Answering Counting Queries over DL-Lite Ontologies (Extended Abstract)

Meghyn Bienvenu[1], Quentin Manière[1], and Michaël Thomazo[2]

[1] University of Bordeaux, CNRS, Bordeaux INP, LaBRI, Talence, France
[2] Inria, DI ENS, ENS, CNRS, University PSL, Paris, France

Abstract. This extended abstract briefly summarizes our recent work [1] on answering counting queries over DL-Lite knowledge bases. Our contributions include the definition of a more general form of counting conjunctive query (CCQ), a detailed study of the data and combined complexity of CCQ answering over DL-Lite$_{\mathsf{core}}$ and DL-Lite$_{\mathcal{R}}$ knowledge bases under various restrictions on the TBox and query, and the precise data complexity of identifying the best certain interval.

Ontology-mediated query answering (OMQA) utilizes ontologies to provide a convenient vocabulary for query formulation and to capture domain knowledge that is exploited during the querying process to obtain more complete sets of answers [9, 2, 10]. Much of the work on OMQA considers ontologies formulated using description logics (DLs), among which the DL-Lite family [4] has received particular attention as it enjoys favorable computational properties.

The vast majority of work on OMQA supposes that user queries are given as conjunctive queries (CQs). However, there are many other kinds of database queries, beyond plain CQs, that are relevant in practice. This motivates research into the feasibility of adopting other database query languages for OMQA. While enriching CQs with either negated atoms or inequalities has been shown to lead to undecidability even in very restricted settings [6], the situation is more positive for navigational queries (like regular path queries), which can be adopted without losing decidability, sometimes even retaining tractable data complexity [3].

Aggregate queries, which use numeric operators (e.g. count, sum, max) to summarize selected parts of a dataset, constitute another prominent class of database queries. Although such queries are widely used for data analysis, they have been little explored in context of OMQA. This may be partly due to the fact that it is not at all obvious how to define the semantics of such queries in the OMQA setting. A first exploration of aggregate queries in OMQA was conducted by Calvanese et al. [5]. They argued that the most straightforward adaptation of classical certain answer semantics to aggregate queries was unsatisfactory, as often values would differ from model to model, leading to no certain answers. For this reason, an epistemic semantics was proposed, in which variables involved in

	Data		Combined	
	Rooted	**Exh. rooted**	**Rooted**	**Exh. rooted**
DL-Lite$_{\mathsf{core}}$	coNP-**c**	TC^0-**c**	Π_2^p-**h** PP-**h** / coNEXP	PP-**c**
DL-Lite$_{\mathcal{R}}$	coNP-**c**	coNP-**c**	coNEXP-**h** / coN2EXP coNEXP-**c** (f.-d. TBox)	Π_2^p-**h** PP-**h** / coNEXP

Table 1. Complexity results for rooted and exhaustive rooted counting CQs ('f.-d.' stands for 'finite-depth', '-h' for '-hard', '-c' for '-complete')

the aggregation are required to match to data constants. However, as discussed in [7], this semantics can also give unintuitive results by ignoring ways of mapping aggregate variables to anonymous elements inferred due the ontology axioms. For instance, if no children of alex are listed in the data, then a query that asks to return the number of children will yield 0 under epistemic semantics, even if it can be inferred (e.g. due to a family tax benefit) that there must be at least 3 children. This led Kostylev and Reutter [7] to define an alternative semantics for two kinds of counting queries (inspired by the COUNT and COUNT DISTINCT in SQL) which adopts a form of certain answer semantics but considers lower and upper bounds on the count value across different models. For the two considered logics (DL-Lite$_{\mathsf{core}}$ and DL-Lite$_{\mathcal{R}}$), only the lower bounds on the count value are non-trivial, and a complexity analysis shows that they are challenging to identify: coNP-data complexity for both logics, and Π_2^p-hard (resp. coNEXP-hard) in combined complexity for DL-Lite$_{\mathsf{core}}$ (resp. DL-Lite$_{\mathcal{R}}$). Several questions were left unanswered by their work, including the difficulty of recognizing the optimal lower bound and the impact of allowing multiple aggregation variables.

In the present work, which is reported in [1], we return to the issue of handling counting queries in OMQA and make several important contributions. We first introduce a new notion of counting CQ that generalizes the two forms of queries from [7] and allows arbitrarily many counting variables. We show that existing complexity results for DL-Lite$_{\mathsf{core}}$ and DL-Lite$_{\mathcal{R}}$ KBs continue to hold for our more general notion of counting CQ, and we further provide an improved coNEXP upper bound for the relevant case of finite-depth TBoxes. We also consider the impact of restricting the query structure, focusing on the class of rooted queries, in which every query variable must be connected to an answer variable or individual in the query graph. A recent result, obtained as part of a study of bag semantics for OMQA [8], identified a case in which rootedness leads to tractable data complexity for counting queries. This motivated us to perform a thorough investigation of rooted counting queries, which yielded several improvements upon existing complexity bounds (see Table 1), including a tight PP-completeness result for the natural subclass of *exhaustive rooted* counting CQs, in which every non-answer variable is a counting variable. As our final contribution, we investigate the problem of recognizing the best certain interval

and show it to be DP-complete in data complexity. Our results close some questions that were left open by the work of Kostylev and Reutter [7] and pave the way for further study of counting and aggregate queries in the OMQA setting.

References

1. Bienvenu, M., Manière, Q., Thomazo, M.: Answering counting queries over DL-Lite ontologies. In: Proceedings of the 29th International Joint Conference on Artificial Intelligence (IJCAI) (2020)
2. Bienvenu, M., Ortiz, M.: Ontology-mediated query answering with data-tractable description logics. In: Tutorial Lectures of the 11th Reasoning Web International Summer School. pp. 218–307 (2015)
3. Bienvenu, M., Ortiz, M., Simkus, M.: Regular path queries in lightweight description logics: Complexity and algorithms. Journal of Artificial Intelligence Research (JAIR) **53**, 315–374 (2015)
4. Calvanese, D., De Giacomo, G., Lembo, D., Lenzerini, M., Rosati, R.: Tractable reasoning and efficient query answering in description logics: The DL-Lite family. Journal of Automated Reasoning (JAR) **39**(3), 385–429 (2007)
5. Calvanese, D., Kharlamov, E., Nutt, W., Thorne, C.: Aggregate queries over ontologies. In: Proceedings of the 2nd International Workshop on Ontologies and Information Systems for the Semantic Web (ONISW). pp. 97–104 (2008)
6. Gutiérrez-Basulto, V., Ibáñez-García, Y.A., Kontchakov, R., Kostylev, E.V.: Queries with negation and inequalities over lightweight ontologies. Journal of Web Semantics (JWS) **35**, 184–202 (2015)
7. Kostylev, E.V., Reutter, J.L.: Complexity of answering counting aggregate queries over DL-Lite. Journal of Web Semantics (JWS) **33**(1), 94–111 (2015)
8. Nikolaou, C., Kostylev, E.V., Konstantinidis, G., Kaminski, M., Grau, B.C., Horrocks, I.: Foundations of ontology-based data access under bag semantics. Artificial Intelligence (AIJ) **274**, 91 – 132 (2019)
9. Poggi, A., Lembo, D., Calvanese, D., De Giacomo, G., Lenzerini, M., Rosati, R.: Linking data to ontologies. Journal of Data Semantics **10**, 133–173 (2008)
10. Xiao, G., Calvanese, D., Kontchakov, R., Lembo, D., Poggi, A., Rosati, R., Zakharyaschev, M.: Ontology-based data access: A survey. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI). pp. 5511–5519 (2018)