) = \emptyset\}. \quad (7)$$

► **Example 26.** In the running example, we would obtain the following values of  $pst'_\Sigma$ :

$$\begin{aligned} pst'_\Sigma(\{r_1, r_2\}) &= \{\{r_1\}, \{r_1, r_2\}\} & pst'_\Sigma(\{r_1\}) &= \emptyset \\ pst'_\Sigma(\{r_2\}) &= \{\{r_2\}\} & pst'_\Sigma(\{\emptyset\}) &= \{\{\emptyset\}\} \end{aligned}$$

Compare the results with the calculation in Equation (3).  $\triangleleft$

Lastly, for every  $rgn_i \in 2^{\mathbf{R}}$  such that  $pst'_\Sigma(rgn_o) \in rgn_i$ , we set  $Transfer(rgn_i) = rgn_o$ .

► **Example 27.** Continuing in the running example, we obtain

$$\begin{aligned} Transfer(\{r_1, r_2\}) &= \{r_1, r_2\} & Transfer(\{r_1\}) &= \{r_1, r_2\} \\ Transfer(\{r_2\}) &= \{r_2\} & Transfer(\{\emptyset\}) &= \{\emptyset\}, \end{aligned}$$

which is the same result as in Example 24. Figure 8 contains the TPN fragment for all TPN transitions constructed from  $\mathcal{A}$ 's transition  $q - [a \mid \emptyset, \emptyset, \{r_1 \mapsto \{r_1, in\}, r_2 \mapsto \{r_1, r_2\}\}] \rightarrow q$ .  $\triangleleft$

The following claim shows that this construction is indeed well defined.

▷ **Claim 28.** The function *Transfer* is well defined.

*Proof.* It is necessary to show that no set of values will be duplicated and assigned to two distinct regions when *Transfer* is calculated. According to the definition of *Transfer*, for all regions  $rgn' \in 2^{\mathbf{R}}$  such that  $pst'_\Sigma(rgn_o) \in rgn'$ , the value of  $Transfer(rgn')$  is set to be

$rgn_o$ , therefore we need to prove that for each pair of distinct regions  $rgn_1$  and  $rgn_2 \in 2^{\mathbf{R}}$  it holds that  $pst'_{\Sigma}(rgn_1) \cap pst'_{\Sigma}(rgn_2) = \emptyset$ . We prove this by contradiction.

Assume that there are two distinct regions  $rgn_1$  and  $rgn_2$  such that there exists a region  $rgn_3 \in pst'_{\Sigma}(rgn_1) \cap pst'_{\Sigma}(rgn_2)$ . According to the construction of  $pst'_{\Sigma}$ , it holds that

$$\begin{aligned} rgn_3 &\in pst_{\Sigma}(rgn_1) \quad \wedge \quad rgn_3 \cap ngt(rgn_1) = \emptyset \quad \text{and} \\ rgn_3 &\in pst_{\Sigma}(rgn_2) \quad \wedge \quad rgn_3 \cap ngt(rgn_2) = \emptyset. \end{aligned}$$

Then, according to the construction of  $pst_{\Sigma}$ ,

$$\begin{aligned} rgn_3 &\in \coprod pst_{\Pi}(rgn_1) \quad \wedge \quad rgn_3 \cap ngt(rgn_1) = \emptyset \quad \text{and} \\ rgn_3 &\in \coprod pst_{\Pi}(rgn_2) \quad \wedge \quad rgn_3 \cap ngt(rgn_2) = \emptyset. \end{aligned}$$

Following the construction of  $pst_{\Pi}$ :

$$\begin{aligned} (\forall r \in rgn_3 \exists P \in pst_{\Pi}(rgn_1): r \in P) \wedge (\forall P' \in pst_{\Pi}(rgn_1) \exists r' \in rgn_3: r' \in P') \wedge (rgn_3 \cap ngt(rgn_1) = \emptyset) \wedge \\ (\forall r \in rgn_3 \exists P \in pst_{\Pi}(rgn_2): r \in P) \wedge (\forall P' \in pst_{\Pi}(rgn_2) \exists r' \in rgn_3: r' \in P') \wedge (rgn_3 \cap ngt(rgn_2) = \emptyset). \end{aligned}$$

According to the construction of  $pst_{\Pi}$ , it holds that if  $P \in pst_{\Pi}(rgn)$  then there exists a register  $r \in rgn$  such that  $P = up(r_i)$ . Therefore, for each  $P \in pst_{\Pi}(rgn)$  there exists a register  $r' \in rgn'$  such that  $r \in P$ . Then, continuing in the proof, we obtain

$$\begin{aligned} (\forall r \in rgn_3 \exists up(r_i): r_i \in rgn_1 \wedge r \in up(r_i)) \wedge (\forall r^{\bullet} \in rgn_1 \exists r' \in rgn_3: r' \in up(r^{\bullet})) \wedge (rgn_3 \cap ngt(rgn_1) = \emptyset) \wedge \\ (\forall r \in rgn_3 \exists up(r_i): r_i \in rgn_2 \wedge r \in up(r_i)) \wedge (\forall r^{\bullet} \in rgn_2 \exists r' \in rgn_3: r' \in up(r^{\bullet})) \wedge (rgn_3 \cap ngt(rgn_2) = \emptyset) \end{aligned}$$

From the construction of  $ngt(rgn)$ , the formula  $rgn_i \cap ngt(rgn_j) = \emptyset$  is equivalent to the formula  $\forall r \in rgn_i \neg \exists r^* \notin rgn_j: r \in up(r^*)$ . Therefore:

$$\begin{aligned} (\forall r \in rgn_3 \exists up(r_i): r_i \in rgn_1 \wedge r \in up(r_i)) \wedge (\forall r^{\bullet} \in rgn_1 \exists r' \in rgn_3: r' \in up(r^{\bullet})) \wedge \\ (\forall r \in rgn_3 \neg \exists r^* \notin rgn_1: r \in up(r^*)) \wedge \\ (\forall r \in rgn_3 \exists up(r_i): r_i \in rgn_2 \wedge r \in up(r_i)) \wedge (\forall r^{\bullet} \in rgn_2 \exists r' \in rgn_3: r' \in up(r^{\bullet})) \wedge \\ (\forall r \in rgn_3 \neg \exists r^* \notin rgn_2: r \in up(r^*)) \end{aligned}$$

Further, we only make use of

$$(\forall r \in rgn_3 \neg \exists r^* \notin rgn_1: r \in up(r^*)) \quad \wedge \quad (\forall r^{\bullet} \in rgn_2 \exists r' \in rgn_3: r' \in up(r^{\bullet})),$$

and the fact that  $rgn_1$  and  $rgn_2$  are distinct. Therefore,  $\exists r^{dist}: r^{dist} \in rgn_2 \wedge r^{dist} \notin rgn_1$  (or vice versa).

By simplifying

$$\begin{aligned} (\exists r^{dist}: r^{dist} \in rgn_2 \wedge r^{dist} \notin rgn_1) \wedge \\ (\forall r \in rgn_3 \neg \exists r^* \notin rgn_1: r \in up(r^*)) \wedge \\ (\forall r^{\bullet} \in rgn_2 \exists r' \in rgn_3: r' \in up(r^{\bullet})), \end{aligned}$$

we obtain

$$\begin{aligned} (\exists r^{dist}: r^{dist} \in rgn_2 \wedge r^{dist} \notin rgn_1) \wedge \\ (\forall r \in rgn_3 \forall r^*: r^* \notin rgn_1 \rightarrow r \notin up(r^*)) \wedge \\ (\exists r' \in rgn_3: r' \in up(r^{dist})), \end{aligned}$$

which is clearly a contradiction, since  $r^{dist} \notin rgn_1 \wedge \exists r' \in rgn_3: r' \in up(r^{dist})$ .  $\triangleleft$

## 23:24 Register Set Automata (Technical Report)

Finally, the marking  $M_F$  to be covered is constructed as  $M_F = \mathbf{u}_{fin}$ . We have finished the construction of  $\mathcal{N}_A$ , now we need to show that it preserves the answer.

▷ **Claim 29.**  $\mathcal{L}(\mathcal{A}) \neq \emptyset$  iff the marking  $M_F$  is coverable in  $\mathcal{N}_A$ .

Proof. ( $\Rightarrow$ ) Let  $w \in (\Sigma \times \mathbb{D})^*$  such that  $w \in \mathcal{L}(\mathcal{A})$ . Moreover, assume that

$$\rho: c_0 \vdash_{t_1}^{w_1} c_1 \vdash_{t_2}^{w_2} \dots \vdash_{t_n}^{w_n} c_n$$

is an accepting run of  $\mathcal{A}$  on  $w$ . We will show that there exists a sequence of firings

$$\rho': M_{init}[t'_{init}] \ M_0[t'_1] M_1[t'_2] \dots [t'_n] M_n \ [t'_{fin}] M_{fin}$$

in  $\mathcal{N}_A$  such that  $M_{fin}$  covers  $M_F$ . In particular, we construct the markings and transitions as follows:

- $M_{init} = \mathbf{u}_{init}$  and  $t'_{init} = \langle \mathbf{u}_{init}, \mathbf{u}_{q_0}, \mathbf{id} \rangle$  for  $c_0 = (q_0, f_0)$ .
- For all  $0 \leq i \leq n$  with  $c_i = (q_i, f_i)$ , we set  $M_i$  as follows:

$$M_i = \{init \mapsto 0, fin \mapsto 0, q_i \mapsto 1\} \cup \{q \mapsto 0 \mid q \in Q \setminus \{q_i\}\} \cup \left\{ rgn \mapsto x \mid rgn \subseteq \mathbf{R}, x = \left| \bigcap_{r \in rgn} f_i(r) \right| \right\}$$

Furthermore,  $t'_i = \gamma(t_i, rgn_g)$  where  $rgn_g = \{r \mid d_i \in f_{i-1}(r)\}$ .

- $M_{fin}$  is as follows:

$$M_{fin} = \{init \mapsto 0, fin \mapsto 1\} \cup \{q \mapsto 0 \mid q \in Q\} \cup \{rgn \mapsto M_n(rgn) \mid rgn \subseteq \mathbf{R}\}$$

and  $t'_{fin} = \langle \mathbf{u}_{q_n}, \mathbf{u}_{fin}, \mathbf{id} \rangle$  for  $c_n = (q_n, f_n)$ .

Note that  $M_F$  is covered by  $M_{fin}$ .

We can now show by induction that  $\rho'$  is valid, i.e., all firings are enabled and respect the transition relation.

( $\Leftarrow$ ) Let  $\rho: M_{init}[t_0] \ M_0[t_1] M_1[t_2] \dots [t_n] M_n \ [t_{fin}] M_{fin}$  be a run of  $\mathcal{N}_A$  such that  $M_{fin} \geq M_F$ , where  $M_F$  is the final marking. We show that there exists a sequence of transitions

$$\rho': c_0 \vdash_{t'_1}^{w_1} c_1 \vdash_{t'_2}^{w_2} \dots \vdash_{t'_n}^{w_n} c_n$$

in  $\mathcal{A}$  on  $w = w_1 w_2 \dots w_n$ , such that  $c_n$  is a final configuration, and, therefore,  $w \in \mathcal{L}(\mathcal{A})$ . First, we notice the following easy to see invariant of  $\mathcal{N}_A$ , which holds for every  $M_i$ :

$$\sum_{p \in Q \cup \{init, fin\}} M_i(p) = 1 \tag{8}$$

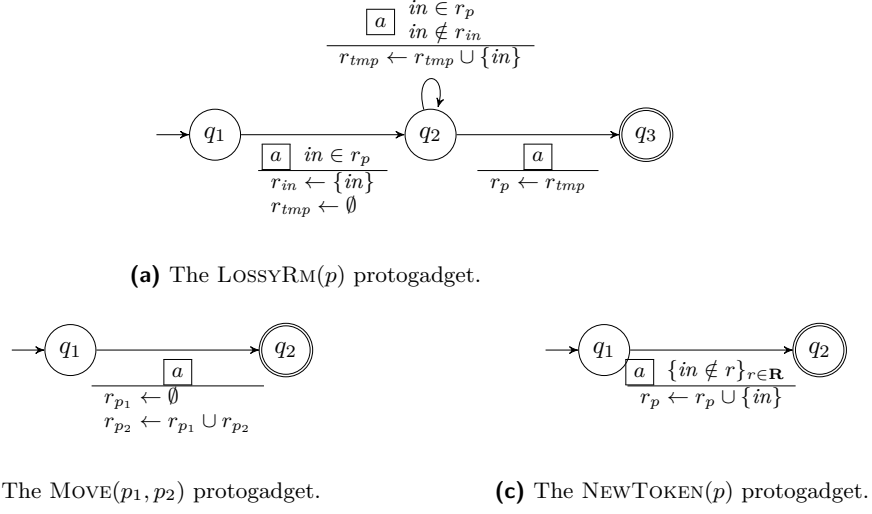
i.e., there is always exactly one token in any of the places in  $Q \cup \{init, fin\}$ .

Let us now construct  $\rho'$  as follows:

- $c_0 = (q_0, f_0)$  is constructed such that  $q_0$  is picked to be the state  $q_0 \in Q$  with  $M_0(q_0) = 1$  (this is well defined due to Equation (8)).
- For all  $1 \leq i \leq n$ , the transition  $t'_i$  is picked to be the transition such that  $t_i \in \gamma(t'_i, rgn_g)$  for some region  $rgn_g$ . The data value  $d_i$  of  $w_i$  is then chosen to be compatible with the guard of  $rgn_g$ , i.e.,  $d_i \in \bigcap_{r \in rgn_g} f_{i-1}(r)$  and  $d_i \notin \bigcup_{r \in \mathbf{R} \setminus rgn_g} f_{i-1}(r)$ .

It can then be shown by induction that, for all  $0 \leq i < n$ , the following holds:

- a.  $c_i = (q_i, f_i)$  where  $q_i$  is the (exactly one) state such that  $M_i(q_i) = 1$  and
- b. the transition  $t_{i+1}$  is enabled.



■ **Figure 9** Protogadgets used in the construction of  $\text{RsA}_{\mathcal{N}}$ .

We can then conclude that, since the last firing in  $\rho$  was  $M_n[t_{fin}]M_{fin}$ , then, from the construction of  $\mathcal{N}_{\mathcal{A}}$ , it holds that  $M_n(q_f) = 1$  for  $q_f \in F$  and so  $\rho'$  is accepting.  $\triangleleft$

Claim 29 and the observation that  $\mathcal{N}_{\mathcal{A}}$  is single-exponentially larger than  $\mathcal{A}$  conclude the proof ( $\mathbf{F}_{\omega}$  is closed under primitive-recursive reductions).  $\blacktriangleleft$

► **Lemma 30.** *The emptiness problem for  $\text{RsA}$  is  $\mathbf{F}_{\omega}$ -hard.*

**Proof.** The proof is based on a reduction of coverability in TPNs (which is  $\mathbf{F}_{\omega}$ -complete) to non-emptiness of  $\text{RsAs}$ . Intuitively, given a TPN  $\mathcal{N}$ , we will construct the  $\text{RsA}$   $\mathcal{A}_{\mathcal{N}}$  simulating  $\mathcal{N}$ , which will have the following structure:

- There will be the state  $q_{main}$ , which will be active before and after simulating the firing of TPN transitions.
- Each place of  $\mathcal{N}$  will be simulated by a register of  $\mathcal{A}_{\mathcal{N}}$ ; every token of  $\mathcal{N}$  will be simulated by a unique data value.
- For every TPN transition,  $\mathcal{A}_{\mathcal{N}}$  will contain a *gadget* that transfers data values between the registers representing the places active in the TPN transition. The gadget will start in  $q_{main}$  and end also in  $q_{main}$ .
- Coverability of a marking will be simulated by another gadget connected to  $q_{main}$  that will try to remove the number of tokens given in the marking from the respective places and arrive at the single final state  $q_{fin}$ .

Formally, let  $\mathcal{N} = (P, T, M_0)$  with  $P = \{p_1, \dots, p_n\}$  be a TPN. W.l.o.g. we can assume that  $M_0$  contains a single token in the place  $p_1$ , i.e.,  $M_0 = \{p_1 \mapsto 1, p_2 \mapsto 0, \dots, p_n \mapsto 0\}$ . We will show how to construct the  $\text{RsA}$   $\mathcal{A}_{\mathcal{N}} = (Q, \mathbf{R}, \Delta, \{q_{init}\}, \{q_{fin}\})$  over the unary alphabet  $\Sigma = \{a\}$  such that a marking  $M_f$  is coverable in  $\mathcal{N}$  iff the language of  $\mathcal{A}_{\mathcal{N}}$  is non-empty. The set of registers of  $\mathcal{A}_{\mathcal{N}}$  will be the set  $\mathbf{R} = \{r_{in}, r_{tmp}\} \cup \{r_p, r_{p'} \mid p \in P\}$ .

Let us now define the following *protogadgets*, which we will later use for creating a *gadget* for each TPN transition and a gadget for doing the coverability test. We define the following protogadgets:

1. The *Lossy Removal* protogadget, which simulates a (lossy) removal of one token from a place  $p$  is the RsA defined as  $\text{LOSSYRM}(p) = (\{q_1, q_2, q_3\}, \mathbf{R}, \Delta', \{q_1\}, \{q_3\})$  where  $\Delta'$  contains the following three transitions (cf. Figure 9a):

$$\Delta' = \left\{ \begin{array}{l} q_1 - \boxed{a \mid \{r_p\}, \emptyset, \{r_{in} \mapsto \{in\}, r_{tmp} \mapsto \emptyset\}} \rightarrow q_2, \\ q_2 - \boxed{a \mid \{r_p\}, \{r_{in}\}, \{r_{tmp} \mapsto \{r_{tmp}, in\}\}} \rightarrow q_2, \\ q_2 - \boxed{a \mid \emptyset, \emptyset, \{r_p \mapsto \{r_{tmp}\}\}} \rightarrow q_3 \end{array} \right\} \quad (9)$$

Intuitively, the protogadget stores the data value to be removed from  $p$  in a special register  $r_{in}$ . Next, it simulates the calculation of the difference of  $r_p$  and  $r_{in}$ . This is done by accumulating the values which are present in  $r_p$  and are not present in  $r_{in}$  into  $r_{tmp}$ . Since some values may get “lost”, and disappear because of not being added to the accumulated difference, this protogadget is considered *lossy*.

2. The *Move* protogadget, which simulates *moving* all tokens from a place  $p_1$  to a place  $p_2$ , is the following RsA (also depicted in Figure 9b):

$$\text{MOVE}(p_1, p_2) = (\{q_1, q_2\}, \mathbf{R}, \{q_1 - \boxed{a \mid \emptyset, \emptyset, \{r_{p_1} \mapsto \emptyset, r_{p_2} \mapsto \{r_{p_1}, r_{p_2}\}\}} \rightarrow q_2\}, \{q_1\}, \{q_2\}).$$

Intuitively, the protogadget empties out the register which represents the place  $p_1$ . Its previous value is assigned to register representing the place  $p_2$  in union with its value.

3. The *New Token* protogadget, which simulates adding a token to a place  $p$ , is defined as follows (depiction is in Figure 9c):

$$\text{NEWTOKEN}(p) = (\{q_1, q_2\}, \mathbf{R}, \{q_1 - \boxed{a \mid \emptyset, \mathbf{R}, \{r_p \mapsto \{r_p, in\}\}} \rightarrow q_2\}, \{q_1\}, \{q_2\}).$$

Intuitively, the protogadget adds the unique data value from the input tape into the register representing the place  $p$ . The uniqueness of the data value is ensured by the  $g^\#$ , which requires that the *in* does not belong to any of the registers from  $\mathbf{R}$ .

For convenience, we will use the following notation. Let  $\mathcal{A}_1 = (Q_1, \mathbf{R}, \Delta_1, \{q_1^I\}, \{q_1^F\})$  and  $\mathcal{A}_2 = (Q_2, \mathbf{R}, \Delta_2, \{q_2^I\}, \{q_2^F\})$  be a pair of RsAs with a single initial state and a single final state. We will use  $\mathcal{A}_1 \cdot \mathcal{A}_2$  to denote the RsA  $(Q_1 \uplus Q_2, \mathbf{R}, \Delta_1 \cup \{q_1^F - \boxed{a \mid \emptyset, \emptyset, \emptyset} \rightarrow q_2^I\} \cup \Delta_2, \{q_1^I\}, \{q_2^F\})$ . Moreover, for  $n \in \mathbb{N}_0$ , we use  $\mathcal{A}_1^{[n]}$  to denote the RsA defined inductively as

$$\begin{aligned} \mathcal{A}_1^{[0]} &= (\{q\}, \mathbf{R}, \emptyset, \{q\}, \{q\}), \\ \mathcal{A}_1^{[i+1]} &= \mathcal{A}_1^{[i]} \cdot \mathcal{A}_1. \end{aligned}$$

Intuitively,  $\mathcal{A}_1^{[n]}$  is a concatenation of  $n$  copies of  $\mathcal{A}_1$ .

For each TPN transition  $t = \langle \text{In}, \text{Out}, \text{Transfer} \rangle$ , we then create the *gadget* RsA  $\mathcal{A}_t$  in several steps.

1. First, we transform *In* into the RsA  $\mathcal{A}_{In} = \mathcal{A}_{In(p_1)} \cdot \dots \cdot \mathcal{A}_{In(p_n)}$  where every  $\mathcal{A}_{In(p_i)}$  is defined as  $\mathcal{A}_{In(p_i)} = \text{LOSSYRM}(p_i)^{[In(p_i)]}$ , i.e., it is a concatenation of  $In(p_i)$  copies of  $\text{LOSSYRM}(p_i)$ .
2. Second, from *Out* we create the RsA  $\mathcal{A}_{Out} = \mathcal{A}_{Out(p_1)} \cdot \dots \cdot \mathcal{A}_{Out(p_n)}$  with  $\mathcal{A}_{Out(p_i)}$  defined as  $\mathcal{A}_{Out(p_i)} = \text{NEWTOKEN}(p_i)^{[Out(p_i)]}$ , i.e., it is a concatenation of  $Out(p_i)$  copies of  $\text{NEWTOKEN}(p_i)$ .
3. Third, from *Transfer* we obtain the RsA  $\mathcal{A}_{Transfer} = \mathcal{A}_{Transfer(p_1)} \cdot \dots \cdot \mathcal{A}_{Transfer(p_n)} \cdot \mathcal{A}_{unprime(p_1)} \cdot \dots \cdot \mathcal{A}_{unprime(p_n)}$  such that  $\mathcal{A}_{Transfer(p_i)} = \text{MOVE}(r_{p_i}, r_{p'_j})$  with  $p_j = \text{Transfer}(p_i)$  and  $\mathcal{A}_{unprime(p_i)} = \text{MOVE}(p'_i, p_i)$ . Intuitively,  $\mathcal{A}_{Transfer}$  first moves the contents of all registers according to *Transfer* to primed instances of the target registers (in order to avoid mix-up) and then unprimes the register names.



4. Finally, we combine the RsAs created above into the single gadget  $\mathcal{A}_t = \mathcal{A}_{In} \cdot \mathcal{A}_{Transfer} \cdot \mathcal{A}_{Out}$ .

The initial marking will be encoded by a gadget that puts one new data value in the register representing the place  $p_1$ . For this, we construct the RsA  $\mathcal{A}_{M_0} = \text{NEWToken}(p_1)$  and rename its initial state to  $q_{init}$ .

The last ingredient we need is to create a gadget that will encode the marking  $M_f$ , whose coverability we are checking. For this, we construct the gadget  $\mathcal{A}_{M_f} = \mathcal{A}_{M_f(p_1)} \dots \mathcal{A}_{M_f(p_n)}$  where every  $\mathcal{A}_{M_f(p_i)}$  is defined as  $\mathcal{A}_{M_f(p_i)} = \text{LOSSYRM}(p_i)^{[M_f(p_i)]}$ , i.e., it is a concatenation of  $M_f(p_i)$  copies of  $\text{LOSSYRM}(p_i)$ . We rename the final state of  $\mathcal{A}_{M_f}$  to  $q_{fin}$ . W.l.o.g. we assume that the set of states of all constructed gadgets are pairwise disjoint.

We can now finalize the construction:  $\mathcal{A}_{\mathcal{N}}$  is obtained as the union of the following RsAs:  $\mathcal{A}_{M_0} = (Q_{M_0}, \mathbf{R}, \Delta_{M_0}, \{q_{init}\}, \{q_{M_0}^F\})$ ,  $\mathcal{A}_{M_f} = (Q_{M_f}, \mathbf{R}, \Delta_{M_f}, \{q_{M_f}^I\}, \{q_{fin}\})$ , and  $\mathcal{A}_t = (Q_t, \mathbf{R}, \Delta_t, \{q_t^I\}, \{q_t^F\})$  for every  $t \in T$ , connected to the state  $q_{main}$ , i.e.,  $\mathcal{A}_{\mathcal{N}} = (Q, \mathbf{R}, \Delta, \{q_{init}\}, \{q_{fin}\})$  where

$$\begin{aligned} \blacksquare Q &= \{q_{main}\} \cup Q_{M_0} \cup Q_{M_f} \cup \bigcup_{t \in T} Q_t \text{ and} \\ \blacksquare \Delta &= \Delta_{M_0} \cup \Delta_{M_f} \cup \{q_{M_0}^F - (a \mid \emptyset, \emptyset, \emptyset) \rightarrow q_{main}, q_{main} - (a \mid \emptyset, \emptyset, \emptyset) \rightarrow q_{M_f}^I\} \cup \\ &\quad \bigcup_{t \in T} (\Delta_t \cup \{q_{main} - (a \mid \emptyset, \emptyset, \emptyset) \rightarrow q_t^I, q_t^F - (a \mid \emptyset, \emptyset, \emptyset) \rightarrow q_{main}\}). \end{aligned}$$

▷ **Claim 31.** The marking  $M_F$  is coverable in  $\mathcal{N}$  iff  $\mathcal{L}(\mathcal{A}_{\mathcal{N}}) \neq \emptyset$ .

Proof. ( $\Rightarrow$ ) Let there be the following run of  $\mathcal{N}$ :

$$\rho: M_0[t_1]M_1[t_2] \dots [t_n]M_n$$

such that  $M_n$  covers  $M_F$ . We will show that there exists a word  $w \in (\Sigma \times \mathbb{D})^*$  and a run

$$\begin{aligned} \rho': c(init, 0) \vdash_{t'_{(init, 1)}}^{w_1} c(init, 1) \vdash_{t'_{(init, 2)}}^{w_2} \dots c(init, k_{init}) \vdash_{t'_{(0, 0)}}^{w_{i_0}} & \quad \text{[initialization]} \\ c(0, 0) \vdash_{t'_{(0, 1)}}^{w_{i_0+1}} c(0, 1) \dots c(0, k_1) \vdash_{t'_{(1, 0)}}^{w_{i_1}} c(1, 0) \vdash_{t'_{(1, 1)}}^{w_{i_1+1}} c(1, 1) \dots c(1, k_1) \dots c(n-1, k_{n-1}) \vdash_{t'_{(n, 0)}}^{w_{i_n}} \\ c(n, 0) \vdash_{t'_{(fin, 1)}}^{w_{i_n+1}} c(fin, 1) \dots \vdash_{t'_{(fin, k_{fin})}}^{w_{i_{fin}}} c(fin, k_{fin}) & \quad \text{[finalization]} \end{aligned}$$

of  $\mathcal{A}_{\mathcal{N}}$  on  $w$  such that  $q \in F$  for  $c(fin, k_{fin}) = (q, \cdot)$ . The run  $\rho'$  will be constructed to preserve the following invariant for each  $0 \leq i \leq n$ :

$$c(i, 0) = (q_{main}, f_i) \quad \text{such that} \quad \forall p \in P: |f_i(r_p)| = M_i(p). \quad (10)$$

Therefore, configurations with state  $q_{main}$  represent the TPN's state after (or before) firing a transition. Firing a transition  $t$  is simulated by going to the gadget for  $t$  in  $\mathcal{A}_{\mathcal{N}}$  and picking input data values such that the run returns to  $q_{main}$  in as many steps as possible (this is to take the run through the  $\text{LOSSYRM}$  protogadgets that preserves the precise value of the marking). By induction on  $0 \leq i \leq n$ , we can show that the invariant in Equation (10) is preserved (the base case is proved by observing that the *initialization* part is correct).

( $\Leftarrow$ ) Let  $w \in \mathcal{L}(\mathcal{A}_{\mathcal{N}})$  and

$$\begin{aligned} \rho: c(init, 0) \vdash_{t'_{(init, 1)}}^{w_1} c(init, 1) \vdash_{t'_{(init, 2)}}^{w_2} \dots c(init, k_{init}) \vdash_{t'_{(0, 0)}}^{w_{i_0}} & \quad \text{[initialization]} \\ c(0, 0) \vdash_{t'_{(0, 1)}}^{w_{i_0+1}} c(0, 1) \dots c(0, k_1) \vdash_{t'_{(1, 0)}}^{w_{i_1}} c(1, 0) \vdash_{t'_{(1, 1)}}^{w_{i_1+1}} c(1, 1) \dots c(1, k_1) \dots c(n-1, k_{n-1}) \vdash_{t'_{(n, 0)}}^{w_{i_n}} \\ c(n, 0) \vdash_{t'_{(fin, 1)}}^{w_{i_n+1}} c(fin, 1) \dots \vdash_{t'_{(fin, k_{fin})}}^{w_{i_{fin}}} c(fin, k_{fin}) & \quad \text{[finalization]} \end{aligned}$$

be an accepting run of  $\mathcal{A}_{\mathcal{N}}$  on  $w$  such that for each  $0 \leq i \leq n$ , it holds that  $c_{(i,0)} = (q_{main}, \cdot)$ —this follows from the structure of  $\mathcal{A}_{\mathcal{N}}$ . We will construct a run

$$\rho': M_0[t_1]M_1[t_2] \dots [t_n]M_n$$

where each  $t_i$  is the TPN's transition corresponding to the gadget which the corresponding part of  $\rho$  traversed. For all  $0 \leq i \leq n$ , the following invariant will hold:

$$\forall p \in P: |f_i(r_p)| \leq M_i(p). \quad (11)$$

We note that the run  $\rho$  might not have been the “*most precise*” run of  $\mathcal{A}_{\mathcal{N}}$ , so the markings in the TPN run overapproximate the contents of  $\mathcal{A}_{\mathcal{N}}$ 's registers in  $\rho$ . The invariant can be proved by induction.  $\triangleleft$

Claim 31 and the observation that  $\mathcal{A}_{\mathcal{N}}$  is single-exponentially larger than  $\mathcal{N}$  (assuming binary encoding of the numbers in  $\mathcal{N}$ ) conclude the proof ( $\mathbf{F}_{\omega}$  is closed under primitive-recursive reductions).  $\blacktriangleleft$

From Lemma 23 and Lemma 30, we immediately obtain Theorem 7.

► **Theorem 7.** *The emptiness problem for RsA is decidable, in particular,  $\mathbf{F}_{\omega}$ -complete.*

## A.2 Proofs of Closure Properties

► **Theorem 9.** *The following closure properties hold for RsA:*

1. RsA is closed under union and intersection.
2. RsA is not closed under complement.

**Proof.** The proofs for closure under union and intersection are standard: for two RsAs  $\mathcal{A}_1 = (Q_1, \mathbf{R}_1, \Delta_1, I_1, F_1)$  and  $\mathcal{A}_2 = (Q_2, \mathbf{R}_2, \Delta_2, I_2, F_2)$  with disjoint sets of states and registers, the RsA  $\mathcal{A}_{\cup}$  accepting the union of their languages is obtained as  $\mathcal{A}_{\cup} = (Q_1 \cup Q_2, \mathbf{R}_1 \cup \mathbf{R}_2, \Delta_1 \cup \Delta_2, I_1 \cup I_2, F_1 \cup F_2)$ . Similarly,  $\mathcal{A}_{\cap}$  accepting their intersection is constructed as the product  $\mathcal{A}_{\cap} = (Q_1 \times Q_2, \mathbf{R}_1 \cup \mathbf{R}_2, \Delta', I_1 \times I_2, F_1 \times F_2)$  where

$$\begin{aligned} (s_1, s_2) - \boxed{a \mid g_1^{\in} \cup g_2^{\in}, g_1^{\notin} \cup g_2^{\notin}, up_1 \cup up_2} \rightarrow (s'_1, s'_2) \in \Delta' \\ \text{iff} \\ s_1 - \boxed{a \mid g_1^{\in}, g_1^{\notin}, up_1} \rightarrow s'_1 \in \Delta_1 \quad \text{and} \quad s_2 - \boxed{a \mid g_2^{\in}, g_2^{\notin}, up_2} \rightarrow s'_2 \in \Delta_2. \end{aligned}$$

Correctness of the constructions is clear.

For showing the non-closure under complement, consider the language  $L_{\neg \forall repeat}$  from Example 5, which can be accepted by RsA. Let us show that for the complement of the language, namely, the language

$$L_{\forall repeat} = \{w \mid \forall i \exists j: i \neq j \wedge \mathbb{D}[w_i] = \mathbb{D}[w_j]\}$$

where all data values appear at least twice, there is no RsA that can accept it.

Our proof is a minor modification of the proof of Proposition 3.2 in [17]. In particular, we show that if  $L_{\neg \forall repeat}$  were expressible using an RsA, then we could construct an RsA encoding accepting runs of a Minsky machine. Since emptiness of an RsA is decidable (cf. Theorem 7) and emptiness of a Minsky machines is not, we would then obtain a contradiction.

Let us consider a Minsky machine  $\mathcal{M}$  with two counters and instructions of the form  $(q, \ell, q')$  where  $q$  and  $q'$  are states of  $\mathcal{M}$  and  $\ell \in \{\text{inc}, \text{dec}, \text{ifzero}\} \times \{1, 2\}$  is the corresponding

counter operation. A run of  $\mathcal{M}$  is a sequence of instructions (which can be viewed as  $\Sigma$ -symbols) together with a data value  $d \in \mathbb{D}$  assigned to every symbol. The data values are used to match increments with decrements of the same counter (intuitively, we are trying to say that “each increment is matched with a decrement”, in order to express that the value of the counter is zero). For instance, consider the following run:

$$\begin{array}{ccccccc} (q_1, \text{inc}_1, q_2) & (q_2, \text{inc}_1, q_3) & (q_3, \text{dec}_1, q_3) & (q_3, \text{inc}_2, q_2) & (q_2, \text{dec}_1, q_1) & (q_1, \text{ifzero}_1, q_4) & \\ 12 & 42 & 12 & 17 & 42 & 7 & \end{array} \quad (12)$$

Here, the first increment of counter 1 is matched with the first decrement of the counter (both having data value 12) and the second increment of counter 1 is matched with the second decrement of the counter (both having data value 42). Since all increments of the counter are uniquely matched with a decrement, the test at the end is satisfied so  $\mathcal{M}$  would accept (we assume  $q_4$  is a final state). To can accept such words, we can construct an automaton that checks the following properties of the input word:

1. The first instruction is of the form  $(q_1, \cdot, \cdot)$  for  $q_1$  being the initial state of  $\mathcal{M}$ .
2. Each instruction of the form  $(\cdot, \cdot, q_i)$  is followed by an instruction of the form  $(q_i, \cdot, \cdot)$ .
3. All increments have different data values and all decrements have different data values.
4. Between every two  $(\cdot, \text{ifzero}_i, \cdot)$  instructions (or between the start and the first such an  $\text{ifzero}_i$  instruction),
  - a. every  $(\cdot, \text{dec}_i, \cdot)$  needs to be preceded by an  $(\cdot, \text{inc}_i, \cdot)$  instruction with the same data value and
  - b. every  $(\cdot, \text{inc}_i, \cdot)$  needs to be followed by a  $(\cdot, \text{dec}_i, \cdot)$  instruction with the same data value.

Properties 1 and 2 can be easily expressed using an NFA and, therefore, also using a DRsA. Property 3 is easily expressible using an RsA (in fact, using a DRsA) that collects data values of increments and decrements of each counter in registers (we need two registers for every counter). Property 4a is also expressible using an RsA (again, using a DRsA) that collects the data values of decrements and whenever it reads an increment, it checks whether it has seen the increment's data value before.

Let us now focus on Property 4b. The negation of this property would be “*there is an increment not followed by a decrement with the same data value*”. This negated property is essentially captured by the language  $L_{\neg \forall \text{repeat}}$  and so it is expressible using RsA (in fact, it can be expressed by an NRA with guessing; or by a simple NRA provided that we change the accepted language to be prepended by a sequence of data values that will be used in the run, separated from the run by a delimiter). Therefore, if an RsA could accept the complement of  $L_{\neg \forall \text{repeat}}$ , i.e., the language  $L_{\forall \text{repeat}}$ , then we would be able to solve the emptiness problem of a Minsky machine, which is a contradiction. ◀

► **Theorem 10.** *For each  $n \in \mathbb{N}$ , the following closure properties hold for  $\text{RsA}_n$ :*

1.  $\text{RsA}_n$  is closed under union.
2.  $\text{RsA}_n$  is not closed under intersection and complement.

**Proof.** The proof of closure under union is the same as in the proof of Theorem 9 with the exception that the results uses only registers  $\mathbf{R}_1$  (we assume  $|\mathbf{R}_1| = n$ ): all references to registers  $r \in \mathbf{R}_2$  are changed to references to  $f(r)$  where  $f: \mathbf{R}_2 \rightarrow \mathbf{R}_1$  is an injection.

To show non-closure under intersection, consider the two languages

$$\mathcal{L}_n^A = \{w \mid \forall i < n: \mathbb{D}[w_i] = \mathbb{D}[w_{|w|-i+1}]\} \text{ and } \mathcal{L}_n^B = \{w \mid \forall n \leq i < 2n: \mathbb{D}[w_i] = \mathbb{D}[w_{|w|-i+1}]\}$$

Intuitively,  $\mathcal{L}_n^A$  is the language of words where the first  $n$  data values in the word are repeated (in the reverse order) at the end of the word and  $\mathcal{L}_n^B$  is the language of words where the  $(n+1)$ -th to  $2n$ -th data values are repeated (also in the reverse order) at the  $2n$ -th to  $(n+1)$ -th position from the end. Both languages can be expressed via  $\text{NRA}_n$ , and therefore also via  $\text{RsA}_n$ . Their intersection is the language  $\mathcal{L}_n^{AB} = \{w \mid \forall i < 2n: \mathbb{D}[w_i] = \mathbb{D}[w_{|w|-i+1}]\}$ , which is the same as  $\mathcal{L}_{2n}^A$  and clearly needs  $2n$  registers.

Non-closure under complement follows from Theorem 9 (its proof uses  $\text{RsA}_1$ ). ◀

► **Theorem 11.** *DRsA is closed under union, intersection, and complement.*

**Proof.** The proof of closure of DRsA under union is standard. Let  $\mathcal{A}_1$  and  $\mathcal{A}_2$  be two complete DRsAs, such that  $\mathcal{A}_1 = (Q_1, \mathbf{R}_1, \Delta_1, I_1, F_1)$  and  $\mathcal{A}_2 = (Q_2, \mathbf{R}_2, \Delta_2, I_2, F_2)$ , and their sets of states and registers are disjoint. The DRsA  $\mathcal{A}_\cup$  accepting the union of their languages is obtained as  $\mathcal{A}_\cup = (Q_1 \times Q_2, \mathbf{R}_1 \cup \mathbf{R}_2, \Delta', I_1 \times I_2, F')$  where

$$(s_1, s_2) \xrightarrow{a \mid g_1^\in \cup g_2^\in, g_1^\notin \cup g_2^\notin, up_1 \cup up_2} (s'_1, s'_2) \in \Delta'$$

iff

$$s_1 \xrightarrow{a \mid g_1^\in, g_1^\notin, up_1} s'_1 \in \Delta_1 \quad \text{and} \quad s_2 \xrightarrow{a \mid g_2^\in, g_2^\notin, up_2} s'_2 \in \Delta_2,$$

and  $F' = \{(q_1, q_2) \mid q_1 \in F_1 \vee q_2 \in F_2\}$ .

The construction of the DRsA  $\mathcal{A}_\cap = (Q_1 \times Q_2, \mathbf{R}_1 \cup \mathbf{R}_2, \Delta', I_1 \times I_2, F'_\cap)$  accepting the intersection of  $\mathcal{L}(\mathcal{A}_1)$  and  $\mathcal{L}(\mathcal{A}_2)$  is similar to the construction of  $\mathcal{A}_\cup$ , with the exception of  $F'_\cap$ , which is obtained as  $F'_\cap = F_1 \times F_2$ .

The complement of DRsA is obtained in the standard way by completing it and swapping final and non-final states. Since the automaton is already deterministic, the correctness of the construction is obvious. ◀

► **Theorem 12.** *For each  $n \in \mathbb{N}$ , the following closure properties hold for  $\text{DRsA}_n$ :*

1.  $\text{DRsA}_n$  is closed under complement.
2.  $\text{DRsA}_n$  is not closed under union and intersection.

**Proof.** The closure under complement is trivial (complete the DRsA and swap final and non-final states; no new register is introduced).

To show that  $\text{DRsA}_n$  is not closed under intersection, we use languages  $\mathcal{L}_n^A$  and  $\mathcal{L}_n^B$  from the proof of Theorem 10. In particular, both these languages are in  $\text{DRA}_n$  (the  $\text{DRA}_n$  needs more states than the corresponding  $\text{NRA}_n$  because it cannot *guess* where the final part of the word starts and needs to consider all possibilities, making the  $\text{DRA}_n$  exponentially larger). Similarly as in the proof of Theorem 10, a DRsA for the intersection of the languages, the language  $\mathcal{L}_n^{AB}$ , would need at least  $2n$  registers.

Non-closure under union follows from De Morgan's laws. ◀

## B Proofs for Section 5

► **Theorem 14.** *When Algorithm 1 returns a DRsA  $\mathcal{A}'$ , then  $\mathcal{L}(\mathcal{A}) = \mathcal{L}(\mathcal{A}')$ .*

**Proof.** ( $\subseteq$ ) Let  $w = \langle a_1, d_1 \rangle \dots \langle a_n, d_n \rangle \in \mathcal{L}(\mathcal{A})$ . Then there is an accepting run  $\rho$  of  $\mathcal{A}$  on  $w$ , such that

$$\rho: (q_0, f_0) \vdash_{t_1}^{\langle a_1, d_1 \rangle} (q_1, f_1) \vdash_{t_2}^{\langle a_2, d_2 \rangle} \dots \vdash_{t_n}^{\langle a_n, d_n \rangle} (q_n, f_n)$$

with  $q_n \in F$ . Furthermore, let

$$\rho': ((S'_0, c'_0), f'_0) \vdash_{t'_1}^{\langle a_1, d_1 \rangle} ((S'_1, c'_1), f'_1) \vdash_{t'_2}^{\langle a_2, d_2 \rangle} \dots \vdash_{t'_n}^{\langle a_n, d_n \rangle} ((S'_n, c'_n), f'_n)$$

be the run of  $\mathcal{A}'$  on  $w$  ( $\mathcal{A}'$  is deterministic and complete, so  $\rho'$  is unique). We will show that  $\rho'$  is accepting, and so  $w \in \mathcal{L}(\mathcal{A}')$ .

Let us show that for all  $0 \leq i \leq n$ , the following conditions hold:

1.  $q_i \in S'_i$ ,
2.  $\forall r \in \mathbf{R}: f_i(r) \neq \perp \implies f_i(r) \in f'_i(r)$ , and
3.  $\forall r \in \mathbf{R}: c'_i(r) = \begin{cases} 0 & \text{iff } f'_i(r) = \emptyset, \\ 1 & \text{iff } |f'_i(r)| = 1, \\ \omega & \text{iff } |f'_i(r)| \geq 2. \end{cases}$

We proceed by induction on  $i$ :

- $i = 0$ : Since the (only) initial state of  $\mathcal{A}'$  is the macrostate  $(I, c_0 = \{r_i \mapsto 0 \mid r_i \in \mathbf{R}\})$  (cf. Line 24), and  $q_0 \in I$ , then condition 1 holds. Moreover,  $f_0(r) = \perp$  for every  $r \in \mathbf{R}$ , so condition 2 holds trivially, and so does condition 3 (the  $c'_i$  of all registers is initialised to zero on Line 1 and all registers are initialised to  $\emptyset$  in a run of an RsA).
- $i = j+1$ : We assume the conditions hold for  $i = j$ . Let there be the following transition from the  $j$ -th to the  $(j+1)$ -th configurations of the runs  $\rho$  and  $\rho'$ :

$$(q_j, f_j) \vdash_{t_{j+1}}^{\langle a_{j+1}, d_{j+1} \rangle} (q_{j+1}, f_{j+1})$$

and

$$((S'_j, c'_j), f'_j) \vdash_{t'_{j+1}}^{\langle a_{j+1}, d_{j+1} \rangle} ((S'_{j+1}, c'_{j+1}), f'_{j+1})$$

and let  $t_{j+1}: q_j \xrightarrow{a_{j+1} \mid g^-, g^\neq, up} q_{j+1}$ . Moreover, let  $g \subseteq \mathbf{R}$  be the set of registers  $r$  such that  $f(r) = d_{j+1}$  ( $g$  here corresponds directly to the  $g$  on Line 5 of Algorithm 1). Transition  $t'_{j+1}$  would then be  $(S'_j, c'_j) \xrightarrow{a_{j+1} \mid g, \mathbf{R} \setminus g, up'} (S'_{j+1}, c'_{j+1})$  where  $S'_{j+1}$  is constructed using  $a_{j+1}$  and  $g$  as on Lines 6–7.

We need to show the following:

- (i) transition  $t'_{j+1}: (S'_j, c'_j) \xrightarrow{a_{j+1} \mid g, \mathbf{R} \setminus g, up'} (S'_{j+1}, c'_{j+1})$  is enabled and
- (ii) conditions 1–3 hold for  $i = j+1$ .

► **Claim 32.** Transition  $t'_{j+1}: (S'_j, c'_j) \xrightarrow{a_{j+1} \mid g, \mathbf{R} \setminus g, up'} (S'_{j+1}, c'_{j+1})$  is enabled.

**Proof.** Since transition  $t_{j+1}$  is enabled in configuration  $(q_j, f_j)$ , it holds that

- (a)  $\forall r \in g^-: f_j(r) = d_{j+1}$  and
- (b)  $\forall r \in g^\neq: f_j(r) \neq d_{j+1}$ .

From (a) and the induction hypothesis (condition 2), we have that  $\forall r \in g: d_{j+1} \in f'_j(r)$ , so the  $g^\in$ -part of  $t'_{j+1}$ 's enabledness holds. Proving the  $g^\neq$ -part (i.e., that  $\forall r \in \mathbf{R} \setminus g: d_{j+1} \notin f'_j(r)$ ) is more difficult. We prove this by contradiction.

For the sake of contradiction, assume that there exists a register  $r \in g^\neq$  such that  $d_{j+1} \in f'_j(r)$  (other registers in  $\mathbf{R} \setminus g$  do not need to be considered because they do not affect the enabledness of  $t_{j+1}$ ). Because (i)  $f_j(r) \neq d_{j+1}$ , (ii)  $f_j(r) \in f'_j(r)$  (from condition 2 of the induction hypothesis), and (iii)  $d_{j+1} \in f'_j(r)$  (from the assumption), we know that  $|f'_j(r)| \geq 2$ . From condition 3 of the induction hypothesis, it holds that  $c'_j(r) = \omega$ . But then, since there is a register  $r \in g^\neq$  such that  $c'_j(r) = \omega$ , Algorithm 1 would on Line 8 return  $\perp$ , which gives us contradiction.  $\triangleleft$

▷ **Claim 33.** The following holds:

1.  $q_{j+1} \in S'_{j+1}$ ,
2.  $\forall r \in \mathbf{R}: f_{j+1}(r) \neq \perp \implies f_{j+1}(r) \in f'_{j+1}(r)$ , and
3.  $\forall r \in \mathbf{R}: c'_{j+1}(r) = \begin{cases} 0 & \text{iff } f'_{j+1}(r) = \emptyset, \\ 1 & \text{iff } |f'_{j+1}(r)| = 1, \\ \omega & \text{iff } |f'_{j+1}(r)| \geq 2. \end{cases}$

**Proof.** 1. Trivial.

2. Follows from the induction hypothesis and the definition of the update function  $up'$  on Lines 9–18.

3. Follows from the induction hypothesis and the definition of  $c'$  on Line 19.  $\triangleleft$

This concludes the first direction of the proof.

( $\supseteq$ ) Let  $w = \langle a_1, d_1 \rangle \dots \langle a_n, d_n \rangle \in \mathcal{L}(\mathcal{A}')$ . Then there is an accepting run  $\rho'$  of  $\mathcal{A}'$  on  $w$ , such that

$$\rho': ((S'_0, c'_0), f'_0) \vdash_{t'_1}^{\langle a_1, d_1 \rangle} ((S'_1, c'_1), f'_1) \vdash_{t'_2}^{\langle a_2, d_2 \rangle} \dots \vdash_{t'_n}^{\langle a_n, d_n \rangle} ((S'_n, c'_n), f'_n)$$

with  $(S'_n, c'_n) \in F'$ . We will construct in a backward manner a sequence of sets of  $\mathcal{A}$ 's configurations (i.e., pairs containing a state and assignment to registers)  $U_0, U_1, \dots, U_{n-1}, U_n$ , such that  $\forall 1 \leq i \leq n: U_i \in \mathcal{Q} \times (\mathbf{R} \rightarrow \mathbb{D})$ , that represents all accepting runs of  $\mathcal{A}$  over  $w$ , and show that  $U_0$  contains a configuration  $(q_0, f_0)$  with  $q_0 \in I$  and  $f_0 = \{r \mapsto \perp \mid r \in \mathbf{R}\}$ . Let us start with  $U_n$ , which we construct as  $U_n = \{(q_n, f_n) \mid q_n \in S'_n \cap F', f_n = \{r \mapsto d \mid d \in f'_n(r)\}\}$ . Now, given  $U_{i+1}$  for  $0 \leq i \leq n-1$ , we construct the set of previous configurations  $U_i$  as the set of pairs  $(q_i, f_i)$  for which the following conditions hold:

1.  $q_i \in S'_i$ ,
2. for every register  $r \in \mathbf{R}$ , it holds that  $f_i(r) \in f'_i(r) \cup \{\perp\}$ ,
3. there is a transition  $t_{i+1}: q_i \xrightarrow{a_{i+1} \mid g^=, g^\neq, up'} q_{i+1} \in \Delta$  such that
  - \*  $(q_{i+1}, f_{i+1}) \in U_{i+1}$  and
  - \*  $(q_i, f_i) \vdash_{t_{i+1}}^{\langle a_{i+1}, d_{i+1} \rangle} (q_{i+1}, f_{i+1})$ .

Let us now show that for all  $0 \leq i \leq n$ , the set  $U_i$  is nonempty. We proceed by backward induction.

- \*  $i = n$  (base case):  $\rho'$  is accepting, so there is at least one state in  $S'_n \cap F'$ .
- \*  $i = j + 1$  (induction hypothesis): we assume that  $U_{j+1} \neq \emptyset$ .
- \*  $i = j$  (induction step): because the transition  $t'_{j+1}: (S'_j, c'_j) \xrightarrow{a_{j+1} \mid g^\in, g^\neq, up'} (S'_{j+1}, c'_{j+1})$  is enabled in configuration  $((S'_j, c'_j), f'_j)$ , it needs to hold that
  - $\forall r \in g^\in: d_{j+1} \in f'_j(r)$  and
  - $\forall r \in g^\neq: d_{j+1} \notin f'_j(r)$ .



Then, for every  $q_{j+1}$  such that  $(q_{j+1}, f_{j+1}) \in U_{j+1}$ , from the construction of  $t'_{j+1}$ , there needs to exist a transition  $t_{j+1}: q_j \xrightarrow{a_{j+1} \mid g^-, g^\neq, up} q_{j+1} \in \Delta$  with  $g^- \subseteq g^\in$  and  $g^\neq \subseteq g^\neq$ . It is left to show that there is an assignment  $f_j$  such that  $(q_j, f_j) \in U_j$  and  $(q_j, f_j) \vdash_{t_{j+1}}^{(a_{j+1}, d_{j+1})} (q_{j+1}, f_{j+1})$ . But this clearly holds since if it did not hold, then the algorithm would fail on Lines 15–17 (when checking whether the update is overapproximating or not) and would not produce  $\mathcal{A}'$ . ◀

- **Theorem 15.** (a) For every  $\text{NRA}_1^-$ , there exists a DRsA accepting the same language.  
 (b) For every  $\text{URA}_1^-$ , there exists a DRsA accepting the same language.

**Proof.** (a) Let  $\mathcal{A}$  be an  $\text{NRA}_1^-$  and  $\mathcal{A}_r$  be its register-local version. Because  $\mathcal{A}_r$  contains no disequality guards, the only way how Algorithm 1 could fail is at Line 17. Since  $\mathcal{A}$  used at most one register  $r$ , then each state  $q$  of  $\mathcal{A}_r$  will also use at most one register  $r_q$  (the copy of  $r$  for  $q$ ). Then  $P$  on Line 15 will only be a set of elements with no dependency. For such a set, the Cartesian abstraction is precise, so the test on Line 17 will never cause an abort.

- (b) Let  $\mathcal{A}$  be a  $\text{URA}_1^-$  such that  $\mathcal{L}(\mathcal{A}) = \mathcal{L}$ . First, using Fact 1, we construct an  $\text{NRA}_1^- \mathcal{A}_N$  accepting  $\bar{\mathcal{L}}$  (note that although the first step in the construction in the proof of Fact 1 is to make  $\mathcal{A}$  complete, we actually do not need to add transitions with missing guards (which might make us add transitions with a  $\neq$  guard), but we can add transitions of the form  $\cdot \xrightarrow{a \mid \emptyset, \emptyset, \emptyset} q_{\text{sink}}$ , which do not introduce  $\neq$  guards). Then, we use (a) to convert  $\mathcal{A}_N$  into a DRsA  $\mathcal{A}_D$  accepting  $\bar{\mathcal{L}}$ . Finally, we complement  $\mathcal{A}_D$  (by completing it and swapping final and non-final states), obtaining a DRsA accepting  $\mathcal{L}$ . ◀

- **Corollary 17.** For any language in  $\mathcal{B}(\text{NRA}_1^-)$ , there exists a DRsA accepting it.

**Proof.** Let  $L \in \mathcal{B}(\text{NRA}_1^-)$  and  $t$  be a term describing  $L$  using unions, intersections, and complements, with atoms being  $\text{NRA}_1^-$  languages. W.l.o.g., we can assume that  $t$  is in the negation normal form (i.e., complements are only over atoms—it is easy to transform any term into this form using De Morgan’s laws and double-complement elimination).

We can construct a DRsA  $\mathcal{A}_D$  accepting  $L$  inductively as follows:

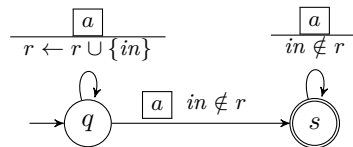
1. *Non-complemented literals:* if the literal is not complemented, we use Theorem 15(a) to obtain a DRsA accepting it.
2. *Complemented literals:* first, we use Fact 1 to obtain the  $\text{URA}_1^-$  accepting the complemented language and then use Theorem 15(b) to convert it into a DRsA.
3. *Unions and intersections:* we simply use Theorem 11 to obtain the resulting DRsA. ◀

## C Other Examples

- **Example 34.** Consider the language  $L_{\text{disjoint}}$  [32, Example 2.5] of words  $w = uv$  where  $u$  and  $v$  are non-empty and the data values in  $u$  and  $v$  are disjoint, i.e.,

$$L_{\text{disjoint}} = \{w \mid \exists u, v \in \Sigma^+ \times \mathbb{D}^+: w = uv \wedge u \neq \epsilon \wedge v \neq \epsilon \wedge \forall i, j: \mathbb{D}[u_i] \neq \mathbb{D}[v_j]\}$$

For instance, the language contains the word *ddee*, but does not contain the word *dede* (we only consider the data values here). An  $\text{RsA}_1$  accepting this language looks as follows:



Intuitively, the RsA stays in state  $q$ , accumulating the so-far seen values in register  $r$ , and at some point, it nondeterministically moves to state  $s$ , where it checks that no data value from the first part of the word appears in the second part.

We note that  $L_{disjoint}$  can be accepted neither by any ARA (intuitively, the different threads in an ARA cannot synchronize on the transition from the first part to the second part of the word), nor by any ARA(guess, spread) (which are strictly more expressive than ARAs). It also cannot be accepted by a DRsA. ◀

## D More Details about Extensions

### D.1 RsAs with Register Emptiness Test

An RsA *with register emptiness test* ( $\text{RsA}^{=\emptyset}$ ) is a tuple  $\mathcal{A}_E = (Q, \mathbf{R}, \Delta^{=\emptyset}, I, F)$  where  $Q, \mathbf{R}, I, F$  are the same as for RsAs and the transition relation  $\Delta^{=\emptyset}$  for  $\text{RsA}^{=\emptyset}$  is defined as  $\Delta^{=\emptyset} \subseteq Q \times \Sigma \times 2^{\mathbf{R}} \times 2^{\mathbf{R}} \times (\mathbf{R} \rightarrow 2^{\mathbf{R} \cup \{in\}}) \times Q$ . The semantics of a transition  $q \xrightarrow{a \mid g^{\in}, g^{\notin}, g^{=\emptyset}, up} s$  is such that  $\mathcal{A}_E$  can move from state  $q$  to state  $s$  if the  $\Sigma$ -symbol at the current position of the input word is  $a$  and the  $\mathbb{D}$ -symbol at the current position is in all registers from  $g^{\in}$  and in no register from  $g^{\notin}$  and all registers from  $g^{=\emptyset}$  are empty; the content of the registers is updated so that  $r_i \leftarrow \bigcup \{x \mid x \in up(r_i)\}$ .

► **Lemma 35.** *For every  $\text{RsA}^{=\emptyset}$   $\mathcal{A}$ , there exists an RsA  $\mathcal{A}'$  with the same language.*

**Proof.** Proof is done by showing the construction of modified RsA  $\mathcal{A}'$  corresponding to any given  $\text{RsA}^{=\emptyset}$   $\mathcal{A}$ .

The modification of  $\mathcal{A}'$  appears in the structure of states, where we code the information about the empty registers into the states themselves, and modify the transition relation accordingly. The information about the emptiness of registers is kept in a form of binary vector, denoted  $\mathbf{v}_\emptyset$ , such that  $\mathbf{v}_\emptyset[r_i] = 0$  iff  $r_i = \emptyset$ , and  $\mathbf{v}_\emptyset[r_i] = 1$  otherwise. Let  $\text{RsA}^{=\emptyset}$  be a tuple  $\mathcal{A}^{=\emptyset} = (Q, \mathbf{R}, \Delta, I, F)$  then the equivalent RsA is a tuple  $\mathcal{A}' = (Q', \mathbf{R}, \Delta', I', F')$  where  $Q' = Q \times \mathbf{v}_\emptyset$ ,  $I' = \{(q_i, \mathbf{v}_\emptyset^0) \mid q_i \in I \wedge \forall r_i \in \mathbf{R} : \mathbf{v}_\emptyset[r_i] = 0\}$ ,  $F' = \{(q_f, \mathbf{v}_\emptyset) \mid q_f \in F\}$ , and  $\Delta' = \{(q_1, \mathbf{v}_\emptyset^1) \xrightarrow{a \mid g^{\in}, g^{\notin}, up} (q_2, \mathbf{v}_\emptyset^2) \mid q_1 \xrightarrow{a \mid g^{\in}, g^{\notin}, S, up} q_2 \in \Delta \wedge \forall r \in S : \mathbf{v}_\emptyset^1[r] = 0\}$ , such that

$$\mathbf{v}_\emptyset^2[r_i] = \begin{cases} 1 & \text{if } (r_i \mapsto \{r_j\} \in up \wedge \mathbf{v}_\emptyset^1[r_j] \neq 0) \vee \\ & (r_i \mapsto \{in\} \in up) \text{ and} \\ 0 & \text{otherwise.} \end{cases} \quad (13)$$

◀

### D.2 RsAs with Register Equality Test

An RsA *with register equality test* ( $\text{RsA}^{=r}$ ) is a tuple  $\mathcal{A}_{Eq} = (Q, \mathbf{R}, \Delta^{=r}, I, F)$  where  $Q, \mathbf{R}, I, F$  are the same as for RsAs and the transition relation  $\Delta^{=r}$  for  $\text{RsA}^{=r}$  is defined as  $\Delta^{=r} \subseteq Q \times \Sigma \times 2^{\mathbf{R}} \times 2^{\mathbf{R}} \times (\mathbf{R} \rightarrow 2^{\mathbf{R}}) \times (\mathbf{R} \rightarrow 2^{\mathbf{R} \cup \{in\}}) \times Q$ . The semantics of a transition  $q \xrightarrow{a \mid g^{\in}, g^{\notin}, g^{=}, up} s$  is such that  $\mathcal{A}_E$  can move from state  $q$  to state  $s$  if the  $\Sigma$ -symbol at the current position of the input word is  $a$  and the  $\mathbb{D}$ -symbol at the current position is in all registers from  $g^{\in}$  and in no register from  $g^{\notin}$  and for all  $r_i \in g^{=}(r)$  it holds that  $r_i = r$ . The content of the registers is updated so that  $r_i \leftarrow \bigcup \{x \mid x \in up(r_i)\}$ .

► **Theorem 36.** *The emptiness problem for  $\text{RsA}^{=r}$  is undecidable.*

**Proof.** The proof is done by reduction from coverability in *Petri nets* with *inhibitor arcs*, which is an undecidable problem [35]. Given a  $\text{PN}_I \mathcal{N}_I$  with inhibitor arcs, we construct a corresponding  $\text{RsA}^{\text{=r}} \mathcal{A}_I^{\text{=r}}$ . The process of construction of  $\mathcal{A}_I^{\text{=r}}$  follows the reduction of TPN to RsA in the proof of Lemma 30 of Appendix A.1. The structure of the resulting automaton differs in *protogadgets* whose concatenation is used for construction of *gadgets* which make up the reduced  $\text{RsA}^{\text{=r}}$ .

The following *protogadgets* are used in the reduction:

1. The *EmptyEq* protogadget (depicted in Figure 10a), which simulates the inhibitor arc leading from place  $p$  is an  $\text{RsA}^{\text{=r}}$  defined as:

$$\text{EMPTYEQ}(p) = (\{q_1, q_2, q_3\}, \mathbf{R}, \{q_1 - \boxed{a \mid \emptyset, \emptyset, \emptyset, \{r_e \mapsto \emptyset\}} \rightarrow q_2, q_2 - \boxed{a \mid \emptyset, \emptyset, \{r_p \mapsto \{r_e\}\}, \emptyset} \rightarrow q_3, \{q_1\}, \{q_3\}\}).$$

Intuitively, the *EMPTYEQ* simulates a register emptiness test. At first, the protogadget explicitly assigns the value of empty set to the register  $r_e$  and then it compares its equality with the content of the register representing the place in which the inhibitor arc originates.

2. The *New Token* (depicted in Figure 10b) which simulates adding a token to a place  $p$ , is an  $\text{RsA}_{=\emptyset}^{\text{rm}}$  defined in the following way:

$$\text{NEWTOKEN}(p) = (\{q_1, q_2\}, \mathbf{R}, \{q_1 - \boxed{a \mid \emptyset, \mathbf{R}, \emptyset, \{r_p \mapsto \{r_p, \text{in}\}\}} \rightarrow q_2, \{q_1\}, \{q_2\}\}).$$

Intuitively, for each arc originating in the transition and ending in a particular place, a token is added to the register representing the destination. The guard ensures that the added value is not already present within the register, so that the number of values actually increases.

3. The *Non-lossy Remove Token* (depicted in Figure 10c) which simulates removal of a token from a place  $p$  is an  $\text{RsA}^{\text{=r}}$  defined in the following way:

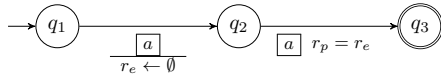
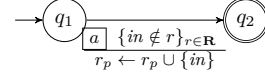
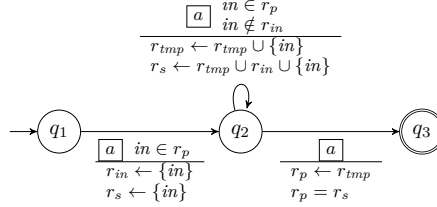
$$\begin{aligned} \text{NONLOSSYRM}(p) = (\{q_1, q_2, q_3\}, \mathbf{R}, \{q_1 - \boxed{a \mid \{r_p\}, \emptyset, \emptyset, \{r_{\text{in}} \mapsto \{\text{in}\}, r_s \mapsto \{\text{in}\}\}} \rightarrow q_2, \\ q_2 - \boxed{a \mid \{r_p\}, \{r_{\text{in}}\}, \emptyset, \{r_{\text{tmp}} \mapsto \{r_{\text{tmp}}, \text{inp}\}, r_s \mapsto \{r_{\text{tmp}}, r_{\text{in}}, \text{in}\}\}} \rightarrow q_2, \\ q_2 - \boxed{a \mid \emptyset, \emptyset, \{r_p \mapsto \{r_s\}\}, \{r_p \mapsto \{r_{\text{tmp}}\}\}} \rightarrow q_3, \{q_1\}, \{q_2\}\}). \end{aligned}$$

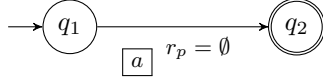
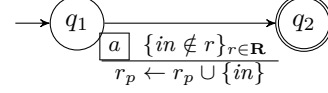
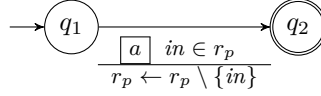
Intuitively, for each arc originating in a place  $p$  and terminating in transition, respective number of values has to be removed from the register representing place  $p$ . Therefore on each *protogadget* of this kind, one value is removed from a register representing the source place in a lossless manner. The quality of being lossless is necessary in order to sustain the semantics of source PN. Otherwise, some transitions may be enabled even though they were not in the source PN.

◀

### D.3 RsAs with Removal and Register Emptiness Test

An RsA with *removal and register emptiness test* ( $\text{RsA}_{=\emptyset}^{\text{rm}}$ ) is a tuple  $\mathcal{A}_{RE} = (Q, \mathbf{R}, \Delta_{=\emptyset}^{\text{rem}}, I, F)$ , where  $Q, \mathbf{R}, I, F$  are the same as for RsAs and the transition relation  $\Delta_{=\emptyset}^{\text{rem}}$  is defined as  $\Delta_{=\emptyset}^{\text{rem}} \subseteq Q \times \Sigma \times 2^{\mathbf{R}} \times 2^{\mathbf{R}} \times 2^{\mathbf{R}} \times 2^{\mathbf{R}} \times (\mathbf{R} \rightarrow 2^{\mathbf{R} \cup \{\text{in}\}}) \times Q$ . The semantics of a transition  $q - \boxed{a \mid g^{\in}, g^{\notin}, g^{\text{=}\emptyset}, \text{rem}, \text{up}} \rightarrow s$  is such that  $\mathcal{A}_{RE}$  can move from state  $q$  to state  $s$  if the  $\Sigma$ -symbol at the current position of the input word is  $a$  and the  $\mathbb{D}$ -symbol at the current


 (a) The EMPTYEQ( $p$ ) protogadget.

 (b) The NEWTOKEN( $p$ ) protogadget.

 (c) The NONLOSSYRM( $p$ ) protogadget.

 ■ **Figure 10** Protogadgets used in the construction of the  $\text{RsA}^{\text{=r}}$  for  $\mathcal{N}_I$ .

 (a) The EMPTY( $p$ ) protogadget.

 (b) The NEWTOKEN( $p$ ) protogadget.

 (c) The TOKENREM( $p$ ) protogadget.

 ■ **Figure 11** Protogadgets used in the construction of the  $\text{RsA}_{=\emptyset}^{\text{rm}}$  for  $\mathcal{N}_I$ .

position is in all registers from  $g^{\in}$  and in no register from  $g^{\notin}$  and all registers from  $g^{\emptyset}$  are empty; with respect to the *rem* and *up*, the content of the registers is updated so that for all  $r_i$  from *rem*,  $r_i \leftarrow r_i \setminus \{\text{in}\}$  and  $r_i \leftarrow \bigcup \{x \mid x \in \text{up}(r_i)\}$ .

► **Theorem 37.** *The emptiness problem for  $\text{RsA}_{=\emptyset}^{\text{rm}}$  is undecidable.*

**Proof.** The proof is done by showing the reducibility from *reachability* in *Petri nets* with *inhibitor arcs*, which is an undecidable problem.

Given a  $\text{PN}_I \mathcal{N}_I$  with inhibitor arc, we construct the  $\text{RsA}_{=\emptyset}^{\text{rm}} \mathcal{A}_{\mathcal{N}_I}$ . The structure of the  $\text{RsA}_{=\emptyset}^{\text{rm}} \mathcal{A}_{\mathcal{N}_I}$  is similar to the structure of  $\text{RsA} \mathcal{A}_{\mathcal{N}}$  in the proof of the  $\mathbf{F}_\omega$ -hardness of emptiness in  $\text{RsA}$ , which can be seen in Lemma 30 of Appendix A.1.

The only difference is in *gadgets* used for simulation of respective transition in  $\text{PN}_I$  and gadget for doing the reachability test. These are created by concatenation of respective *protogadgets*, which are defined in the following way:

1. The *Empty* protogadget (depicted in Figure 11a), which simulates the inhibitor arc leading from the place  $p$  is an  $\text{RsA}_{=\emptyset}^{\text{rm}}$  defined as:

$$\text{EMPTY}(p) = (\{q_1, q_2\}, \mathbf{R}, \{q_1 \xrightarrow{a \mid \emptyset, \emptyset, \{r_p\}, \emptyset, \{r_p \mapsto \emptyset\}} q_2\}, \{q_1\}, \{q_2\}).$$

Intuitively, since the inhibitor arc enables its transition only if the source place is empty, the respective *protogadget* checks on its guard whether the register representing the source place is empty as well.

2. The *New Token* protogadget (depicted in Figure 11b) which simulates adding a token to a place  $p$ , is an  $\text{RsA}_{=\emptyset}^{rm}$  defined in the following way:

$$\text{NEWTOKEN}(p) = (\{q_1, q_2\}, \mathbf{R}, \{q_1 - \boxed{a \mid \emptyset, \mathbf{R}, \emptyset, \emptyset, \{r_p \mapsto \{r_p, in\}\}} \rightarrow q_2\}, \{q_1\}, \{q_2\}).$$

Intuitively, for each arc originating in the transition and ending in a particular place, a token is added to the register representing the destination. The guard ensures that the added value is not already present within the register, so that the number of values actually increases.

3. The *Remove Token* protogadget (depicted in Figure 11c) which simulates the removal of a token from a place  $p$ , is an  $\text{RsA}_{=\emptyset}^{rm}$  defined in the following way:

$$\text{TOKENREM}(p) = (\{q_1, q_2\}, \mathbf{R}, \{q_1 - \boxed{a \mid \emptyset, \{r_p\}, \emptyset, \{r_p\}, \{r_p \mapsto \{r_p\}\}} \rightarrow q_2\}, \{q_1\}, \{q_2\}).$$

Intuitively, for each token removed from the source place, one value is removed from the register which represents that place. ◀