Nested Regular Path Queries in Description Logics*
(Extended Abstract)

Meghyn Bienvenu¹, Diego Calvanese², Magdalena Ortiz³, and Mantas Šimkus³

¹ LRI - CNRS & Université Paris Sud
² KRDB Research Centre, Free University of Bozen-Bolzano
³ Institute of Information Systems, Vienna University of Technology

1 Introduction

Both in knowledge representation and in databases, there has been great interest recently in expressive mechanisms for querying data, while taking into account complex domain knowledge [9]. Description Logics (DLs), which on the one hand underlie the W3C standard Web Ontology Language (OWL), and on the other hand are able to capture at the intensional level conceptual modeling formalisms like UML and ER, are considered particularly well suited for representing a domain of interest [7]. In DLs, instance data, stored in a so-called ABox, is constituted by ground facts over unary and binary predicates (concepts and roles, respectively), and hence resembles data stored in graph databases [12,4]. There is a crucial difference, however, between answering queries over graph databases and over DL ABoxes. In the former, the data is assumed to be complete, hence query answering amounts to the standard database task of query evaluation. In the latter, it is typically assumed that the data is incomplete and additional domain knowledge is provided by the DL ontology (or TBox). Hence query answering amounts to the more complex task of computing certain answers, i.e., those answers that are obtained from all databases that both contain the explicit facts in the ABox and satisfy the TBox constraints. This difference has driven research in different directions.

In databases, expressive query languages for querying graph-structured data have been studied, which are based on the requirement of relating objects by flexibly navigating the data. The main querying mechanism that has been considered for this purpose is that of one-way and two-way regular path queries (RPQs and 2RPQs) (cf. [13]). Conjunctive 2RPQs (C2RPQs) [10] are a significant extension of such queries that add to the navigational ability the possibility of expressing arbitrary selections, projections, and joins over objects related by 2RPQs, in line with conjunctive queries (CQs) over relational databases. Two-way RPQs are present in the property paths in SPARQL 1.1 [15], the new standard RDF query language, and in XPath as well. An additional construct that is present in XPath is the possibility of using test operators, also known as nesting, to express sophisticated conditions along navigation paths. This construct has been advocated for querying RDF graphs in the extension of SPARQL called nSPARQL [17], and it has been added to RPQs in the language of nested regular expressions for querying graph databases [34]. It is important to notice that existential tests in general

* This work has been supported by ANR project PAGODA (ANR-12-JS02-007-01), EU IP Project FP7-318338 OPTIQUE, FWF projects T515 and P25518, and WWTF project ICT12-015.
cannot be captured even by C2RPQs, hence adding nesting effectively increases the expressive power of 2RPQs and of C2RPQs.

In the DL community, query answering has been investigated extensively for a wide range of DLs, but most work has been devoted to CQs and unions thereof (see [5] for discussion and references). C2RPQs have been explored for very expressive DLs [11], and recently also for the so called lightweight DLs, which are popular for query answering and data access [6]. Here we consider the extensions 2RPQs and C2RPQs with nesting, obtaining the complexity bounds summarized in Table 1. For DLs containing at least EL, we are able to encode nesting away, thus showing that the worst-case complexity of query answering is not affected by this construct. By contrast, for lightweight DLs (starting already from DL-Lite!), we are able to show that adding nesting to 2RPQs leads to a surprising jump in combined complexity, from P-complete to EXP-complete. Via a sophisticated rewriting-based technique, we prove that for DL-Lite the problem remains in NL in data complexity. We thus demonstrate that adding nesting to (C)2RPQs does not affect the worst-case data complexity of query answering for lightweight DLs.

See [5] for the full version of this paper.

2 Preliminaries

We briefly recall the syntax and semantics of description logics (DLs). As usual, we assume countably infinite, mutually disjoint sets \(N_C\), \(N_R\), and \(N_I\) of concept names, role names, and individuals. We typically use \(A\) for concept names, \(p\) for role names, and \(a, b\) for individuals. An inverse role takes the form \(p^-\) where \(p \in N_R\). We let \(N_R^+ = N_R \cup \{p^- \mid p \in N_R\}\) and denote by \(r\) elements of \(N_R^+\).

A DL knowledge base (KB) consists of a TBox and an ABox, whose forms depend on the DL in question. For example, in the DL ELH\(I\), a TBox is a set of (positive) role inclusions of the form \(r \sqsubseteq r'\) and (negative) role inclusions of the form \(r \sqcap r' \sqsubseteq \bot\) with \(r, r' \in N_R^+\), and concept inclusions of the form \(C \sqsubseteq D\), where \(C\) and \(D\) are complex concepts formed according to the following syntax, with \(A \in N_C\) and \(r \in N_R^+\):

\[
C ::= \top \mid \bot \mid A \mid \exists r.C \mid C \sqcap C.
\]

4 We slightly generalize the usual ELH\(I\) by allowing for negative role inclusions.
\(\mathcal{ELHI}_\perp\) is a Horn DL. In contrast, expressive DLs (such as ALC and SHIQ) allow disjunction \(C \cup C\) and universal restrictions \(\forall r.C\) in complex concepts. We refer the reader to [2] for their definition. The so-called lightweight DLs can be defined as sublogics of \(\mathcal{ELHI}_\perp\). \(\mathcal{ELHI}\) is the fragment of \(\mathcal{ELHI}_\perp\) that has no \(\perp\). \(\mathcal{ELH}\) and \(\mathcal{ELI}\) are obtained by additionally disallowing inverse roles and role inclusions, respectively. DL-Lite\(_R\) is also a fragment of \(\mathcal{ELHI}_\perp\), in which concept inclusions can only take the forms \(B_1 \sqsupseteq B_2\) and \(B_1 \sqsubseteq B_2\), for \(B_i\) a concept name or a concept of the form \(\exists r.\top\) with \(r \in N^r_R\). DL-Lite is the fragment of DL-Lite\(_R\) that disallows role inclusions.

An ABox is a set of assertions of the form \(C(a)\) or \(r(a, b)\), where \(C\) is a complex concept, \(r \in N^r_R\), and \(a, b \in N\). We use Ind(\(A\)) to refer to the set of individuals in \(A\).

**Semantics** The semantics of DL KBs is based upon interpretations, which take the form \(\mathcal{I} = (\Delta^\mathcal{I}, \cdot^\mathcal{I})\), where \(\Delta^\mathcal{I}\) is a non-empty set and \(\cdot^\mathcal{I}\) maps each \(a \in N\) to \(a^\mathcal{I} \in \Delta^\mathcal{I}\), each \(A \in N_C\) to \(A^\mathcal{I} \subseteq \Delta^\mathcal{I}\), and each \(p \in N_R\) to \(p^\mathcal{I} \subseteq \Delta^\mathcal{I} \times \Delta^\mathcal{I}\). The function \(\cdot^\mathcal{I}\) can be straightforwardly extended to complex concepts and roles. In the case of \(\mathcal{EL}(\mathcal{I}, \cdot)\), this is done as follows: \(\top^\mathcal{I} = \Delta^\mathcal{I}\), \(\bot^\mathcal{I} = \emptyset\), \((p^-)^\mathcal{I} = \{(c, d) \mid (d, c) \in p^\mathcal{I}\}\), \((\exists r.C)^\mathcal{I} = \{c \mid \exists d : (c, d) \in r^\mathcal{I}, d \in C^\mathcal{I}\}\), and \((C \cap D)^\mathcal{I} = C^\mathcal{I} \cap D^\mathcal{I}\). An interpretation \(\mathcal{I}\) satisfies an inclusion \(G \sqsubseteq H\) if \(G^\mathcal{I} \subseteq H^\mathcal{I}\), and it satisfies an assertion \(C(a)\) (resp. \(r(a, b)\)) if \(a^\mathcal{I} \in A^\mathcal{I}\) (resp. \((a^\mathcal{I}, b^\mathcal{I}) \in r^\mathcal{I}\)). A model of a KB \((\mathcal{T}, \mathcal{A})\) is an interpretation \(\mathcal{I}\) which satisfies all inclusions in \(\mathcal{T}\) and assertions in \(\mathcal{A}\).

## 3 Nested Regular Path Queries

We now introduce our query languages. In RPQs, nested RPQs and their extensions, atoms are given by (nested) regular expressions whose symbols are roles. The set Roles of roles contains \(N^r_R\), and all test roles of the forms \(\{a\}\) and \(\{a\}\) with \(a \in N\) and \(A \in N_C\). They are interpreted as \(\{a\}\) and \(\{a\}\) with \(a \in N^r_R\) and \(A \in N^C\).

**Definition 1.** A nested regular expression (NRE), denoted by \(E\), is constructed according to the following syntax, where \(\sigma \in \text{Roles}\):

\[
E ::= \sigma \mid E \cdot E \mid E \cup E \mid E^* \mid \langle E \rangle.
\]

We assume a countably infinite set \(N_\eta\) of variables (disjoint from \(N_C, N^r_R, \text{ and } N\)). Each \(t \in N_\eta \cup N\) is a term. An atom is either a concept atom of the form \(A(t)\), with \(A \in N_C\) and \(t\) a term, or a role atom of the form \(E(t, t')\), with \(E\) an NRE and \(t, t'\) two (possibly equal) terms.

A nested two-way regular path query (N2RPQ) \(q(x, y)\) is an atom of the form \(E(x, y)\), where \(E\) is an NRE and \(x, y\) are two distinct variables. A conjunctive N2RPQ (CN2RPQ) \(q(x)\) with answer variables \(x\) has the form \(\exists y.\varphi\), where \(\varphi\) is a conjunction of atoms whose variables are among \(x, y\).

A (plain) regular expression (RE) is an NRE that has no subexpressions of the form \(\langle E \rangle\). Two-way regular path queries (2RPQs) and conjunctive 2RPQs (C2RPQs) are defined analogously to N2RPQs and CN2RPQs but allowing only plain REs in atoms.

\(^5\) N2RPQs coincide with the queries called simply NREs in [34].
Given an interpretation $I$, the semantics of an NRE $E$ is defined inductively:

\[
(E_1 \cdot E_2)^I = E_1^I \circ E_2^I, \quad (E_1^*)^I = (E_1^I)^*, \\
(E_1 \cup E_2)^I = E_1^I \cup E_2^I, \quad \langle E \rangle^I = \{(o, o') \mid \text{there is } o' \in \Delta^I \text{ s.t. } (o, o') \in E^I\}.
\]

Assume a C2NRQP $q(x) = \exists y. \varphi$. A match for $q$ in an interpretation $I$ is a mapping from the terms in $\varphi$ to $\Delta^I$ such that (i) $\pi(a) = a^I$ for every individual $a$ of $\varphi$, (ii) $\pi(x) \in A^I$ for every concept atom $A(x)$ of $\varphi$, and (iii) $\langle \pi(x), \pi(y) \rangle \in E^I$ for every role atom $E(x, y)$ of $\varphi$. Let $\text{ans}(q, I) = \{\pi(x) \mid \pi \text{ is a match for } q \text{ in } I\}$. An individual tuple $a$ with the same arity as $x$ is called a certain answer to $q$ over a KB $\langle T, A \rangle$ if $\langle a \rangle^I \in \text{ans}(q, I)$ for every model $I$ of $\langle T, A \rangle$. We use $\text{ans}(q, \langle T, A \rangle)$ to denote the set of all certain answers to $q$ over $\langle T, A \rangle$. By query answering, we mean the problem of deciding whether $a \in \text{ans}(q, \langle T, A \rangle)$.

**Example 1.** We consider an ABox of advisor relationships of PhD holders. We assume an advisor relation between nodes representing academics. There are also nodes for theses, universities, research topics, and countries, related in the natural way via roles wrote, subm(itted), topic, and loc(ation). We give two queries over this ABox.

\[
q_1(x, y) = (\text{advisor} \cdot (\text{wrote} \cdot \text{topic} \cdot \text{Physics}))^* (x, y)
\]

Query $q_1$ is an N2RPQ that retrieves pairs of a person $x$ and an academic ancestor $y$ of $x$ such that all people on the path from $x$ to $y$ (including $y$) wrote a thesis in Physics.

\[
q_2(x, y, z) = \text{advisor}^* (x, z), \quad \text{advisor}^* (x, w), \\
\text{advisor} \cdot (\text{wrote} \cdot \langle \text{topic} \cdot \text{DBs}\rangle \cdot \text{subm} \cdot \text{loc} \cdot \langle \text{USA}\rangle) (y, z), \\
\text{advisor} \cdot (\text{wrote} \cdot \langle \text{topic} \cdot \text{Logic}\rangle \cdot \text{subm} \cdot \text{loc} \cdot \langle \text{EUcountry}\rangle) (y, w)
\]

Query $q_2$ is a CN2RPQ that looks for triples of individuals $x, y, z$ such that $x$ and $y$ have both supervised $z$, who wrote a thesis on Databases and who submitted this thesis to a university in the USA. Moreover, $x$ and $y$ have a common ancestor $w$, and all people on the path from $x$ to $w$, including $w$, must have written a thesis in Logic and must have submitted this thesis to a university in an EU country.

### 4 Complexity of Query Answering

For the lightweight DLs DL-Lite and $\mathcal{EL}$, a P upper bound in combined complexity for answering 2RPQs and a PSPACE upper bound for C2RPQs are known [6]. However, the addition of nesting causes a significant increase in complexity: already evaluating one N2RPQ in the presence of a DL-Lite or $\mathcal{EL}$ TBox is EXP-hard [5].

**Theorem 1.** N2RPQs in DL-Lite and $\mathcal{EL}$ are EXP-hard in combined complexity.

The above lower bound for answering N2RPQs hinges on the support for existential concepts in the right-hand-side of inclusions. If they are disallowed, then one can find a polynomial-time algorithm [17]. To our knowledge, it was open until now whether the polynomial-time upper bound is optimal. We prove P-hardness of the problem, already for plain graph databases. The proof is by a logspace reduction from the classical P-complete problem of checking entailment in propositional definite Horn theories.

---

6 The examples are inspired by the MGP project [http://genealogy.math.ndsu.nodak.edu/].
Theorem 2. Given as input an $N2RPQ q$, a finite interpretation $I$ and a pair $(o, o') \in \Delta^I \times \Delta^I$, it is $P$-hard to check whether $(o, o') \in \text{ans}(q, I)$.

For $ALC$ and all the expressive DLs that extend it, answering $C2RPQ$s is $2\text{Exp}$-hard. Indeed, the $2\text{Exp}$ hardness proof for conjunctive queries in $SH$ by [14] can be adapted to use an $ALC$ TBox and a $C2RPQ$. We show that this bound and the one in Theorem 1 are tight. This is a consequence of the fact that answering $CN2RPQ$s can be polynomially reduced to answering non-nested $C2RPQ$s using TBox axioms that employ inverses, conjunction on the left, and qualified existential restrictions.

Proposition 1. For each $CN2RPQ q$, we can compute in polynomial time an $ELI$ TBox $T'$ and $C2RPQ q'$ such that $\text{ans}(q, \langle T, A \rangle) = \text{ans}(q', \langle T \cup T', A \rangle)$ for every $\langle T, A \rangle$.

It follows that in every DL that contains $ELI$, answering $CN2RPQ$s is no harder than answering $C2RPQ$s. From existing upper bounds for $C2RPQ$s [11,16], we obtain:

Corollary 1. Answering $CN2RPQ$s is in $2\text{Exp}$ in combined complexity for all DLs contained in $SHIQ$, $SHOI$, $ZIQ$, or $ZOI$; and in $\text{Exp}$ in combined complexity and in $P$ in data complexity for all DLs contained in Horn-$SHOIQ$.

We point out that the $2\text{Exp}$ upper bound for expressive DLs can also be inferred, without using the reduction above, from the existing results for answering $C2RPQ$s in $ZIQ$ and $ZOI$ [11]. Indeed, these DLs support regular role expressions as concept constructors, and a nested expression $\langle E \rangle$ in a query can be replaced by a concept $\exists E. \top$. Hence, in $ZIQ$ and $ZOI$, nested expressions provide no additional expressiveness and $CN2RPQ$s and $C2RPQ$s coincide.

The construction used in Proposition 1 also allows us to reduce the evaluation of an $N2RPQ$ to standard reasoning in any DL that contains $ELI$.

Proposition 2. For every $N2RPQ q$ and every pair of individuals $a, b$, one can compute in polynomial time an $ELI$ TBox $T'$, and a pair of assertions $A_b(b)$ and $A_s(a)$ such that $(a, b) \in \text{ans}(q, \langle T, A \rangle) \iff \langle T \cup T', A \cup \{A_b(b)\} \rangle \models \text{ans}(q', \langle T \cup T', A \cup \{A_b(b)\} \rangle)$, for every DL $\langle T, A \rangle$.

From this and existing upper bounds for instance checking in DLs, we easily obtain:

Corollary 2. Answering $N2RPQ$s is in $\text{Exp}$ in combined complexity for every DL that contains $ELI$ and is contained in $SHIQ$, $SHOI$, $ZIQ$, or $ZOI$.

We note that the $\text{Exp}$ bounds in Corollaries 1 and 2 are optimal for all DLs that contain $ELI$, because standard reasoning tasks like satisfiability checking are already $\text{Exp}$-hard in this logic [11]. For the same reasons, the $P$ bound for data complexity in Corollary 1 is tight for $ELI$ and its extensions [8].

The results stated so far leave a gap for the data complexity of the $DL-Lite$ family: we inherit NL-hardness from plain RPQs, but we only have the $P$ upper bound stemming from Proposition 1. This gap can be closed showing an NL upper bound, by extending to $CN2RPQ$s an algorithm for answering $C2RPQ$s due to Bienvenu et al. [6]). The algorithm uses a sophisticated rewriting technique, and has the additional advantage of being more likely to serve as a basis for practicable techniques than the reductions sketched above. Please consult [5] for details.

7 For (1-way) CRPQs, which contain no inverse roles, the same applies to $ZOQ$ and its sublogics.
5 Conclusions and Future Work

We have studied the extension of (C)2RPQs with a nesting construct inspired by XPath, and have characterized the data and combined complexity of answering nested 2RPQs and C2RPQs for a wide range of DLs. In light of the surprising jump from P to EXP in the combined complexity of answering nested 2RPQs in lightweight DLs, a relevant problem is to identify classes that exhibit better computational properties.

References