Finding Good Proofs for Answers to Conjunctive Queries Mediated by Lightweight Ontologies (Extended Abstract)*

Christian Alrabbaa[0000−0002−2925−1765], Stefan Borgwardt[0000−0003−0924−8478], Patrick Koopmann[0000−0001−5999−2583], and Alisa Kovtunova[0000−0001−9936−0943]

Institute of Theoretical Computer Science, Technische Universität Dresden, Germany
firstname.lastname@tu-dresden.de

1 Introduction

Description logics (DLs) are a family of knowledge representation formalisms that can be seen as decidable fragments of first-order logic [4]. In this paper we focus on two light-weight members of this family *DL-Lite*$_R$ [7] and $\mathcal{EL}$ [3], which underlies the profiles QL and EL of the standardized Web Ontology Language[1] OWL 2. Theories $\mathcal{T} \cup \mathcal{A}$ are called *ontologies* or *knowledge bases* and are composed of a TBox $\mathcal{T}$ and an ABox $\mathcal{A}$. Many DL ontologies can be equivalently expressed using the formalism of existential rules [6]. Existential rules are first-order sentences of the form $\forall \vec{y}, \vec{z}.\, \psi(\vec{y}, \vec{z}) \to \exists \vec{u}.\, \chi(\vec{z}, \vec{u})$, with the *body* $\psi(\vec{y}, \vec{z})$ and the *head* $\chi(\vec{z}, \vec{u})$ being conjunctions of atoms of the form $A(x)$ or $P(x_1, x_2)$, for a concept name A, role name P and terms x, x_1 and x_2, which are individual names or variables from $\vec{z}$, $\vec{u}$ and $\vec{y}$. We usually omit the universal quantification. In particular, $\mathcal{EL}$ and *DL-Lite*$_R$ can be equivalently expressed as sets of existential rules and we write all further DL expressions in the rule syntax.

Ontology-mediated query answering (OMQA) is a popular reasoning problem for DLs, it generalizes query answering over databases by allowing to query implicit knowledge that is inferred by an ontology [7]; this is called *open-world* reasoning. A *conjunctive query (CQ)* $\mathbf{q}(\vec{x})$ is an expression of the form $\exists \vec{y}.\, \phi(\vec{x}, \vec{y})$, where $\phi(\vec{x}, \vec{y})$ is a conjunction of atoms using *answer variables* $\vec{x}$ and *existentially quantified variables* $\vec{y}$. If $\vec{x} = ()$, then $\mathbf{q}(\vec{x})$ is called *Boolean*. As a running example, consider $\mathcal{A} = \{B(\mathsf{b})\}$,

$$\begin{aligned}\mathcal{T} = \{&R(x,y) \to \exists z.\, T(y,z),\ B(x) \to \exists y.\, P(x,y),\\ &P(x,y) \to \exists z.S(y,z),\ P(x,y) \to R(y,x)\}, \text{ and}\end{aligned}$$

$\mathbf{q}(y'') = \exists x', x'', y, z, z'.\, R(x'', y) \wedge T(y, z) \wedge S(x', z') \wedge S(x'', z') \wedge P(y'', x'')$. In the following, we investigate explanations for ontology-mediated CQ answers, i.e.

* This is an extended abstract of a paper accepted at the 35th International Workshop on Description Logics; see its technical report [2]

[1] https://www.w3.org/TR/owl2-profiles/

why the entailment $\mathcal{T} \cup \mathcal{A} \models \mathbf{q}(\mathsf{b})$ holds for a particular instantation (*answer*) b of the answer variable(s) y''. In particular, we are interested in *proofs* [1], i.e. step-wise derivations of the answer from the input. Our techniques are inspired by [5,10,8], but we additionally consider the problem of generating *good* proofs according to some quality measures and provide a range of complexity results focusing on *DL-Lite*$_R$ and $\mathcal{EL}$. We consider the *size* of a proof as well as its *tree size*, which corresponds to the size when the proof is presented in a tree-shaped way (possibly repeating subproofs). To compute proofs, we assume that there is a so-called *deriver*, i.e. some reasoning system or calculus, that defines a derivation structure for a given CQ entailment, and the structure may contain several proofs for that entailment. Given a measure, the quest for good proofs outlined in Section 3 is formalized as a search problem in a derivation structure. We consider two different kinds of derivers for generating proofs for CQ answers; their descriptions are in Section 2. In addition to classical OMQA, in Section 4 we also have a brief look at explaining inferences over temporal data using a query language incorporating metric temporal operators.

2 Deriving Query Answers

In the first deriver, $\mathfrak{D}_{\mathsf{cq}}$, inferences always produce Boolean CQs, which is similar to the approach used in [10,8]. It is defined by the inference schemas in Figure 1 and a resulting proof for our example is shown in Figure 2. **(MP)** describes how to apply an existential rule to a given CQ. It is admissible if there is a substitution π with $\pi(\psi(\vec{y},\vec{z})) \subseteq \phi(\vec{x})$, and $\rho(\vec{w})$ is the result of replacing *any subset of* $\pi(\psi(\vec{y},\vec{z}))$ in $\phi(\vec{x})$ by *any subset of* $\pi(\chi(\vec{z},\vec{u}'))$, where the variables $\vec{u}$ are renamed into fresh variables $\vec{u}'$. In **(C)**, we allow to merge two CQs and also rename $\vec{y}$ to fresh variables $\vec{y}'$. **(T)** allows us, together with **(MP)**, to introduce copies of variables in CQs. Finally, **(E)** transforms individual names into existentially quantified variables.

$$\frac{\exists\vec{x}.\,\phi(\vec{x}) \qquad \psi(\vec{y},\vec{z}) \to \exists\vec{u}.\,\chi(\vec{z},\vec{u})}{\exists\vec{w}.\rho(\vec{w})}\ \textbf{(MP)} \qquad \frac{\exists\vec{x}.\,\phi(\vec{x}) \qquad \exists\vec{y}.\,\psi(\vec{y})}{\exists\vec{x},\vec{u}.\phi(\vec{x}) \wedge \psi(\vec{u})}\ \textbf{(C)}$$

$$\frac{}{\phi(\vec{x},\vec{y}) \to \exists\vec{x}.\,\phi(\vec{x},\vec{y})}\ \textbf{(T)} \qquad \frac{\exists\vec{x}.\,\phi(\vec{x},\vec{a})}{\exists\vec{x},\vec{y}.\,\phi(\vec{x},\vec{y})}\ \textbf{(E)}$$

Fig. 1. Inference schemas for $\mathfrak{D}_{\mathsf{cq}}$.

We can show that these inferences are sound and complete for CQ entailment. However, they can be hard to follow due to the scope of existential quantification in a CQ, which forces atoms connected by the same variables to be carried along inferences they are not relevant for. In our example, x'' and z' in Figure 2 are

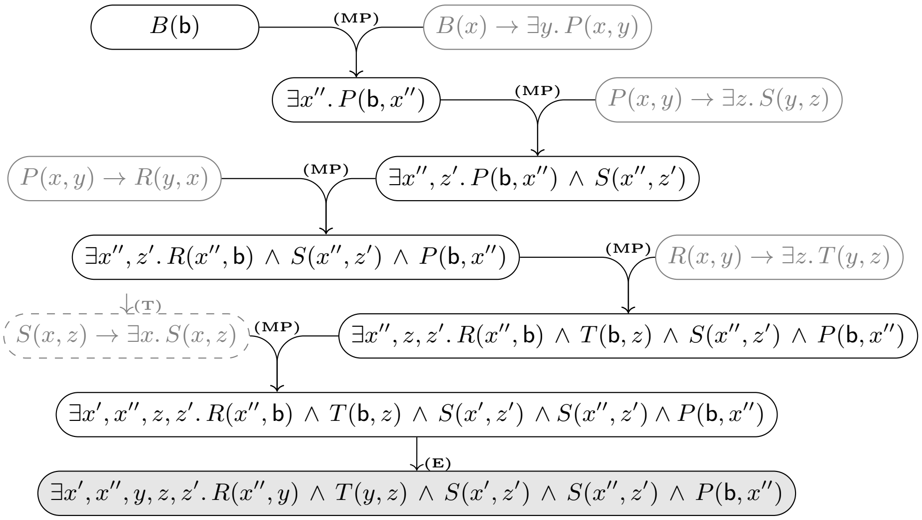

Fig. 2. A CQ proof for the example. Trivially, a "worse" proof can be one e.g. with the fourth **(MP)** replacing the CQ completely and **(C)** conjuncting $\exists z\, T(\mathsf{b}, z)$ back.

connected to each other and to the constant b, and thus have to be kept together: although $\exists x'', z'.P(\mathsf{b}, x'') \wedge S(x'', z')$ implies $\exists x''.P(\mathsf{b}, x'')$ and $\exists x'', z'.S(x'', z')$, those two CQs do not imply the original CQ anymore. To overcome this issue, we consider a second deriver, $\mathfrak{D}_{\mathsf{sk}}$, which is inspired by an approach from [5]. It mainly operates on ground CQs, and requires the ontology to be *Skolemized*; for example $P(x, y) \to \exists z.S(y, z)$ becomes $P(x, y) \to S(y, g(y))$, where g is a fresh unary function symbol. This means that we eliminate existential quantification at the cost of introducing function symbols. $\mathfrak{D}_{\mathsf{sk}}$ now operates on CQs in which ground function terms may appear. For example, instead of $\exists x'', z'.\, P(\mathsf{b}, x'') \wedge S(x'', z')$ in Figure 2 we now use $P(\mathsf{b}, f(\mathsf{b})) \wedge S(f(\mathsf{b}), g(f(\mathsf{b})))$. Since these atoms do not share variables, in our derivation structure we mainly need to consider inferences on single atoms, which allows for more fine-grained proofs. The inference schemas and an example proof can be seen in Figures 3 and 4, respectively.

$$\frac{\alpha_1(\vec{t}_1) \quad \dots \quad \alpha_n(\vec{t}_n) \quad \psi(\vec{y}, \vec{z}) \to \chi(\vec{z})}{\beta(\vec{s})}\ \textbf{(MP}_\mathbf{s}\textbf{)}$$

$$\frac{\alpha_1(\vec{t}_1) \quad \dots \quad \alpha_n(\vec{t}_n)}{\alpha_1(\vec{t}_1) \wedge \dots \wedge \alpha_n(\vec{t}_n)}\ \textbf{(C}_\mathbf{s}\textbf{)} \qquad \frac{\phi(\vec{t})}{\exists \vec{x}.\phi(\vec{x})}\ \textbf{(E}_\mathbf{s}\textbf{)}$$

Fig. 3. Inference schemas for $\mathfrak{D}_{\mathsf{sk}}$.

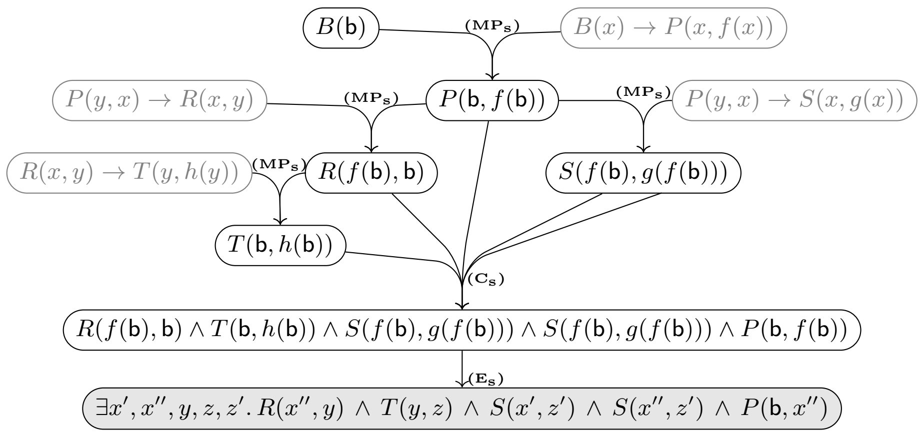

Fig. 4. A Skolemized proof for the example

In $\mathbf{(MP_s)}$, $\alpha_i(\vec{t_i})$ and $\beta(\vec{s})$ are ground atoms with terms composed from individual names and Skolem functions, and likewise $\chi(\vec{z})$ may contain Skolem functions; similar to **(MP)**, we require that there is a substitution π such that $\pi(\psi(\vec{y}, \vec{z})) = \{\alpha_1(\vec{t_1}), \ldots, \alpha_n(\vec{t_n})\}$ and $\beta(\vec{s}) \in \pi(\chi(\vec{z}))$. In $\mathbf{(E_s)}$, $\vec{t}$ is now a vector of ground terms that may contain function symbols. Moreover, we do not need a version of **(T)** here since it would be trivial for ground atoms. Its effects in $\mathfrak{D}_{\mathsf{cq}}$ can be simulated here due to the fact that the same atom can be used several times as a premise for $\mathbf{(MP_s)}$ or $\mathbf{(C_s)}$. Note that the most important inferences using $\mathbf{(MP_s)}$ operate only on ground atoms, and $\mathbf{(C_s)}$ and $\mathbf{(E_s)}$ are only needed at the end of a proof to obtain the desired CQ (see Figure 4).

3 Finding Good Proofs

We consider the following decision problem: given an entailment $\mathcal{T} \cup \mathcal{A} \models \mathbf{q}(\vec{a})$ and a natural number n is there a proof for $\mathbf{q}(\vec{a})$ in $\mathfrak{D}(\mathcal{T} \cup \mathcal{A}, \mathbf{q}(\vec{a}))$ for a deriver $\mathfrak{D} \in \{\mathfrak{D}_{\mathsf{cq}}, \mathfrak{D}_{\mathsf{sk}}\}$ whose (tree) size is $\leq n$? To better distinguish the complexity of finding small proofs from that of query answering, we assume $\mathcal{T} \cup \mathcal{A} \models \mathbf{q}(\vec{a})$ as precondition, which fits the intuition that users would request an explanation only after they know that $\vec{a}$ is an answer. Nevertheless, our results show that regardless of $\mathfrak{D}$ for *DL-Lite*$_R$ ontologies, this problem is in AC^0 in data complexity and NP-complete in combined complexity, and therefore of the same complexity as query answering itself [7]. Moreover, for tree-shaped queries one can compute proofs of minimal tree size in polynomial time. For $\mathcal{EL}$, the data complexity rises to P, again matching the complexity of query answering [9]. The combined complexity of NP also falls to P in case the CQ $\mathbf{q}$ contains only a single atom.

4 Metric Temporal CQs

To generalize these results to *temporal* query answering, we assume that *i)* TBox axioms hold globally, i.e. at all time points, *ii)* the ABox contains information about the state of the world in different time intervals, and *iii)* the query contains (metric) temporal operators such as $\boxminus_I$, $\boxplus_I$, $\mathcal{U}_I$, $\mathcal{S}_I$ defined for finite, non-negative time intervals I. The operators are interpreted over an integer timeline as follows. For a temporal interpretation $\mathfrak{I} = (\Delta^{\mathfrak{I}}, (\mathcal{I}_i)_{i\in\mathbb{Z}})$, where $\mathcal{I}_i$ are DL interpretations, and $i \in \mathbb{Z}$, we have $\mathfrak{I}, i \models \boxplus_I \phi$ iff $\forall k \in I$ we have $\mathcal{I}_{i+k} \models \phi$ and $\mathfrak{I}, i \models \phi \mathcal{U}_I \psi$ iff $\exists k \in I$ such that $\mathcal{I}_{i+k} \models \psi$ and $\forall j : 0 \leq j < k : \mathcal{I}_{i+j} \models \phi$ (operators $\mathcal{S}_I$ and $\boxminus_I$ are symmetric). In addition to domain variables as before, metric temporal queries contain an interval variable.

In order to extend the schemas from Figures 1 and 3 to the temporal dimension, we annotate every CQ premise and conclusion of the inference schemas with the same interval variable. In Figure 5, we show some additional schemas needed for dealing with time annotations, where $\nu := (\iota + 1) \cap \iota'$ (all time points where ψ-s are immediately preceded by ϕ-s) and $[w_1, w_2] - [r_1, r_2] := [w_1 - r_2, w_2 - r_1]$, and none of the involved intervals should be empty. Inferences for $\boxminus$ and $\mathcal{S}$ are similar. Altogether, this results in two temporal derivers $\mathfrak{D}_{\mathsf{tcq}}$ and $\mathfrak{D}_{\mathsf{tsk}}$.

$$\frac{\exists \vec{x}_1.\, \phi(\vec{x}_1)@\iota_1 \qquad \ldots \qquad \exists \vec{x}_n.\, \phi(\vec{x}_n)@\iota_n}{(\exists \vec{x}.\, \phi(\vec{x}))@\bigcup_{i=1}^n \iota_i} \qquad\qquad \frac{\exists \vec{x}.\, \phi(\vec{x})@\iota}{\exists \vec{x}.\, \phi(\vec{x})@\iota',\ \text{for } \iota' \subseteq \iota}$$

$$\frac{\phi(\vec{x})@\iota \qquad \psi(\vec{y})@\iota'}{\phi(\vec{x})\,\mathcal{U}_{[r_1,r_2]}\psi(\vec{y})@(\nu - [r_1, r_2]) \cap \iota}\ (\mathcal{U}) \qquad\qquad \frac{\phi(\vec{x})@[t_1, t_2]}{\boxplus_{[r_1,r_2]}\phi(\vec{x})@[t_1 - r_1, t_2 - r_2]}\ (\boxplus)$$

Fig. 5. Additional time-related inference schemas.

For *DL-Lite*$_R$ ontologies, we can lift some of the previous results to the temporal setting. Namely, given a temporal ABox and a number n, the problems if there exists a proof for $\mathbf{q}(\vec{a}, \iota)$ in $\mathfrak{D}(\mathcal{T} \cup \mathcal{A}, \mathbf{q}(\vec{a}, \iota))$ of (tree) size $\leq n$, where ι is a time interval and $\mathfrak{D} \in \{\mathfrak{D}_{\mathsf{tcq}}, \mathfrak{D}_{\mathsf{tsk}}\}$, is in AC^0 in data complexity and NP-complete in combined complexity.

Acknowledgements This work was supported by DFG in grant 389792660, TRR 248 (`https://perspicuous-computing.science`), and QuantLA, GRK 1763 (`https://lat.inf.tu-dresden.de/quantla`).

References

1. Alrabbaa, C., Baader, F., Borgwardt, S., Koopmann, P., Kovtunova, A.: Finding small proofs for description logic entailments: Theory and practice. In: Albert, E., Kovacs, L. (eds.) LPAR-23: 23rd International Conference on Logic for Programming, Artificial Intelligence and Reasoning. EPiC Series in Computing, vol. 73, pp. 32–67. EasyChair (2020). https://doi.org/10.29007/nhpp
2. Alrabbaa, C., Borgwardt, S., Koopmann, P., Kovtunova, A.: Finding good proofs for answers to conjunctive queries mediated by lightweight ontologies (technical report) (2022). https://doi.org/10.48550/ARXIV.2206.09758, https://arxiv.org/abs/2206.09758
3. Baader, F.: Terminological cycles in a description logic with existential restrictions. In: Gottlob, G., Walsh, T. (eds.) IJCAI-03, Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence. pp. 325–330. Morgan Kaufmann (2003), http://ijcai.org/Proceedings/03/Papers/048.pdf
4. Baader, F., Horrocks, I., Lutz, C., Sattler, U.: An Introduction to Description Logic. Cambridge University Press (2017). https://doi.org/10.1017/9781139025355, http://dltextbook.org/
5. Borgida, A., Calvanese, D., Rodriguez-Muro, M.: Explanation in the *DL-Lite* family of description logics. In: Meersman, R., Tari, Z. (eds.) On the Move to Meaningful Internet Systems: OTM 2008. Lecture Notes in Computer Science, vol. 5332, pp. 1440–1457. Springer (2008). https://doi.org/10.1007/978-3-540-88873-4_35
6. Calì, A., Gottlob, G., Lukasiewicz, T.: A general datalog-based framework for tractable query answering over ontologies. J. Web Semant. **14**, 57–83 (2012). https://doi.org/10.1016/j.websem.2012.03.001
7. Calvanese, D., De Giacomo, G., Lembo, D., Lenzerini, M., Rosati, R.: Tractable reasoning and efficient query answering in description logics: The DL-Lite family. J. of Automated Reasoning **39**(3), 385–429 (2007). https://doi.org/10.1007/s10817-007-9078-x
8. Croce, F., Lenzerini, M.: A framework for explaining query answers in DL-Lite. In: Faron-Zucker, C., Ghidini, C., Napoli, A., Toussaint, Y. (eds.) Knowledge Engineering and Knowledge Management - 21st International Conference, EKAW 2018. Lecture Notes in Computer Science, vol. 11313, pp. 83–97. Springer (2018). https://doi.org/10.1007/978-3-030-03667-6_6
9. Rosati, R.: On conjunctive query answering in EL. In: Calvanese, D., Franconi, E., Haarslev, V., Lembo, D., Motik, B., Turhan, A., Tessaris, S. (eds.) Proceedings of the 2007 International Workshop on Description Logics (DL2007). CEUR Workshop Proceedings, vol. 250. CEUR-WS.org (2007), http://ceur-ws.org/Vol-250/paper_83.pdf
10. Stefanoni, G.: Explaining query answers in lightweight ontologies. Diploma thesis, Technische Universität Wien (2011), http://www.cs.ox.ac.uk/files/7942/thesis.pdf