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

Meghyn Bienvenu[1], Diego Calvanese[2], Magdalena Ortiz[3], and Mantas Šimkus[3]

[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 [3,4]. 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.

	2RPQ		C2RPQ		N2RPQ / CN2RPQ	
	data	combined	data	combined	data	combined
Graph DBs	NL-c	NL-c	NL-c	NP-c	NL-c	**P-c** / NP-c
DL-Lite	NL-c	P-c	NL-c	PSPACE-c	**NL-c**	**EXP-c**
Horn DLs ($\mathcal{EL}$, Horn-$\mathcal{SHIQ}$)	P-c	P-c	P-c	PSPACE-c	**P-c**	**EXP-c**
Expressive DLs ($\mathcal{ALC}$, $\mathcal{SHIQ}$, $\mathcal{ZIQ}$)	coNP-h	EXP-c	coNP-h	2EXP-c	coNP-h	**2EXP-c**

Table 1. Complexity of query answering. The 'c' indicates completeness, the 'h' hardness. New results are marked in bold. For references to existing results, consult [5].

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 $\mathcal{ELI}$, 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 $\mathsf{N_C}$, $\mathsf{N_R}$, and $\mathsf{N_I}$ 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 $p \in \mathsf{N_R}$. We let $\mathsf{N_R^\pm} = \mathsf{N_R} \cup \{p^- \mid p \in \mathsf{N_R}\}$ and denote by r elements of $\mathsf{N_R^\pm}$.

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 $\mathcal{ELHI}_\bot$, 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 $r, r' \in \mathsf{N_R^\pm}$, and *concept inclusions* of the form $C \sqsubseteq D$, where C and D are *complex concepts* formed according to the following syntax, with $A \in \mathsf{N_C}$ and $r \in \mathsf{N_R^\pm}$:[4]

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

[4] We slightly generalize the usual $\mathcal{ELHI}_\bot$ by allowing for negative role inclusions.

$\mathcal{ELHI}_\bot$ is a *Horn* DL. In contrast, expressive DLs (such as $\mathcal{ALC}$ and $\mathcal{SHIQ}$) allow disjunction $C \sqcup C$ 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 $\mathcal{ELHI}_\bot$. $\mathcal{ELHI}$ is the fragment of $\mathcal{ELHI}_\bot$ that has no $\bot$. $\mathcal{ELH}$ and $\mathcal{ELI}$ are obtained by additionally disallowing inverse roles and role inclusions, respectively. *DL-Lite*$_\mathcal{R}$ is also a fragment of $\mathcal{ELHI}_\bot$, in which concept inclusions can only take the forms $B_1 \sqsubseteq B_2$ and $B_1 \sqcap B_2 \sqsubseteq \bot$, for B_i a concept name or a concept of the form $\exists r.\top$ with $r \in \mathsf{N}_\mathsf{R}^\pm$. *DL-Lite* is the fragment of *DL-Lite*$_\mathcal{R}$ 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, $r \in \mathsf{N}_\mathsf{R}^\pm$, and $a, b \in \mathsf{N}_\mathsf{I}$. We use $\mathsf{Ind}(\mathcal{A})$ to refer to the set of individuals in $\mathcal{A}$.

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 $a \in \mathsf{N}_\mathsf{I}$ to $a^\mathcal{I} \in \Delta^\mathcal{I}$, each $A \in \mathsf{N}_\mathsf{C}$ to $A^\mathcal{I} \subseteq \Delta^\mathcal{I}$, and each $p \in \mathsf{N}_\mathsf{R}$ 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 $\mathcal{ELHI}_\bot$, 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 $(C \sqcap D)^\mathcal{I} = C^\mathcal{I} \cap D^\mathcal{I}$. 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 $\mathsf{N}_\mathsf{R}^\pm$, and all *test roles* of the forms $\{a\}?$ and $A?$ with $a \in \mathsf{N}_\mathsf{I}$ and $A \in \mathsf{N}_\mathsf{C}$. They are interpreted as $(\{a\}?)^\mathcal{I} = \{(a^\mathcal{I}, a^\mathcal{I})\}$ and $(A?)^\mathcal{I} = \{(o, o) \mid o \in A^\mathcal{I}\}$.

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

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

We assume a countably infinite set N_V of variables *(disjoint from N_C, N_R, and N_I). Each $t \in \mathsf{N}_\mathsf{V} \cup \mathsf{N}_\mathsf{I}$ is a* term*. An* atom *is either a* concept atom *of the form $A(t)$, with $A \in \mathsf{N}_\mathsf{C}$ 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.*[5] *A* conjunctive N2RPQ *(CN2RPQ) $q(\boldsymbol{x})$ with* answer variables *$\boldsymbol{x}$ has the form $\exists \boldsymbol{y}.\varphi$, where φ is a conjunction of atoms whose variables are among $\boldsymbol{x} \cup \boldsymbol{y}$.*

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 [3,4].

Given an interpretation $\mathcal{I}$, the semantics of an NRE E is defined inductively:

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

Assume a C2NRPQ $q(\boldsymbol{x}) = \exists \boldsymbol{y}.\varphi$. A *match* for q in an interpretation $\mathcal{I}$ is a mapping from the terms in φ to $\Delta^{\mathcal{I}}$ such that (i) $\pi(a) = a^{\mathcal{I}}$ for every individual a of φ, (ii) $\pi(x) \in A^{\mathcal{I}}$ for every concept atom $A(x)$ of φ, and (iii) $(\pi(x), \pi(y)) \in E^{\mathcal{I}}$ for every role atom $E(x,y)$ of φ. Let $\mathsf{ans}(q, \mathcal{I}) = \{\pi(\boldsymbol{x}) \mid \pi \text{ is a match for } q \text{ in } \mathcal{I}\}$. An individual tuple $\boldsymbol{a}$ with the same arity as $\boldsymbol{x}$ is called a *certain answer* to q over a KB $\langle \mathcal{T}, \mathcal{A} \rangle$ if $(\boldsymbol{a})^{\mathcal{I}} \in \mathsf{ans}(q, \mathcal{I})$ for every model $\mathcal{I}$ of $\langle \mathcal{T}, \mathcal{A} \rangle$. We use $\mathsf{ans}(q, \langle \mathcal{T}, \mathcal{A} \rangle)$ to denote the set of all certain answers to q over $\langle \mathcal{T}, \mathcal{A} \rangle$. By *query answering*, we mean the problem of deciding whether $\boldsymbol{a} \in \mathsf{ans}(q, \langle \mathcal{T}, \mathcal{A} \rangle)$.

Example 1. We consider an ABox of advisor relationships of PhD holders[6]. 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.

$$q_1(x,y) = (advisor \cdot \langle wrote \cdot topic \cdot Physics? \rangle)^* \, (x,y)$$

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.

$$\begin{aligned} q_2(x,y,z) = {} & advisor^-(x,z), \;\; advisor^*(x,w), \\ & advisor^- \cdot \langle wrote \cdot \langle topic \cdot DBs? \rangle \cdot subm \cdot loc \cdot \{USA\}? \rangle (y,z), \\ & \big(advisor \cdot \langle wrote \cdot \langle topic \cdot Logic? \rangle \cdot subm \cdot loc \cdot EUcountry? \rangle\big)^* \, (y,w) \end{aligned}$$

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*$_{\mathcal{R}}$ 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 $\mathcal{I}$ and a pair $(o, o') \in \Delta^{\mathcal{I}} \times \Delta^{\mathcal{I}}$, it is* P*-hard to check whether $(o, o') \in \mathsf{ans}(q, \mathcal{I})$.*

For $\mathcal{ALC}$ and all the expressive DLs that extend it, answering C2RPQs is 2EXP-hard. Indeed, the 2EXP hardness proof for conjunctive queries in $\mathcal{SH}$ by [14] can be adapted to use an $\mathcal{ALC}$ 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 CN2RPQs can be polynomially reduced to answering non-nested C2RPQs 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 $\mathcal{ELI}$ TBox $\mathcal{T}'$ and C2RPQ q' such that $\mathsf{ans}(q, \langle \mathcal{T}, \mathcal{A} \rangle) = \mathsf{ans}(q', \langle \mathcal{T} \cup \mathcal{T}', \mathcal{A} \rangle)$ for every $\langle \mathcal{T}, \mathcal{A} \rangle$.*

It follows that in every DL that contains $\mathcal{ELI}$, answering CN2RPQs is no harder than answering C2RPQs. From existing upper bounds for C2RPQs [11,16], we obtain:

Corollary 1. *Answering CN2RPQs is in* 2EXP *in combined complexity for all DLs contained in $\mathcal{SHIQ}$, $\mathcal{SHOI}$, $\mathcal{ZIQ}$, or $\mathcal{ZOI}$; and in* EXP *in combined complexity and in* P *in data complexity for all DLs contained in Horn-$\mathcal{SHOIQ}$.*

We point out that the 2EXP upper bound for expressive DLs can also be inferred, without using the reduction above, from the existing results for answering C2RPQs in $\mathcal{ZIQ}$ and $\mathcal{ZOI}$ [11].[7] 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 $\mathcal{ZIQ}$ and $\mathcal{ZOI}$, nested expressions provide no additional expressiveness and CN2RPQs and C2RPQs 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 $\mathcal{ELI}$.

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

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

Corollary 2. *Answering N2RPQs is in* EXP *in combined complexity for every DL that contains $\mathcal{ELI}$ and is contained in $\mathcal{SHIQ}$, $\mathcal{SHOI}$, $\mathcal{ZIQ}$, or $\mathcal{ZOI}$.*

We note that the EXP bounds in Corollaries 1 and 2 are optimal for all DLs that contain $\mathcal{ELI}$, because standard reasoning tasks like satisfiability checking are already EXP-hard in this logic [1]. For the same reasons, the P bound for data complexity in Corollary 1 is tight for $\mathcal{EL}$ 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 CN2RPQs an algorithm for answering C2RPQs 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 $\mathcal{ZOQ}$ 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

1. F. Baader, S. Brandt, and C. Lutz. Pushing the $\mathcal{EL}$ envelope further. In *Proc. of the 5th Int. Workshop on OWL: Experiences and Directions (OWLED 2008)*, 2008.
2. F. Baader, D. Calvanese, D. McGuinness, D. Nardi, and P. F. Patel-Schneider, editors. *The Description Logic Handbook: Theory, Implementation and Applications*. Cambridge University Press, 2003.
3. P. Barceló, J. Pérez, and J. L. Reutter. Relative expressiveness of nested regular expressions. In *Proc. of AMW'12*, CEUR Workshop Proceedings 866, pages 180–195, 2012.
4. P. Barceló Baeza. Querying graph databases. In *Proc., of PODS'13*, pages 175–188, 2013.
5. M. Bienvenu, D. Calvanese, M. Ortiz, and M. Šimkus. Nested regular path queries in description logics. In *Proc. of KR 2014*. AAAI Press, 2014.
6. M. Bienvenu, M. Ortiz, and M. Simkus. Conjunctive regular path queries in lightweight description logics. In *Proc. of IJCAI 2013*, 2013.
7. A. Borgida and R. J. Brachman. Conceptual modeling with description logics. In Baader et al. [2], chapter 10, pages 349–372.
8. D. Calvanese, G. De Giacomo, D. Lembo, M. Lenzerini, and R. Rosati. Data complexity of query answering in description logics. In *Proc. of KR 2006*, pages 260–270, 2006.
9. D. Calvanese, G. De Giacomo, and M. Lenzerini. Conjunctive query containment and answering under description logics constraints. *ACM TOCL*, 9(3):22.1–22.31, 2008.
10. D. Calvanese, G. De Giacomo, M. Lenzerini, and M. Y. Vardi. Containment of conjunctive regular path queries with inverse. In *Proc. of KR 2000*, pages 176–185, 2000.
11. D. Calvanese, T. Eiter, and M. Ortiz. Regular path queries in expressive description logics with nominals. In *Proc. of IJCAI 2009*, pages 714–720, 2009.
12. M. P. Consens and A. O. Mendelzon. GraphLog: a visual formalism for real life recursion. In *Proc. of the 9th ACM SIGACT SIGMOD SIGART Symp. on Principles of Database Systems (PODS'90)*, pages 404–416, 1990.
13. I. F. Cruz, A. O. Mendelzon, and P. T. Wood. A graphical query language supporting recursion. In *Proc. of the ACM SIGMOD Int. Conf. on Management of Data*, pages 323–330, 1987.
14. T. Eiter, C. Lutz, M. Ortiz, and M. Simkus. Query answering in description logics with transitive roles. In *Proc. of IJCAI 2009*, pages 759–764, 2009.
15. S. Harris and A. Seaborne. SPARQL 1.1 Query Language. W3C Recommendation, World Wide Web Consortium, Mar. 2013.
16. M. Ortiz, S. Rudolph, and M. Simkus. Query answering in the Horn fragments of the description logics $\mathcal{SHOIQ}$ and $\mathcal{SROIQ}$. In *Proc. of IJCAI 2011*, pages 1039–1044, 2011.
17. J. Pérez, M. Arenas, and C. Gutierrez. nSPARQL: A navigational language for RDF. *J. of Web Semantics*, 8(4):255–270, 2010.