Optimizing FOL reducible query answering

Damian Bursztyn  
INRIA & U. Paris-Sud, France  
damian.burstzyn@inria.fr

François Goasdoué  
U. Rennes 1 & INRIA, France  
fg@irisa.fr

Ioana Manolescu  
INRIA & U. Paris-Sud, France  
ioana.manolescu@inria.fr

ABSTRACT

We devise a query optimization framework for formalisms enjoying FOL reducibility of query answering, for which it reduces to the evaluation of a FOL query against facts. This framework allows searching within a set of alternative equivalent FOL queries, i.e., FOL reformulations, one with minimal evaluation cost when evaluated through a relational database system. We provide two algorithms, an exhaustive and a greedy, for exploring this space of alternatives. We apply this framework to the DL-Lite\textsubscript{R} description logic underpinning the W3C’s OWL2 QL profile; an experimental evaluation validates the interest and applicability of our technique.

1. INTRODUCTION

Answering queries under deductive constraints is a longstanding database research topic.\footnote{http://www.w3.org/2001/sw/Specs.html} It recently gained the interest of the database community at large, with the advent of ontology-based data access (OBDA)\footnote{http://www.w3.org/TR/owl2-overview}. The latter consists in querying facts described by an application domain conceptualization, in the spirit of an ER models or UML class diagram, expressed by deductive constraints. Several languages have been investigated for expressing such constraints, e.g., Datalog$^+$\footnote{1} [9], Description Logics\footnote{2} and Existential Rules\footnote{3} for logical facts, or RDF Schema for RDF graphs.\footnote{4} In this context, query answering denotes the task of computing the answer to the given query, by taking into account both the facts and the constraints holding on them, in contrast with the more common query evaluation task, which just needs to examine the stored facts.

A prominent trend in the OBDA literature is to identify dialects of the above mentioned languages, enjoying first-order logic (FOL) reducibility (or rewritability) of query answering, e.g., \cite{10,9,20}. For such languages, query answering reduces to the evaluation of a FOL query against the facts, obtained by compiling the domain constraints into the query. The so-called FOL reformulations thus obtained, correspond to SQL statements, which can be handed over for evaluation to an efficient relational database management system (RDBMS) storing the facts. However, FOL reformulations encoding domain constraints tend to be (significantly) more complex than the typical queries that RDBMS optimizers were tuned for, in particular due to repeated sub-queries, thus RDBMSs perform overall poorly at evaluating them. Therefore, OBDA optimizations investigated so far mainly aim to produce fast, for specific logical query, constraint and reformulation dialects, FOL reformulations as minimal as possible to limit redundancy, e.g., \cite{24,14,23,11,29,17,20,28,15}.

In this paper, our approach is broader, and performance-oriented. We devise a query optimization framework for any logical OBDA setting enjoying FOL reducibility of query answering.\footnote{We also apply this framework to the lightweight DL-Lite\textsubscript{R} description logic\footnote{5} underpinning the popular W3C’s OWL2 QL standard for rich Semantic Web applications. Query answering in DL-Lite\textsubscript{R} has received significant attention in the literature, notably techniques based on FOL reducibility, e.g., \cite{10,1,24,26,11,29}.} We reduce query answering to the evaluation of alternative FOL queries belonging to richer languages than those considered so far in the literature; in particular, this may allow several (equivalent) FOL reformulations for the input query. This contrasts with related works (cf. Section\footnote{6}), which aim at a single FOL reformulation (modulo minimization). Allowing a variety of reformulations is crucial for efficiency, as such alternatives, while computing the same answers, may have very different evaluation costs when evaluated through an RDBMS. Therefore, instead of having a single default choice which may perform poorly, we select one with lowest estimated evaluation cost among the possible alternatives. Figure\footnote{7} illustrates our query optimization framework.

We also apply this framework to the lightweight DL-Lite\textsubscript{R} description logic\footnote{8} underpinning the popular W3C’s OWL2 QL standard for rich Semantic Web applications. Query answering in DL-Lite\textsubscript{R} has received significant attention in the literature, notably techniques based on FOL reducibility, e.g., \cite{10,1,24,26,11,29}. The contributions we bring to the problem of optimizing FOL reducible query answering can be outlined as follows:

1. For logical formalisms enjoying FOL reducibility of query answering, we provide a general optimization framework that reduces query answering to searching among a set of alternative equivalent FOL queries, i.e., FOL reformulations, one with minimal evaluation cost in an RDBMS (Section\footnote{9}).

2. We characterize interesting spaces of such alternative equivalent FOL queries for DL-Lite\textsubscript{R} (Section\footnote{10}), called JUCQ or JUSCQ reformulations.

3. We then optimize query answering in the setting of DL-Lite\textsubscript{R} by searching within one of these spaces and picking a reformulation with lowest (estimated) evaluation cost w.r.t. an RDBMS cost model. We provide two algorithms, an exhaustive and a greedy, for selecting a low-cost reformulation (Section\footnote{11}).

Contributions. The contributions we bring to the problem of optimizing FOL reducible query answering can be outlined as follows:

(c) 2015, Copyright is with the authors. Informally presented at the BDA 2015 Conference (September 29-October 2, 2015, Île de Porquerolles, France).

(c) 2015, Droits restant aux auteurs. Présenté à la conférence BDA 2015 (29 Septembre-02 Octobre 2015, Île de Porquerolles, France).

BDA 29 septembre 2015, Île de Porquerolles, France.
4. Finally, we demonstrate experimentally the effectiveness and the efficiency of our query answering technique for DL-Lite\(\forall\), by deploying it on top of the Postgres RDBMS (Section 6).

In the sequel, Section 2 recalls the basics of DL-Lite\(\forall\) description logic [10], on which the OWL2 QL W3C standard for managing semantic-rich Web data is built. DL-Lite\(\forall\) is a fragment of Datalog\(±\) and of Existential Rules; it largely extends the subset of RDF (comprising RDF Schema constraints) which translates into description logics.

## 2. PRELIMINARIES

We recall the basics of the DL-Lite\(\forall\) description logic [10], on which the OWL2 QL W3C standard for managing semantic-rich Web data is built. DL-Lite\(\forall\) is a fragment of Datalog\(±\) and of Existential Rules; it largely extends the subset of RDF (comprising RDF Schema constraints) which translates into description logics.

### 2.1 DL-Lite knowledge bases

A knowledge base (KB) \(K\) consists of a TBox \(T\) (ontology, or axiom set) and an ABox \(A\) (fact set, or dataset), denoted \(K = \langle T, A \rangle\), with \(T\) expressing constraints on \(A\).

\(T\) and \(A\) are built from a set \(N_C\) of concept names (unary predicates), a set \(N_R\) of role names (binary predicates), and a set \(N_I\) of individuals (constants). The ABox consists of a finite number of concept assertions of the form \(A(a)\) with \(A \in N_C\) and \(a \in N_I\), and of role assertions of the form \(R(a, b)\) with \(R \in N_R\) and \(a, b \in N_I\). The TBox is a set of axioms that can be concept inclusions of the form \(B \sqsubseteq C\), or role inclusions of the form \(Q \subseteq S\), formed using the following grammar (where \(A \in N_C\) and \(R \in N_R\)):

\[
B := A \mid \exists Q, C := B \mid \neg B, Q := R \mid R^-, S := Q \mid \neg Q
\]

Remark that negation can only be used in the right-hand side of inclusion constraints. Constraints featuring such negation are called disjunctive (or exclusion) constraints.

The semantics of inclusion constraints is defined, as customary, in terms of their FOL interpretations. Tables 1 and 2 provide the FOL and relational notations expressing these constraints equivalently.

An interpretation \(I\) has the form \(I = (\Delta^A, \Delta^R)\), where \(\Delta^A\) is a non-empty set and \(\Delta^R\) is a function mapping each \(\alpha \in N_I\) to \(\alpha^I \in \Delta^A\), each \(A \in N_C\) to \(A^I \subseteq \Delta^A\), and each \(R \in N_R\) to \(R^I \subseteq \Delta^A \times \Delta^A\). The function \(-^I\) is straightforwardly extended to concepts and roles:

- projection: \(\exists Q)^I = \{a_1 \mid \exists a_2 (o_1, a_2) \in Q^I\}\)
- inverse: \(R^-)^I = \{(o_1, o_2) \mid (a_1, a_2) \in R^I\}\)
- negation: \(\neg B)^I = \Delta^A \setminus B^I\) and \((\neg Q)^I = (\Delta^A \times \Delta^A) \setminus Q^I\)

An interpretation \(I\) satisfies an inclusion \(B \subseteq C\) (resp. \(Q \subseteq S\)) if \(B^I \subseteq C^I\) (resp. \(Q^I \subseteq S^I\)), it satisfies \(A(a)\) (resp. \(R(a, b)\)) if \(a^I \in A^I\) (resp. \(a^I, b^I \in R^I\)).

An interpretation \(I\) is a model of \(K = \langle T, A \rangle\) if \(I\) satisfies all inclusions in \(T\) and assertions in \(A\). A KB \(K = \langle T, A \rangle\) is consistent if it has a model. In this case, we say that \(A\) is \(T\)-consistent.

Finally, an inclusion or assertion \(\alpha\) is entailed by a KB \(K\), written \(K \models \alpha\), if \(\alpha\) is satisfied in all the models of \(K\).

**Example 1 (DL-Lite\(\forall\) data model).** Consider the DL-Lite\(\forall\) TBox \(T\) below expressing constraints on the atomic concepts Researcher and PhDStudent, and atomic roles worksWith and supervisedBy.

\[
\begin{align*}
(T1) & \quad \text{PhDStudent} \sqsubseteq \text{Researcher} \\
(T2) & \quad \exists \text{worksWith} \sqsubseteq \exists \text{Researcher} \\
(T3) & \quad \text{worksWith} \sqsubseteq \text{Researcher} \\
(T4) & \quad \text{worksWith} \sqsubseteq \text{worksWith}^- \\
(T5) & \quad \text{supervisedBy} \sqsubseteq \text{worksWith}^- \\
(T6) & \quad \text{supervisedBy} \sqsubseteq \text{PhDStudent} \\
(T7) & \quad \text{PhDStudent} \sqsubseteq \neg \exists \text{supervisedBy}^-
\end{align*}
\]

It states that PhD students are researchers (T1), researchers work with researchers (T2)/(T3), working with someone is a symmetric relation (T4), being supervised by someone implies working with her/him (T5), only PhD students are supervised (T6) and they cannot supervise (T7).

Consider now the following ABox for the same atomic concepts and roles:

\[
\begin{align*}
(A1) & \quad \text{worksWith}(\text{IOANA, FRANCOIS}) \\
(A2) & \quad \text{supervisedBy}(\text{DAMIAN, IOANA}) \\
(A3) & \quad \text{supervisedBy}(\text{DAMIAN, FRANCOIS})
\end{align*}
\]

It states that Ioana works with François (A1), Damian is supervised by Ioana (A2) and François (A3).

It is worth noting that the knowledge base \(K = \langle T, A \rangle\) entails many constraints and assertions. For instance:

- \(K \models \exists \text{supervisedBy}^\setminus \neg \exists \text{supervisedBy}^-,\) i.e., the two attributes of supervisedBy are disjoints, due to (T6) + (T7);
- \(K \models \exists \text{worksWith}(\text{FRANCOIS, IOANA}),\) i.e., François works with Ioana, due to (T4) + (A1);
- \(K \models \text{PhDStudent}(\text{DAMIAN}),\) i.e., Damian is a PhD student, due to (A2) + (T6);
- \(K \models \exists \text{worksWith}(\text{FRANCOIS, DAMIAN}),\) i.e., François works with Damian, due to (A3) + (T5) + (T4).

Finally remark that \(A\) is \(T\)-consistent, i.e., there is no violation of (T7), since the KB \(K\) does not entail that some PhD student supervises another.
<table>
<thead>
<tr>
<th>DL constraint</th>
<th>FOL constraint</th>
<th>Relational constraint (under Open World Assumption)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A \subseteq A'$</td>
<td>$\forall x \left( A(x) \rightarrow A'(x) \right)$</td>
<td>$A \subseteq A'$</td>
</tr>
<tr>
<td>$A \subseteq R R$</td>
<td>$\forall x \left( A(x) \rightarrow \exists y R(x,y) \right)$</td>
<td>$A \subseteq \Pi_1(R)$</td>
</tr>
<tr>
<td>$A \subseteq R R'$</td>
<td>$\forall x \left( A(x) \rightarrow \exists y R(y,x) \right)$</td>
<td>$A \subseteq \Pi_2(R)$</td>
</tr>
<tr>
<td>$R \subseteq A$</td>
<td>$\forall A \exists y R(y,x) \Rightarrow A(x)$</td>
<td>$\Pi_1(R) \subseteq A$</td>
</tr>
<tr>
<td>$R \subseteq R A$</td>
<td>$\forall A \exists y R(y,x) \Rightarrow A(x)$</td>
<td>$\Pi_2(R) \subseteq A$</td>
</tr>
<tr>
<td>$R \subseteq R \exists R A$</td>
<td>$\forall A \exists y R(y,x) \Rightarrow \exists z R(x,z)$</td>
<td>$\Pi_1(R) \subseteq \Pi_1(R)$</td>
</tr>
<tr>
<td>$R \subseteq R \exists R R A$</td>
<td>$\forall A \exists y R(y,x) \Rightarrow \exists z R(y,x)$</td>
<td>$\Pi_2(R) \subseteq \Pi_1(R)$</td>
</tr>
<tr>
<td>$R \subseteq R$ or $R \subseteq R'$</td>
<td>$\forall A \exists y R(y,x) \Rightarrow R'(y,x)$</td>
<td>$R \subseteq \Pi_2(R)$ or $\Pi_1(R) \subseteq R'$</td>
</tr>
<tr>
<td>$R \subseteq R'$ or $R' \subseteq R$</td>
<td>$\forall A \exists y R(y,x) \Rightarrow R'(y,x)$</td>
<td>$R \subseteq R'$ or $\Pi_1(R) \subseteq \Pi_2(R')$</td>
</tr>
</tbody>
</table>

Table 1: DL-Lite$_\mathbb{E}$ inclusion constraints without negation in FOL, and in relational notation; $A, A'$ are concept names while $R, R'$ are role names. For the relational notation, we use $\bot$ to designate the first attribute of any atomic role, and $\bot$ for the second.

<table>
<thead>
<tr>
<th>DL constraint</th>
<th>FOL constraint</th>
<th>Relational constraint (under Open World Assumption)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A \subseteq \neg A'$</td>
<td>$\forall x \left( A(x) \Rightarrow \neg A'(x) \right)$</td>
<td>$A \cap A' \subseteq \bot$</td>
</tr>
<tr>
<td>$A \subseteq \neg R R$</td>
<td>$\forall x \left( A(x) \Rightarrow \neg \exists y R(x,y) \right)$</td>
<td>$A \cap \Pi_1(R) \subseteq \bot$</td>
</tr>
<tr>
<td>$A \subseteq \neg R R'$</td>
<td>$\forall x \left( A(x) \Rightarrow \neg \exists y R(y,x) \right)$</td>
<td>$A \cap \Pi_2(R) \subseteq \bot$</td>
</tr>
<tr>
<td>$\exists R \subseteq \neg A$</td>
<td>$\forall A \exists y R(y,x) \Rightarrow \neg A(x)$</td>
<td>$A \cap \Pi_1(R) \subseteq \bot$</td>
</tr>
<tr>
<td>$\exists R \subseteq \neg R A$</td>
<td>$\forall A \exists y R(y,x) \Rightarrow \neg A(x)$</td>
<td>$A \cap \Pi_2(R) \subseteq \bot$</td>
</tr>
<tr>
<td>$\exists R \subseteq \neg R R A$</td>
<td>$\forall A \exists y R(y,x) \Rightarrow \neg \exists z R(x,z)$</td>
<td>$\Pi_1(R) \cap \Pi_1(R) \subseteq \bot$</td>
</tr>
<tr>
<td>$\exists R \subseteq \neg R R R A$</td>
<td>$\forall A \exists y R(y,x) \Rightarrow \neg \exists z R(x,z)$</td>
<td>$\Pi_2(R) \cap \Pi_2(R) \subseteq \bot$</td>
</tr>
<tr>
<td>$\exists R \subseteq \neg R A$</td>
<td>$\forall A \exists y R(y,x) \Rightarrow \neg A(x)$</td>
<td>$\Pi_1(R) \cap \Pi_1(R) \subseteq \bot$</td>
</tr>
<tr>
<td>$\exists R \subseteq \neg R R A$</td>
<td>$\forall A \exists y R(y,x) \Rightarrow \neg A(x)$</td>
<td>$\Pi_2(R) \cap \Pi_2(R) \subseteq \bot$</td>
</tr>
<tr>
<td>$R \subseteq R$ or $R \subseteq R'$</td>
<td>$\forall A \exists y R(y,x) \Rightarrow R'(y,x)$</td>
<td>$R \cap \Pi_1(R) \subseteq \bot$ or $\Pi_1(R) \cap R' \subseteq \bot$</td>
</tr>
<tr>
<td>$R \subseteq R'$ or $R' \subseteq R$</td>
<td>$\forall A \exists y R(y,x) \Rightarrow R'(y,x)$</td>
<td>$R \cap \Pi_2(R) \subseteq \bot$ or $\Pi_2(R) \cap R' \subseteq \bot$</td>
</tr>
</tbody>
</table>

Table 2: DL-Lite$_\mathbb{E}$ inclusion constraints with negation, i.e., disjointness constraints, in FOL and relational notation. $\bot$ denotes the empty relation.

2.2 Queries

A FOL query is a FOL formula $\phi(x)$ whose free variables are $x$. The arity of a query is the number of its free variables, e.g., 0 for a Boolean query. Given a FOL interpretation $I = (\Delta^I, ^I\cdot)$, the semantics of a Boolean query is defined as true if $\phi(x)^I = true$ and false otherwise, while the semantics of a non-Boolean query of arity $n \geq 1$ is the relation of arity $n$ defined on $\Delta^I$ as follows: $\{ o \in (\Delta^I)^n \mid \phi(o)^I = true \}$. An interpretation that makes true a Boolean FOL query, or non-empty a non-Boolean FOL query, is a model of that query.

Given a query $q$ asked against a KB $K$:

- If $q$ is Boolean, the answer set of $q$ over $K$ is defined as: $\text{ans}(q,K) = \{\text{false}\}$ if $K \not\models q$, with $\{\text{true}\}$ the empty tuple, and $\text{ans}(q,K) = \{\text{true}\}$ otherwise.

- If $q$ is a non-Boolean query of arity $n$, the answer set of $q$ against $K$ is: $\text{ans}(q,K) = \{ t \in (N_I)^n \mid K \models q(t) \}$ where $K \models q(t)$ means as usual that every model of $K$ is a model of $q(t)$.

Example 2 (Query answering). Consider the FOL query asking for the PhD students with whom someone works:

$$\exists y \text{PhDStudent}(x) \land \text{worksWith}(y,x)$$

Given the KB $K$ of the preceding example, the answer set of this query is $\{\text{Damian}\}$, since $K \models \text{PhDStudent}(\text{Damian})$ and $K \models \text{worksWith}(\text{Francois}, \text{Damian})$ hold.

FOL query dialects. In this paper, we focus on the following FOL query sublanguages. Conjunctive Queries (CQs), the core database queries a.k.a. Select-Project-Join queries, are conjunctions of atoms whose bound variables are existentially quantified, and whose atoms are of the forms $A(t)$ or $R(t, t')$ with $t, t'$ variables or individuals. Semi-Conjunctive Queries (SCQs) are conjunctions of disjunctions of single-atom CQs with the same arity, where the atom is either of the form $A(t)$ or of the form $R(t, t')$ as above; the bound variables of SCQs are also existentially quantified. Unions of CQs (UCQs) are disjunctions of CQs with same arity; similarly Unions of SCQs (USCQs) are disjunctions of SCQs with same arity. Finally, Joins of CQs (JUCQs) are conjunctions of UCQs; similarly Joins of USCQs (JSUCQs) are conjunctions of USCQs. All the abovementioned query languages are easily translated into SQL and thus can be evaluated by an RDBMS.

Notations. We write a FOL query $q(x) \leftarrow \phi(x)$ and we refer to it by its name $q$ or its head $q(x)$ when we also need to refer to its free variables; $\phi(x)^I$ is the body of $q$. In the queries we consider, existential variables in $\phi(x)$ are not explicitly quantified; they correspond to those which are not in $x$.

Unless otherwise specified, we systematically use $q$ to refer to the input CQ, $a_1, \ldots, a_n$ to designate the atoms in the body of $q$, $T$ to designate a DL-Lite$_\mathbb{E}$ TBox, and $A$ for the ABox.

FOL-reducibility of data management. DL-Lite$_\mathbb{E}$ has been designed so that the core data management tasks of deciding KB consistency and of query answering are FOL-reducible [19]. This property allows reducing a data management task over a KB $K = (T, A)$ to the evaluation against $A$ only of a FOL query computed using $T$ only.
The interest of FOL-reducibility is to perform a data management task in two separate steps: a first reasoning step that begets the FOL query, and a second step which evaluates that query in a purely relational fashion, i.e., within an RDBMS after translation to SQL. These systems benefit from decades of research and development invested in optimizing data stores, and query evaluation engines; thus, they hold great potential for efficient FOL query evaluation.

The literature provides techniques for computing FOL reformulations of a CQ in settings related to DL-Lite\(\mathbb{Q}\) \(\forall\). They produce (i) a UCQ w.r.t. a DL-Lite\(\mathbb{Q}\) TBox, e.g., \(\mathcal{T}_{\mathcal{U}}\) \[12\], \[21\], \[22\], \[21\], \[20\], or extensions thereof using existential rules \[17\] or Datalog \[12\], \[24\]. (ii) a USCQ \[10\] w.r.t. a set of existential rules generalizing a DL-Lite\(\mathbb{Q}\) TBox, and (iii) a set of alternative equivalent JUCQs \[8\] w.r.t. an RDF dataset \[15\], whose RDF Schema constraints is the following subset of the DL-Lite\(\mathbb{Q}\) TBox ones: \(A_1 \subseteq A_2, R_1 \subseteq R_2, \exists R \subseteq A\) and \(\exists H \subseteq A\).

Finally, we present the pioneering CQ-to-UCQ reformulation technique for DL-Lite\(\mathbb{Q}\) \[10\], on which we rely in order to establish our results. These results extend to any other FOL reformulation techniques for DL-Lite\(\mathbb{Q}\), e.g., optimized CQ-to-UCQ or CQ-to-USCQ reformulation techniques, since they produce equivalent FOL queries.

The technique of \[10\] relies on two operations: the specialization of an atom into another through backward chaining on a TBox constraint (recall that such constraints are particular FOL rules, see Tables \[1\] and \[2\], and the specialization of two atoms into one through their most general unifier. The technique consists in exhaustively applying these two operations to the input CQ, each operation generating a CQ subsumed by the input CQ. The same is then recursively applied on the subsumed CQs thus obtained, until reaching a fixpoint for the set of generated CQs. The finite union of the input CQ and generated ones forms the UCQ reformulation of the input CQ w.r.t. the TBox. The next example illustrates this.

Example 3 (CQ-to-UCQ reformulation technique). Consider the query \(q(x) \leftarrow \text{PhDStudent}(x) \land \text{worksWith}(y,x)\) and KB \(K\) of the preceding examples. The UCQ reformulation of \(q\) is

\[
q^{\text{UCQ}}(x) \leftarrow q^1(x) \leftarrow \text{PhDStudent}(x) \land \text{worksWith}(y,x) \\
\lor q^2(x) \leftarrow \text{PhDStudent}(x) \land \text{worksWith}(x,y) \\
\lor q^3(x) \leftarrow \text{PhDStudent}(x) \land \text{supervisedBy}(y,x) \\
\lor q^4(x) \leftarrow \text{PhDStudent}(x) \land \text{supervisedBy}(x,y) \\
\lor q^5(x) \leftarrow \text{supervisedBy}(x,y) \land \text{worksWith}(x,y) \\
\lor q^6(x) \leftarrow \text{supervisedBy}(x,y) \land \text{supervisedBy}(y,x) \\
\lor q^7(x) \leftarrow \text{supervisedBy}(x,y) \\
\lor q^8(x) \leftarrow \text{supervisedBy}(x,y) \\
\lor q^9(x) \leftarrow \text{PhDStudent}(x) \land \text{supervisedBy}(x,y)
\]

where the first disjunct is \(q\)’s body, the second is obtained from the first through a backward chaining application of \(\mathcal{T}_A\) on the atom \(\text{worksWith}(y,x)\), the third (resp. forth) is obtained from the first (resp. second) through a backward chaining application of \(\mathcal{T}_B\) on the atom \(\text{worksWith}(y,x)\) (resp. on \(\text{worksWith}(x,y)\)). The next four disjuncts, \(q^5\) to \(q^8\), are obtained from the above four through a backward chaining application of \(\mathcal{T}_B\) on the atom \(\text{PhDStudent}(x)\). Finally, the ninth disjunct is obtained from the seventh through the most general unifier of \(\text{supervisedBy}(x,y)\) and \(\text{supervisedBy}(y,x)\), while the tenth is obtained similarly from the eighth.

It is worth noting that the (naive) exhaustive application of specialization steps leads, in general, to highly redundant reformulations w.r.t. the containment of their disjuncts. For instance, minimizing \(q^{\text{UCQ}}\) by eliminating disjuncts contained in another leads to

\[
q_{\text{min}}^{\text{UCQ}}(x) \leftarrow q^1(x) \land \text{PhDStudent}(x) \land \text{worksWith}(y,x) \\
\lor q^2(x) \land \text{PhDStudent}(x) \land \text{worksWith}(x,y) \\
\lor q^3(x) \land \text{PhDStudent}(x) \land \text{supervisedBy}(y,x) \\
\lor q^4(x) \land \text{PhDStudent}(x) \land \text{supervisedBy}(x,y)
\]

in which all the disjuncts removed from \(q^{\text{UCQ}}\) are contained in \(q^1\).

Still, while minimal UCQ reformulations can be obviously processed more efficiently than non-minimal ones, query engines overall perform poorly at evaluating them. This is mainly due to the repeated data accesses that are hardly optimized across union operators, e.g., above for \(\text{PhDStudent} \land \text{worksWith} \lor \text{supervisedBy}\). CQs within a UCQ are basically processed independently one from another.

3. OPTIMIZATION FRAMEWORK

In this section, we recall the notion of FOL reformulation which underpins FOL reducibility of query answering (Section \[11\]). Then, for formalisms enjoying FOL reducibility of query answering, e.g., some DLs or some existential rules and Datalog \(\pm\) fragments, we lay the principles of query answering based on query covers, which define a space of alternative equivalent FOL reformulations for a given query (Section \[3.2\]), wherein we can search for a reformulation leading to the minimal evaluation cost in an RDBMS (Section \[3.3\]).

3.1 FOL reducibility of query answering

FOL reducibility of query answering relies on FOL reformulations of queries.

Definition 1 (FOL reformulation). A FOL reformulation \(q_{\text{FOL}}\) of \(q\) w.r.t. \(\mathcal{T}\) is a FOL query such that \(\text{ans}(q,(\mathcal{T},\mathcal{A})) = \text{ans}(q_{\text{FOL}},(\emptyset,\mathcal{A}))\) for any \(\mathcal{T}\)-consistent ABox \(\mathcal{A}\).

Observe that FOL reformulation reduces query answering against a KB to FOL query answering against a database storing the KB’s ABox. The latter can hence be delegated to the query evaluation engine of an RDBMS.

3.2 Cover-based query answering

RDBMS query optimizers consider a set of evaluation alternatives (a.k.a. logical and physical plans), and select the one minimizing a cost estimation function. Since the number of alternatives is in \(O(2^n \times n!)\) for a CQ of \(n\) atoms \[22\], modern optimizers rely on heuristics to explore only a few alternatives; this works (very) well for small-to-moderate size CQs. However, FOL reformulations go beyond CQs in general, and may be extremely large (see Section \[6\]), leading the RDBMS to perform poorly.

To work around this limitation, we introduce the cover-based query answering technique to define a space of equivalent FOL reformulations of a CQ. A cover defines how the query is split into subqueries, that may overlap, called fragment queries, such that substituting each subquery with its FOL reformulation (obtained from any state-of-the-art technique) and joining the corresponding (reformulated) subqueries, may yield a FOL reformulation for the query to answer. Indeed, as we shall see soon, not every cover of a query leads to a FOL reformulation. However, every cover which does, yields an alternative cover-based FOL reformulation of the original query. Crucially for our problem, a smart cover choice may lead to a cover-based reformulation whose evaluation is more efficient. Thus, the cover-based technique amounts to circumventing the difficulty of modern RDBMSs to efficiently evaluate FOL reformulations in general.
Definition 2 (CQ cover). A cover of a query $q$, whose atoms are \{a_1, \ldots, a_n\}, is a set $C = \{f_1, \ldots, f_m\}$ of non-empty subsets of atoms of $q$, called fragments, such that (i) $\bigcup_{i=1}^{m} f_i = \{a_1, \ldots, a_n\}$ and (ii) no fragment is included into another.

Definition 3 (Fragment queries of a CQ). Let $C = \{f_1, \ldots, f_m\}$ be a cover of $q$. A fragment query $q_{f_i}^{\text{frag}}$ of $q$ w.r.t. $C$ is the subquery whose body consists of the atoms in $f_i$ and whose free variables are the free variables $\bar{x}$ of $q$ appearing in the atoms of $f_i$, plus the existential variables in $f_i$ that are shared with another fragment $f_j, 1 \leq j \leq m, j \neq i$, i.e., on which the two fragments join.

Definition 4 (Cover-based FOL reformulation). Let $C = \{f_1, \ldots, f_m\}$ be a cover of $q$, and $q^{\text{FOL}}_f(x) \leftarrow \bigwedge_{i=1}^{m} q_{f_i}^{\text{frag}}$, a FOL query, where $q_{f_i}^{\text{frag}}$, for $1 \leq i \leq m$, is a FOL reformulation w.r.t. $C$ of the fragment query $q_{f_i}$ of $q$. $q^{\text{FOL}}_f$ is a cover-based FOL reformulation of $q$ w.r.t. $C$ and $q$ if it is a FOL reformulation of $q$ w.r.t. $T$.

3.3 Cover-based query optimization

We assume given a query cost estimation function $\epsilon$ which, for any FOL query $q$, returns the cost of evaluating it through an RDBMS storing the ABox. Thus, $\epsilon$ reflects the operations (selections, projections, joins, unions, scans etc) applied on the ABox to compute the result of a $q^{\text{FOL}}$ reformulation. The cost estimation $\epsilon$ also accounts for the effort needed to join the reformulated fragment query results. Note that any RDBMS storing the ABox includes such a cost estimation function, accessible for instance using the SQL explain directive. While the RDBMS optimizer explores several reformulation techniques, since they produce possibly syntactically different FOL queries.

Problem 1 (Optimization problem). Given a CQ $q$ and a KB $K$, the cost-driven cover-based query answering problem consists of finding a cover-based reformulation of $q$ based on $K$ with lowest (estimated) evaluation cost.

4. COVER-BASED QUERY ANSWERING IN DL-Lite$^R$

We now apply our general optimization framework to the particular case of the DL-Lite$^R$ description logic. As stated in Definition 2, a query cover may not lead to a cover-based query reformulation. This is true even in the lightweight DL-Lite$^R$ setting, as we shall see in the running example below, while it is not in the simpler RDF setting of [8]. We therefore provide a condition ensuring that a cover is safe for query answering, i.e., it leads to a FOL query reformulation. This results in a cover-based technique for computing a set of JUCQ or of JUSQ reformulations of a CQ, when querying DL-Lite$^R$ KBs. This extends [8] from RDF databases to DL-Lite$^R$ and builds on the aforementioned state-of-the-art on UCQ and USQ reformulations.

As already mentioned in Section 2, we use the simple CQ-to-UCQ reformulation technique of [10] for establishing our results, as well as for the running example for this section and the next one. Our results trivially extend to any other FOL reformulation techniques for DL-Lite$^R$, e.g., optimized CQ-to-UCQ or CQ-to-USQ reformulation techniques, since they produce equivalent (though possibly syntactically different) FOL queries.

Example 4 (Running example). Consider the KB $K = \langle T = \{b \sqsubseteq \exists R, R \sqsubseteq \Box\}, A = \{a(b), A(a)\}\rangle$ and the query $q(x) \leftarrow A(x) \wedge R(x, y) \wedge R'(z, y)$, whose only answer is $a$. The UCQ reformulation of $q$ is:

\[
q^{\text{UCQ}}_f(x) \leftarrow \begin{align*}
qu_1(x) & \leftarrow (A(x) \wedge R(x, y) \wedge R'(z, y)) \\
qu_2(x) & \leftarrow (A(x) \wedge R'(x, y)) \\
qu_3(x) & \leftarrow (A(x) \wedge B(x)) \\
qu_4(x) & \leftarrow (A(x) \wedge R'(z, y)) \\
qu_5(x) & \leftarrow (A(x) \wedge B(z)) \\
qu_6(x) & \leftarrow (A(x) \wedge B(z))
\end{align*}
\]

where the first disjunct is $q$’s body, the second is obtained from the first by specializing the atom $B(z, y)$ through a backward-chaining application of $R' \sqsubseteq \Box \sqsubseteq \exists \bowtie \setminus T$, the third is obtained from the second through the most general unifier of $R'(x, y)$ and $R'(z, y)$, and the fourth results from the third by specializing $R'(x, y)$ through the backward-chaining application of $B \sqsubseteq \exists \bowtie \setminus T$.

Following Definition 2, the FOL query obtained from $q$, its cover $C_1 = \{\{A(x), R'(x, y)\}, \{R'(z, y)\}\}$, $T$ and the CQ-to-UCQ reformulation technique is the JUCQ:

\[
q^{\text{UCQ}}_f(x) \leftarrow \begin{align*}
q_1^{\text{UCQ}}(x, y) & \leftarrow (A(x) \wedge R'(x, y)) \\
q_2^{\text{UCQ}}(y) & \leftarrow (A(x) \wedge R'(y, z) \wedge R'(z, y)) \\
q_3^{\text{UCQ}}(x) & \leftarrow (A(x) \wedge R'(x, y)) \\
q_4^{\text{UCQ}}(y) & \leftarrow (A(x) \wedge B(y))
\end{align*}
\]

where $q^{\text{UCQ}}_f(x)$ is not a reformulation of $q$ w.r.t. $T$: $\text{ans}(q^{\text{UCQ}}_f, (\emptyset, A)) = \emptyset \neq \{a\}$

The example above shows that the chosen cover $C_1$ prevents the unification of $R'(z, y)$ and $R'(y, z)$ performed by the plain CQ-to-UCQ reformulation, as these two atoms are split across two subqueries that are reformulated separately, while this unification is required for the correctness (actually the completeness) of the technique: it is the only way to trigger the use of the constraint $B \sqsubseteq \exists \bowtie \setminus T$.

More generally, given an input CQ and a TBox, each unification necessary for the correctness of the CQ-to-UCQ technique defines a set of input CQ atoms that cannot be separated within the cover: those which beget this unification.

We therefore provide a sufficient condition for a cover to be safe for query answering, i.e., to lead to a cover-based FOL reformulation. The main idea for this condition is to have a cautious approximation of the query atoms which are interdependent w.r.t. reformulation, i.e., which (directly or after specialization) unify through the CQ-to-UCQ reformulation technique, and keep them in the same cover fragment.

Only atoms with the same predicate may unify. Thus, we identify for each query atom predicate (i.e., concept or role name), the set of all TBox predicates in which the atom may turn through some sequence of atom specializations, i.e., backward chaining and/or unifications, the two operations applied by the technique of [10]. We say the query atom predicate depends on such TBox predicates, and we capture them in a conservative fashion, based on the syntax of the TBox constraints, in the following recursive definition.

Definition 5 (Concept and role dependencies w.r.t. a TBox). Given a TBox $T$, a concept or role name $N$ depends w.r.t. $T$ the set of concept and role names denoted $\text{dep}(N)$ and defined as the fixpoint of:

\[
\begin{align*}
\text{dep}^0(N) &= \{N\} \\
\text{dep}^{n+1}(N) &= \text{dep}^n(N) \cup \{cr(Y) \mid Y \subseteq X \in T \text{ and } cr(X) \in \text{dep}^n(N)\}
\end{align*}
\]

where the cr function returns the concept or role name involved in any DL-Lite$^R$ concept or role defined as input.

Example 5 (Dependencies (cont.)).

\[
\begin{align*}
\text{dep}(A) &= \{A\} \\
\text{dep}(B) &= \{B\} \\
\text{dep}(R) &= \{R, R', B\} \\
\text{dep}(R') &= \{R', B\}
\end{align*}
\]
Definition 6 (Safe cover for query answering). A cover \( C \) of \( q \) is safe for query answering w.r.t. \( T \) (or safe in short) if it is a partition of \( q \)'s atoms such that two atoms whose predicates depend on a common concept or role name w.r.t. \( T \) are in the same fragment.

Note that while Definition 6 requires covers to be partitions, we will relax this restriction in Section 5.2.

Example 6 (Covers for JUCQ reformulation (cont.)). The FOL query obtained with the safe cover \( C_2 = \{ \{ A(x) \}, \{ R(x, y), R'(z, y) \} \} \) is the JUCQ reformulation:

\[
q_{\text{JUCQ}}(x) \leftarrow \begin{align*}
q_1(x) & \leftarrow A(x) \\
q_2(x) & \leftarrow (R(x, y) \land R'(z, y)) \\
& \quad \lor (R'(x, y) \land R'(z, y)) \\
& \quad \lor (R'(x, y)) \lor (B(x))
\end{align*}
\]

Observe that \( \text{ans}(q_{\text{JUCQ}}, (\emptyset, A)) = \{ a \} = \text{ans}(q, (T, A)) \).

Theorem 1 (Cover-based query answering). Applying Definition 4 on a safe cover \( C \) of \( q \) w.r.t. \( T \), and any CQ-to-JUCQ or CQ-to-USCQ reformulation technique, yields a cover-based JUCQ (resp. JUSCQ) reformulation \( q_{\text{PB}} \) of \( q \) w.r.t. \( T \).

Proof. The proof follows that of correctness of the CQ-to-USCQ reformulation technique in [10] for query answering. It directly extends to the use of any CQ-to-JUCQ or CQ-to-USCQ reformulation technique for DL-Lite\( \_2 \) as, for any CQ and TBox, the FOL queries they compute are equivalent to that of [10].

Soundness: for any \( T \)-consistent Abox \( A \), \( \text{ans}(q_{\text{PB}}, (\emptyset, A)) \subseteq \text{ans}(q, (T, A)) \).

Let \( t \) be a tuple in \( \text{ans}(q_{\text{PB}}, (\emptyset, A)) \). From Definition 4, \( q_{\text{PB}}(x) \leftarrow \bigwedge_{i=1}^{m} q_{\text{PB}}^i(x) \), thus \( t \) results from \( t_i \in \text{ans}(q_{\text{PB}}^i, (\emptyset, A)) \), for \( 1 \leq i \leq m \). Therefore, for \( 1 \leq i \leq m \), \( t_i \in \text{ans}(q_{\text{PB}}^i, (T, A)) \) holds, because of the soundness of the CQ-to-USCQ reformulation technique. Hence, from Definition 4, \( t \in \text{ans}(q, (T, A)) \) holds.

Completeness: for any \( T \)-consistent Abox \( A \), \( \text{ans}(q, (T, A)) \subseteq \text{ans}(q_{\text{PB}}, (\emptyset, A)) \).

Let \( t \) be a tuple in \( \text{ans}(q, (T, A)) \). Let \( q_{\text{PB}} \) be its reformulation using the CQ-to-USCQ technique. From the completeness of this technique, \( t \in \text{ans}(q_{\text{PB}}, (\emptyset, A)) \) holds. Let \( q_{\text{PB}}^i \) be \( \bigvee_{j=1}^{m_i} q_{\text{PB}}^j \), then necessarily for some \( l : t \in \text{ans}(q_{\text{PB}}^l, (\emptyset, A)) \) holds [10].

Let \( q_{\text{PB}}^j \) be \( \bigwedge_{k=1}^{n_j} q_{\text{PB}}^k \), \( q_{\text{PB}}^i \). Since Definition 6 makes the reformulation of each fragment independent from another w.r.t. the Q-to-USCQ reformulation technique, for any \( c_{\text{PB}} \) in \( q_{\text{PB}} \) holds. Hence, \( t \in \text{ans}(q_{\text{PB}}, (\emptyset, A)) \) holds.

Remark. That the trivial one-fragment cover (comprising all query atoms) is always safe; in this case, our Q-to-JUCQ (resp. CQ-to-JUSCQ) reformulation techniques reduce to the CQ-to-USCQ (resp. CQ-to-USCQ) one.

5. COVER-BASED QUERY OPTIMIZATION IN DL-Lite\( \_2 \)

We study now the query answering optimization problem of Section 3.3 for DL-Lite\( \_2 \). We analyze a first optimization space in Section 5.1 before extending our discussion to a larger space in Section 5.2. Finally, we provide search algorithms in Section 5.3.

5.1 Safe cover optimization space

We show next that the set of all safe covers for a query \( q \) form a lattice \( L_q \), whose precedence relationship is denoted \( \prec \), where \( C_1 \prec C_2 \) if each fragment of \( C_2 \) is a union of some fragments of \( C_1 \). The lattice has as lower bound (minimum) the single-fragment cover, and as upper bound (supremum) a specific cover \( C_{\text{root}} \), termed the root cover, and defined below:

Definition 7 (Root cover). We term root cover for a query \( q \) and TBox \( T \) the cover \( C_{\text{root}} \) obtained as follows. Start with a cover \( C_1 \) where each atom is in a fragment. Then, for any pair of fragments \( f_1, f_2 \in C_1 \) and atoms \( a_1 \in f_1, a_2 \in f_2 \) such that there exists a predicate on which those of \( a_1 \) and \( a_2 \) depend w.r.t. \( T \), create a fragment \( f' \equiv f_1 \cup f_2 \) and a new cover \( C_2 = (C_1 \setminus \{ f_1, f_2 \}) \cup \{ f' \} \). Repeat the above until the cover is stationary; this is the root cover, denoted \( C_{\text{root}} \).

It is easy to see that \( C_{\text{root}} \) is safe, and does not depend on the order in which the fragments are considered.

Example 7 (Root cover (cont.)). Recall the query and TBox from Example 7. The root cover \( C_{\text{root}} \) in this case is the cover \( C_2 \) from Example 7.

Proposition 1 states that \( C_{\text{root}} \) has the maximal number of smallest fragments. Its proof is based on the following lemma:

Lemma 1 (\( C_{\text{root}} \) fragment structure). A fragment \( f \) in the root cover \( C_{\text{root}} \) is of one of the following two forms:

1. a singleton, i.e., \( f = \{ a \} \) for some query atom \( a \);
2. \( f = \{ a_1, \ldots, a_n \} \), for \( n \geq 2 \), and for every atom \( a_i \in f \), there exists one atom \( a_i \in f \), and a predicate \( b_i \) in the TBox such that both predicates of \( a_i \) and \( a_j \) depend on \( b_j \).

Proof. The lemma follows directly from the definition of \( C_{\text{root}} \). Those atoms that do not share a dependency with any other atom appear in singleton fragments (case 1 above, as they never get fused). Atoms which share some dependencies (i.e., atoms whose predicates depend on one another) get fused in fragments of the form 2 above.

Figure 2: Sample \( C_{\text{root}} \) cover.
what we call an extended path $e$, going from $a_i$ to $a_j$ following the dependency edges either from source to target, or in the opposite direction; in other words, $e$ alternately moves “up (or down) then down (or up)” a certain number of times in the fragment.

If $e$ only contains edges in the same direction (either all are $\rightarrow$ or all are $\leftarrow$), it follows immediately that $a_i$ and $a_j$ are in the same fragment of $C$.

In the contrary case, there must exist some predicates in the $\text{TBox} b_1, \ldots, b_m, m \geq 1$, and some $f$ atoms $a_1, \ldots, a_{m-1}$ defining an extended path $e$ from $a_i$ to $a_j$ in $f$, as follows:

1. $a_i \to \cdots \to b_1$ (a $\to$ path segment), or $b_1$ is the predicate used in $a_i$;
2. $b_k (1 \leq k < m - 1)$, is the predicate used in $a_i (k \leq l < m - 1)$, or $b_k \leftarrow \cdots \leftarrow a_l$ (a $\leftarrow$ path segment);
3. $b_k \leftarrow \cdots \leftarrow a_l$ (a $\leftarrow$ path segment), with $1 \leq k < m - 1$ and $k \leq l < m - 1$, and $a_i \to \cdots \to b_{k+1}$;
4. $b_{m-1}$ is the predicate used in $a_{m-1}$ or $b_{m-1}$ $\leftarrow \cdots \leftarrow a_{m-1}$, then $a_{m-1} \to \cdots \to b_m$;
5. $b_m$ is the predicate used in $a_j$ or $b_m \leftarrow \cdots \leftarrow a_j$.

Observe that items (2) and (3) can repeat (alternately) until $b_{m-1}$ is reached.

Since $C$ is safe, $a_i$ and $a_j$ must appear in the same fragment in $C$ (and only in that fragment), because they both depend on $b_1$.

For $1 \leq i \leq m - 2$, $a_i$ must appear in the same fragment as $a_{i+1}$ (and only there), given that they both depend on $b_i$.

Since $C$ is safe, $a_i$ and $a_{m-1}$ must appear in the same fragment of $C$ (and only there).

From the above follows that $\{a_1, a_2, \ldots, a_{m-1}, a_j\}$ are all in the same fragment of $C$, which contradicts our hypothesis.

Finally, from the above result, we derive easily that the space of safe covers of a query $q$ is a lattice, denoted $\mathcal{L}_q$, with $C_{\text{root}}$ as supremum and the single-fragment cover as infimum:

**Theorem 2 (Safe cover space).** Let $C$ be a safe cover and $f$ one of its fragment. Then, $f$ is the union of some fragments from $C_{\text{root}}$.

**Proof.** Suppose that $f$ is not a union of some fragments from $C_{\text{root}}$, and let us show a contradiction. In this case, $f$ necessarily contains a strict, non-empty subset of a fragment of $C_{\text{root}}$. It follows that there are two atoms whose predicates depend on a common concept or role name w.r.t. $\mathcal{R}$ (as they were together in the fragment of $C_{\text{root}}$) that are not in a same fragment of $C$. Therefore $C$ is not a safe cover, a contradiction.

5.2 Optimization Beyond Safe Covers

A dependency-rich TBox leads to few, large fragments in $C_{\text{root}}$, and thus to a relatively small lattice. In this section, we identify a set of covers which are not safe (in particular, they are not partitions), yet from which we still derive FOL reformulations of the query; this enlarges our space of alternatives and thus gives more optimization opportunities.

An extended fragment $f\parallel g$ of $q$ is a pair of atom sets such that $f = \{a_1, \ldots, a_q\}$ is a set of atoms from $q$, and $g \subseteq f$. When $f = g$, extended fragments coincide with simple fragments, thus we generalize query covers (Definition 1) to consider that they only consist of extended fragments, while preserving that no fragment must be contained in another.

To an extended fragment we associate:

**Definition 8 (Extended fragment query of a CQ).** Let $f\parallel g \subseteq C$ be an extended fragment of $q$. The extended fragment query $q_{f\parallel g}$ of $q$ w.r.t. $C$ is the subquery whose body consists of the atoms in $f$, and whose free variables are the free variables of $g$ appearing in the atoms of $g$, plus the variables appearing in an atom of $g$ that are shared with some atom in $q$, for some fragment $f'\parallel g'[q]$ of $C$.

In an extended fragment query, atoms from $f \setminus g$ only reduce (filter) the answers, without adding variables to the head.

Given a cover of extended fragments, the extended cover-based reformulation of a query $q$ is the FOL query:

$$q^e(\bar{x}) \leftarrow \bigwedge_{i=1}^{m} q_{f_i\parallel b_i}^\text{FOL}$$

**Example 8 (Extended cover-based reformulation).** Recall the query $q$ and KB from Example 4. Let $f_0 = \{A(x)\}$ and $f_1 = \{R(x, y), R'(z, y)\}$ be the two fragments of the root cover $C$. Consider also the cover $C_3 = \{f_1\parallel f_2\parallel f_3\}$, with $f_2 = \{A(x), R(x, y)\}$.

The extended fragment $f_1\parallel f_1$ coincide with the simple fragment $f_1$, therefore the extended fragment query $q_{f_1\parallel f_1}$ of $q$ w.r.t. $C_3$ is the subquery $q_{f_1\parallel f_1}(x) \leftarrow R(x, y) \land R'(z, y)$. Observe that $y$ is not a free variable of $q_{f_1\parallel f_1}$, as it is not a free variable of $q$, and moreover, it does it appear in $f_0$, whereas $f_2\parallel f_3$ is the only other fragment in the cover $C_3$.

The extended fragment query $q_{f_2\parallel f_3}$ of $q$ w.r.t. $C_3$ is the subquery $q_{f_2\parallel f_3}(x) \leftarrow A(x) \land R(x, y)$. Again, note that $y$ is not a free variable of $f_0$, and therefore its not a free variable of $q_{f_2\parallel f_3}$.

Then, the extended cover-based reformulation corresponding to $C_3$ is the FOL query:

$$q^e(x) \leftarrow q_{f_1\parallel f_1}^\text{FOL}(x) \land q_{f_2\parallel f_3}^\text{FOL}(x)$$

where:

$$q_{f_1\parallel f_1}^\text{FOL}(x) \leftarrow (\langle R(x, y) \land R'(z, y) \rangle \lor R'(x, y) \lor B(x))$$

$$q_{f_2\parallel f_3}^\text{FOL}(x) \leftarrow (\langle A(x) \land R(x, y) \rangle \lor (A(x) \land R'(x, y)) \lor (A(x) \land B(x)))$$

Observe that applying the TBox constraint $R' \subseteq R$ to $q_{f_1\parallel f_1}(x)$ leads to:

$$\langle (R(x, y) \land R'(z, y)) \lor (R'(x, y) \land R'(z, y)) \rangle$$

Then, applying the TBox constraint $B \subseteq R$ we obtain the reformulation of $q_{f_1\parallel f_1}$ w.r.t. TBox $T$, i.e., $q_{f_1\parallel f_1}^\text{FOL}(x)$. Similarly, applying to $q_{f_2\parallel f_3}$ the constraint $R' \subseteq R$ and subsequently $B \subseteq R$ leads to $q_{f_2\parallel f_3}^\text{FOL}(x)$. Note that $\text{ans}(q^e, (\emptyset, A)) = \{a\} = \text{ans}(q, (T, A))$.

The introduction of extra atoms in extended fragments is reminiscent of the classical semijoin reducer optimization technique, which we briefly recall. Consider a CQ $q(x) \leftarrow R(x, y) \land S(y, z)$, to which corresponds the algebra expression $R(x, y) \bowtie_{y} S(y, z)$. The semijoin reducer technique consists, instead, of computing the result of $q$ by the equivalent expression:

$$(R(x, y) \bowtie_{y} S(y, z)) \bowtie_{y} S(y, z)$$

where $\pi_y(S(y, z))$ is the first projection of $S$, and $\pi_y$ denotes the semijoin relational operator, returning every tuple from the left-hand side input that joins with the right-hand input. The semijoin filters (“reduces”) the $R$ relation to only those tuples having a match in $S$. If there are few distinct values of $y$ in $S$, $\pi_y(S(y, z))$ is small, thus it can be loaded in memory and the $\bowtie_{y}$ operator can be evaluated very efficiently. Further, if only few $R$ tuples survive
the κq, the cost of the ∨q operator is reduced, given that there are less tuple comparisons to be made.

While the benefits of semijoins are well-known, there are many ways to introduce them in a given query, further increasing the space of alternative plans to be considered by an optimizer. While some heuristics have been proposed to explore only some carefully chosen semijoin plans [27], we noted that RDBMS optimizers do not explore semijoin options, in particular for the very large queries resulting from the FOL reformulations of CQs. *Extended fragments mitigate this problem by intelligently using semijoin reducers to fasten the evaluation of the FOL reformulation by the RDBMS.*

We now define an extended search space $E_q$ of covers for a given query $q$, based on the safe cover lattice $C_q$.

1. First, any $C_q$ cover is also in $E_q$.
2. Second, from any cover $C \in E_q$ and fragment $f \mid g \in C$, a cover $C'$ defined as follows is in $E_q$:
   (a) pick a query atom $a \notin f$, which shares a variable with an atom in $f$;
   (b) let the new extended fragment $f' \mid g \in f \cup \{a\} \mid g$; let $C'$ be $(C \setminus \{f\mid g\}) \cup \{f'\mid g\}$.

Since the newly added atom was already in a fragment of $C_{root}$, $C'$ is not a partition.

In the above, $E_q$ covers are “grown” starting from the $E_q$ ones. To any fragment $f$ of a cover $C \in E_q$, one can add an atom as described above, leading to an extended fragment $f'\mid f$: replacing $f$ with $f'$ in $C$ leads to $C'$. Note that any subsequent addition of an atom to $f' \in C'$ is not allowed to change the head of the extended fragment query thus obtained, i.e., the latter will be of the form $f'\mid f'$ etc.

The core result allowing us to benefit of the performance savings of extended covers in order to efficiently answer queries is:

**Theorem 3** ($E_q$ cover-based query answering). Applying Definition 2 together with Definition 3 on a CQ $q$, a TBox $T$, a cover $C \in E_q$, and any CQ-to-UCQ (resp. CQ-to-UCQ) reformulation technique, yields a JUCQ (resp. JUSCQ) reformulation of $q$ w.r.t. $T$.

**Proof.** The proof follows from that of Theorem 1. It relies on the fact that, given a safe cover $C = \{g_1, \ldots, g_m\}$ of $q$ and an extended cover $C' = \{f_1, \ldots, f_m \mid g\}$ of $q$, the queries $q(x) \leftarrow \bigwedge_{i=1}^m q_i(x)$ and $q'(x) \leftarrow \bigwedge_{i=1}^m q_i'(x)$ are equivalent, though each $q_i(x)$ subsumes $q_i'(x)$. Indeed, $q'$ is obtained from $q$ by duplicating atoms already present in $q$, thus $q'$ only adds redundancy w.r.t. $q$, hence remains equivalent to it.

### 5.3 Cost-based cover search algorithms

Our first algorithm, EC-DL (Exhaustive Covers), starts from $C_{root}$ and builds all $E_q$ covers by fusing fragments, and all $E_q$ covers by adding atoms, as explained previously. The second one, GC-DL (Greedy Covers), explores $E_q$ partially, in greedy fashion. It uses an explored cover set initialized with $\{C_{root}\}$. GC-DL picks from this set a cover $C$ inducing a $q^{\text{deg}}$ reformulation with minimum evaluation cost, and attempts to build from it a cover $C'$, through fragment fusion or atom addition. GC-DL only adds atoms to the explored set if $\epsilon(C') < \epsilon(C)$, thus it only explores a small part of the search space. Both algorithms return a cover-based reformulation with the minimum estimated evaluation cost w.r.t. the explored space.

When fusing two fragments, $\epsilon$ decreases if the resulting fragment is more selective than the two fragments it replaces. Therefore, the RDBMS may find a more efficient way to evaluate the query of this fragment, and/or its result may be smaller, making the evaluation of $q^{\text{deg}}$ based on the new cover $C'$ faster. When adding an atom to an extended fragment, $\epsilon$ decreases if the conditions are met for the semijoin reducer to be effective (Section 5.2). Such performance benefits can be very significant, as we show below.

### 6. EXPERIMENTAL EVALUATION

We implemented our reformulation-based query answering cover-based approach in Java 7, on top of PostgreSQL v9.3.2, running on an 8-core Intel Xeon (E5506) 2.13 GHz machine with 16GB RAM, using Mandriva Linux r2010.0. We used the LUBM_B0 DL-LiteR TBox and associated EUDG data generator [19]; LUBM_B0 consists of 34 roles, 128 concepts and 212 constraints; the generated ABox comprises 15 million facts.

We store the ABox in a binary table for each role and a unary one for each concept; as customary in efficient Semantic Web data management systems, e.g., [21], each identifier or literal occurring in a fact is encoded into an integer prior to storing them in the RDBMS. Also in keeping with the best practices in efficient Semantic Web stores [21], each concept table is indexed, and similarly each role table is indexed by each of its two attributes. The TBox and predicates dependencies are stored in memory. For our experiments, we relied on a CQ-to-UCQ reformulation tool, namely RAPID [11] (recall that any CQ-to-UCQ reformulation technique could have been used instead). For the cost estimation function $\epsilon$, we relied on Postgres’ own estimation, obtained using the `explain directive`.

We devised a set of 13 CQs (cf. Appendix) against this KB: the queries have between 2 and 10 atoms, with an average of 5.77 atoms. Their UCQ reformulations have 35 to 667 CQs, 290.2 on average. We depict below $Q_9$, one of the most complex, whose UCQ reformulation is a union of 368 CQs:

$q_9(n, e) \leftarrow \text{Student}(x) \land \text{TakesCourse}(x, c) \land \text{Advisor}(x, a) \land \text{memberOf}(x, \text{http://dep0.univ0.edu}) \land \text{Phone}(x, \text{"xxx"}) \land \text{MemberOf}(a, \text{http://dep0.univ0.edu}) \land \text{TeacherOf}(p, c) \land \text{Email}(p, e) \land \text{ResearchInterest}(a, \text{"res7"}) \land \text{Name}(c, n)$

where quoted strings stand for constants produced by the data generator (shortened for the presentation).

Figure 2 depicts the evaluation time through Postgres, for our 13 queries, of four FOL reformulations: (i) the UCQ produced by RAPID [11]; (ii) the JUCQ reformulation based on $C_{root}$; (iii) the JUCQ reformulation corresponding to the best-performing cover found by our algorithm EC-DL, and (iv) the JUCQ reformulation based on the best-performing cover found by GC-DL. First, the figure shows that fixed FOL reformulations are not efficiently evaluated, e.g., the UCQs for $Q_1$, $Q_2$, and $Q_3$, and the JUCQs based on $C_{root}$ for $Q_4$, $Q_9$, and $Q_{13}$. This poor performance correlates with the large size of the reformulations: they feature very large unions of CQs, very poorly handled by current RDBMS optimizers, which are designed and tuned for small CQs. This finding is confirmed in [8] for two other leading RDBMSs, thus it is not specific to the Postgres system used here. Second, the reformulation based on the cover returned by EC-DL is always more efficient than UCQ reformulation (more than one order of magnitude for $Q_9$), respectively, $C_{root}$-based JUCQ reformulation (up to a factor of 230 for $Q_9$). Third, in our experiments, the GC-DL-chosen cover leads to a JUCQ reformulation as efficient as the EC-DL one, demonstrating that even a partial, greedy cover search leads to good performance. (In general, though, the optimality of GC-DL is not guaranteed.)

---

1 See [http://www.postgresql.org/docs/9.1/static/sql-explain.html](http://www.postgresql.org/docs/9.1/static/sql-explain.html)
For $Q_3$, $C_{\text{root}}$ is $\{\{a_5\}, \{a_6\}, \{a_9\}, \{a_{10}\}, \{a_1, a_2, a_3, a_4, a_6, a_7\}\}$ (where the atoms are ordered as shown above), while the cover chosen by EC-DL and GC-DL fuses the largest fragment of $C_{\text{root}}$ with $\{a_{10}\}$. For $Q_7$ and $Q_9$-$Q_{13}$, the best cover we found is safe; for all the others, this is not the case, confirming the interest of extended fragments within covers.

Discussion: implementation alternatives. Recall that all the above experiments were made with a simple data layout. Adopting another layout (as in [6, 8]) is orthogonal to our technique, though obviously the above-mentioned times would change. Indeed, our cost-based approach will get the best performance of any layout.

Finally, the choice of the cost model used in our approach deserves a discussion. We have used Postgres’ own cost model; one could alternatively rely on an externally implemented cost estimation, as we did in our previous work [8]. There, we have shown that a relatively straightforward cost model estimator based on textbook material leads to estimations quite similar to those of Postgres, even though, clearly, Postgres’ internal optimization strategy is much more elaborate than the simple assumptions our cost model made. This may be due to the fact that the biggest impact of the cost model is to allow the optimizer to avoid very bad plans – this is far more important than to pick the best one in a set of similar-performance efficient plans. Our findings in [8] (where, furthermore, we calibrated our own cost model to suit three different RDBMSs) make us confident that the benefits of our cost-based approach can carry over when other RDBMSs are used, too.

7. RELATED WORK AND CONCLUSION

We proposed a novel framework for any OBDA setting enjoying FOL reducibility of query answering, for which we studied a space of alternative FOL reformulations to evaluate through an RDBMS. This space is crucial, as it allows choosing a FOL reformulation most efficiently evaluated by the RDBMS. We applied this framework to the DL-Lite$_R$ description logic underpinning the W3C’s OWL2 QL standard for the Semantic Web, and experimentally demonstrated its performance benefits.

Our approach departs from the literature focused on a single FOL query reformulation, where optimization mainly reduces to producing fast a UCQ reformulation as minimized as possible: [24][1][12][29][17][30][14][2][15] consider DL-Lite$_X$, existential rules and Datalog+; [28] studies CQ-to-USCQ reformulation for existential rules encompassing DL-Lite$_X$; USCQ reformulations are shown to perform overall better than UCQ ones in an RDBMS. We build on these works to devise CQ-to-JUCQ and CQ-to-JUSCQ reformulation techniques, further used for cost-based query answering optimization. Our semi-join-inspired technique goes against minimization, as it adds redundant atoms, but does so with the help of a cost model only when this improves performance. [26] proposes a CQ-to-Datalog reformulation technique; a produced non-recursive Datalog program corresponds to a JUCQ, which is not chosen according to a cost-model. Finally, [8] is an instance of our general framework for the strictly less expressive case of RDF databases, where any cover leads to a reformulation. In general, and in particular for DL-Lite$_R$ as we have shown, this is not the case, therefore the core contribution of the present work is an analysis of safe covers, guaranteed to lead to reformulations, and of a larger carefully chosen space of covers which do not satisfy the safety criterion, yet still lead to equivalent FOL reformulations of the incoming query.

We plan to extend our framework to efficient query answering using views, i.e., through a set of predefined CQs, which is typical of the database query optimization and information integration settings [16].

Acknowledgements This work has been partially funded by the PIA Datalyse project and the ANR PAGODA project.

8. REFERENCES


APPENDIX

\[ q_1(u, i, n, e, t) \leftarrow \text{Professor}(x) \land \text{degreeFrom}(x, u) \land \text{researchInterest}(x, t) \land \text{name}(x, n) \land \text{emailAddress}(x, e) \land \text{telephone}(x, t) \]

\[ q_2(e, t) \leftarrow \text{Professor}(x) \land \text{degreeFrom}(x, \text{"http://www.University870.edu")} \land \text{researchInterest}(x, \text{"Research21")} \land \text{name}(x, \text{"AssociateProfessor2")} \land \text{emailAddress}(x, e) \land \text{telephone}(x, t) \]

\[ q_3(x) \leftarrow \text{Professor}(x) \land \text{degreeFrom}(x, \text{"http://www.University870.edu")} \land \text{researchInterest}(x, \text{"Research21")} \land \text{name}(x, \text{"AssociateProfessor2")} \land \text{emailAddress}(x, \text{" AssociateProfessor2@Department1.University0.edu")} \land \text{telephone}(x, \text{"xxx – xxx – xxx")} \]

\[ q_4(x, y) \leftarrow \text{Professor}(x) \land \text{teacherOf}(x, y) \land \text{worksFor}(x, z) \land \text{degreeFrom}(x, \text{"http://www.University870.edu")} \land \text{researchInterest}(x, \text{"Research21")} \land \text{name}(x, \text{"AssociateProfessor2")} \land \text{emailAddress}(x, \text{"AssociateProfessor2@Department1.University0.edu")} \land \text{telephone}(x, \text{"xxx – xxx – xxx")} \]

\[ q_5(y, z) \leftarrow \text{Professor}(x) \land \text{teacherOf}(x, y) \land \text{worksFor}(x, z) \land \text{degreeFrom}(x, \text{"http://www.University870.edu")} \land \text{researchInterest}(x, \text{"Research21")} \land \text{name}(x, \text{"AssociateProfessor2")} \land \text{emailAddress}(x, \text{"AssociateProfessor2@Department1.University0.edu")} \land \text{telephone}(x, \text{"xxx – xxx – xxx")} \]

\[ q_6(x, n) \leftarrow \text{Faculty}(x) \land \text{publicationAuthor}(y, x) \land \text{researchInterest}(x, \text{"Research16")} \land \text{name}(y, n) \land \text{emailAddress}(x, \text{"AssociateProfessor0@Department0.University0.edu")} \land \text{telephone}(x, \text{"xxx – xxx – xxx")} \]

\[ q_7(n) \leftarrow \text{Professor}(x) \land \text{teacherOf}(x, c) \land \text{memberOf}(x, \text{"http://www.Department0.University0.edu")} \land \text{name}(x, n) \land \text{emailAddress}(x, \text{"FullProfessor0@Department0.University0.edu")} \land \text{telephone}(x, \text{"xxx – xxx – xxx")} \]

\[ q_8(x, n) \leftarrow \text{Faculty}(x) \land \text{publicationAuthor}(y, x) \land \text{researchInterest}(x, \text{"Research16")} \land \text{name}(y, n) \land \text{telephone}(x, \text{"xxx – xxx – xxx")} \]

\[ q_9(n, e) \leftarrow \text{Student}(x) \land \text{takeCourse}(x, c) \land \text{advisor}(x, a) \land \text{memberOf}(x, \text{"http://www.Department0.University0.edu")} \land \text{telephone}(x, \text{"xxx – xxx – xxx")} \land \text{memberOf}(p, c) \land \text{emailAddress}(p, e) \land \text{researchInterest}(a, \text{"Research7")} \land \text{telephone}(x, \text{"xxx – xxx – xxx")} \land \text{memberOf}(a, \text{"http://www.Department0.University0.edu")} \land \text{name}(c, n) \]

\[ q_{10}(n, e) \leftarrow \text{Professor}(x) \land \text{takeCourse}(x, c) \land \text{advisor}(x, a) \land \text{memberOf}(x, \text{"http://www.Department0.University0.edu")} \land \text{telephone}(x, \text{"xxx – xxx – xxx")} \land \text{memberOf}(p, c) \land \text{emailAddress}(p, e) \land \text{researchInterest}(a, \text{"Research7")} \land \text{telephone}(x, \text{"xxx – xxx – xxx")} \land \text{memberOf}(a, \text{"http://www.Department0.University0.edu")} \land \text{name}(c, n) \]

\[ q_{11}(x) \leftarrow \text{Professor}(x) \land \text{Student}(x) \]

\[ q_{12}(x) \leftarrow \text{Professor}(x) \land \text{Department}(x) \]

\[ q_{13}(x) \leftarrow \text{Publication}(x) \land \text{Department}(x) \]

**Table 3:** Queries used in our experiments.