FO-Rewritability of Expressive Ontology-Mediated Queries

Cristina Feier, Antti Kuusisto, Carsten Lutz

Fachbereich Informatik, Universität Bremen, Germany

Abstract. We show that FO-rewritability of OMQs based on (extensions of) $\mathcal{ALC}$ and unions of conjunctive queries (UCQs) is decidable and 2NExpTime-complete. Previously, decidability was only known for atomic queries (AQs). On the way, we establish the same results also for monotone monadic SNP without inequality (MMSNP) and for monadic disjunctive Datalog (MDDLog). We also analyze the shape of FO-rewritings, thus making a step towards their actual computation.

1 Introduction

Query rewriting is a widely used technique for efficiently answering ontology-mediated queries (OMQs) using off-the-shelf database systems. In particular, if an OMQ is rewritable into a first-order (FO) query, then it can be answered using a relational database system. It is thus a fundamental problem to design algorithms that compute an FO-rewriting of a given OMQ. For OMQs based on expressive DLs such as $\mathcal{ALC}$, finding complete such algorithms turns out to be a technically very challenging problem; moreover, the desired rewritings are not always guaranteed to exist. A natural first step is thus to find an algorithm that decides the existence of a rewriting; in fact, any complete and terminating algorithm for computing rewritings will implicitly also solve the decision problem and one can expect to learn important lessons already from the latter case.

An *OMQ* is a triple $Q = (\mathcal{T}, \Sigma, q)$ with $\mathcal{T}$ a TBox, Σ an ABox signature, and q a query [5]. We use $(\mathcal{L}, \mathcal{Q})$ to denote the *OMQ language* that consists of all OMQs where $\mathcal{T}$ is formulated in the DL $\mathcal{L}$ and q in the query language $\mathcal{Q}$. It has been shown in [5] that decidability results and tight NExpTime complexity bounds for FO-rewritability in $(\mathcal{ALC}, \text{AQ})$ and related OMQ languages can be obtained by translating the input OMQ Q into a constraint satisfaction problem (CSP) whose complement is equivalent to Q and then applying known algorithms that decide FO-rewritability of the CSP [15]. When atomic queries (AQs) are replaced with conjunctive queries (CQs) or unions thereof (UCQs), such equivalence-preserving translation to CSP is no longer possible; instead and as also shown in [5], one can translate Q into a formula φ of the strictly more expressive logical generalization MMSNP of CSP such that $\neg\varphi$ is equivalent to Q [11]. In contrast to the CSP case, though, it was not known whether FO-rewritability of MMSNP formulas is decidable. In this abstract, we show that this is the case and that the complexity is 2NExpTime-complete, with the lower bounds coming from [7].

We then lift this result to FO-rewritability of OMQs formulated in any OMQ language between $(\mathcal{ALCI}, \text{UCQ})$ and $(\mathcal{SHI}, \text{UCQ})$. Technically, our approach consists in a reduction of the MMSNP case to the CSP case. We also analyze the shape of the rewritings, which we hope will provide guidance for finding algorithms that compute actual rewritings. One can prove analogous results for rewritability into monadic Datalog in a very similar way. We do not report about details here for lack of space. We are currently working on the unrestricted Datalog case, which is more challenging.

Related work. FO-rewritability was first studied in an OMQ context for the inexpressive DL-Lite family of DLs [1, 9, 14, 17]. FO-rewritability in OMQ languages based on more expressive Horn DLs has been investigated in [3, 4, 12]. Rewritability of Horn DLs into Datalog was considered in [10, 13, 18–20].

2 Results

Instead of working with MMSNP, we prefer monadic disjunctive Datalog, MDDLog. It was observed in [5] that these have the same expressive power up to complementation and can be mutually translated in polynomial time. We refer to [5] for full definitions and notation. An MDDLog program Π is *Boolean* if it only returns yes/no answers. The diameter of Π is the maximum number of variables occuring in a rule in Π. A *generalized CSP* is defined by a set of templates S instead of a single template and asks for a homomorphism from the input I to at least one template $T \in S$ (a template is simply a finite relational structure). With *co*CSP, we mean the complement of a (potentially generalized) CSP. The *girth* of a structure is the length of a smallest cycle in it and ∞ if there is no cycle, see e.g. [7] for details.

We start with Boolean MDDLog programs. Our proofs make use of the classical translation of MMSNP sentences into generalized CSPs first described by Feder and Vardi and stated in the following in terms of MDDLog.

Theorem 1 ([11]). *Given a Boolean MDDLog program Π over EDB schema $\mathbf{S}_E$ of diameter k, one can effectively construct a set of templates S_Π over a different EDB schema $\mathbf{S}'_E$ such that*

1. *every finite $\mathbf{S}_E$-structure I can be converted in polytime into an $\mathbf{S}'_E$-structure I' such that $I \models \Pi$ iff $I' \notin \mathit{CSP}(S_\Pi)$;*
2. *every finite $\mathbf{S}'_E$-structure I' of girth exceeding k can be converted in polytime into an $\mathbf{S}_E$-instance I such that $I \models \Pi$ iff $I' \notin \mathit{CSP}(S_\Pi)$.*

Exact size bounds on the set S_Π and its elements are given in [7]; although they are needed for our final result, we omit them from this abstract for brevity.

We next observe that the reduction described in Theorem 1 preserves FO-rewritability, up to a certain gap related to inputs of small girth. In the following, we will consider UCQ-rewritability instead of FO-rewritability since for MDDLog, coCSP, and all other formalisms considered here FO-rewritability implies UCQ-rewritability.

Proposition 1. *Let Π be a Boolean MDDLog program of diameter k and let S_Π be as in Theorem 1. Then*

1. *every UCQ-rewriting of coCSP(S_Π) can be converted into a UCQ-rewriting of Π in polynomial time;*
2. *every UCQ-rewriting of Π can effectively be converted into a UCQ-rewriting of coCSP(S_Π) on structures of girth exceeding k.*

Thus, FO-rewritability of Π is equivalent to FO-rewritability of coCSP(S_Π) on structures of sufficiently high girth. The 'high girth' qualification is removed by the following observation.

Lemma 1. *Let S be a set of templates over schema $\mathbf{S}_E$ and $g \geq 0$. Then, if coCSP(S) is UCQ-definable on finite $\mathbf{S}_E$-structures of girth exceeding g, it is UCQ-definable on (unrestricted) finite $\mathbf{S}_E$-structures.*

Analyzing the involved blowups and exploiting that FO-rewritability of generalized CSPs can be decided in NP [5,15] and that FO-rewritability of Boolean MDDLog programs is 2NExpTime-hard [7], we obtain the following.

Theorem 2. *FO-rewritability of Boolean MDDLog programs is decidable in 2*NExpTime*, thus 2*NExpTime*-complete.*

The limitation to Boolean programs can be lifted by a simple reduction which replaces answer variables with fresh monadic relation symbols. Moreover, it was shown in [7] that OMQs formulated in $(\mathcal{SHI}, \text{UCQ})$ can be translated into MDDLog with a blowup that is double exponential, but does not add up with the blowup caused by the translation of MDDLog programs to generalized CSPs. Finally, [7] also establishes a 2NExpTime lower bound for FO-rewritability in $(\mathcal{ALCI}, \text{CQ})$ and $(\mathcal{ALC}, \text{UCQ})$. We can thus extend Theorem 2 as follows.

Theorem 3. *FO-rewritability is 2*NExpTime*-complete in MMSNP, in MDDLog, and in OMQ languages between ($\mathcal{ALC}$, UCQ) and ($\mathcal{SHI}$, UCQ) as well as between ($\mathcal{ALCI}$, CQ) and ($\mathcal{SHI}$, UCQ) (without transitive roles in UCQs).*

We now consider the shape of rewritings. For brevity, we concentrate on Boolean queries. An analysis of the proof of Proposition 1 yields the following.

Proposition 2. *Every FO-rewritable Boolean MDDLog program of diameter k has a rewriting of the form $q_1 \vee \cdots \vee q_n$ with each q_i a CQ of treewidth bounded by $(1, k)$. The same holds for Boolean OMQs from the OMQ languages in Theorem 3.*

See [6] for a definition of treewidth $(1, k)$. We believe that Proposition 2 is interesting for at least two reasons. Firstly, it says that when constructing rewritings, it is enough to look for a structurally very restricted type of UCQ instead of for an unrestricted FO formula. And secondly because it clarifies the shape of obstructions of FO-rewritable MMSNP sentences in the spirit of CSP obstructions [8]. In particular, it is interesting to contrast Proposition 2 with the fact that if a CSP is FO-rewritable, then it has a finite set of obstructions that are finite trees, that is, its complement is rewritable into a UCQ that consists of tree-shaped CQs [2,16].

Acknowledgements. We thank Manuel Bodirsky and Florent Madelaine for inspiring discussions and the ERC for funding this work under grant 647289.

References

1. Alessandro Artale, Diego Calvanese, Roman Kontchakov, and Michael Zakharyaschev. The DL-Lite family and relations. *J. of Art. Int. Res. (JAIR)*, 36:1–69, 2009.
2. Albert Atserias. On digraph coloring problems and treewidth duality. *Eur. J. Comb.*, 29(4):796–820, 2008.
3. Meghyn Bienvenu, Carsten Lutz, and Frank Wolter. Deciding FO-rewritability in $\mathcal{EL}$. In *Proc. of DL*, pages 70–80, 2012.
4. Meghyn Bienvenu, Carsten Lutz, and Frank Wolter. First-order rewritability of atomic queries in Horn description logics. In *Proc. of IJCAI*, 2013.
5. Meghyn Bienvenu, Balder ten Cate, Carsten Lutz, and Frank Wolter. Ontology-based data access: A study through disjunctive Datalog, CSP, and MMSNP. *ACM Trans. Database Syst.*, 39(4):33:1–33:44, 2014.
6. Manuel Bodirsky and Víctor Dalmau. Datalog and constraint satisfaction with infinite templates. *J. Comput. Syst. Sci.*, 79(1):79–100, 2013.
7. Pierre Bourhis and Carsten Lutz. Containment in monadic disjunctive Datalog, MMSNP, and expressive description logics. In *Proc. of KR*, 2016.
8. A.A. Bulatov, A. Krokhin, and B. Larose. Dualities for constraint satisfaction problems. *Complexity of Constraints*, 5250 LNCS:93–124, 2008.
9. Diego Calvanese, Giuseppe De Giacomo, Domenico Lembo, Maurizio Lenzerini, and Riccardo Rosati. Tractable reasoning and efficient query answering in description logics: The DL-Lite family. *J. Autom. Reasoning*, 39(3):385–429, 2007.
10. Thomas Eiter, Magdalena Ortiz, Mantas Simkus, Trung-Kien Tran, and Guohui Xiao. Query rewriting for Horn-SHIQ plus rules. In *Proc. of AAAI*. 2012.
11. Tomás Feder and Moshe Y. Vardi. The computational structure of monotone monadic SNP and constraint satisfaction: A study through datalog and group theory. *SIAM J. Comput.*, 28(1):57–104, 1998.
12. Peter Hansen, Carsten Lutz, İnanç Seylan, and Frank Wolter. Efficient query rewriting in the description logic el and beyond. In *Proc. of IJCAI*, 2015.
13. Mark Kaminski, Yavor Nenov, and Bernardo Cuenca Grau. Computing datalog rewritings for disjunctive datalog programs and description logic ontologies. In *Proc. of RR*, pages 76–91, 2014.
14. Stanislav Kikot, Roman Kontchakov, Vladimir V. Podolskii, and Michael Zakharyaschev. Exponential lower bounds and separation for query rewriting. In *Proc. of ICALP (2)*, pages 263–274, 2012.
15. Benoit Larose, Cynthia Loten, and Claude Tardif. A characterisation of first-order constraint satisfaction problems. *Logical Methods in Comp. Sci.*, 3(4), 2007.
16. Jaroslav Nešetřil and Claude Tardif. Duality theorems for finite structures (characterising gaps and good characterisations). *J. Comb. Theory, Ser. B*, 80(1):80–97, 2000.
17. Héctor Pérez-Urbina, Boris Motik, and Ian Horrocks. A comparison of query rewriting techniques for DL-Lite. In *Proc. of DL*, 2009.
18. Héctor Pérez-Urbina, Boris Motik, and Ian Horrocks. Tractable query answering and rewriting under description logic constraints. *J. of Applied Logic*, 8(2):186–209, 2010.
19. Riccardo Rosati. On conjunctive query answering in $\mathcal{EL}$. In *Proc. of DL*, pages 451–458, 2007.
20. Despoina Trivela, Giorgos Stoilos, Alexandros Chortaras, and Giorgos B. Stamou. Optimising resolution-based rewriting algorithms for OWL ontologies. *J. Web Sem.*, 33:30–49, 2015.