Expressiveness and Static Analysis of Extended Conjunctive Regular Path Queries☆

Dominik D. Freydenberger, Nicole Schweikardt

Institut für Informatik, Goethe-Universität, Frankfurt am Main, Germany

Abstract

We study the expressiveness and the complexity of static analysis of *extended conjunctive regular path queries (ECRPQs)*, introduced by Barceló et al. (PODS '10). ECRPQs are an extension of *conjunctive regular path queries (CRPQs)*, a well-studied language for querying graph structured databases. Our first main result shows that query containment and equivalence of a CRPQ in an ECRPQ is undecidable. This settles one of the main open problems posed by Barceló et al. As a second main result, we prove a non-recursive succinctness gap between CRPQs and the CRPQ-expressible fragment of ECRPQs. Apart from this, we develop a tool for proving inexpressibility results for CRPQs and ECRPQs. In particular, this enables us to show that there exist queries definable by regular expressions with backreferencing, but not expressible by ECRPQs.

1. Introduction

Many application areas (e. g., concerning the Semantic Web or biological applications) consider *graph structured data*, where the data consists of a finite set of nodes connected by labeled edges. For querying such data, one usually needs to specify types of paths along which nodes are connected. A widely studied class of queries for graph structured databases are the *conjunctive regular path queries (CRPQs)* (cf., e. g., [4, 6, 7]), where types of paths can be described by regular expressions specifying labels along the paths. For modern applications, however, also more expressive query languages are desirable, allowing not only to specify regular properties of path labels, but also to compare paths based on, e. g., their lengths, labels, or similarity.

To start a formal investigation of such concepts, Barceló et al. [3] introduced the class of *extended conjunctive regular path queries (ECRPQs)*, allowing to use not only regular languages to express properties of individual paths, but also *regular relations* among several paths, capable of expressing certain associations

☆The present paper is a full version of the conference contributen [9].

Email addresses: freydenberger@em.uni-frankfurt.de (Dominik D. Freydenberger), schweika@informatik.uni-frankfurt.de (Nicole Schweikardt)

between paths. The authors of [3] investigated the complexity of query evaluation and static analysis of ECRPQs. While query containment is known to be decidable and EXPSPACE-complete for CRPQs [7, 4], it was shown to be undecidable for ECRPQs [3]. However, checking containment of an ECRPQ in a CRPQ still is decidable and EXPSPACE-complete [3]. (Un)Decidability of checking containment (or, equivalence) of a CRPQ in an ECRPQ was posed as an open question in [3].

In the present paper, we answer this question by showing that containment of a CRPQ in an ECRPQ is undecidable — even if the ECRPQ is, in fact, a CRPQ extended only by relations for checking equality of path labels (or, similarly, equal lengths of paths). Our proof proceeds by (a) simulating Turing machine runs by so-called *H-systems*, a concept from formal language theory generalizing pattern languages, and (b) using CRPQs and ECRPQs to represent languages described by H-systems. Our proof generalizes to (i) the case where one of the two queries is fixed, (ii) the case where all queries are Boolean and acyclic, and (iii) the problem of deciding equivalence rather than containment of CRPQs and ECRPQs.

Apart from the static analysis of queries, the present paper also investigates the expressiveness and succinctness of ECRPQs. Using the machinery developed for proving our undecidability results concerning static analysis, we show that CRPQ-definability of a given ECRPQ is undecidable, and that there is no recursive function f such that every CRPQ-definable ECRPQ of length n is equivalent to a CRPQ of length $f(n)$.

Concerning the expressivity of (E)CRPQs, to the best of our knowledge, tools for showing inexpressibility results have not been presented in the literature yet. We develop such tools, enabling us to show, for example, that no ECRPQ-query can return exactly those tuples of nodes between which there is a path whose length is a composite number (i. e., a number of the form nm for $n, m \geq 2$). Since these paths can be easily described by a *regular expression with backreferencing* (cf. [1, 8]) of the form $(a\,a^+)\%x\,x^+$, this refutes a claim of [3] stating that all regular expressions with backreferencing can be expressed by ECRPQs.

Structure of the paper. We start with the necessary notations and definitions in Section 2 where, in particular, the syntax and semantics of ECRPQs (and restrictions thereof) are defined. Section 3 is devoted to the static analysis of ECRPQs and CRPQs, showing that containment and equivalence of CRPQs in ECRPQs are undecidable. Section 4 investigates the relative succinctness between CRPQs and CRPQ-expressible ECRPQs and provides tools for proving limitations to the expressive power of CRPQs and ECRPQs.

2. Preliminaries

Let $\mathbb{N}$ denote the set of non-negative integers. We denote the *empty word* by ε. Let A, B be alphabets. A *morphism* (between A^* and B^*) is a function $h : A^* \to B^*$ with $h(uv) = h(u)h(v)$ for all $u, v \in A^*$. For every word $w \in A^*$,

$|w|$ stands for the length of w, and for every letter $a \in A$, $|w|_a$ denotes the number of occurrences of a in w.

2.1. DB-Graphs and Queries.

A Σ-*labeled db-graph* is a directed graph $G = (V, E)$, where V is a finite set of nodes, and $E \subseteq V \times \Sigma \times V$ is a finite set of directed edges with labels from Σ. A *path* ρ between two nodes v_0 and v_n in G with $n \geq 0$ is a sequence $v_0 a_1 v_1 \cdots v_{n-1} a_n v_n$ with $v_0, \ldots, v_n \in V$, $a_1, \ldots, a_n \in \Sigma$, and $(v_i, a_{i+1}, v_{i+1}) \in E$ for $0 \leq i < n$. We define the *label* $\lambda(\rho)$ of the path ρ by $\lambda(\rho) := a_1 \cdots a_n$. Furthermore, for every $v \in V$, we define the *empty path* $v\varepsilon v$, with $\lambda(v\varepsilon v) = \varepsilon$.

A central concept considered in the present paper are *regular relations* (cf. [3] and the references therein). Let Σ be a finite alphabet, let $\perp$ be a new symbol with $\perp \notin \Sigma$, and let $\Sigma_\perp := \Sigma \cup \{\perp\}$. Let $\overline{w} = (w_1, \ldots, w_k) \in (\Sigma^*)^k$, where $w_i = a_{i,1} \cdots a_{i,|w_i|}$ (and all $a_{i,j} \in \Sigma$). We define the string $[\overline{w}] \in (\Sigma_\perp^*)^k$ by $[\overline{w}] := b_1 \cdots b_n$, where n is the maximum of all $|w_i|$, and $b_j := (b_{j,1}, \ldots, b_{j,k})$, with $b_{j,i} = a_{i,j}$ if $j \leq |w_i|$, and $b_{j,i} = \perp$ if $j > |w_i|$. In other words, $[\overline{w}]$ is obtained by aligning all w_i to the left, and padding the unfilled space with $\perp$ symbols. A k-ary relation $R \subseteq (\Sigma^*)^k$ is called *regular* if the language $\{[\overline{r}] \mid \overline{r} \in R\}$ is regular.

Obviously, every regular language is a (unary) regular relation. In addition to this, the present paper focuses on the following k-ary regular relations ($k \geq 2$):

1. the *equality relation* $\mathrm{eq} := \{(w_1, \ldots, w_k) \mid w_1 = \ldots = w_k\}$,
2. the *length equality relation* $\mathrm{el} := \{(w_1, \ldots, w_k) \mid |w_1| = \ldots = |w_k|\}$.

Note that each of these relations needs to be defined w. r. t. a finite alphabet Σ, which we usually omit for the sake of brevity.

We now define ECRPQs and CRPQs, following the definitions from [3]. Fix a countable set of *node variables* and a countable set of *path variables*. Let Σ be a finite alphabet. An *extended conjunctive regular path query (ECRPQ) Q* over Σ is an expression of the form

$$\mathrm{Ans}(\overline{z}, \overline{\chi}) \leftarrow \bigwedge_{1 \leq i \leq m} (x_i, \pi_i, y_i), \bigwedge_{1 \leq j \leq l} R_j(\overline{\omega}_j), \tag{1}$$

such that $m \geq 1$, $l \geq 0$, and

1. each R_j is a regular expression that defines a regular relation over Σ,
2. $\overline{x} = (x_1, \ldots, x_m)$ and $\overline{y} = (y_1, \ldots, y_m)$ are tuples of (not necessarily distinct) node variables,
3. $\overline{\pi} = (\pi_1, \ldots, \pi_m)$ is a tuple of distinct path variables,
4. $\overline{\omega}_1, \ldots, \overline{\omega}_l$ are tuples of path variables, such that each $\overline{\omega}_j$ is a tuple of variables from $\overline{\pi}$, of the same arity as R_j,
5. $\overline{z}$ is a tuple of node variables among $\overline{x}$, $\overline{y}$, and

6. $\overline{\chi}$ is a tuple of path variables among those in $\overline{\pi}$.

The expression $\mathrm{Ans}(\overline{z}, \overline{\chi})$ is the *head*, and the expression to the right of $\leftarrow$ is the *body* of Q. If $\overline{z}$ and $\overline{\chi}$ are the empty tuple (i. e., the head is of the form $\mathrm{Ans}()$), Q is a *Boolean query.* The *relational* part of an ECRPQ Q is $\bigwedge_{1 \leq i \leq m}(x_i, \pi_i, y_i)$, and the *labeling* part is $\bigwedge_{1 \leq j \leq l} R_j(\overline{\omega}_j)$. We denote the set of node variables in Q by $\mathrm{nvar}(Q)$.

Intuitively, all variables are quantified existentially, and the words formed by the labels along the paths have to satisfy the respective relations. Formally, for every Σ-labeled *db*-graph G, every ECRPQ Q (of the form described in (1)) over Σ, every mapping σ from the node variables of Q to nodes in G, and every mapping μ from the path variables of Q to paths in G, we write $(G, \sigma, \mu) \models Q$ if

1. $\mu(\pi_i)$ is a path from $\sigma(x_i)$ to $\sigma(y_i)$ for every $1 \leq i \leq m$,
2. for each $\overline{\omega}_j = (\pi_{j_1}, \ldots, \pi_{j_k})$, $1 \leq j \leq l$, the tuple $(\lambda(\mu(\pi_{j_1})), \ldots, \lambda(\mu(\pi_{j_k})))$ belongs to the relation R_j.

Finally, we define the output of Q (of the form described in (1)) on G by

$$Q(G) \quad := \quad \{ \ \big(\sigma(\overline{z}), \mu(\overline{\chi})\big) \ \mid \ \sigma, \mu \text{ such that } (G, \sigma, \mu) \models Q \ \}.$$

As usual, if Q is Boolean, we model the Boolean constants `true` and `false` by the empty tuple () and the empty set $\emptyset$, respectively. In other words, $Q(G) = \texttt{true}$ iff there exist assignments σ and μ with $(G, \sigma, \mu) \models Q$.

Two queries Q and Q' are called *equivalent* ($Q \equiv Q'$, for short) if $Q(G) = Q'(G)$ for all *db*-graphs G. A query Q is said to be *contained* in a query Q' ($Q \subseteq Q'$, for short) if $Q(G) \subseteq Q'(G)$ for all *db*-graphs G.

With an ECRPQ Q we associate an edge-labeled directed graph H_Q^{lab} whose vertex set is the set of node variables occurring in Q, and where there is an edge from x to y labeled π iff (x, π, y) occurs in the relational part of Q. As in [3], we write H_Q to denote the (unlabeled) directed graph obtained from H_Q^{lab} by deleting the edge-labels (and removing duplicate edges). A query Q is called *acyclic* if H_Q is acyclic.

In accordance with [3], a *conjunctive regular path query (CRPQ)* Q over Σ is an ECRPQ over Σ of the form described in (1), where all relations R_j are unary relations, and (hence), all tuples $\overline{\omega}_j$ are singletons.

Thus, CRPQs can only refer to the languages that are allowed to occur along the paths, while ECRPQs can also describe relations between different paths.

The present paper devotes special attention to two classes of queries with an expressive power that lies strictly between CRPQs and ECRPQs: A *CRPQ with equality relations* is an ECRPQ where every relation in the labeling part is either of arity 1 (i. e., a regular language), or a k-ary eq-relation for some $k \geq 2$. Analogously, a *CRPQ with equal length relations* is an ECRPQ where every relation in the labeling part is either of arity 1, or a k-ary el-relation.

It is easy to see that ECRPQs and CRPQs can be transformed into queries in the following normal forms (note, though, that these transformations might increase the size of the queries):

Lemma 2.1. *For every ECRPQ* $Q = \mathrm{Ans}(\overline{z}, \overline{\chi}) \leftarrow \bigwedge_{1 \leq i \leq m} (x_i, \pi_i, y_i),\ \bigwedge_{1 \leq j \leq l} R_j(\overline{\omega}_j)$, *there exists a regular relation* R *of arity* m *such that* Q *is equivalent to the ECRPQ* $Q' := \mathrm{Ans}(\overline{z}, \overline{\chi}) \leftarrow \bigwedge_{1 \leq i \leq m} (x_i, \pi_i, y_i),\ R(\pi_1, \ldots, \pi_m)$.

Proof. As every relation R_i of arity m_i can be interpreted as a regular language over the alphabet $\{a, \perp\}^{m_i}$ that is recognized by some finite automaton M_i, one can obtain the relation R from these R_i by letting m be the maximum of all the m_i and by constructing a finite automaton M over the alphabet $\{a, \perp\}^m$ that simulates all M_i in parallel. □

Lemma 2.2. *For every CRPQ* $Q = \mathrm{Ans}(\overline{z}, \overline{\chi}) \leftarrow \bigwedge_{1 \leq i \leq m} (x_i, \pi_i, y_i),\ \bigwedge_{1 \leq j \leq l} L_j(\pi_{i_j})$ *(where* $i_j \in \{1, \ldots, m\}$*), there exist regular languages* $L'_1, \ldots, L'_m \subseteq \Sigma^*$ *such that* Q *is equivalent to the CRPQ* $Q' := \mathrm{Ans}(\overline{z}, \overline{\chi}) \leftarrow \bigwedge_{1 \leq i \leq m} (x_i, \pi_i, y_i),\ \bigwedge_{1 \leq i \leq m} L'_i(\pi_i)$.

Proof. Let Q be a CRPQ over Σ. For every path variable π_i, $1 \leq i \leq m$, we define $I_i := \{j \mid i_j = i\}$. We construct the labeling part Q' by defining atoms $L'_i(\pi_i)$ for $1 \leq i \leq m$ in the following way:

1. If I_i is empty, let $L'_i := \Sigma^*$,
2. if I_i contains exactly one element j, let $L'_i := L_j$,
3. if I_i contains more than one element, let $L'_i := \bigcap_{j \in I_i} L_j$. As every language L_j is regular, L'_i is also regular.

The relational part of Q' is identical to the relational part of Q; and it is easy to see that $Q \equiv Q'$ holds. □

Hence, for ECRPQs it suffices to consider just one regular relation of arity m; and for CRPQs, it suffices to consider just one regular language per path variable.

2.2. Turing Machines and H-Systems

Let $\mathcal{M}$ be a (deterministic) Turing machine with state set Q, initial state $q_0 \in Q$, halting state $q_H \in Q$, tape alphabet Γ (including the blank symbol), such that $Q \cap \Gamma = \emptyset$, and an input alphabet $\Gamma_I \subset \Gamma$ that does not include the blank symbol. We adopt the conventions that $\mathcal{M}$ accepts by halting, and does not halt in the first step (i. e., $q_0 \neq q_H$).

A *configuration* of $\mathcal{M}$ is a word $w_1 q w_2$, with $w_1, w_2 \in \Gamma^*$ and $q \in Q$. We interpret $w_1 q w_2$ as $\mathcal{M}$ being in state q, while the tape contains w_1 on the left side, and w_2 on the right side. The head is on the position of the first (leftmost) letter of w_2 (if $w_2 = \varepsilon$, $\mathcal{M}$ reads the blank symbol). We denote the successor relation on configurations of $\mathcal{M}$ by $\vdash_{\mathcal{M}}$. An *accepting run* of $\mathcal{M}$ is a sequence $C_0, \ldots, C_n$ of configurations of $\mathcal{M}$ (with $n \geq 1$), such that $C_0 \in q_0 \Gamma_I^*$ (C_0 is an initial configuration), $C_n \in \Gamma^* q_H \Gamma^*$ (C_n is an accepting configuration), and $C_i \vdash_{\mathcal{M}} C_{i+1}$ holds for all $0 \leq i < n$. Let $\Sigma := \Gamma \cup Q \cup \{\#\}$, where $\#$ is a new letter that does not occur in Γ or Q. We define the set of *valid computations* of $\mathcal{M}$ by $\mathrm{VALC}(\mathcal{M}) := \{\#C_0\# \cdots \#C_n\# \mid C_0, \ldots, C_n$ is an accepting run of $\mathcal{M}\}$, and denote its complement by $\mathrm{INVALC}(\mathcal{M}) := \Sigma^* \setminus \mathrm{VALC}(\mathcal{M})$. Finally, we define $\mathrm{dom}(\mathcal{M})$ to be the set of all $w \in \Gamma_I^*$ such that $\mathcal{M}$ halts after a finite number of steps when started in the configuration $q_0 w$.

By definition, $\mathrm{INVALC}(\mathcal{M}) = \Sigma^*$ holds if and only if $\mathrm{dom}(\mathcal{M}) = \emptyset$; and note that (given $\mathcal{M}$), the question if $\mathrm{dom}(\mathcal{M}) = \emptyset$ is undecidable.

As a technical tool for our proofs, we use the notion of *H-systems* to describe the sets $\mathrm{INVALC}(\mathcal{M})$ for Turing machines $\mathcal{M}$. Our notion of H-systems can be viewed as a generalization of pattern languages (cf. Salomaa [15]), or as a restricted version of the H-systems introduced by Albert and Wegner [2].

Definition 2.3. *An* H-system *(over the alphabet Σ) is a 4-tuple $H := (\Sigma, X, \mathcal{L}, \alpha)$, where (i) X and Σ are finite, disjoint alphabets, (ii) $\mathcal{L}$ is a function that maps every $x \in X$ to a regular language $\mathcal{L}(x) \subseteq \Sigma^*$ with $\varepsilon \in \mathcal{L}(x)$, and (iii) $\alpha \in (X \cup \Sigma)^+$.*

*A morphism $h : (\Sigma \cup X)^* \to \Sigma^*$ is H-*compatible *if $h(a) = a$ for every $a \in \Sigma$, and $h(x) \in \mathcal{L}(x)$ for every $x \in X$. We then define the language $L(H)$ that is generated by $H = (\Sigma, X, \mathcal{L}, \alpha)$ as $L(H) := \{h(\alpha) \mid h$ is an H-compatible morphism$\}$.*

For every finite, nonempty set of H-systems $\mathcal{H} = \{H_1, \ldots, H_k\}$, we define $L(\mathcal{H}) = \bigcup_{i=1}^{k} L(H_i)$.

In other words, the letters from Σ are constants, the letters from X are variables, and $L(H)$ is obtained from α by uniformly replacing every variable x with a word from $\mathcal{L}(x)$. We assume w.l.o.g. that X is chosen minimally; i. e., every $x \in X$ occurs in α. It is easy to see that H-systems are able to generate non-regular languages; e. g., the system $H = (\Sigma, \{x\}, \mathcal{L}, xx)$ with $\mathcal{L}(x) = \Sigma^*$ generates the language of all ww, $w \in \Sigma^*$.We use unions of H-system languages to describe the sets $\mathrm{INVALC}(\mathcal{M})$:

Lemma 2.4. *Given a Turing machine $\mathcal{M}$, one can effectively construct a set $\mathcal{H} = \{H_1, \ldots, H_k\}$ of H-systems (for some $k \geq 1$) such that* $\mathrm{INVALC}(\mathcal{M}) = L(\mathcal{H})$.

Proof. Let $\mathcal{M}$ be a Turing machine with state set Q and tape alphabet Γ, and define $\Sigma := Q \cup \Gamma \cup \{\#\}$. We approach the process of defining $\mathcal{H}$ from the following angle: Every word $w \in \mathrm{INVALC}(\mathcal{M})$ contains at least one error that prevents w from being an element of $\mathrm{VALC}(\mathcal{M})$. Most of these conditions can be described by a regular languages; e. g, if

$$w \;\notin\; \# q_0 (\Gamma_I)^* (\# \Gamma^* Q \Gamma^*)^* \# \Gamma^* q_H \Gamma^* \#,$$

w is not an encoding of a sequence of configurations of $\mathcal{M}$, or it is such an encoding, but the first configuration is not an initial configuration, or the last configuration is not a halting configuration. Hence, we can define a H-System

$H_1 := (\Sigma, \{x\}, \mathcal{L}_1, x)$, where $\mathcal{L}_1$ maps x to the complement of the language

$$\#q_0(\Gamma_I)^*(\#\Gamma^*Q\Gamma^*)^*\#\Gamma^*q_H\Gamma^*\#.$$

Thus, if $w \notin L(H_1)$, we know that w is an encoding of configurations $C_0, \ldots, C_n$ for some $n \geq 1$, such that C_0 is an initial configuration, and C_n is a halting configuration. All that remains is to describe all possible transition errors, i. .e, C_i, C_{i+1} for which $C_i \vdash_{\mathcal{M}} C_{i+1}$ does not hold. Most of these errors can be described using only regular languages, e. g., if when reading some $a \in \Gamma$ in a state $q \in Q$, $\mathcal{M}$ is supposed to enter a state $p \in Q$, we can describe all errors in the transition of states using an H-system $H = (\Sigma, \{x\}, \mathcal{L}, x)$, where

$$\mathcal{L}(x) \;:=\; \Sigma^*\#\Gamma^*qa\Gamma^*\#\Gamma^*(Q \setminus \{p\})\Sigma^*.$$

It is easy to see that $w \in L(H) \setminus L(H_1)$ if and only if w includes a sequence of configurations that contains a transition with the aforementioned error. All other state transition errors can be described analogously, as can be all errors regarding the symbols that $\mathcal{M}$ is supposed to write. For example, if $\mathcal{M}$ reads some $a \in \Gamma$ while in state $q \in Q$ and is supposed to write some $b \in \Gamma$, move the head to the right, and enter some state $p \in Q$, the regular language

$$\Sigma^*\#\Gamma^*qa\Gamma^*\#\Gamma^*(\Gamma \setminus \{b\})p\Sigma^*$$

describes all errors where a symbol other than b was written.

Of course, as INVALC($\mathcal{M}$) can be non-regular (if dom($\mathcal{M}$) is infinite), regular languages alone are not sufficient to describe all possible errors in a run of $\mathcal{M}$. More specifically, we cannot handle arbitrary errors in the preservation of the tape contents from one configuration to the other. Again, assume $\mathcal{M}$ reads some $a \in \Gamma$ while in state $q \in Q$ and is supposed to write some $b \in \Gamma$, move the head to the right, and enter some state $p \in Q$. In all these cases, a configuration $C = w_1qaw_2$ with $w_1, w_2 \in \Gamma^*$ is followed by the configuration $C' = w_1bpw_2$.

Our goal is to define H-expressions that capture all cases where the encoding of configurations $C_0, \ldots, C_n$ contains configurations $C_i = w_1qaw_2$, $C_{i+1} = w_3bpw_4$ where $w_1 \neq w_3$, or $w_2 \neq w_4$ holds (with $w_1, \ldots, w_4 \in \Gamma^*$). Note that, for all words $w, w' \in \Gamma^*$, $w \neq w'$ holds if and only if there exist words $u, v, v' \in \Gamma^*$ and letters $c, d \in \Gamma$ with $c \neq d$, $w = ucv$, and $w' = udv'$, or exactly one of w, w' is the empty word.

As errors described in the latter case (i. e., that exactly one of w_1, w_3 or of w_2, w_4 is empty) can be expressed using regular languages, we focus our explanation on the former case. In order to express these errors, for every $c \in \Gamma$, we define languages

$$\begin{aligned} L_{c,1} \;&:=\; \bigcup_{v \in \Gamma^*} \Sigma^*\#\, v\, c\Gamma^*qa\Gamma^*\#\, v\, (\Gamma \setminus \{c\})\Sigma^*, \\ L_{c,2} \;&:=\; \bigcup_{v \in \Gamma^*} \Sigma^*\#\Gamma^*qa\, v\, c\Gamma^*\#\Gamma^*bp\, v\, (\Gamma \setminus \{c\})\Sigma^*. \end{aligned}$$

These languages can be generated by H-systems, e.g., the system $(\Sigma, X, \mathcal{L}, \alpha)$ with $X = \{x_1, x_2, x_3, x_4\}$, $\alpha = x_1 x_2 x_3 x_2 x_4$ and

$$\mathcal{L}(x_1) = \Sigma^* \#, \qquad \mathcal{L}(x_2) = \Gamma^*,$$
$$\mathcal{L}(x_3) = c\Gamma^* qa\Gamma^* \#, \qquad \mathcal{L}(x_4) = (\Gamma \setminus \{c\})\Sigma^*$$

generates $L_{c,1}$. If $\mathcal{M}$ is supposed to move to the left instead of to the right, the corresponding H-expressions can be defined analogously. Hence, by defining appropriate H-expressions for all possible tape letters $a \in \Gamma$ and states $q \in Q$ and the corresponding actions of $\mathcal{M}$, $\mathcal{H}$ can be constructed effectively. $\square$

As we shall see in the next section, it is possible to reduce decision problems on finite unions of H-systems (and, hence, on the domains of Turing machines) to decision problems on CRPQs and ECRPQs.

3. Query Containment and Equivalence

3.1. Query Containment

The *query containment problem* is the problem to decide for two input queries Q and Q' whether $Q \subseteq Q'$.

The containment of CRPQs in CRPQs and of ECRPQs in CRPQs is known to be decidable and EXPSPACE-complete (cf. [7, 4] and [3], resp.). In [3], the authors proved the undecidability of the containment problem for ECRPQs, and mentioned the decidability of containment of CRPQs in ECRPQs as an important open problem. Our first main result states that this problem is undecidable, even if the ECRPQs are of a comparatively restricted form:

Theorem 3.1. *For every alphabet Σ with $|\Sigma| \geq 2$, the containment problem of CRPQs in CRPQs with equality relations over Σ is undecidable.*

The proof is a consequence of Lemma 2.4, the undecidability of the emptiness of $\mathrm{dom}(\mathcal{M})$ for Turing machines $\mathcal{M}$, and the following lemma:

Lemma 3.2. *Let Σ be an alphabet. For every set $\mathcal{H} = \{H_1, \ldots, H_k\}$ of H-systems over Σ, one can effectively construct an alphabet Σ', a CRPQ Q_1 over Σ', and a CRPQ with equality relations Q_2 over Σ' such that $Q_1 \subseteq Q_2$ if and only if $L(\mathcal{H}) = \Sigma^*$.*

Proof. Let $\Sigma = \{a_1, \ldots, a_s\}$ for some $s \geq 1$. Let $\mathcal{H}$ be a set of k H-systems $\mathcal{H} = \{H_1, \ldots, H_k\}$ over Σ (with $k \geq 1$). We define $\Sigma' := \Sigma \cup \{\bigstar, \$\}$, where $\bigstar$ and \$ are distinct letters that do not occur in Σ. Next, we define

$$Q_1 \ := \ \mathrm{Ans}() \leftarrow (x, \pi, y), L(\pi),$$

where $L := \$\bigstar a_1 \cdots a_s \bigstar \$ \bigstar \Sigma^* \bigstar \$$, and x and y are distinct variables. Thus, $Q_1(G) = \texttt{true}$ if and only if G contains a path ρ with $\lambda(\rho) \in L$.

The definition of Q_2 is more involved. Informally explained, Q_2 uses the structure provided by Q_1 to implement the union of the languages $L(H_i)$. We define Q_2 such that, for every *db*-graph G with $Q_1(G) = \mathtt{true}$, $Q_2(G) = \mathtt{true}$ holds if and only if there is a path ρ in G with $\lambda(\rho) = \$\bigstar a_1 \cdots a_s \bigstar\$\bigstar w \bigstar\$$, where $w \in L(\mathcal{H})$ (i. e., $w \in L(H_i)$ for some $H_i \in \mathcal{H}$).

Note that the paths ρ described by Q_1 contain exactly three occurrences of the \$ symbol, which can be understood to divide ρ into two parts, where the left part is labeled $\bigstar a_1 \cdots a_s \bigstar$. Likewise, the query Q_2 can be understood as consisting of two parts, which are to be defined in the subqueries $\bigwedge_{1 \leq i \leq k} \phi_i^{sel}$ and $\bigwedge_{1 \leq i \leq k} \phi_i^{cod}$, respectively. Our goal is to construct Q_2 in such a way that, when matching Q_2 to ρ, the ϕ_i^{sel} are used to *select* which H-system H_i is simulated in Q_2, while the actual *encoding* of that H-system is achieved by ϕ_i^{cod} (hence, the superscripts *sel* and *cod*). We define Q_2 as

$$\begin{aligned} Q_2 \;:=\; & \mathrm{Ans}() \leftarrow (x_0, c_1^{\$}, x_1), (x_{k+1}, c_2^{\$}, \hat{x}_1), (\hat{x}_{k+1}, c_3^{\$}, \hat{x}_{k+2}), \\ & L_{\$}(c_1^{\$}), L_{\$}(c_2^{\$}), L_{\$}(c_3^{\$}), \bigwedge_{1 \leq i \leq k} \phi_i^{sel}, \bigwedge_{1 \leq i \leq k} \phi_i^{cod} \end{aligned}$$

where $L_{\$} = \{\$\}$, and the ϕ_i^{sel} and ϕ_i^{cod} consist of relational and labeling atoms that shall be defined further down. As explained above, the subqueries ϕ_i^{sel} are used to select which H-system is active when matching Q_2 to a graph. These queries are defined by

$$\begin{aligned} \phi_i^{sel} \;:=\; & (x_i, c_{i,1}^{\bigstar}, y_{i,1}), (y_{i,1}, c_i^{a_1}, y_{i,2}), \ldots, (y_{i,s}, c_i^{a_s}, y_{i,s+1}), (y_{i,s+1}, c_{i,2}^{\bigstar}, x_{i+1}), \\ & L_{\bigstar}(c_{i,1}^{\bigstar}), L_{a_1}(c_i^{a_1}), \ldots, L_{a_s}(c_i^{a_s}), L_{\bigstar}(c_{i,2}^{\bigstar}), \mathrm{eq}(c_{i,1}^{\bigstar}, c_{i,2}^{\bigstar}) \end{aligned}$$

where $L_a := \{\varepsilon, a\}$ for each $a \in \{\bigstar, a_1, \ldots, a_s\}$.

In order to define each ϕ_i^{cod}, we need to consider the respective H-system H_i: Let $H_i = (\Sigma, X_i, \mathcal{L}_i, \alpha_i)$, where $\alpha_i = \beta_{i,1} \cdots \beta_{i,m_i}$ for some $m_i \geq 1$ and $\beta_{i,1}, \ldots, \beta_{i,m_i} \in (X \cup \Sigma)$. We define the relational part of ϕ_i^{cod} to be

$$(\hat{x}_i, c_{i,3}^{\bigstar}, z_{i,1}), (z_{i,1}, d_{i,1}, z_{i,2}), \ldots, (z_{i,m_i}, d_{i,m_i}, z_{i,m_i+1}), (z_{i,m_i+1}, c_{i,4}^{\bigstar}, \hat{x}_{i+1}),$$

where $c_{i,3}^{\bigstar}$, $c_{i,4}^{\bigstar}$, and all $d_{i,j}$ are (pairwise distinct) new path variables. We start the construction of the labeling part of ϕ_i^{cod} with the labeling atoms $L_{\bigstar}(c_{i,3}^{\bigstar})$, $L_{\bigstar}(c_{i,4}^{\bigstar})$, $\mathrm{eq}(c_{i,1}^{\bigstar}, c_{i,3}^{\bigstar})$, and $\mathrm{eq}(c_{i,1}^{\bigstar}, c_{i,4}^{\bigstar})$. Furthermore, we define a regular language $L_{i,j}$ for every $1 \leq j \leq m_i$ by $L_{i,j} := \mathcal{L}_i(\beta_{i,j})$ if $\beta_{i,j} \in X$, and $L_{i,j} := \{\varepsilon, \beta_{i,j}\}$ if $\beta_{i,j} \in \Sigma$. In addition to this, we add a label atom $\mathrm{eq}(c_i^{\beta_{i,j}}, d_{i,j})$ for every j with $\beta_{i,j} \in \Sigma$. Finally, for every j with $\beta_{i,j} \in X$ such that $\beta_{i,j}$ occurs more than once in α_i, we add a relation $\mathrm{eq}(d_{i,j}, d_{i,l})$ for every $l \neq j$ with $\beta_{i,l} = \beta_{i,j}$.

Note that the relation graph H_{Q_2} consists only of a path from x_0 to $\hat{x}_{k+1}$, where each node (except $\hat{x}_{k+1}$, the last node) has exactly one successor. Thus, the relation graph is acyclic and has no branches.

We claim that $L(\mathcal{H}) = \Sigma^*$ holds if and only if $Q_1 \subseteq Q_2$, which completes the proof of Lemma 3.2.

"$\Longrightarrow$": Assume that $L(\mathcal{H}) = \Sigma^*$, and let $G = (V, E)$ be a *db*-graph over Σ' with $Q_1(G) = \mathtt{true}$. By definition of Q_1, G contains a path ρ with $\lambda(\rho) =$

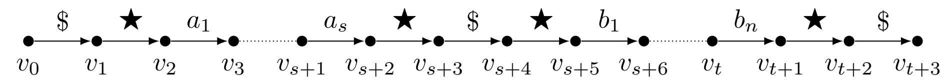

Figure 1: A graphic representation of the path ρ that is characteristic for all graphs G with $Q_1(G) = \texttt{true}$. To increase readability, this figure uses $t := s + n + 4$.

$\$\bigstar a_1 \cdots a_s \bigstar \$ \bigstar w \bigstar \$$ for some $w \in \Sigma^*$. Let $w = b_1 \cdots b_n$ with $n \geq 0$ and $b_j \in \Sigma$ for $1 \leq j \leq n$. Accordingly, there are nodes $v_0, \ldots, v_{s+n+7} \in V$ such that

$$\begin{aligned}\rho = & v_0 \,\$\, v_1 \bigstar v_2\, a_1\, v_3 \ldots v_{s+1}\, a_s\, v_{s+2} \bigstar v_{s+3} \,\$\, v_{s+4} \bigstar \\ & v_{s+5}\, b_1\, v_{s+6} \cdots v_{s+n+4}\, b_n\, v_{s+n+5} \bigstar v_{s+n+6} \,\$\, v_{s+n+7}.\end{aligned}$$

See Figure 1 for a graphic representation of this path. Although this does not matter for our considerations, note that these v_i are not necessarily distinct.

In order to show that $Q_2(G) = \texttt{true}$, we construct a node mapping σ and a path mapping μ such that $(G, \sigma, \mu) \models Q_2$. We first define

$$\begin{aligned} \sigma(x_0) &:= v_0, & \mu(c_1^{\$}) &:= v_0 \,\$\, v_1, \\ \sigma(x_1) &:= v_1, & & \\ \sigma(x_{k+1}) &:= v_{s+3}, & \mu(c_2^{\$}) &:= v_{s+3} \,\$\, v_{s+4}, \\ \sigma(\hat{x}_1) &:= v_{s+4}, & & \\ \sigma(\hat{x}_{k+1}) &:= v_{s+n+6}, & \mu(c_3^{\$}) &:= v_{s+n+6} \,\$\, v_{s+n+7}, \\ \sigma(\hat{x}_{k+2}) &:= v_{s+n+7}. & & \end{aligned}$$

As $L(\mathcal{H}) = \Sigma^*$, there is an i with $1 \leq i \leq k$ such that $w \in L(H_i)$. We now want to map the path described in ϕ_i^{sel} to the path between v_1 and v_{s+3} (for an illustration, see Figure 2). In order to achieve this, we define

$$\begin{aligned} \sigma(x_i) &:= v_1, & \mu(c_{i,1}^{\bigstar}) &:= v_1 \bigstar v_2, \\ \sigma(y_{i,1}) &:= v_2, & \mu(c_i^{a_1}) &:= v_2\, a_1\, v_3, \\ &\vdots & &\vdots \\ \sigma(y_{i,s+1}) &:= v_{s+2}, & \mu(c_i^{a_s}) &:= v_{s+1}\, a_s\, v_{s+2}, \\ \sigma(x_{i+1}) &:= v_{s+3}, & \mu(c_{i,2}^{\bigstar}) &:= v_{s+2} \bigstar v_{s+3}. \end{aligned}$$

As all other ϕ_j^{sel} are not needed, we define

$$\sigma(x_j) := \begin{cases} v_1 & \text{if } 1 \leq j < i, \\ v_{s+3} & \text{if } i < j \leq k, \end{cases}$$

and $\sigma(y_{j,l}) := \sigma(x_j)$ for all $j \neq i$, $1 \leq j \leq k$ and $1 \leq l \leq s+1$. Accordingly, for all $\pi_j \in \{c_{j,1}^{\bigstar}, c_{j,2}^{\bigstar}, c_j^{a_1}, \ldots, c_j^{a_s}\}$ with $j \neq i$, we define

$$\mu(\pi_j) := \begin{cases} v_1\, \varepsilon\, v_1 & \text{if } 1 \leq j < i, \\ v_{s+3}\, \varepsilon\, v_{s+3} & \text{if } i < j \leq k. \end{cases}$$

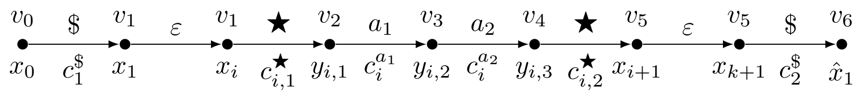

Figure 2: An illustration of the first half of the path ρ, compared to Q_2 under the assignments σ and μ, for the special case $s = 2$. The bottom row shows the node and path variables, while the top row contains the respective nodes and path labels. See also Figure 3 for an illustration of the second half.

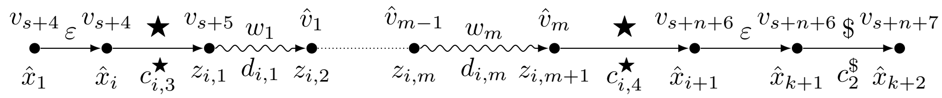

Figure 3: A graphic representation of the assignments σ and μ that are defined in the only-if-direction of the proof of Lemma 3.2. As in Figure 2, the bottom row shows the node and path variables, while the top row contains the respective nodes and path labels.

We can already observe that $(G, \sigma, \mu) \models Q_2$ holds modulo the subquery $\bigwedge_{1 \leq j \leq k} \phi_j^{cod}$, using the following reasoning: As $\lambda(\mu(c_j^{\$})) = \$$ for $j \in \{1, 2, 3\}$, $\lambda(\mu(c^{\$})) \in L_{\$} = \{\$\}$ is true, and $L_{\$}(c_j^{\$})$ is satisfied. Furthermore, for every $j \neq i$ with $1 \leq j \leq k$, we observe

$$\lambda(\mu(c_{j,1}^{\bigstar})) = \lambda(\mu(c_j^{a_1})) = \ldots = \lambda(\mu(c_j^{a_s})) = \lambda(\mu(c_{j,2}^{\bigstar})) = \varepsilon.$$

Due to $\varepsilon \in L_{\bigstar}, L_{a_1}, \ldots L_{a_s}$, each of

$$L_{\bigstar}(c_{j,1}^{\bigstar}), L_{a_1}(c_j^{a_1}), \ldots, L_{a_s}(c_j^{a_s}), L_{\bigstar}(c_{j,2}^{\bigstar}), \mathrm{eq}(c_{j,1}^{\bigstar}, c_{j,2}^{\bigstar})$$

is satisfied. Similarly, we observe

$$\begin{aligned}
\lambda(\mu(c_{i,1}^{\bigstar})) = \lambda(\mu(c_{i,2}^{\bigstar})) &= \bigstar \in L_{\bigstar}, \\
\lambda(\mu(c_i^{a_1})) &= a_1 \in L_{a_1}, \\
&\vdots \\
\lambda(\mu(c_i^{a_s})) &= a_s \in L_{a_s},
\end{aligned}$$

which demonstrates that ϕ_i^{sel} is satisfied as well.

All that remains is to find a proper assignment of the variables in ϕ_i^{cod} that describes the second half of ρ, while all other variables describe only the empty path. A graphic representation of the underlying idea can be found in Figure 3.

Accordingly, we define

$$\sigma(\hat{x}_j) := \begin{cases} v_{s+4} & \text{if } 1 \leq j < i, \\ v_{s+n+6} & \text{if } i < j \leq k, \end{cases}$$

and, likewise,

$$\sigma(z_{j,l}) := \begin{cases} v_{s+4} & \text{if } 1 \leq j < i, \\ v_{s+n+6} & \text{if } i < j \leq k, \end{cases}$$

for all l such that $z_{j,l}$ occurs in Q_2. Consequently, we define

$$\mu(c^{\bigstar}_{j,3}) = \mu(c^{\bigstar}_{j,4}) = \mu(d_{j,l}) = \sigma(\hat{x}_j)\, \varepsilon\, \sigma(\hat{x}_j)$$

for all $1 \leq j \leq k$, $j \neq i$.

We observe that for all $j \neq i$ with $1 \leq j \leq k$, ϕ^{cod}_j is satisfied: First, observe that

$$\lambda(\mu(c^{\bigstar}_{j,3})) = \lambda(\mu(c^{\bigstar}_{j,4})) = \varepsilon$$

holds. As $\lambda(\mu(c^{\bigstar}_{j,1})) = \varepsilon$, all of

$$L_{\bigstar}(c^{\bigstar}_{j,3}), L_{\bigstar}(c^{\bigstar}_{j,4}), \mathrm{eq}(c^{\bigstar}_{j,1}, c^{\bigstar}_{j,3}), \mathrm{eq}(c^{\bigstar}_{j,1}, c^{\bigstar}_{j,4})$$

are satisfied. Furthermore, for all $d_{j,l}$ that occur in ϕ^{cod}_j, $\lambda(\mu(d_{j,l})) = \varepsilon$. Therefore, every $L_{j,l}(d_{j,l})$ is satisfied, as $\varepsilon \in L_{j,l}$ holds by Definition 2.3. Moreover, as each of these paths is an empty path, all relations $\mathrm{eq}(d_{j,l}, d_{j,l'})$ in ϕ^{cod}_j are satisfied as well, which means that ϕ^{cod}_j is satisfied.

As the last remaining task, we need to complete the definition of σ and μ such that ϕ^{cod}_i is satisfied. In order to examine H_i in detail, assume that $H_i = (\Sigma, X_i, \mathcal{L}_i, \alpha_i)$, and let $\alpha_i = \beta_{i,1} \cdots \beta_{i,m_i}$ for some $m_i \geq 0$ with $\beta_{i,j} \in (X_i \cup \Sigma)$ for $1 \leq j \leq m_i$. By definition of $w \in L(H_i)$, there is an H_i-compatible morphism $h : (X_i \cup \Sigma)^* \to \Sigma^*$.

As $w = h(\alpha_i)$, there is a natural decomposition of w into factors $w_1 \cdots w_{m_i}$, which are defined by $w_j := h(\beta_j)$ for $1 \leq j \leq m_i$. We take special note of the subpaths of ρ that can be derived from these w_j, and define

$$\begin{aligned} n_0 &:= s + 5, & \hat{v}_0 &:= v_{n_0} = v_{s+5}, \\ n_j &:= s + 5 + |w_1 \cdots w_j|, & \hat{v}_j &:= v_{n_j} = v_{s+5+|w_1 \cdots w_j|} \end{aligned}$$

for each $1 \leq j \leq m_i$. Hence, for each j, the subpath of ρ between $\hat{v}_{j-1}$ and $\hat{v}_j$ is labeled with w_j. Note that $\hat{v}_{j-1} = \hat{v}_j$ might hold, in particular if $w_j = \varepsilon$. Also note that, by definition, $\hat{v}_{m_i} = v_{s+n+5}$.

Hence, we define

$$\begin{aligned} \sigma(\hat{x}_i) &:= v_{s+4}, & \mu(c^{\bigstar}_{i,3}) &:= v_{s+4} \bigstar v_{s+5}, \\ \sigma(z_{i,1}) &:= \hat{v}_0 = v_{s+5}, & & \\ &\vdots & & \\ \sigma(z_{i,m_i}) &:= \hat{v}_{m_i-1}, & & \\ \sigma(z_{i,m_i+1}) &:= \hat{v}_{m_i} = v_{s+n+5} & \mu(c^{\bigstar}_{i,4}) &= v_{s+n+5} \bigstar v_{s+n+6}, \\ \sigma(\hat{x}_{i+1}) &:= v_{s+n+6}. & & \end{aligned}$$

Finally, we define each $\mu(d_{i,j})$ with $1 \leq j \leq m_i$ to correspond to the subpath of ρ between $\hat{v}_{j-1}$ and $\hat{v}_j$ that is labeled with w_j.

We now prove that $L_{i,j}(d_{i,j})$ is satisfied for every $1 \leq j \leq m_i$. As in the definition of $L_{i,j}$, we distinguish the following cases:

1. If $\beta_{i,j} \in X_i$, $L_{i,j} = \mathcal{L}_i(\beta_{i,j})$. As $w_j = h(\beta_{i,j})$, and due to $h(\beta_{i,j}) \in \mathcal{L}_i(\beta_{i,j})$, $\lambda(\mu(d_{i,j})) \in L_{i,j}$,
2. if $\beta_{i,j} \in \Sigma$, $L_{i,j} = L_{\beta_{i,j}}$. As $w_j = h(\beta_{i,j}) = \beta_{i,j}$, $\lambda(\mu(d_{i,j})) \in L_{i,j}$ holds.

This also proves that, for every j with $\beta_{i,j} \in \Sigma$, $\lambda(\mu(d_{i,j})) = \beta_{i,j} = \lambda(\mu(c_i^{\beta_{i,j}}))$. Hence, these relations $\mathrm{eq}(c_i^{\beta_{i,j}}, d_{i,j})$ are satisfied as well. Finally, for every $\beta_{i,j} \in X_i$ that occurs more than once in α_i, we need to consider the relations $\mathrm{eq}(d_{i,j}, d_{i,l})$ for all l, j with $l \neq j$ and $\beta_{i,l} = \beta_{i,j}$. As h is a morphism, $\beta_{i,j} = \beta_{i,l}$ implies $h(\beta_{i,j}) = h(\beta_{i,l})$, and thus,

$$\lambda(\mu(d_{i,j})) = w_j = w_l = \lambda(\mu(d_{i,l})).$$

Obviously, $\mathrm{eq}(d_{i,j}, d_{i,l})$ is satisfied. We now have demonstrated that $(\sigma, \mu, G) \models Q_2$. Hence, $Q_2(G) = \texttt{true}$, and as G was chosen arbitrarily with $Q_1(G) = \texttt{true}$, $Q_1 \subseteq Q_2$ follows.

"$\Longleftarrow$": We prove this direction through its contraposition; i.e., we show that $L(\mathcal{H}) \neq \Sigma^*$ implies $Q_1 \not\subseteq Q_2$. Assume there is a $w \in \Sigma^*$ with $w \notin L(\mathcal{H})$. Let $w = b_1 \cdots b_n$ for some $n \geq 0$ with $b_i \in \Sigma$ for all $1 \leq i \leq k$. We define $G := (V, E)$, where $V := \{v_0, \ldots, v_{s+n+6}\}$ (and all elements of V are pairwise distinct), and

$$\begin{aligned}E := \{&(v_0, \$, v_1), (v_1, \bigstar, v_2), (v_2, a_1, v_3), \ldots, (v_{s+1}, a_s, v_{s+2}), (v_{s+2}, \bigstar, v_{s+3}),\\ &(v_{s+3}, \$, v_{s+4}), (v_{s+4}, \bigstar, v_{s+5}), (v_{s+5}, b_1, v_{s+6}), \ldots, (v_{s+n+4}, b_n, v_{s+n+5}),\\ &(v_{s+n+5}, \bigstar, v_{s+n+6}), (v_{s+n+6}, \$, v_{s+n+7})\}.\end{aligned}$$

In other words, G is an acyclic graph that consists solely of a path from v_0 to v_{s+n+7} labeled $\$\bigstar a_1 \cdots a_s \bigstar \$ \bigstar w \bigstar \$$. As $w \in \Sigma^*$, $Q_1(G) = \texttt{true}$ holds. For convenience, we denote this path by ρ.

For the sake of contradiction, assume $Q_1(G) \subseteq Q_2(G)$, which necessarily implies $Q_2(G) = \texttt{true}$. Thus, there are assignments σ, μ such that $(\sigma, \mu, G) \models Q_2$. As $\lambda(\rho)$ contains exactly three occurrences of \$, and as $L_\$(c_i^\$)$ occurs in Q_2 for $1 \leq i \leq 3$, we know that σ and μ must satisfy the following conditions:

$$\begin{aligned}
\sigma(x_0) &= v_0, & \mu(c_1^\$) &= v_0 \,\$\, v_1,\\
\sigma(x_1) &= v_1, & &\\
\sigma(x_{k+1}) &= v_{s+3}, & \mu(c_2^\$) &= v_{s+3} \,\$\, v_{s+4},\\
\sigma(\hat{x}_1) &= v_{s+4}, & &\\
\sigma(\hat{x}_{k+1}) &= v_{s+n+6}, & \mu(c_3^\$) &= v_{s+n+6} \,\$\, v_{s+n+7},\\
\sigma(\hat{x}_{k+2}) &= v_{s+n+7}. & &
\end{aligned}$$

As $\text{eq}(c^{\bigstar}_{i,1}, c^{\bigstar}_{i,2})$ needs to be satisfied for all $1 \leq i \leq k$, and as the subpath between v_1 and v_5 contains exactly two occurrences of $\bigstar$, there must be exactly one i with $\lambda(\mu(c^{\bigstar}_{i,1})) = \bigstar$ (although this i is not necessarily uniquely defined). We shall see that our assumption allows us to conclude that $w \in L(H_i)$, which leads to the intended contradiction. Due to our previous observations, the following must hold:

$$\begin{aligned} \sigma(x_i) &= v_1, & \mu(c^{\bigstar}_{i,1}) &= v_1 \bigstar v_2, \\ \sigma(y_{i,1}) &= v_2, & \mu(c^{a_1}_i) &= v_2\, a_1\, v_3, \\ &\vdots & &\vdots \\ \sigma(y_{i,s+1}) &= v_{s+2}, & \mu(c^{a_s}_i) &= v_{s+1}\, a_s\, v_{s+2}, \\ \sigma(x_{i+1}) &= v_{s+3}. & \mu(c^{\bigstar}_{i,2}) &= v_{s+2} \bigstar v_{s+3}. \end{aligned}$$

Now, note that Q_2 is acyclic. Therefore, the structure of G permits no other assignments than

$$\sigma(x_j) = \begin{cases} v_1 & \text{if } 1 \leq j < i, \\ v_5 & \text{if } i < j \leq k, \end{cases}$$

and $\sigma(y_{j,l}) = \sigma(x_j)$ for all $j \neq i$ and all $1 \leq l \leq s+1$. Accordingly,

$$\mu(c^{\bigstar}_{j,1}) = \mu(c^{a_1}_j) = \ldots = \mu(c^{a_s}_j) = \mu(c^{\bigstar}_{j,2}) = \sigma(x_j)\, \varepsilon\, \sigma(x_j)$$

holds for all these j. Thus, only the path variables from ϕ^{sel}_i are mapped to a nonempty path. The same phenomenon occurs for the variables of ϕ^{cod}_i: As Q_2 contains relations $\text{eq}(c^{\bigstar}_{i,1}, c^{\bigstar}_{i,3})$ and $\text{eq}(c^{\bigstar}_{i,1}, c^{\bigstar}_{i,4})$, we conclude that

$$\begin{aligned} \sigma(\hat{x}_i) &= v_{s+4}, & \mu(c^{\bigstar}_{i,3}) &= v_{s+4} \bigstar v_{s+5}, \\ \sigma(z_{i,1}) &= v_{s+5}, & & \\ \sigma(z_{i,m+1}) &= v_{s+n+5}, & \mu(c^{\bigstar}_{i,4}) &= v_{s+n+5} \bigstar v_{s+n+6}, \\ \sigma(\hat{x}_{i+1}) &= v_{s+n+6} & & \end{aligned}$$

holds. Let the H-system H_i be defined by $H_i = (\Sigma, X_i, \mathcal{L}_i, \alpha_i)$, where $\alpha_i = \beta_{i,1} \cdots \beta_{i,m_i}$ for some $m_i \geq 0$. By definition of Q_2, we know that ϕ^{cod}_i contains the path variables $d_{i,1}, \ldots, d_{i,m_i}$ (in addition to $c^{\bigstar}_{i,3}$ and $c^{\bigstar}_{i,4}$). This implies

$$\lambda(\mu(d_{i,1}))\, \lambda(\mu(d_{i,2})) \cdots \lambda(\mu(d_{i,m_i})) = b_1 \cdots b_n.$$

We define words $w_j := \lambda(\mu(d_{i,j}))$ for $1 \leq j \leq m_i$. In order to prove $w \in L(H_i)$, we show that these words can be used to define an H_i-compatible morphism h with $h(\alpha_i) = w$. First, we distinguish two possible cases for every $1 \leq j \leq m_i$:

1. If $\beta_{i,j} \in X_i$, $L_{i,j} = \mathcal{L}_i(\beta_{i,j})$ holds by definition of Q_2, which implies $w_j \in \mathcal{L}_i(\beta_{i,j})$,

2. if $\beta_{i,j} \in \Sigma$, $w_j = \beta_{i,j}$ must hold, as Q_2 contains label atoms $L_{\beta_{i,j}}(d_{i,j})$ and $\mathrm{eq}(c_i^{\beta_{i,j}}, d_{i,j})$, and $\lambda(\mu(c_i^{\beta_{i,j}})) = \beta_{i,j}$.

Furthermore, for all $j \neq l$ with $\beta_{i,j}, \beta_{i,l} \in X_i$ and $\beta_{i,j} = \beta_{i,l}$, Q_2 contains a label atom $\mathrm{eq}(d_{i,j}, d_{i,l})$; hence, $w_j = w_l$ holds. This allows us to define a morphism $h : (X_i \cup \Sigma)^* \to \Sigma^*$ by $h(\beta_{i,j}) := w_j$ for all $1 \leq j \leq m_i$.

Furthermore, as shown above for the two possible cases, h is H_i-compatible. Finally, $h(\alpha_i) = h(\beta_{i,1} \cdots \beta_{i,m_i}) = w$ holds by definition.

Thus, $w \in L(H_i) \subseteq L(\mathcal{H})$, which contradicts our initial assumption. This concludes the if-direction of the proof. □

By using standard encoding techniques for representing arbitrary finite alphabets by an alphabet of size 2, the proof of Theorem 3.1 now easily follows from Lemma 2.4, the undecidability of the emptiness of $\mathrm{dom}(\mathcal{M})$ for Turing machines $\mathcal{M}$, and Lemma 3.2. By using universal Turing machines instead of arbitrary Turing machines, we also obtain the following strengthening of Theorem 3.1:

Theorem 3.3. *For every alphabet Σ with $|\Sigma| \geq 2$, there are a fixed CRPQ Q_1 over Σ and a fixed CRPQ with equality relations Q_2 over Σ such that (i) the containment problem of Q_1 in CRPQs with equality relations, and (ii) the containment problem of CRPQs in Q_2 are both undecidable. This holds even if all queries are Boolean and acyclic.*

Proof. The first claim follows from the proof of Theorem 3.1, as Q_1 is fixed. In order to prove the second claim, we choose $\mathcal{M}$ to be a certain kind of universal Turing machine, and use Q_1 to choose the program number of the universal machine we want to simulate.

More precisely, let $\Psi : \mathbb{N} \times \mathbb{N} \to \mathbb{N}$ be a universal partially recursive function, i. e., for every partially recursive function $\phi : \mathbb{N} \to \mathbb{N}$, there is an $m \geq 0$ such that $\Psi_m(n) := \Psi(m, n) = \phi(n)$ for every $n \geq 0$. It is an elementary fact of recursion theory that such a function exists (cf. Kozen [13]), and moreover, there is a Turing machine $\mathcal{U}$ over some tape alphabet Γ such that

$$\mathrm{dom}(\mathcal{U}) = \{\mathtt{a}^m\, \mathtt{b}^n \mid \Psi(m, n) \text{ is defined}\},$$

where $\mathtt{a}, \mathtt{b}$ are two distinct letters in the input alphabet of $\mathcal{U}$. The machine $\mathcal{U}$ might be understood as simulating partial recursive function number Ψ_m (in the numbering that is defined by Ψ) on the input n.

We define $\Sigma := \Gamma \cup Q \cup \{\#\} = \{a_1, \ldots, a_s\}$ (for some $s \geq 2$) and $\Sigma' := \Sigma \cup \{\bigstar, \$\}$, and construct Q_2 and $\mathcal{H}$ from $\mathcal{U}$ as in the proof of Lemma 3.2. For every $m \geq 0$, we define a CRPQ $Q_{1,m}$ by

$$Q_{1,m} := \mathrm{Ans}() \leftarrow (x, \pi, y), L_m(\pi),$$

where

$$L_m := \$\bigstar a_1 \cdots a_s \bigstar \$ \bigstar \# q_0 \mathtt{a}^m \mathtt{b}^* \# \Sigma^* \bigstar \$,$$

and proceeding as in the proof of Lemma 3.2, *mutatis mutandis.* In other words, Q_1 does not generate arbitrary sequences, but sequences that start with the encoding of all possible initial sequences of simulations of the function Ψ_m.

Then, $Q_{1,m} \subseteq Q_2$ holds if and only if $\mathrm{dom}(\mathcal{U}) \cap \mathtt{a}^m\mathtt{b}^*$ is empty, which holds if and only if $\Psi_m(n)$ is undefined on all inputs n. As decidability of this problem would allow to decide the emptiness of the domain of partial-recursive functions (an undecidable problem), the second claim follows. Again, the common encoding techniques can be used to replace the alphabet Σ' with a binary alphabet.

Furthermore, all queries used are Boolean and acyclic by definition. □

Applying slight modifications to the proof of Lemma 3.2, we observe the same situation for ECRPQs that use length equality instead of equality relations:

Theorem 3.4. *For every alphabet Σ with $|\Sigma| \geq 2$, there are a fixed CRPQ Q_1 over Σ and a fixed CRPQ with length equality relations Q_2 over Σ such that (i) the containment problem of Q_1 in CRPQs with length equality relations, and (ii) the containment problem of CRPQs in Q_2, are both undecidable. This holds even if all queries are Boolean and acyclic.*

Proof. As we shall see, it suffices to replace all occurrences of eq in the queries that are constructed according to the proof of Theorem 3.3 (and the other proofs referenced therein) with el.

Note that, in order to prove Theorem 3.1, we do not need to express all possible unions of H-systems, but only those $\mathcal{H}$ that are derived from a Turing machine $\mathcal{M}$ as explained in Lemma 2.4. Furthermore, the construction in Lemma 3.2 uses the eq-predicates in two different contexts: First, on path variables that are associated with languages $\{a, \varepsilon\}$ for some $a \in \Sigma$, and second, on path variables that simulate variables x in H-expressions $(\Sigma, X, \mathcal{L}, \alpha)$ such that $|\alpha|_x \geq 2$.

For the first case, we can simply replace eq with el without changing the behavior of the query: Obviously, for all $w, w' \in \{a, \varepsilon\}$, $w = w'$ holds if and only if $|w| = |w'|$.

Regarding the second case, note that almost all H-expressions that are derived from the proof of Lemma 2.4 describe regular languages. The only non-regular languages that are constructed describe cases where $w \neq w'$ holds for certain words $w, w' \in \Gamma^*$, and characterizes this relation by $w \neq w'$ if and only if

1. there exist words $u, v, v' \in \Gamma^*$ and letters $c, d \in \Gamma$ with $c \neq d$, $w = ucv$, and $w' = udv'$, or

2. exactly one of w, w' is the empty word.

The first condition holds if and only if there exist words $u, u'v, v' \in \Gamma^*$ and letters $c, d \in \Gamma$ with $c \neq d$, $w = ucv$, $w' = u'dv'$, and $|u| = |u'|$, which demonstrate that in this case, the replacement of eq with el leads to the same results. □

3.2. Query Equivalence

The *query equivalence problem* is the problem to decide for two input queries Q and Q' whether $Q \equiv Q'$.

Another question specifically posed in [3] is whether the equivalence problem for CRPQs and ECRPQs is decidable. Using a variant of the proof of Theorem 3.3, we can answer this question negatively:

Theorem 3.5. *For every alphabet Σ with $|\Sigma| \geq 2$, there are a fixed CRPQ Q_1 over Σ and a fixed ECRPQ Q_2 over Σ such that (i) the equivalence problem of Q_1 and ECRPQs, and (ii) the equivalence problem of CRPQs and Q_2, are both undecidable. This holds even if all queries are Boolean and acyclic.*

Theorem 3.5 can be obtained from the proof of Theorem 3.3 by using the following lemma instead of Lemma 3.2:

Lemma 3.6. *Let Σ be an alphabet. For every regular language $L \subseteq \Sigma^*$ and every set $\mathcal{H} = \{H_1, \ldots, H_k\}$ of H-systems over Σ, one can effectively construct a CRPQ Q_1 and an ECRPQ Q_2 such that $Q_1 \equiv Q_2$ if and only if $L(\mathcal{H}) = L$.*

Proof. Let Σ, Σ', and Q_1 be defined as in the proof of Lemma 3.2, and let $L \subseteq \Sigma^*$ be a regular language.

We define $Q_1 := \mathrm{Ans}() \leftarrow (x, \pi, y), L'(\pi)$, where $L' := \$\bigstar a_1 \cdots a_s \bigstar\$\bigstar L \bigstar\$$ (as L is regular, L' is regular as well). In order to define Q_2, we introduce the regular k-ary relation $\mathrm{xor}(w_1, \ldots, w_k)$, which is defined by

$$\mathrm{xor} := \{(w_1, \ldots, w_k) \mid \text{there is exactly one } 1 \leq i \leq k \text{ with } w_i \neq \varepsilon\}.$$

We now obtain Q_2 by adding $\mathrm{xor}(c^{\bigstar}_{1,1}, \ldots, c^{\bigstar}_{k,1})$ to the query Q_2 used in the proof of Lemma 3.2. Then $Q_2(G) = \texttt{true}$ holds if and only if G contains a path ρ with $\lambda(\rho) \in \$\bigstar\, a_1 \cdots a_s \,\bigstar\$\bigstar L(\mathcal{H})\bigstar\$$. $\square$

If query equivalence were decidable, we could use Lemma 3.6 to decide whether $\mathrm{INVALC}(\mathcal{M}) = L$ for every Turing machine $\mathcal{M}$ and every regular language L. As this problem is undecidable, query equivalence must be undecidable. Hence, Theorem 3.5 follows.

Note that the ECRPQs in the proof use only one relatively simple relation in addition to the equality relations that are from the proof of Theorem 3.1. As in the proof of Theorem 3.4, this construction can be adapted to use length equality relations instead of equality relations.

4. Expressiveness and Relative Succinctness

4.1. Expressiveness of (E)CRPQs

In this section, we examine the expressive power of CRPQs and ECRPQs. In particular, we give a classes of query functions for which we characterize expressibility in CRPQS, and in ECRPQs over unary alphabets.

We say that a query function F is *CRPQ-expressible* (or *ECRPQ-expressible*) if there is a CRPQ (or ECRPQ, resp.) Q such that $Q(G) = F(G)$ for every Σ-labeled *db*-graph G.

For every language $L \subseteq \Sigma^*$, we define a query function F_L by

$$F_L(G) \ := \ \{(x, y) \mid G \text{ contains a path } \rho \text{ from } x \text{ to } y \text{ with } \lambda(\rho) \in L\}$$

for every Σ-labeled *db*-graph G. Analogously, we define a Boolean query function F_L^B by $F_L^B(G) := \texttt{true}$ if and only if $F_L(G) \neq \emptyset$.

The proofs presented in this section will use specific *db*-graphs G_w representing strings $w \in \Sigma^*$ as follows: If $w = b_1 \cdots b_{|w|}$ (with all $b_i \in \Sigma$), we define the *db*-graph $G_w := (V_w, E_w)$ by $V_w := \{v_0, \ldots, v_{|w|}\}$ (where all v_i are distinct nodes), and $E_w = \{(v_i, b_{i+1}, v_{i+1}) \mid 0 \leq i < |w|\}$. Thus, G_w consists of a path from v_0 to $v_{|w|}$ that is labeled with w.

Clearly, if $L \subseteq \Sigma^*$ such that F_L is expressible by an ECRPQ Q_L, then for all words $w \in \Sigma^*$ we have $w \in L$ iff $(v_0, v_{|w|}) \in Q_L(G_w)$.

Lemma 4.1. *Let Σ be an alphabet, let $L \subseteq \Sigma^*$. Then F_L is CRPQ-expressible if and only if L is regular.*

Proof. The *if*-direction is obvious: If L is regular, the CRPQ $Q := \mathrm{Ans}(x, y) \leftarrow (x, \pi, y), L(\pi)$ expresses F_L.

To prove the *only if*-direction, let $L \subseteq \Sigma^*$, and assume there exists a CRPQ

$$Q_L = \mathrm{Ans}(x, y) \leftarrow \bigwedge_{1 \leq i \leq m} (x_i, \pi_i, y_i), \bigwedge_{1 \leq i \leq m} L_i(\pi_i)$$

with $Q_L(G) = F_L(G)$ for every Σ-labeled *db*-graph G.

We will show that L is regular by considering the restricted class of *db*-graphs G_w for words $w \in \Sigma^*$.

Obviously, $Q_L(G_w)$ contains the pair $(v_0, v_{|w|})$ if and only if $w \in L$. The main idea of the proof is as follows: First, we construct a CRPQ Q'_L that is of a certain normal form, and satisfies $Q'_L(G_w) = Q_L(G_w)$ for all $w \in \Sigma^*$. Then, we show that the existence of Q'_L allows us to construct an NFA M with $L(M) = L$, showing that L is regular.

In order to construct Q'_L, we define a graph $H := (V, E)$, where V is the set of all node variables in Q_L, while E is the set of all (x_i, L_i, y_i) such that (x_i, π_i, y_i) occurs in the relational part of Q_L. In other words, we make use of the normal form for CRPQs, and interpret $H_{Q_L}^{lab}$ as being labeled with the languages L_i instead of the path variables π_i.

We now define the relation $\rightarrow$ on V by $z_1 \rightarrow z_2$ if, for some j, there is an edge $(z_1, L_j, z_2) \in E$, and define $\xrightarrow{*}$ as the reflexive transitive closure of $\rightarrow$. Our goal is to construct Q'_L by turning H into a graph that is acyclic, satisfies $x \xrightarrow{*} z \xrightarrow{*} y$ for every node z, has x as maximum of $\xrightarrow{*}$, and has y as minimum of $\xrightarrow{*}$.

We achieve this by executing the following modifications to V and E in order:

1. For all $z \in V$ with $z \xrightarrow{*} x$, remove z from V, and replace all occurrences of z in elements of E with x,
2. for all $z \in V$ with $y \xrightarrow{*} z$, remove z from V, and replace all occurrences of z in elements of E with y,
3. for all $z_1, z_2 \in V$ with $z_1 \xrightarrow{*} z_2 \xrightarrow{*} z_1$, remove z_2 from V, and replace all occurrences of z_2 in elements of E with z_1. Repeat this step as long as such z_1, z_2 exist.
4. remove all remaining loops from z, i. e., all (z, L_i, z).

Note that, at any point of the construction, we can interpret (V, E) as a CRPQ by "reversing" the construction; i. e., each edge is interpreted as an atom of the relational part, while each edge label corresponds to an atom in the labeling part that expresses the corresponding regular language.

We now prove that for the CRPQ Q'_L that is derived according to these removals, $Q'_L(G_w) = Q_L(G_w)$ holds for all $w \in \Sigma^*$. First, as Q'_L is obtained from Q_L by removing relations, $Q'_L \supseteq Q_L$ holds by definition. For the other direction, we first make the following basic observations. Assume that σ, μ are assignments with $(G_w, \sigma, \mu) \models Q_L$, $\sigma(x) = v_0$, and $\sigma(y) = v_{|w|}$ for some $w \in \Sigma^*$

As v_0 is of in-degree 0, we know that $\sigma(z) = \sigma(x) = v_0$ must hold for all z with $z \xrightarrow{*} x$. Likewise, as $v_{|w|}$ is of out-degree 0, $\sigma(z) = \sigma(y) = v_{|w|}$ holds for all z with $y \xrightarrow{*} z$. Furthermore, as G_w is acyclic, $\sigma(z_1) = \sigma(z_2)$ holds for all z_1, z_2 with $z_1 \xrightarrow{*} z_2 \xrightarrow{*} z_1$. Hence, for all languages L_i on the edges of $H^{lab}_{Q_L}$ that were removed during the construction process of Q'_L, $\varepsilon \in L_i$ must hold.

Now assume that σ', μ' are assignments such that $(G_{w'}, \sigma', \mu') \models Q'_L$, $\sigma'(x) = v_0$, and $\sigma'(y) = v_{|w'|}$ hold for some $w' \in \Sigma^*$ (in other words, $(v_0, v_{|w'|}) \in Q'_L(G_{w'})$ holds). In order to prove $(v_0, v_{|w'|}) \in Q_L(G_{w'})$, we define $\sigma(z) := \sigma'(z)$ for every node variable z that occurs in Q'_L (and, hence, also in Q_L), and $\mu(\pi) := \mu'(\pi)$ for every path variable π that occurs in Q'_L (and, hence, also in Q'_L). For the remaining node variables $z_1 \in \mathrm{nvar}(Q_L)$ that do not occur in Q'_L, define $\sigma(z_1) := \sigma(z_2)$, where z_2 is a variable with $z_2 \in \mathrm{nvar}(Q_L)$ (such a variable must exist, according to the construction procedure). Finally, as explained above, all remaining path variables can be assigned to the empty path for the appropriate node. Hence, $(v_0, v_{|w'|}) \in Q'_L(G_{|w'|})$ holds. As $w' \in \Sigma^*$ was chosen freely, this proves $Q_L(G_{w'}) = Q'_L(G_{w'})$ for all $w' \in \Sigma^*$.

Note that there might still exist some $z \in V$ such that $x \xrightarrow{*} z$ or $z \xrightarrow{*} y$ does not hold. In order to simplify our technical construction further down, for each such z, we add an edge (x, Σ^*, z) if $x \xrightarrow{*} z$ does not hold, and (z, Σ^*, y) if $z \xrightarrow{*} y$ does not hold. Again, this does not change the result of the corresponding query on all graphs G_w.

As a last step in our construction of the NFA M with $L(M) = L$, let $e_1, \ldots, e_k$ be any numbering of the edges in E. For every $z \in V$, let

$$\mathrm{in}(z) := \{i \mid e_i \in E,\ e_i \text{ ends in } z\},$$
$$\mathrm{out}(z) := \{i \mid e_i \in E,\ e_i \text{ starts in } z\}.$$

Furthermore, for every $e_i \in E$, let $M_i = (Q_i, \Sigma, \delta_i, q_{0,i}, F_i)$ be a DFA for the language that labels e_i. We now construct an NFA $M = (Q, \Sigma, \delta, q_0, F)$ that imitates the matching of Q'_L to graphs G_w by simulating all M_i in parallel.

In principle, M shall guess nondeterministically how the nodes of V are assigned to nodes of G_w, and processes the respective edges that are active at a certain point, simulating all possible paths through (V, E) in parallel. If (and only if) there is an assignment of nodes in V to nodes in G_w such that all edges have been processed correctly and all paths end at y at the same time, M accepts.

We define $Q := (Q_1 \cup \{w, d\}) \times \ldots \times (Q_k \cup \{w, d\})$. At every point of the simulation, each edge is either active (then the respective automaton M_i is in some state from Q_i), is *waiting* to be activated (represented by w), or is *done* (represented by d).

Hence, the initial state q_0 of M is defined by $q_0 := (q_1, \ldots, q_k)$, where $q_i := q_{0,i}$ for all $i \in \mathrm{out}(x)$, and $q_i := w$ for all $i \notin \mathrm{out}(x)$.

In the same spirit, we define $\delta(q, a)$ for each $q = (q_1, \ldots, q_k) \in Q$ and each $a \in \Sigma$ according to the following rules:

1. If $q_i = d$ for some automaton M_i, that automaton stays in d.
2. For every $z \in V$, M can nondeterministically decide that z has been reached if the following conditions are met:
   (a) $\delta_i(q_i, a) \in F$ for all $i \in \mathrm{in}(z)$ (all ingoing edges are allowed to end), and
   (b) $q_j = w$ for all $j \in \mathrm{out}(z)$ (all outgoing edges are ready).

   Then, every M_i with $i \in \mathrm{in}(z)$ enters d (the "done state") instead of $\delta(q_i, a)$, and every M_j with $j \in \mathrm{out}(z)$ enters its initial state $q_{0,j}$.
3. All M_i with $q_i \neq d$ that are not affected by such a change of active edges stay in the state w if $q_i = w$, or advance to the respective successor state $\delta_i(q_i, a)$ if $q_i \in Q_i$.

The construction already suggests that $L(M) = L$.

We illustrate this by examining the behavior of $Q'(L)$ on all graphs G_w, as $(v_0, v_{|w|}) \in Q'_L(G_w)$ holds if and only if $w \in L$.

First, assume $w \in L$. Then there exist assignments σ, μ with $(G_w, \sigma, \mu) \models Q'_L$, $\sigma(x) = v_0$, and $\sigma(\mu) = v_{|w|}$. As there is exactly one path ρ from v_0 to $v_{|w|}$, and as $\lambda(\rho) = w$, M is able to process w according to the assignments $\sigma(z)$ for all $z \in \mathrm{nvar}(Q'_L)$. This leads to $w \in L(M)$.

For the opposite direction, assume $w \in L(M)$. The node assignment σ can be derived from the non-deterministic guesses of M, as every change in active edges corresponds to a node in G_w. Then μ can be assigned accordingly, and $\lambda(\mu(\pi_i))$ holds for all path variables π_i in Q'_L. Consequently, $(G_w, \sigma, \mu) \models Q'_L$, $\sigma(x) = v_0$, and $\sigma(y) = v_{|w|}$ hold, and as the only possible path from v_0 to $v_{|w|}$ is labeled w, we conclude $w \in L$ (by $Q'_L(G_w) = Q_L(G_w) = F_L(G_w)$). Hence, L is regular. $\square$

The situation is not strictly the same for Boolean queries (e. g., if L contains every single letter of Σ, $F^B_L(G) = \texttt{true}$ holds for all non-empty *db*-graphs G); but a similar result can be observed:

Lemma 4.2. *Let Σ be an alphabet with $|\Sigma| \geq 2$, let $a \in \Sigma$, and let $L \subseteq (\Sigma \setminus \{a\})^*$. Then F^B_{aLa} is CRPQ-expressible if and only if L is regular.*

For alphabets Σ of size ≥ 2, ECRPQs can express queries F_L for non-regular $L \subseteq \Sigma^*$ which, according to Lemma 4.1, are not CRPQ-expressible. For example, for $L := \{a^n b^n \mid n \in \mathbb{N}\}$, F_L is not CRPQ-expressible, but is expressed by the ECRPQ $\text{Ans}(x, y) \leftarrow (x, \pi_1, z), (z, \pi_2, y), L_1(\pi_1), L_2(\pi_2), \text{el}(\pi_1, \pi_2)$, where $L_1 := a^*$ and $L_2 := b^*$. For *unary* alphabets (i. e., alphabets of size 1), however, we can show the following:

Lemma 4.3. *Let Σ be a unary alphabet, let $L \subseteq \Sigma^*$. Then F_L is ECRPQ-expressible if and only if it is CRPQ-expressible.*

Before giving a proof of this lemma, let us note that, in spite of Lemma 4.3, there exist ECRPQ-queries over unary alphabets that are *not* CRPQ-expressible. For example, consider the ECRPQ

$$Q \ := \ \text{Ans}(x, y) \leftarrow (x, \pi_1, z), (y, \pi_2, z), \text{el}(\pi_1, \pi_2),$$

selecting all pairs of nodes (u, v) in a db-graph G, for which there exists a node w such that there are paths from u to w and from v to w of the same length. It should be not too difficult to see that this query is not CRPQ-expressible.

Proof (Lemma 4.3). The *if*-direction holds by definition, as every CRPQ is an ECRPQ. Before we proceed to the proof of the other direction, we introduce some basic definitions. For every $k \geq 1$ and every vector $a \in \mathbb{N}^k$, define $a\mathbb{N} := \{ai \mid i \in \mathbb{N}\}$. For all sets $A, B \subseteq \mathbb{N}^k$, let $A + B := \{a + b \mid a \in A, b \in B\}$. A set $A \subseteq \mathbb{N}^k$ is *linear* if there exist $a_0, \ldots, a_n \in \mathbb{N}^k$ for some $n \geq 0$ such that $A = a_0 + a_1\mathbb{N} + \ldots + a_n\mathbb{N}$. A set is *semi-linear* if it is a finite union of linear sets.

Let $A = \{a_1, \ldots, a_k\}$ be a (finite) alphabet. The *Parikh mapping* (for A) is the function $\psi : A^* \to \mathbb{N}^k$ that is defined as $\psi(w) := (|w|_{a_1}, \ldots, |w|_{a_k})$ for all $w \in A^*$. We extend this to the Parikh image of a language by $\psi(L) := \{\psi(w) \mid w \in L$ for all $L \subseteq A^*$, and say that a language L is semi-linear if $\psi(L)$ is semi-linear.

Let A be any set, and let $k \geq 1$. For every $(a_1, \ldots, a_k) \in A^k$, we define functions $\text{proj}_i(a_1, \ldots, a_k) := a_i$ for all $1 \leq i \leq k$. In other words, the function proj_i projects an element of A^k to its i-th component.

For the *only if*-direction, let $\Sigma := \{a\}$, and assume there is a language $L \subseteq \Sigma^*$ such that F_L is ECRPQ-expressible, but not CRPQ-expressible. Then Lemma 4.1 implies that L is not a regular language.

Assume that Q_L is an ECRPQ with $Q_L(G) = F_L(G)$ for every Σ-labeled *db*-graph G, and assume (recalling Lemma 2.1) that

$$Q_L = \text{Ans}(x, y) \leftarrow \bigwedge_{1 \leq i \leq k} (x_i, \pi_i, y_i), R(\pi_1, \ldots, \pi_k),$$

We interpret the regular relation R as a regular language over the alphabet $\{a, \perp\}^k$. Let $\psi : (\{a, \perp\}^k)^* \to \mathbb{N}$ denote the Parikh mapping for $A := \{a, \perp\}^k$. As R is regular, its Parikh set $\psi(R) \subseteq \mathbb{N}^{2^k}$ is semi-linear (cf. Harrison [11]).

We define the set $R_{\text{len}} \subseteq \mathbb{N}^k$ by

$$R_{\text{len}} := \{(|w_1|, \ldots, |w_k|) \mid (w_1, \ldots, w_k) \in R\}.$$

In order to show that R_{len} is semi-linear, let $b_1, \ldots, b_{2^k}$ be the enumeration of $\{a, \perp\}^k$ that corresponds to ψ (i.e., for every $1 \leq i \leq 2^k$, $b_i \in \{a, \perp\}^k$, $\text{proj}_i(\psi(b_i)) = 1$, and all other positions of $\psi(b_i)$ are 0). We define functions $f_i : \mathbb{N}^{2^k} \to \mathbb{N}$ with $1 \leq i \leq k$ by

$$f_i(n_1, \ldots, n_{2^k}) := \sum_{j:\text{proj}_i(b_j)=a} n_j,$$

and extend this to a function $f : \mathbb{N}^{2^k} \to \mathbb{N}^k$ by

$$f(\overline{n}) := (f_1(\overline{n}), \ldots, f_k(\overline{n}))$$

for every $\overline{n} \in \mathbb{N}^{2^k}$. It is easy to see that $R_{\text{len}} = f(\psi(R))$. As $\psi(R)$ is semi-linear, there exist an $m \geq 1$ and linear sets $R_1, \ldots, R_m \subseteq \mathbb{N}$ such that $\psi(R) = \bigcup_{i=1}^m R_i$, which leads to $R_{\text{len}} = \bigcup_{i=1}^m f(R_i)$.

Every R_i is a linear set, hence, for every $1 \leq i \leq m$, there exist an $n \geq 0$ and $c_0, \ldots, c_n \subseteq \mathbb{N}^{2^k}$ with $R_i = c_0 + c_1\mathbb{N} + \ldots + c_n\mathbb{N}$. Therefore, $f(R_i) = f(c_0) + f(c_1)\mathbb{N} + \ldots + f(c_n)\mathbb{N}$, which demonstrates that $f(R_i)$ is a linear subset of $\mathbb{N}^k$. Hence, R_{len} is semi-linear.

The next step is the construction of a relation that extends R_{len} by not only describing the lengths of paths that are obtained from a single path variable, but to paths that are formed by connecting these single paths.

A *label sequence (in Q_L)* is a sequence $i_1, \ldots, i_m$ with $m \geq 1$, and

1. $1 \leq i_j \leq k$ for all $1 \leq j \leq m$ (every i_j corresponds to the path variable π_{i_j} in Q_L),
2. $i_j \neq i_{j'}$ if $j \neq j'$,
3. there exist $z_0, \ldots, z_m \in \text{nvar}(Q_L)$ such that $(z_j, \pi_{i_{j+1}}, z_{j+1})$ is an atom in Q_L for every $0 \leq j < m$.

Hence, every label sequence describes an non-empty, acyclic path trough the relation graph $H^{lab}_{Q_L}$; moreover, for every labeling sequence, the corresponding node variables $z_0, \ldots, z_m$ are uniquely defined, as every path variable occurs exactly once in the relational part of Q_L.

For every label sequence p with corresponding node variables $z_0, \ldots, z_{m_p}$, we define $\text{start}(p) := z_0$, $\text{end}(p) := z_{m_p}$, and let $\text{lab}(p) \subseteq \{1, \ldots, k\}$ denote all i_j that occur in p.

Let $\mathcal{P} = \{p_1, \ldots, p_l\}$ with $l \geq 1$ denote the set of all label sequences in Q_L (as there is only a finite number of path variables in Q_L, and no index i_j

occurs twice in a labeling sequence, $\mathcal{P}$ is finite by definition). Without loss of generality, assume that $\mathrm{start}(p_1) = x$ and $\mathrm{end}(p_1) = y$ hold; i. e., p_1 corresponds to a path from x to y in $H^{lab}_{Q_L}$.

For each p_i in $\mathcal{P}$, we define a function $\hat{p}_i : \mathbb{N}^k \to \mathbb{N}$ by

$$\hat{p}_i(r_1, \dots, r_k) := \sum_{j \in \mathrm{lab}(p_i)} r_j.$$

Hence, if $r \in \psi(R)$ (and, hence, corresponds to the path lengths in an assignment that satisfies Q_L), $\hat{p}_i(r)$ is the length of the path between the $\mathrm{start}(p_i)$ and $\mathrm{end}(p_i)$ along the edges labeled with π_j for $j \in \mathrm{lab}(p_i)$.

We combine these functions $\hat{p}_i$ to a function $\hat{p} : \mathbb{N}^k \to \mathbb{N}^l$ by

$$\hat{p}(\overline{r}) := (\hat{p}_1(\overline{r}), \dots, \hat{p}_l(\overline{r}))$$

for all $\overline{r} \in \mathbb{N}^k$, and define

$$\hat{p}(R_{\mathrm{len}}) := \{\hat{p}(\overline{r}) \mid \overline{r} \in R_{\mathrm{len}}\}.$$

Using the same approach as for R_{len}, we can conclude that $\hat{p}(R_{\mathrm{len}})$ is semi-linear.

Projecting $\hat{p}(R_{\mathrm{len}})$ to its first component does not yield the intended contradiction, as R might permit assignments where the path corresponding to p_1 is not labeled with a word from L, as long as there exists a different path between the same two nodes, but with a correct length. The problem holds for all other pairs of nodes that are connected non-uniquely. To overcome this problem, we need to enforce that for every pair of nodes, all paths between these nodes have the same length.

We now define the equivalence relation $\equiv$ on $\mathcal{P}$ by $p_i \equiv p_j$ if $\mathrm{start}(p_i) = \mathrm{start}(p_j)$ and $\mathrm{end}(p_i) = \mathrm{end}(p_j)$. For every $1 \leq i \leq l$, let

$$S_i := \{(s_1, \dots, s_l) \in \mathbb{N}^l \mid s_j = s_i \text{ for all } j \text{ with } p_i \equiv p_j\},$$

and define

$$B := \{(s_1, \dots, s_l) \in \mathbb{N}^l \mid s_j \leq s_1 \text{ for all } j\},$$

$$T := \hat{p}(R_{\mathrm{len}}) \cap B_m \cap \bigcap_{i=1}^{l} S_i.$$

First, note that B and all S_i are linear. Due to the closure of the class of semi-linear sets under intersection (cf. Ginsburg and Spanier [10]), T is semi-linear.

Intuitively, the sets S_i enforce that all paths with the share the same exterior nodes are assigned paths of the same lengths. Furthermore, the set B ensures that no path is longer than the path described by p_1. We are now able to state the claim that shall allow us to finish the proof:

Claim: Let $T_1 := \{\mathrm{proj}_1(t) \mid t \in T\}$. Then $T_1 = \psi_a(L)$, where $\psi_a : \{a\}^* \to \mathbb{N}$ denotes the Parikh mapping of $\{a\}$.

As semi-linear sets are closed under projection, this implies that $\psi(L) \subseteq \mathbb{N}$ is semi-linear, which implies that L is regular, which shall yield the contradiction.

In order to prove the claim, we examine the behavior of Q_L on a restricted class of *db*-graphs, an approach that is similar to the proof of Lemma 4.1. For every $n \geq 0$, we define the *db*-graph $G_n := G_w$ with $w = a^n$.

Proof of $\psi_a(L) \subseteq T_1$: Assume $a^n \in L$ for some $n \geq 0$. Then $(v_0, v_n) \in Q_L(G_n)$ holds by definition, and there exist assignments σ, μ such that $(G_n, \sigma, \mu) \models Q_L$, $\sigma(x) = v_0$, and $\sigma(y) = v_n$ hold. We define

$$\overline{r} := (\lambda(\mu(\pi_1)), \ldots, \lambda(\mu(\pi_k))),$$

and observe that $\overline{r} \in R$ holds by definition. Hence, for

$$\overline{r}_{\text{len}} := (|\lambda(\mu(\pi_1))|, \ldots, |\lambda(\mu(\pi_k))|),$$

we observe $\overline{r}_{\text{len}} \in R_{\text{len}}$, and, consequently, $\hat{p}(\overline{r}_{\text{len}}) \in \hat{p}(R_{\text{len}})$. As the path that corresponds to p_1 (the path from v_0 to v_n) is the longest possible path in G_n,

$$\text{proj}_i(\hat{p}(\overline{r}_{\text{len}})) \leq \text{proj}_1(\hat{p}(\overline{r}_{\text{len}}))$$

holds for all $1 \leq i \leq l$. This allows us to conclude $\hat{p}(\overline{r}_{\text{len}}) \in B$.

Furthermore, for all $p_i, p_j \in \mathcal{P}$ with $p_i \equiv p_j$, there is exactly one path in G_n between $\sigma(\text{start}(p_i))$ and $\sigma(\text{end}(p_i))$. Hence, the two paths that result from the assignment of paths to their paths variables under μ are identical, which means that $\hat{p}(\overline{r}_{\text{len}}) \subset S_i$ holds for all $1 \leq i \leq l$.

Thus, $\hat{p}(\overline{r}_{\text{len}}) \in T$, and $\psi_a(a^n) = n = \text{proj}_1(\hat{p}(\overline{r}_{\text{len}})) \in T_1$.

Proof of $\psi_a(L) \supseteq T_1$: Assume to the contrary that there exists an $n \in T_1$ with $n \notin \psi_a(L)$. By definition, there exist a $\overline{t} \in T$ with $n = \text{proj}_1(\overline{t})$ and a $\overline{r}_{\text{len}} \in R_{\text{len}}$ with $\overline{t} = \hat{p}(\overline{r}_{\text{len}})$.

We now use $\overline{r}_{\text{len}}$ to define assignments σ, μ with $(G_n, \sigma, \mu) \models Q_L$, $\sigma(x) = v_0$, $\sigma(y) = v_n$ as follows: First, we choose $\sigma(x) := v_0$ and $\sigma(y) := v_n$. We then follow p_1 and assign paths and nodes according to the respective path lengths in $\overline{r}_{\text{len}}$. We then proceed analogously for all other $p_i \in \mathcal{P}$ with $p_i \equiv p_1$. As $\overline{t} \in S_j$ holds for all $1 \leq j \leq l$, this process is well-defined.

In order to assign the remaining variables and paths, we first process all $p_i \in \mathcal{P}$ that start at x, but end in variables z such that there is no $p_j \in \mathcal{P}$ with $\text{start}(p_j) = z$ and $\text{end}(p_j) = y$. Again, we assign node variables and path variables accordingly. As $\overline{t} \in B$, we know that the resulting paths cannot have a length of more than n; hence, these assignments are possible. Analogously, we work backwards from y, and process all remaining variables that lead to y.

Next, observe that for all label sequences $p_i \in \mathcal{P}$ with $\text{end}(p_i) = x$ or $\text{start}(p_i) = y$, $\text{proj}_i(\overline{t}) = 0$ must hold, as otherwise, this label sequence and p_1 could be concatenated to form a label sequence $p_j \in \mathcal{P}$ with $\text{proj}_j(\overline{t}) > \text{proj}_1(\overline{t})$, which would contradict $\overline{t} \in B$. Hence, all respective node variables can be assign to x or y, and all these paths are assigned the empty path.

In terms of the relation graph $H^{lab}_{Q_L}$, this process yields assignments for all node variables $z \in \text{nvar}(Q)$ that are connected to x (or y), and the respective path variables that occur on the edges. Any unassigned variable must occur in a subgraph of $H^{lab}_{Q_L}$ that is disconnected from the subgraph that contains x. For each such subgraph, pick a node of in-degree 0 and treat it like x, or a node of out-degree 0 and treat it like y, again working forwards or backwards. Again, $\bar{t} \in B$ ensures that such an assignment is possible, and $\bar{t} \in \bigcap_{i=1}^{l} S_i$ prevents inconsistencies as well as problems with cycles.

As σ and μ were derived from R_{len} (and, hence, R), (G_n, σ, μ) holds. Hence, $(v_0, v_n) \in Q_L(G_n)$, which contradicts $Q_L(G_n) = F_L(G_n)$, as $a^n \notin L$. □

In Section 3.1 of [3], Barceló et al. mention that ECRPQs are able to express queries corresponding to *regular expressions with backreferencing* (or *extended regular expressions*) (cf. Aho [1], Freydenberger [8]). These expressions extend the regular expressions with variable binding and repetition operators; e. g., for every expression α, the extended expression $(\alpha)\%x\, xx$ generates the language of all www with $w \in L(\alpha)$ (α generates some $w \in L(\alpha)$, $\%x$ assigns that w to x, and the subsequent uses of x repeat this w – hence, xx generates ww).

Let $L := \{a^n \mid n \geq 4,\ n \text{ is a composite number}\}$. According to Lemma 4.3, F_L is not ECRPQ-expressible (as L is not regular). On the other hand, L is generated by the extended regular expression $(a\,a^+)\%x\ x^+$ (cf. Câmpeanu et al. [5]). This demonstrates that ECRPQs are not able to express all queries that correspond to extended regular expressions.

4.2. Relative Succinctness

In this section, we first obtain an undecidability result on the CRPQ-expressibility of ECRPQs. From this result, we derive a statement of the relative succinctness of ECPRQs in comparison to CRPQs.

We can adapt Lemma 3.2 to observe the following result on the decidability of expressibility:

Theorem 4.4. *CRPQ-expressibility for ECRPQs is not co-semi-decidable.*

Proof. This follows from the proof of Theorem 3.5, a variation of Lemma 4.2, and the observation that $\text{INVALC}(\mathcal{M})$ is regular iff $\text{dom}(\mathcal{M})$ is finite. Regarding the latter, note that if $\text{dom}(\mathcal{M})$ is finite, $\text{INVALC}(\mathcal{M})$ is co-finite; if $\text{dom}(\mathcal{M})$ is infinite, non-regularity of $\text{INVALC}(\mathcal{M})$ can be established using standard tools. This allows us to effectively construct an ECRPQ Q from a Turing machine $\mathcal{M}$ such that Q is CRPQ-expressible if and only if $\text{dom}(\mathcal{M})$ is finite.

Finiteness of $\text{dom}(\mathcal{M})$ is a Σ^0_2-complete problem in the arithmetical hierarchy (cf. Kozen [13]); hence, CRPQ-expressibility is Σ^0_2-hard, which means that this problem is neither semi-decidable, nor co-semi-decidable. □

Using Theorem 4.4 in conjunction with a technique that is due to Hartmanis [12] and has been widely used in Formal Language Theory (cf. Kutrib [14]), we obtain a result on the relative succinctness of ECRPQs and CRPQs. One of the benefits of that technique is that it applies to a wide range of different reasonable definitions of the size of an ECRPQ.

In order to be as general as possible, we define a *complexity measure* for ECRPQs as a computable function c from the set of all ECRPQs to $\mathbb{N}$, such that for every finite alphabet Σ, the set of all ECRPQs Q over Σ (i) can be effectively enumerated in order of increasing $c(Q)$, and (ii) does not contain infinitely many ECRPQs with the same value $c(Q)$. As the following theorem demonstrates, no matter which complexity measure we choose, the size tradeoff between ECRPQs and CRPQs is not bounded by any recursive function:

Theorem 4.5. *Let Σ be a finite alphabet with $|\Sigma| \geq 2$. For every recursive function $f : \mathbb{N} \to \mathbb{N}$ and every complexity measure c, there exists an ECRPQ Q over Σ such that Q is CRPQ-expressible, but for every CRPQ Q' with $Q' \equiv Q$, $c(Q') > f(c(Q))$.*

Proof. Let Σ be a finite alphabet with $|\Sigma| \geq 2$, and let c be a complexity measure for ECRPQs. Assume to the contrary that there exists a recursive function $f_c : \mathbb{N} \to \mathbb{N}$ such that, for every CRPQ-expressible ECRPQ Q over Σ, there is a CRPQ Q' with $Q \equiv Q'$ and $c(Q') \leq f(c(Q))$. We shall now demonstrate that this implies that the set

$$\Delta := \{Q \mid Q \text{ is an ECRPQ over } \Sigma \text{ that is not CRPQ-expressible}\}$$

is semi-decidable. This, in turn, would imply that CRPQ-expressibility for ECRPQs is co-semi-decidable, and contradict Theorem 4.4.

Under our assumptions, the semi-decision procedure for Δ can be defined as follows: Given an ECRPQ Q, compute $n := f_c(c(Q))$, and let F_n be the set of all CRPQs Q' over Σ with $c(Q') \leq n$. As c is a complexity measure, F_n is finite. Furthermore, as we can decide whether an ECRPQ is a CRPQ, we can compute a list of all elements of F_n (as we can effectively enumerate all ECRPQs Q'' with $c(Q'') \leq n$).

For every $Q' \in F_n$, we semi-decide $Q \neq Q'$ by searching for a Σ-labeled *db*-graph $G_{Q'}$ with $Q(G_{Q'}) \neq Q'(G_{Q'})$. If $Q' \neq Q$ holds, such a $G_{Q'}$ can be found in finite time, and if we have found a graph $G_{Q'}$ for *every* $Q' \in F_n$, we let the procedure return 1.

By our choice of f_c (and, hence, F_n), Q is not CRPQ-expressible if and only if $Q \neq Q'$ holds for every $Q' \in F_n$. Hence, this procedure is a semi-decision procedure for Δ, which implies that CRPQ-expressibility for ECRPQs over Σ is co-semi-decidable. This contradicts Theorem 4.4. $\square$

5. Conclusion

TODO

Acknowledgements

We thank Joachim Bremer for helpful comments on an earlier version of this article.

References

[1] A. Aho. Algorithms for finding patterns in strings. In J. van Leeuwen, editor, *Handbook of Theoretical Computer Science*, volume A. Elsevier, 1990.

[2] J. Albert and L. Wegner. Languages with homomorphic replacements. *Theoretical Computer Science*, 16:291–305, 1981.

[3] P. Barceló, C. Hurtado, L. Libkin, and P. Wood. Expressive languages for path queries over graph-structured data. In *Proc. PODS '10*, pages 3–14, 2010.

[4] D. Calvanese, G. D. Giacomo, M. Lenzerini, and M. Y. Vardi. Containment of conjunctive regular path queries with inverse. In *KR'00*, pages 176–185, 2000.

[5] C. Câmpeanu, K. Salomaa, and S. Yu. A formal study of practical regular expressions. *International Journal of Foundations of Computer Science*, 14:1007–1018, 2003.

[6] A. Deutsch and V. Tannen. Optimization properties for classes of conjunctive regular path queries. In *Proc. DBPL'01*, volume 2397 of *LNCS*, pages 21–39. Springer, 2001.

[7] D. Florescu, A. Y. Levy, and D. Suciu. Query containment for conjunctive queries with regular expressions. In *Proc. PODS'98*, pages 139–148, 1998.

[8] D. D. Freydenberger. Extended regular expressions: Succinctness and decidability. In *Proc. STACS 2011*, volume 9 of *LIPIcs*, pages 507–518, 2011.

[9] D. D. Freydenberger and N. Schweikardt. Expressiveness and static analysis of extended conjunctive regular path queries. In *Proc. AMW 2011*, volume 749 of *CEUR Workshop Proceedings*, 2011.

[10] S. Ginsburg and E. Spanier. Bounded ALGOL-like languages. *Transactions of the American Mathematical Society*, 113(2):333–368, 1964.

[11] M. Harrison. *Introduction to Formal Language Theory*. Addison Wesley Publishing Company, 1978.

[12] J. Hartmanis. On Gödel speed-up and succinctness of language representations. *Theoretical Computer Science*, 26(3):335–342, 1983.

[13] D. Kozen. *Theory of Computation*. Springer-Verlag, London, 2006.

[14] M. Kutrib. The phenomenon of non-recursive trade-offs. *International Journal of Foundations of Computer Science*, 16(5):957–973, 2005.

[15] K. Salomaa. Patterns. In *Formal Languages and Applications*, number 148 in Studies in Fuzziness and Soft Computing, pages 367–379. Springer, 2004.