Classifying the Complexity of Ontology-Mediated Queries in $\mathcal{EL}$: From Atomic Queries to Conjunctive Queries*

Carsten Lutz and Leif Sabellek

Department of Computer Science, University of Bremen, Germany

Substantial research efforts have been invested into understanding the computational properties and expressive power of ontology-mediated queries (OMQs). Two important topics are the *data complexity* of OMQ evaluation [10, 12, 17, 5, 1], where only the data is considered to be the input while the OMQ is fixed, and the *rewritability* of OMQs into more standard database query languages such as SQL (which in this context is often equated with first-order logic) and Datalog [7, 3, 1, 11, 9]. Both subjects are thoroughly intertwined. In particular, rewritability into first-order logic (FO) is closely related to AC^0 data complexity while rewritability into Datalog is closely related to PTime data complexity.

Traditionally, data complexity and rewritability have mostly been studied on the level of OMQ languages $(\mathcal{L}, \mathcal{Q})$ defined by an ontology language $\mathcal{L}$ such as $\mathcal{EL}$ and a query language $\mathcal{Q}$ such as conjunctive queries (CQs). A more fine-grained analysis has been initiated in [16, 1], the aim being to understand the exact complexity and rewritability status of *every* OMQ from relevant OMQ languages $(\mathcal{L}, \mathcal{Q})$, see also [18, 15]. For expressive DLs, this turns out to be closely related to the complexity classification of constraint satisfaction problems (CSPs) with a fixed template [8]. Very important progress has recently been made in this area with the proof that CSPs enjoy a dichotomy between PTime and NP [4, 19]. Via the results in [1], this implies that OMQ evaluation in languages such as $(\mathcal{ALCI}, \mathrm{UCQ})$ enjoys a dichotomy between PTime and coNP. However, the complexity of CSPs and OMQs is still far from being fully understood. For example, neither in CSP nor in expressive OMQ languages it is known whether there is a dichotomy between NL and PTime, and whether containment in NL coincides with rewritability into linear Datalog, a well-known fragment of Datalog that allows only linear recursion. It has been conjectured that both of this is the case.

We report on [14], which carries out a fine-grained analysis of the data complexity and rewritability of OMQs from $(\mathcal{EL}, \mathrm{CQ})$, achieving a complete classification. This extends the results of [13] where the same was achieved for the OMQ language $(\mathcal{EL}, \mathrm{AQ})$ that only admits atomic queries, that is, conjunctive queries of the very simple form $A(x)$. For every $Q \in (\mathcal{EL}, \mathrm{CQ})$, let $\textsc{eval}(Q)$ be the problem to decide, given an ABox $\mathcal{A}$ and a tuple of individuals $\mathbf{a}$, whether $\mathbf{a}$ is a certain answer to Q on $\mathcal{A}$. Our main result is the following trichotomy.

Theorem 1. *For every OMQ $Q \in (\mathcal{EL}, \mathrm{CQ})$, one of the following is true:*

1. $\textsc{eval}(Q)$ *is in* AC^0 *and Q is rewritable into FO (actually into a UCQ with equality atoms);*
2. $\textsc{eval}(Q)$ *is* NL*-complete and Q is rewritable into linear Datalog;*
3. $\textsc{eval}(Q)$ *is* PTime*-complete and Q is rewritable into monadic Datalog.*

Consequently, $\textsc{eval}(Q)$ *is in* NL *if and only if Q is rewritable into linear Datalog unless* $\mathrm{NL} = \textsc{PTime}$.

An important aspect of the proof of Theorem 1 are semantic characterizations of the three cases, which are also interesting in their own right. The characterizations are easiest to state for AQs. For $Q \in (\mathcal{EL}, \mathrm{AQ})$, let $\mathcal{M}_Q$ denote the set of all tree shaped ABoxes whose root is a certain answer to Q and that are minimal with that property regarding set inclusion. We say that Q has *bounded depth* if there is a constant bound on the depth of the trees in $\mathcal{M}_Q$ and Q has *bounded pathwidth* if there is a constant bound on the pathwidth of the trees in $\mathcal{M}_Q$; recall that pathwidth measures the similarity of a structure to a path with lower number meaning more similar. For CQs, the definitions are similar, but for defining $\mathcal{M}_Q$ we consider ABoxes of a certain pseudo-tree shape instead of tree shaped ones.

While it is known that bounded depth of an OMQ from $(\mathcal{EL}, \mathrm{CQ})$ implies FO-rewritability [2], we prove that

(1) if Q does not have bounded depth, then $\textsc{eval}(Q)$ is NL-hard;
(2) if Q has bounded pathwidth, then Q is rewritabe into linear Datalog (and thus, $\textsc{eval}(Q)$ is in NL);
(3) if Q does not have bounded pathwidth, then $\textsc{eval}(Q)$ is PTime-hard and Q is not rewritable into linear Datalog.

The proofs are technically subtle, especially the hardness proofs (1) and (3) for Boolean CQs. Furthermore, the proof of (2) yields a way to construct linear Datalog rewritings if they exist, and the rewritings have smaller width (number of variables in rule heads) than the linear Datalog rewritings given in [13]. We also show that all relevant decision problems, such as whether a given OMQ is rewritable into linear Datalog and whether it is PTime-hard, are ExpTime-complete. The upper bounds are proved by reduction to the emptiness problem of tree automata.

There are several natural directions into which our results can be generalized. One is to consider the OMQ language $(\mathcal{EL}, \mathrm{UCQ})$. We conjecture that this generalization is not difficult. In fact, we only refrained from covering it since it makes all proofs more technical and distracts from the main ideas. An important direction for future work is to extend our analysis to $\mathcal{ELI}$, that is, to add inverse roles. Even the case of $(\mathcal{ELI}, \mathrm{AQ})$ appears to be challenging. In fact, we observe that a complexity classification of $(\mathcal{ELI}, \mathrm{AQ})$ is equivalent to a complexity classification of all CSPs that have tree obstructions, which is an open problem. It is easy to see that $(\mathcal{ELI}, \mathrm{AQ})$ contains OMQs that are L-complete and we conjecture that AC^0, L, NL, and PTime are the only complexities that

occur. We also conjecture that L-completeness coincides with rewritability into symmetric Datalog [6].

Acknowledgments. This research was supported by ERC consolidator grant 647289 CODA.

References

1. Bienvenu, M., ten Cate, B., Lutz, C., Wolter, F.: Ontology-based data access: A study through disjunctive datalog, CSP, and MMSNP. ACM Trans. Database Syst. **39**(4), 33:1–33:44 (2014)
2. Bienvenu, M., Hansen, P., Lutz, C., Wolter, F.: First order-rewritability and containment of conjunctive queries in Horn description logics. In: Proc. of IJCAI. pp. 965–971. IJCAI/AAAI Press (2016)
3. Bienvenu, M., Lutz, C., Wolter, F.: First order-rewritability of atomic queries in horn description logics. In: Proc. of IJCAI. pp. 754–760. IJCAI/AAAI (2013)
4. Bulatov, A.A.: A dichotomy theorem for nonuniform csps. In: Proc. of FOCS. pp. 319–330 (2017)
5. Calvanese, D., De Giacomo, G., Lembo, D., Lenzerini, M., Rosati, R.: Data complexity of query answering in description logics. Artif. Intell. **195**, 335–360 (2013)
6. Egri, L., Larose, B., Tesson, P.: Symmetric datalog and constraint satisfaction problems in LogSpace. Electronic Colloquium on Computational Complexity (ECCC) **14**(024), 1 (2007)
7. Eiter, T., Ortiz, M., Simkus, M., Tran, T., Xiao, G.: Query rewriting for Horn-$\mathcal{SHIQ}$ plus rules. In: Proc. of AAAI. AAAI Press (2012)
8. Feder, T., Vardi, M.Y.: 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)
9. Feier, C., Kuusisto, A., Lutz, C.: Rewritability in monadic disjunctive datalog, MMSNP, and expressive description logics. Logical Methods in Computer Science (2019)
10. Hustadt, U., Motik, B., Sattler, U.: Data complexity of reasoning in very expressive description logics. In: Proc. of IJCAI. pp. 466–471. Professional Book Center (2005)
11. Kaminski, M., Nenov, Y., Grau, B.C.: Datalog rewritability of disjunctive datalog programs and its applications to ontology reasoning. In: Proc. of AAAI. pp. 1077–1083. AAAI Press (2014)
12. Krisnadhi, A., Lutz, C.: Data complexity in the $\mathcal{EL}$ family of description logics. In: Proc. of LPAR. LNAI, vol. 4790, pp. 333–347. Springer (2007)
13. Lutz, C., Sabellek, L.: Ontology-mediated querying with the description logic el: Trichotomy and linear datalog rewritability. In: Proc. of IJCAI. pp. 1181–1187. ijcai.org (2017)
14. Lutz, C., Sabellek, L.: A complete classification of the complexity and rewritability of ontology-mediated queries based on the description logic EL. submitted to Journal of Artificial Intelligence Research (JAIR) (2019), http://arxiv.org/abs/1904.12533
15. Lutz, C., Seylan, I., Wolter, F.: Ontology-mediated queries with closed predicates. In: Proc. of IJCAI. pp. 3120–3126. AAAI Press (2015)
16. Lutz, C., Wolter, F.: Non-uniform data complexity of query answering in description logics. In: Proc. of KR. AAAI Press (2012)

17. Rosati, R.: The limits of querying ontologies. In: Proc. of ICDT. LNCS, vol. 4353, pp. 164–178. Springer (2007)
18. Zakharyaschev, M., Kikot, S., Gerasimova, O.: Towards a data complexity classification of ontology-mediated queries with covering. In: Proc. of DL. CEUR Workshop Proceedings, vol. 2211. CEUR-WS.org (2018)
19. Zhuk, D.: A proof of CSP dichotomy conjecture. In: Proc. of FOCS. pp. 331–342 (2017)