Conjunctive Regular Path Queries in Lightweight Description Logics

Meghyn Bienvenu
Laboratoire de Recherche en Informatique
CNRS & Université Paris Sud, France

Magdalena Ortiz and Mantas Šimkus
Institute of Information Systems
Vienna University of Technology, Austria

Abstract

Conjunctive regular path queries are an expressive extension of the well-known class of conjunctive queries and have been extensively studied in the database community. Somewhat surprisingly, there has been little work aimed at using such queries in the context of description logic (DL) knowledge bases, and all existing results target expressive DLs, even though lightweight DLs are considered better-suited for data-intensive applications. This paper aims to bridge this gap by providing algorithms and tight complexity bounds for answering two-way conjunctive regular path queries over DL knowledge bases formulated in lightweight DLs of the DL-Lite and $\mathcal{EL}$ families.

1 Introduction

Recent years have seen a rapidly growing interest in using description logic (DL) ontologies to query instance data. In databases, similar attention has been paid to the related problem of querying graph databases which, like DL instance data, are sets of ground facts using only unary and binary predicates, i.e., node- and edge-labeled graphs [Consens and Mendelzon, 1990; Barceló et al., 2010]. The relevance of both problems lies in the fact that in many application areas, data can be naturally represented in such form. This applies, in particular, to XML and RDF data. While the DL and database communities share some common research goals, the research agendas they have pursued differ significantly. In DLs, the focus has been on designing efficient algorithms for answering (plain) conjunctive queries in the presence of ontological constraints. By contrast, work on graph databases typically does not consider ontological knowledge, but instead aims at supporting expressive query languages, like regular path queries (RPQs) and their extensions, which enable sophisticated navigation of paths. Such path navigation has long been considered crucial for querying data on the web. Indeed, it lies at the core of the XPath language for querying XML data, and of the property paths feature of SPARQL 1.1, the language recently recommended as the new standard for querying RDF data.

In this paper, we are interested in the problem of querying DL knowledge bases using various kinds of regular path queries. We mainly focus on conjunctive (two-way) regular path queries (C(2)RPQs), which are one of the most expressive and popular languages for querying graph databases. CRPQs simultaneously extend plain conjunctive queries (CQs) and basic RPQs: they allow conjunctions of atoms that can share variables in arbitrary ways, where the atoms may contain regular expressions that navigate the arcs of the database (or roles, in DL parlance). In the case of 2RPQs and C2RPQs, roles can be navigated in both directions. C2RPQs have already been studied for DLs, but all existing results concern expressive DLs for which reasoning is provably intractable. In particular, algorithms have been proposed for $\mathcal{ZIQ}$, $\mathcal{ZIO}$, and $\mathcal{ZOQ}$ [Calvanese et al., 2007b; 2009], for which query answering is 2-ExpTime hard. Even in data complexity, that is, when the query and ontology are assumed fixed, these algorithms need exponential time. More recently, algorithms for answering C2RPQs in Horn-$\mathcal{SHOIQ}$ and Horn-SROIQ were proposed [Ortiz et al., 2011]. These algorithms run in polynomial time in the size of the data, but still require exponential time in the size of the ontology. By contrast, to the best of our knowledge, path queries have not yet been considered for the lightweight DLs of the DL-Lite [Calvanese et al., 2007a] and $\mathcal{EL}$ [Baader et al., 2005] families, which underly the OWL 2 QL and EL profiles. This is surprising given that the low complexity of these DLs makes them better suited for data-intensive applications. This paper aims to remedy this situation by providing algorithms and precise complexity bounds for answering (C)2RPQs in the $\mathcal{EL}$ and DL-Lite families of lightweight DLs. We show that in data complexity, the query answering problem for CR2PQs is NL-complete for DL-Lite and P-complete for $\mathcal{EL}$, which in both cases is the lowest complexity that could be expected. For combined complexity, we prove PSPACE-completeness for both DL-Lite and $\mathcal{EL}$, but somewhat surprisingly obtain a tractability result for 2RPQs for both DLs. All of our upper bounds apply to extensions of $\mathcal{EL}$ and DL-Lite with role inclusions. Full proofs of all results can be found in a technical report [Bienvenu et al., 2013].

2 Preliminaries

We briefly recall the syntax of DL-Lite $\mathcal{R}$ [Calvanese et al., 2007a] and $\mathcal{EL}$ [Baader et al., 2005] (and relevant sublogics). As usual, we assume sets $\mathbb{N}_C$, $\mathbb{N}_R$, and $\mathbb{N}_I$ of concept names, role names, and individuals. We will use $\mathbb{N}_R$ to refer
to $N_R \cup \{r^{-} | r \in N_R \}$, and if $R \in \overline{N_R}$, we use $R^{-}$ to mean $r^{-}$ if $R = r$ and $r$ if $R = r^{-}$. An $ABox$ is a set of assertions of the form $A(b)$ or $r(b, c)$, where $A \in N_C$, $r \in N_R$, and $b, c \in N_U$. We use $\text{Ind}(A)$ to refer to the set of individuals in $A$. A $TBox$ is a set of inclusions, whose form depends on the DL in question. In DL-Lite, inclusions take the form $B_1 \subseteq (\neg)B_2$, where each $B_i$ is either $A$ (where $A \in N_C$) or $\exists R$ (where $R \in \overline{N_R}$). DL-Lite$_\mathcal{R}$ additionally allows role inclusions of the form $R_1 \subseteq (\neg)R_2$, where $R_1, R_2 \in N_R$. DL-Lite$_\mathcal{RD}$ is obtained from DL-Lite$_\mathcal{R}$ by disallowing inclusions which contain negation or have existent concepts ($\exists R$) on the right-hand side. In $\mathcal{E}L$, inclusions have the form $C_1 \sqsubseteq C_2$, where $C_1, C_2$ are complex concepts constructed as follows: $C := \top \lor A \lor (C \cap C) \lor \exists r.C$. The DL $\mathcal{EL}$ additionally allows role inclusions of the form $r \subseteq s$, where $r, s \in N_R$. Note that in $\mathcal{EL}(H)$, $TBoxes$, inverse roles are not permitted. A knowledge base (KB) $K = (T, A)$ consists of a TBox $T$ and an ABox $A$.

As usual, the semantics is based upon interpretations, which take the form $I = (\Delta^2, \cdot)$, where $\Delta^2$ is a non-empty set and $\cdot$ maps each $a \in N_I$ to $a^2 \in \Delta^2$, each $A \in N_C$ to $A^2 \subseteq \Delta^2$, and each $r \in N_R$ to $r^2 \subseteq \Delta^2 \times \Delta^2$. The function $\cdot$ is straightforwardly extended to general concepts and roles, e.g. $(\neg A)^2 = \Delta^2 \setminus A^2$, $(\exists r.C)^2 = \{(c, d) | (c, e) \in r^2, (e, d') \in C^2\}$, and $(P^{-1})^2 = \{(d, c) | (c, d') \in P^2\}$. An interpretation $I$ satisfies $G \subseteq H$ if $G^2 \subseteq H^2$; it satisfies $A(a)$ (resp. $r(a, b)$) if $a^2 \in A^2$ (resp. $(a^2, b^2) \in r^2$). $I$ is a model of $(T, A)$ if $I$ satisfies all inclusions in $T$ and assertions in $A$.

To simplify the presentation, we will assume that $\mathcal{EL}H$ TBoxes are in normal form, meaning that all concept inclusions are of one of the following forms:

$$A \subseteq B \quad A \sqcap B \sqsubseteq C \quad A \sqsubseteq \exists r.B \quad \exists r.B \sqsubseteq A$$

with $A, A_1, A_2, B \in N_C \cup \{\top\}$. It is well-known (cf. [Baader et al., 2005]) that for every $\mathcal{EL}H$ TBox $T$, one can construct in polynomial time an $\mathcal{EL}H$ TBox $T'$ in normal form (possibly using new concept names) such that $T' \models T$ and every model of $T$ can be expanded to a model of $T'$.

We use $\text{sig}(T)$ to denote the set of all concept and role names appearing in a TBox $T$. For ease of reference, we also define sets $\text{BC}_T$ of basic concepts and $\text{TC}_T$ of tail concepts for $T$ as follows: $\text{BC}_T = \text{TC}_T = \{N_C \cap \text{sig}(T) \mid T \text{ is an } \mathcal{EL}H \text{ TBox, and } \text{TC}_T = \{\exists r, \forall r \mid r \in N_R \cap \text{sig}(T)\} \text{ and } \text{BC}_T = (N_C \cap \text{sig}(T)) \cup \text{TC}_T$ for a DL-Lite$_\mathcal{R}$ TBox $T$.

Canonical Models We recall the definition of canonical models for DL-Lite$_\mathcal{R}$ and $\mathcal{EL}H$ KBs. For both DLs, the domain of the canonical model $\mathfrak{I}_{T,A}$ for a KB $(T, A)$ consists of paths of the form $a_{R_1} c_1 \ldots R_n c_n$ ($n \geq 0$), where $a \in \text{Ind}(A)$, each $c_i$ is a tail concept, and each $R_i$ a possibly inverse role. When $T$ is a DL-Lite$_\mathcal{R}$ TBox, the domain $\Delta^2_{T,A}$ contains exactly those paths $a_{R_1} \subseteq \cdots \subseteq R_n \subseteq \subseteq r_{n+1}$ which satisfy:

- if $n \geq 1$, then $T, A \models \exists R_1(a)$;
- for $1 \leq i < n$, $T \models \exists R_{i+1} \subseteq \subseteq R_{i+1}$ and $R_{i+1} \neq R_i$.

When $T$ is an $\mathcal{EL}H$ TBox in normal form, the domain $\Delta^2_{T,A}$ contains exactly those paths $a_{R_1} \subseteq \cdots \subseteq r_{n} A_n$ for which each $r_i \in N_R$, and:

- if $n \geq 1$, then $T, A \models \exists r_1, A_1(a)$;
- for $1 \leq i < n$, $T \models A_i \subseteq \exists r_{i+1} A_{i+1}$.

We denote the last concept in a path $p$ by tail$(p)$, and define $\mathfrak{I}_{T,A}$ by taking:

$$a^{\mathfrak{I}_{T,A}} = a \text{ for all } a \in \text{Ind}(A)$$

$$A^{\mathfrak{I}_{T,A}} = \{a \in \text{Ind}(A) \mid T, A \models A(a)\}$$

$$\{p \in \Delta^2_{T,A} \mid \text{Ind}(A) \models T \models \text{tail}(p) \subseteq A\}$$

$$r^{\mathfrak{I}_{T,A}} = \{a, b \mid r(a, b) \in A\} \cup \{(p_1, p_2) \mid p_2 = p_1 S C \text{ and } T \models S \subseteq r \} \cup \{(p_2, p_1) \mid p_1 = p_2 S C \text{ and } T \models S \subseteq r^-*\}$$

Note that $\mathfrak{I}_{T,A}$ is composed of a core consisting of the ABox individuals and an anonymous part consisting of (possibly infinite) trees rooted at ABox individuals. We will use $\mathfrak{I}_{T,A}|_e$ to denote the submodel of $\mathfrak{I}_{T,A}$ obtained by restricting the universe to paths having $e$ as a prefix.

Regular Languages We assume the reader is familiar with regular languages, represented either by regular expressions or nondeterministic finite state automata (NFAs). An NFA over an alphabet $\Sigma$ is a tuple $\alpha = (S, \Sigma, \delta, s_0, F)$, where $S$ is a finite set of states, $\delta \subseteq S \times \Sigma \times S$ the transition relation, $s_0 \in S$ the initial state, and $F \subseteq S$ the set of final states. We use $L(\alpha)$ to denote the language defined by an NFA $\alpha$, and when the way a regular language is represented is not relevant, we denote it simply by $L$.

3 Path Queries

We now introduce the query languages studied in this paper.

Definition 1. A conjunctive (two-way) regular path query (C2RPQ) has the form $q(\vec{x}) = \exists y \varphi$ where $\vec{x}$ and $\vec{y}$ are tuples of variables, and $\varphi$ is a conjunction of atoms of the forms:

(i) $A(t)$, where $A \in N_C$ and $t \in N_I \cup \vec{x} \cup \vec{y}$, and

(ii) $L(t, t')$, where $L$ is (an NFA or regular expression defining) a regular language over $\overline{N_R} \cup \{A? \mid A \in N_C\}$, and $t, t' \in N_I \cup \vec{x} \cup \vec{y}$.

As usual, variables and individuals are called terms, and the variables in $\vec{x}$ are called answer variables. A query with no answer variables is called a Boolean query.

Conjunctive (one-way) regular path queries (C1RPQs) are obtained by disallowing symbols from $\overline{N_R} \setminus N_R$ in atoms of type (ii), and conjunctive queries (CQs) result from only allowing type-(ii) atoms of the form $r(t, t')$ with $r \in N_R$. Two-way regular path queries (2RPs) consist of a single atom of type (ii), and regular path queries (RPQs) further disallow symbols from $\overline{N_R} \setminus N_R$. Finally, instance queries (IQs) take the form $A(x)$ with $A \in N_C$, or $r(x, y)$ with $r \in N_R$.

We now define the semantics of C2RPQs. For a regular language $L$ over the alphabet $\overline{N_R} \cup \overline{N_R} \cup \{A? \mid A \in N_C\}$, we call $d_1$ an $L$-successor of $d_1$ in $I$ if there is some $w = u_1 \ldots u_n \in L$ and some sequence $e_0, \ldots, e_n$ of elements of $\Delta^2$ such that $e_0 = d_1$, $e_n = d_2$, and, for all $1 \leq i \leq n$:

- if $i \leq n$, then $e_{i+1} = e_i$ in $A^2$;
- if $i > n$, then $e_{i+1} = R \in N_R \cup \overline{N_R}$, then $e_{i+1} \in R^2$.

A match for a Boolean C2RPQ $q$ in an interpretation $I$ is a mapping $\pi$ from the terms in $q$ to elements in $\Delta^2$ such that:
Theorem 3. The complexity results in Figure 1 hold.

We split the proof of this theorem into parts, with the lower bounds shown in the next section, and the (more involved) proofs of the upper bounds outlined in Section 5.

4 Lower Bounds

We start by establishing the required lower bounds.

Proposition 4. Boolean CRPQ entailment is

1. NL-hard in data complexity for DL-LiteRDFS;
2. P-hard in data complexity for \( \mathcal{EL} \);
3. NP-hard in combined complexity for DL-LiteRDFS;
4. PSPACE-hard in combined complexity for DL-Lite & \( \mathcal{EL} \).

Statements (1) and (2) hold even for RPQs.

Proof. Statement (1) follows from the analogous result for graph databases [Consens and Mendelzon, 1990]. It can be shown by a simple reduction from the NL-complete directed reachability problem: \( y \) is reachable from \( x \) in a directed graph \( G \) if and only if \((x, y)\) is an answer to \( r^*(x, y)\) w.r.t. the ABox \( A_G \) encoding \( G \). Statement (2) is immediate given the P-hardness in data complexity of instance checking in \( \mathcal{EL} \) [Calvanese et al., 2006], and (3) follows from the well-known NP-hardness in combined complexity of CQ entailment for databases [Abiteboul et al., 1995].

For statement (4), we give a reduction from the problem of emptiness of the intersection of an arbitrary number of regular languages, which is known to be PSPACE-complete [Kozen, 1977]. Let \( L_1, \ldots, L_n \) be regular languages over alphabet \( \Sigma \). We will use the symbols in \( \Sigma \) as role names, and we add a concept name \( A \). Let \( A = \{ A(a) \} \) and \( q = \exists x \ L_1(a, x) \land \ldots \land L_n(a, x) \). For DL-Lite, we will use the following TBox: \( T = \{ A \subseteq \exists r \ | \ r \in \Sigma \} \cup \{ \exists r \subseteq \exists s \ | \ r, s \in \Sigma \} \). For \( \mathcal{EL} \), we can use \( T = \{ A \subseteq \exists r.A \ | \ r \in \Sigma \} \). Notice that in both cases the canonical model \( \mathcal{I}_{T,A} \) consists of an infinite tree rooted at \( a \) such that every element in the interpretation has a unique \( r \)-child for every \( r \in \Sigma \) and no other children. Thus, we can associate to every domain element the word over \( \Sigma \) given by the unique path from \( a \), and moreover, for every word \( w \in \Sigma^* \) we can find an element \( e_w \) whose path from \( a \) is exactly \( w \). This means that if \( w \in L_1 \land \ldots \land L_n \), we obtain a match for \( q \) in the canonical model by mapping \( x \) to \( e_w \). Conversely, if \( q \) is entailed, then any match in the canonical model defines a word which belongs to every \( L_i \), which means \( L_1 \land \ldots \land L_n \) is non-empty.

For (2)RPQs in DL-LiteRDFS and \( \mathcal{EL} \), we inherit combined complexity lower bounds of NL and P respectively from IQs. For DL-Lite, we establish a P lower bound for 2RPQs, which contrasts with the NL-completeness of instance checking.

Proposition 5. Boolean 2RPQ entailment in DL-Lite is P-hard in combined complexity, assuming an NFA representation of the regular language.

Proof sketch. Consider the P-complete entailment problem in which one is given a propositional formula \( T = \rho_1 \land \ldots \land \rho_m \land \forall v_i \) over variables \( v_1, \ldots, v_n \) with \( \rho_i = v_{i_1} \land v_{i_2} \rightarrow v_{i_3} \), and the problem is to decide whether \( T \models v_n \). We construct a DL-Lite TBox \( T \) and 2RPQ \( q \) such that \( T, \{ A(a) \} \models q \) if and only if \( T \models v_n \). We let \( T \) consist of the axioms:

- \( A \subseteq \exists r_{i,j}, \) for \( 1 \leq i \leq m, j \in \{ 1, 2 \} \)
- \( \exists r_{i_1,i_2} \subseteq \exists r_{i_2,i,j}, \) for \( 1 \leq i_1, i_2 \leq m \) and \( i_1, j \in \{ 1, 2 \} \)
- \( q = \exists x \alpha(x, x), \) where \( \alpha = (S, \Sigma, \delta, s_0, \{ v_{i_1}^{out} \} ) \) is the NFA defined as follows:
  - \( S = \{ s_0 \} \cup \{ v_1 \} \cup \{ v_i^{in}, v_i^{out} \ | \ 2 \leq i \leq n \} \cup \{ \rho_i \ | \ 1 \leq i \leq m \} \)
\(- \Sigma = \{ A? \} \cup \{ r_{i,j}, r_{i,j}^\rightarrow \mid 1 \leq i \leq m, 1 \leq j \leq 2 \} \)

\(- \delta \) contains \((s_0, A?, v_0)\), and for each \(\rho_i = v_j \land v_k \rightarrow v_{\ell} \), the following transitions: \((v_{\ell}^\rightarrow, r_{i,1}, v_j^\rightarrow)\), \((v_{\ell}^\rightarrow, r_{i,1}, \rho_i)\), \((\rho_i, r_{i,2}, v_{\ell}^\rightarrow)\), and \((v_k^\rightarrow, r_{i,2}, v_{\ell}^\rightarrow)\). Note: we use \(v_1\) in place of \(v_{\ell}^\rightarrow\) and \(v_{\ell}^\rightarrow\).

The first transition in \(\delta\) enforces that \(x\) must be mapped to \(a\) and that there must be a loop at \(a\) from state \(v_{\ell}^\rightarrow\) to \(v_{\ell}^\rightarrow\). Intuitively, a state \(v_{\ell}^\rightarrow\) indicates that \(v_\ell\) needs to be proven, and \(v_{\ell}^\rightarrow\) signals that \(v_\ell\) has been successfully derived. From a state \(v_{\ell}^\rightarrow\), the available transitions correspond to the rules in \(T\) which conclude on \(v_i\): selecting transition \((v_{\ell}^\rightarrow, r_{i,1}, v_j^\rightarrow)\) means choosing to use \(\rho_i\) to derive \(v_\ell\). The transitions \((v_{\ell}^\rightarrow, r_{i,2}, v_{\ell}^\rightarrow)\) allow us to move to the second variable of \(\rho_i\) once the first variable of \(\rho_i\) has been derived. When both variables have been proven, the transition \((v_k^\rightarrow, r_{i,2}, v_{\ell}^\rightarrow)\) allows us to exit the derivation of \(v_\ell\). Thus, any loop from \(v_{\ell}^\rightarrow\) to \(v_{\ell}^\rightarrow\) in \(I_{A,t,A}^{(\alpha)}\) corresponds to a derivation of \(v_\ell\), and conversely, any derivation of \(v_\ell\) yields a path witnessing the entailment of \(q\).

As a corollary, we get P-hardness of RPQs in DL-Lite\(_R\), by using role inclusions to simulate the inverse roles in the query. We leave open whether the preceding hardness result applies when regular languages are given as regular expressions.

### 5 Upper Bounds

The main objective of this section will be to define a procedure for deciding \(I_{\mathcal{T},\mathcal{A}} \models q\) for a KB \((T, \mathcal{A})\) and C2RQP \(q\). The procedure comprises two main steps. First, we rewrite \(q\) into a set \(Q\) of C2RQPs such that \(I_{\mathcal{T},\mathcal{A}} \models q\) if and only if \(I_{\mathcal{T},\mathcal{A}} \models q'\) for some \(q' \in Q\). The advantage of the rewritten queries is that in order to decide whether \(I_{\mathcal{T},\mathcal{A}} \models q'\), we will only need to consider matches which map the variables to \(\text{Ind}(\mathcal{A})\). The second step decides the existence of such restricted matches for the rewritten queries.

In order to more easily manipulate regular languages, it will prove convenient to use NFAs rather than regular expressions. Thus, in what follows, we assume all binary atoms take the form \(\alpha(t, t')\), where \(\alpha\) is an NFA over \(\Sigma_R \cup \Sigma_R \cup \{ A? \mid A \in \mathcal{N}_R \}\). Given \(\alpha = (S, \Sigma, \delta, s_0, F)\), we use \(\alpha_{s,G}\) to denote the NFA \((S, \Sigma, \delta, s, G)\), i.e., the NFA with the same states and transitions as \(\alpha\) but with initial state \(s\) and final states \(G\).

#### 5.1 Loop Computation

A key to defining our rewriting procedure will be to understand how an atom \(L(t, t')\) can be satisfied in the anonymous part of the canonical model \(I_{\mathcal{T},\mathcal{A}}\). A subtlety arises from the fact that the path witnessing the satisfaction of an atom \(L(t, t')\) may be quite complicated: it may move both up and down, passing by the same element multiple times, and possibly descending below \(t\). This will lead us to decompose an atom \(L(t, t')\) into multiple “smaller” atoms corresponding to segments of the \(L\)-path which are situated wholly above or below an element. Importantly, we know that the canonical model displays a high degree of regularity, since whenever two elements \(p_1\) and \(p_2\) in the anonymous part end with the same concept (i.e., \(\text{Tail}(p_1) = \text{Tail}(p_2)\)), the submodels \(I_{\mathcal{T},\mathcal{A}|p_1}\) and \(I_{\mathcal{T},\mathcal{A}|p_2}\) are isomorphic. In particular, this means that if \(\text{Tail}(p_1) = \text{Tail}(p_2)\), then \(p_1\) is an \(L\)-successor of itself in the interpretation \(I_{\mathcal{A},\mathcal{T}|p_2}\), just in the case that \(p_2\) is an \(L\)-successor of itself in the interpretation \(I_{\mathcal{A},\mathcal{T}|p_2}\).

We will require a way of testing for a given TBox \(T\) and NFA \(\alpha\) with states \(s, s'\) whether \(\text{Tail}(\epsilon) = C\) ensures that there is a loop from \(\epsilon\) to itself, situated wholly within \(I_{\mathcal{T},\mathcal{A}|e}\), which takes \(\alpha\) from state \(s\) to state \(s'\). To this end, we construct a table \(\text{Loop}_{\alpha}\) which contains for each pair \((s, s')\) of states in \(\alpha\), a subset of \(\text{TC}_T\). If \(T\) is a DL-Lite\(_R\) TBox, then \(\text{Loop}_{\alpha}\) is defined inductively using the following rules:

1. For every \(s \in S\): \(\text{Loop}_{\alpha}[s, s] = \text{TC}_T\)
2. If \(C \in \text{Loop}_{\alpha}[s_1, s_2]\) and \(C \in \text{Loop}_{\alpha}[s_2, s_3]\), then \(C \in \text{Loop}_{\alpha}[s_1, s_3]\)
3. If \(C \in \text{TC}_T\), \(T \models C \subseteq A \land (s_1, A?, s_2) \in \delta\), then \(C \in \text{Loop}_{\alpha}[s_1, s_2]\)
4. If \(T \models C \subseteq \exists R, T \models R \subseteq R'\), \(T \models R' \subseteq R''\), \((s_1, R', s_2) \in \delta, \exists R' \subseteq \text{Loop}_{\alpha}[s_2, s_3], (s_3, R''\subseteq s_4) \in \delta, C \in \text{TC}_T\), and \(C \not\models R\), then \(C \in \text{Loop}_{\alpha}[s_1, s_4]\)

For \(E\mathcal{L}\mathcal{H}\), we replace the last rule by:

4. If \(T \models C \subseteq \exists R, T \models R \subseteq R'\), \((s_1, R', s_2) \in \delta, D \in \text{Loop}_{\alpha}[s_2, s_3], (s_3, R''\subseteq s_4) \in \delta, C \in \text{TC}_T\), then \(C \in \text{Loop}_{\alpha}[s_1, s_4]\)

**Example 6** Consider a DL-Lite\(_R\) TBox \(T\) containing the inclusions \(B \subseteq \exists R, \exists R \subseteq B, \exists \exists T_1\), and \(t_1 \subseteq t_2\), and consider the query \(q = \exists y y, \alpha(x, y)\), \(B(y)\), or equivalently, \(q = \exists y x, \alpha(y, x)\), \(B(y)\), where \(\alpha = \{(s_0, s_1, s_2, s_3), \{r, t_1, t_2, t_3\}, \delta, \delta, s_0, s_3)\} and \(\delta = \{(s_0, r, s_0), (s_0, t_1, s_1), (s_1, t_2, s_2), (s_2, t_3, s_3)\}\). In the first step of the loop computation, we infer that \(\text{Loop}_{\alpha}[s_1, s_1]\) is the set of all tail concepts for \(0 \leq i \leq 4\). Next, by rule 4, and using \(T \models B \subseteq \exists T_1, T \models \exists R \subseteq \exists T_1, T \models t_1 \subseteq t_2, (s_0, t_1, s_1), \exists T_1 \subseteq \text{Loop}_{\alpha}[s_1, s_1]\), and \(s_1, t_2, s_2)\), we can infer that \(B \subseteq \text{Loop}_{\alpha}[s_0, s_2]\) and \(\exists R \subseteq \text{Loop}_{\alpha}[s_0, s_2]\). In a further step, we can use \(T \models B \subseteq \exists R, (s_0, r, s_0), \exists R \subseteq \text{Loop}_{\alpha}[s_0, s_2]\), and \(s_2, r, s_3)\) to obtain \(B \subseteq \text{Loop}_{\alpha}[s_0, s_3]\).

Note that the table \(\text{Loop}_{\alpha}\) can be constructed in polynomial time in \(|T|\) and \(|\alpha|\) since entailment of inclusions is polynomial for both DL-Lite\(_R\) and \(E\mathcal{L}\mathcal{H}\). The following lemma shows that \(\text{Loop}_{\alpha}\) has the desired meaning:

**Lemma 7.** For every element \(p \in \Delta_{\mathcal{A},\mathcal{T},\alpha} \setminus \text{Ind}(\mathcal{A})\) \(\text{Tail}(p) \in \text{Loop}_{\alpha}[s, s']\) if and only if \(p\) is an \(L(\alpha_{s,s'})\)-successor of itself in the interpretation \(I_{\mathcal{A},\mathcal{T},\alpha}\).
PROCEDURE rewrite($q, T$) 
1. Choose either to output $q$ or to continue.
2. Choose a non-empty set Leaf $\subseteq$ vars($q$) and $y \in$ Leaf.
   Rename all variables in Leaf to $y$.
3. Choose $C \in TC_T$ such that $T \models C \subseteq B$ whenever $B(y)$ is an atom of $q$. Drop all such atoms from $q$.
4. For each atom $\alpha(t, t')$ where $\alpha = (S, \Sigma, \delta, s, F)$ is an NFA and $y \in \{t, t\}$,
   - choose a sequence $s_1, \ldots, s_n$ of distinct states from $S$ such that $s_n \in F$.
   - replace $\alpha(t, t')$ by the atoms $\alpha_{s_1,s_2}(t, y), \alpha_{s_1,s_2}(y, y), \ldots, \alpha_{s_{n-1},s_n}(y, t')$.
5. Drop all atoms $\alpha_{s',s}(y, y)$ such that $C \in Loop_\alpha[s, s']$.
6. Choose some $D \in BC_T$ and $R_1, R_2 \in \overline{\text{IR}}$ such that:
   a. $C = \exists R^-$ and $T \models D \models \exists R$ for $\text{DL-Lite}_k$, or $R \in N_2$ and $T \models D \models \exists R.C$ for $\text{EL}_H$.
   b. $T \models R \equiv R_1$ and $T \models R \equiv R_2$.
   c. For each atom $\alpha(y, x)$ with $\alpha = (S, \Sigma, \delta, s, F)$, there exists $s' \in S$ such that $(s, R_1, s') \in \delta$.
   d. For each atom $\alpha(x, y)$ with $\alpha = (S, \Sigma, \delta, s, F)$, there exists $s'' \in S$, $s_f \in F$ with $(s', R_2, s_f) \in \delta$.
   For atoms of the form $\alpha(y, y)$, both (c) and (d) apply.
7. Replace
   - each atom $\alpha(y, x)$ with $x \neq y$ by $\alpha_{s',F}(y, x)$
   - each atom $\alpha(x, y)$ with $x \neq y$ by $\alpha_{s',s''}(x, y)$
   - each atom $\alpha(y, y)$ by atoms $\alpha_{s',s''}(y, y)$
   with $s, s', s''$, $F$ as in Step 6.
8. If $D \in NC$ is the concept chosen in Step 6, add $D(y)$ to $q$. If $D = \exists P^-$, add $\alpha_P(z, y)$ to $q$, where $z$ is a fresh variable and $L(\alpha_P) = \{P\}$. Go to Step 1.

Figure 2: Query rewriting procedure rewrite.

Example 8. We illustrate two different ways to apply the rewriting step to the query $q = \exists y.\exists z. t_1 t_2 r^-(x, y), B(y)$ in Example 6. First, let Leaf $= \{x\}$ be the set chosen in step 2. There is no renaming to do, so we proceed to step 3 and choose $\exists r^-$. In step 4, we choose the sequence $s_2, s_3$, and replace $\alpha(x, y)$ by $\alpha_{s_2,s_3}(x, x), \alpha_{s_2,s_3}(x, y)$. Since $\exists r^- \in Loop_\alpha[s_0, s_2]$, we drop the first atom and keep only $\alpha_{s_2,s_3}(x, y)$. In step 6, we can choose $B$ for $D$, and $r$ for the roles $R, R_1, R_2$. This ensures (a) and (b). For (c), we can take $s_3$ since $(s_2, r^-, s_3) \in \delta$. In step 7, we replace $\alpha_{s_2,s_3}(x, y)$ by $\alpha_{s_3,s_3}(x, y)$. At the end of step 8, we are left with the query $q' = \exists y.\exists z. x r^- y, \alpha_{s_3,s_3}(x, y), B(y)$, which is output as a rewriting when we return to step 1. We remark that $q'$ is equivalent modulo $T$ to the simpler $\exists y.\exists z. r(x, y)$ since by choosing $x = y$, the atom $\alpha_{s_3,s_3}(x, y)$ is trivially satisfied, $B(y)$ is enforced by $r(x, y)$ and the inclusion $\exists r^- \subseteq B$. Intuitively, this rewriting captures the fact that, whenever we have an element $e$ in an interpretation that satisfies $\exists r^-$, then we can map $x$ to $e$, thereby ensuring that the initial segment $r^* t_1 t_2$ is satisfied below $e$. Moreover, by mapping $y$ to the $r$-predecessor of $e$, we satisfy the remaining $r^-$. 

As further illustration, suppose that in step 2, we choose Leaf $= \{x, y\}$, and let $y$ be the selected variable. After renaming, we obtain $q = \exists y.\alpha(y, y), B(y)$. In step 3, we choose $\exists r^-$, which leads us to drop the atom $B(y)$, leaving us with $\exists y.\alpha(y, y)$. In step 4, we choose the sequence that contains only $s_3$, so the atom $\alpha(y, y)$ is left untouched. Since $\exists r^- \in Loop_\alpha[s_0, s_3]$, we can drop this atom in step 5, obtaining the empty query. In step 6, we choose $D = B$ and $R = R_1 = R_2 = r$. Step 7 is inapplicable since there are no binary atoms. Finally, in step 8, we add $B(y)$ to obtain the query $q = \exists y.\alpha(y, y)$, $B(y)$. Intuitively, this rewriting captures that if some element $e$ satisfies $B$, then we can map both $x$ and $y$ to it to obtain a query match in which the regular expression $r^* t_1 t_2 r^-$ is fully satisfied below $e$.

The next lemma shows that using rewrite($q, T$), we can reduce the problem of finding an arbitrary query match to finding a match involving only ABox individuals.

**Lemma 9.** $T, A \models q$ if and only if there exists a match $\pi$ for some query $q \in$ rewrite($q, T$) in $I_{A,T}$ such that $\pi(t) \in L(\alpha)$ for every term $t$ in $q$.

We remark that the number of variables and atoms in each query in rewrite($q, T$) is linearly bounded by $|q|$. This is the key property used to show the following:

**Lemma 10.** There are only exponentially many queries in rewrite($q, T$) (up to equivalence), each having size polynomial in $|q|$.

### 5.3 Query Evaluation

Even when all terms are mapped to ABox individuals, the paths between them may need to pass by the anonymous part in order to satisfy the regular expressions in the query. This leads us to define a relaxed notion of query entailment, which exploits the fact that if all variables are mapped to $\text{Ind}(A)$, only loops (that is, paths from an individual $a$ to itself in $I_{A,T} | a$) may participate in the paths between them. Hence, we look for paths in the ABox that may use such loops to skip states in the query automata.

As part of our query evaluation procedure, we will need to decide for a given individual $a$ whether $a$ is an $L(\alpha_{s,s'})$-successor of itself in $I_{A,T} | a$. We cannot use $\text{Loop}_\alpha$ directly, since it does not take into account the concepts which are
entailed due to ABox assertions. We note however that the set of loops starting from a given individual is fully determined by the set of basic concepts which the individual satisfies. We thus define a new table \(\text{ALoop}_\alpha\) such that \(\text{ALoop}_\alpha[s, s']\) contains all subsets \(G \subseteq BCT\) such that \(\alpha\) is an \(L(\alpha, s')\)-successor of itself in \(I_\alpha\) whenever \(G = \{C \in BCT | \alpha \in C^{\text{THT}}\}\). Note that the table \(\text{ALoop}_\alpha\) is exponential in \(|T|\), but the associated decision problem is in P.

**Lemma 11.** It can be decided in polytime in \(|T|\) and \(|\alpha|\) whether \(G \in \text{ALoop}_\alpha[s, s']\).

We use \(\text{ALoop}_\alpha\) to define a relaxed notion of query match.

**Definition 12.** We write \(T, A \models q\) if there is a mapping \(\pi\) from the terms in \(q\) to \(\text{Ind}(A)\) such that:

(a) \(\pi(c) = c\) for each \(c \in N_c\),
(b) \(T, A \models A(\pi(t))\) for each atom \(A(t)\) in \(q\), and
(c) for each \(\alpha(t, t') \in q\) with \(\alpha = (S, \Sigma, \delta, s, F)\), there is a sequence \((a_0, s_0), \ldots, (a_n, s_n)\) of distinct pairs from \(\text{Ind}(A) \times S\) such that \(a_0 = \pi(t)\), \(a_n = \pi(t')\), \(s_0 = s\), \(s_n \in F\), and for every \(0 \leq i < n\), either:

(i) \(a_i = a_{i+1}\) and \(C \in BCT \mid T, A \models C(a_i)\) \(\in \text{ALoop}_\alpha[s_i, s_{i+1}]\), or

(ii) \(T, A \models R(a_i, a_{i+1})\) and \((s_i, R, s_{i+1}) \in \delta\) for some \(R\).

The following lemma characterizes C2RPQ entailment in terms of relaxed matches.

**Lemma 13.** \(T, A \models q\) if and only if \(T, A \models q'\) for some \(q' \in \text{rewrite}(q, T)\).

By applying the preceding characterization, we obtain our C2RPQ upper bounds:

**Proposition 14.** Boolean C2RPQ entailment is

1. NL in data complexity for DL-Lite\(R\) and DL-Lite\(RDFS\);
2. P in data complexity for \(\mathcal{ELH}\);
3. NP in combined complexity for DL-Lite\(RDFS\);
4. \(\text{PSPACE}\) in combined complexity for DL-Lite\(R\) and \(\mathcal{ELH}\).

**Proof sketch.** By Lemmas 9 and 13, \(T, A \models q\) just in the case that \(T, A \models q'\) for some \(q' \in \text{rewrite}(q, T)\). For statements 1 and 2, if \(T, A \models q\), then computing \(\text{rewrite}(q, T)\) requires only constant time in \(|A|\). To decide whether \(T, A \models q'\) for \(q' \in \text{rewrite}(q, T)\), we guess a mapping \(\pi\) from the terms in \(q'\) to \(\text{Ind}(A)\) and verify that it satisfies the conditions in Definition 12. Note that for condition (c), we cannot keep the whole sequence \((a_0, s_0), \ldots, (a_n, s_n)\) in memory at once, so we use a binary counter that counts up to \(|\text{Ind}(A) \times S|\) and store only one pair of nodes \((a_i, s_i), (a_{i+1}, s_{i+1})\) at a time.

The data complexity of verifying conditions (b) and (c) is the same as for instance checking in the corresponding DL: \(A\text{C}_\alpha\) for DL-Lite\(R\), and P for \(\mathcal{ELH}\). This yields the desired upper bounds of NL and NL\(P = P\), respectively.

For statement 4, instead of building the whole set \(\text{rewrite}(q, T)\), which can be exponential, we generate a single \(q' \in \text{rewrite}(q, T)\) non-deterministically. By Lemma 10, every query in \(\text{rewrite}(q, T)\) can be generated after at most exponentially many steps, so we can use a polynomial-size counter to check when we have reached this limit. Since each rewritten query is of polynomial size (Lemma 10), and we keep only one query in memory at a time, the generation of a single query in \(\text{rewrite}(q, T)\) requires only polynomial space. We can then use the same strategy as above to decide in polynomial space whether \(T, A \models q'\). This yields a non-deterministic polynomial space procedure for deciding \(T, A \models q\). Using the well-known fact that \(\text{NPSpace} = \text{PSPACE}\), we obtain the desired upper bound.

For statement 3, we note that if \(T\) is an DL-Lite\(RDFS\) TBox, \(\text{rewrite}(q, T) = \{q\}\). Thus, it suffices to decide \(T, A \models q\), which can be done by guessing a mapping \(\pi\) and verifying in polytime that \(\pi\) satisfies the conditions of Definition 12. □

By moving to 2RPQs, we can achieve tractability even in combined complexity.

**Proposition 15.** Boolean 2RPQ entailment is

1. NL in combined complexity for DL-Lite\(RDFS\);
2. P in combined complexity for DL-Lite\(R\) and \(\mathcal{ELH}\).

**Proof sketch.** For (1), we can iterate over all mappings \(\pi\) of the (at most two) query variables, and for each mapping, we check whether the conditions of Definition 12 are met using the same strategy as in the proof of point 1 of Proposition 14. Recall that in DL-Lite\(RDFS\), instance checking is in NL w.r.t. combined complexity [Calvanese et al., 2007a].

For (2), we first give a polynomial reduction to the problem of deciding whether \(T, A \models q'\) with \(q'\) a 2RPQ. When \(q = \exists xy L(x, y)\) with \(x \neq y\), we can simply replace \(q\) by \(q' = \exists xy \Sigma^* \cdot L \cdot \Sigma^*(x, y)\), where \(\Sigma = N_R \cup N_B \cup \{A | A \in N_C\}\), since \(T, A \models q\) if and only if \(T, A \models q'\). For queries of the form \(\exists x L(x, x)\), the proof is more involved and passes by the definition of an alternative rewriting procedure for 2RPQs, which is similar in spirit to rewrite but is guaranteed to run in polynomial time. We can then check for a match of a 2RPQ in the ABox using essentially the same strategy as for (1), except that we must now perform some polynomial-time \(\text{ALoop}_\alpha\) tests to verify condition (c) of Definition 12. □

6 Conclusion and Future Work

In this paper, we established tight complexity bounds for answering various forms of regular path queries over knowledge bases formulated in lightweight DLs from the DL-Lite and \(\mathcal{EL}\) families. Our results demonstrate that the query answering problem for these richer query languages is often not much harder than for the CQs and IQs typically considered. Indeed, query answering remains tractable in data complexity for the highly expressive class of C2RPQs, and for 2RPQs, we even retain polynomial combined complexity.

In future work, we plan to explore other useful extensions of regular path queries, such as nested path expressions (along the lines of Pérez et al., 2010), and the addition of path variables (recently explored in Barceló et al., 2010).

**Acknowledgements** The authors were supported by a Université Paris-Sud Attractivité grant and the ANR project PAGODA ANR-12-JS02-007-01 (Bienvenue), the FWF project T515-N23 (Ortiz), and the FWF project P25518-N23 and the WWTF project ICT12-015 (Simkus).
References


