The Complexity of Answer Counting for Ontology-Mediated Queries Based on Guarded TGDs (Extended Abstract)*

Cristina Feier, Carsten Lutz, and Marcin Przybyłko

Universität Bremen, Germany

Evaluating ontology-mediated queries (OMQs) is computationally intractable. On the one hand, this is due to the complexity of reasoning with the ontology alone, which is ExpTime-hard even for relatively simple description logics such as $\mathcal{ELI}$. On the other hand, it is due to evaluating the actual queries, which is NP-hard in combined complexity for the widely used class of conjunctive queries (CQs). As a consequence, the complexity of evaluating ontology-mediated queries has been analyzed from several angles. Traditionally, one fixes an ontology language $\mathcal{L}$ and a query language $\mathcal{Q}$ and studies the complexity of evaluating OMQs from the resulting OMQ language $(\mathcal{L}, \mathcal{Q})$, inspecting combined complexity, data complexity, or parameterized complexity [5,8,9,15,16,19,20,21,24]. Recently, this approach has been complemented by more fine-grained analyses. A fine-grained approach to data complexity is pursued in [3,23,22], where the aim is to identify for every OMQ $Q \in (\mathcal{L}, \mathcal{Q})$, the precise data complexity of evaluating Q. A fine-grained study of parameterized complexity is initiated in [2,1], aiming to identify for every class of OMQs $\mathbb{C} \subseteq (\mathcal{L}, \mathcal{Q})$, the precise complexity of evaluating OMQs from $\mathbb{C}$ when the parameter is the size of the OMQ. A main aim in this approach is to delineate classes $\mathbb{C}$ for which evaluation is fixed-parameter tractable (FPT) from classes for which evaluation is W[1]-hard, and it turns out that *bounded treewidth* is an essential property in this context. In all of these studies, OMQ evaluation means that an answer candidate is given as part of the input, in the form of a tuple of constants $\bar{a}$, and then the aim is to decide whether $\bar{a}$ is indeed an answer.

There are, however, many other natural modes of evaluating OMQs such as computing all answers and counting the number of answers (of which there can be exponentially many). This abstract reports on work concerned with counting the number of answers to OMQs. This is important, for example, when there are too many answers to compute all of them, and it is also a fundamental operation in data analytics and in decision support. In fact, answer counting is supported by almost every data management system. Despite its relevance, however, the problem has received little attention so far in ontology-mediated querying. Notable exceptions are [18,17,4,7], mostly in the context of the DL-Lite family of description logics.

Our aim is to study the parameterized complexity of counting the number of answers to OMQs formulated in the language $(\mathbb{G}, \mathbb{UCQ})$ where $\mathbb{UCQ}$ denotes unions of conjunctive queries and $\mathbb{G}$ denotes guarded TGDs used as an ontology language, also known as guarded Datalog$^\pm$ [6]. Paralleling the work in [2,1], our aim is to identify for every class of OMQs $\mathbb{C} \subseteq (\mathbb{G}, \mathbb{UCQ})$, the precise complexity of counting the number of answers to OMQs from $\mathbb{C}$ when the parameter is the size of the OMQ. Note that we count the number of answers according to the traditional definition of certain answers in ontology-mediated querying, unlike [18,4,7] which aim to count the number of homomorphisms that support a given answer or the number of instantiations of specific counting variables.

Before we state our results, let us briefly describe the situation in answer counting for classes of conjunctive queries and UCQs in the classical setting, i.e. without ontologies. In [12,14,10], it is shown that for a class of CQs $\mathbb{C}$, answer counting is in FPT if and only if (1) the CQs in $\mathbb{C}$ have bounded treewidth and (2) the same holds for the so-called contracts of CQs in $\mathbb{C}$.[1] Informally, the contract of a CQ is its hypergraph restricted to the answer variables and extended by adding an edge between any two answer variables x, y that are connected by a path that uses only quantified variables. We refer to [10] for a formal definition. This dichotomy is under the assumptions that the arity of relation symbols is bounded by a constant, $\mathbb{C}$ is recursively enumerable, and FPT $\neq$ W[1]. We generally make the same assumptions in what follows.

It is also shown in [10] that unbounded treewidth and bounded treewidth of contracts a class of CQs $\mathbb{C}$ coincides with W[1]-equivalence of answer counting for $\mathbb{C}$, defined via parameterized Turing reductions, and that unbounded contract treewidth implies #W[1]-hardness. In [13], the latter is strengthened to #W[1]-equivalence when an additional structural measure called *dominating starsize* is bounded for CQs from $\mathbb{C}$, and to #W[2]-hardness if the dominating starsize is unbounded. Informally, the dominating starsize measures how many answer variables are connected by a connected component of the CQ that consists only of quantified variables, see [13] for a formal definition. In [11,13], the above classification is lifted to UCQs using a non-trivial construction. Instead of considering the treewidth / contract treewidth / dominating starsize of CQs $q \in \mathbb{C}$, one now needs to consider these structural measures for a certain (exponential size) set of CQs derived from q on the basis of the inclusion-exclusion principle; we refer to this set as the *Chen-Mengel closure* of q, denoted $\mathsf{cl}_{\mathsf{CM}}(q)$. A formal definition is in [11].

We establish a similar classification for classes of OMQs from the OMQ language $(\mathbb{G}, \mathbb{UCQ})$ which we state in the following. For $Q \in (\mathbb{G}, \mathbb{UCQ})$, we use $Q^\exists$ to denote the *existential rewriting* of Q which is equivalent to Q and does not use existential quantifiers in TGD heads in the ontology. Such a rewriting can be effectively constructed [1]. For $\mathbb{C} \subseteq (\mathbb{G}, \mathbb{UCQ})$, define

$$\mathsf{cl}^{\exists}_{\mathsf{CM}}(\mathbb{C}) = \{\mathsf{core}(\mathsf{ch}_{\mathcal{O}^\exists}(p)) \mid \exists Q \in \mathbb{C} : Q^\exists = (\mathcal{O}^\exists, \mathbf{S}, q^\exists) \text{ and } p \in \mathsf{cl}_{\mathsf{CM}}(q^\exists)\}$$

[1] Here and in what follows, we generally assume that all CQs are homomorphism cores, up to the answer variables. Without this assumptions, one would have to require that each CQ in $\mathbb{C}$ is equivalent to a CQ that has bounded treewidth and bounded treewidth of the contract.

where $\mathsf{ch}_{\mathcal{O}}(q)$ denotes the result of chasing q with $\mathcal{O}$. An OMQ has *full data schema* if every relation symbol can occur in the data.

Theorem 1. *Let $\mathbb{C} \subseteq (\mathbb{G}, \mathbb{UCQ})$ be a recursively enumerable class of OMQs with full data schema and relation symbols of bounded arity. Then the following hold:*

1. *If the treewidths and contract treewidths of CQs in $\mathsf{cl}^{\exists}_{\mathsf{CM}}(\mathbb{C})$ are bounded, then $\mathsf{AnswerCount}(\mathbb{C})$ is in FPT.*
2. *If the treewidths of CQs in $\mathsf{cl}^{\exists}_{\mathsf{CM}}(\mathbb{C})$ are unbounded and the contract treewidths of CQs in $\mathsf{cl}^{\exists}_{\mathsf{CM}}(\mathbb{C})$ are bounded, then $\mathsf{AnswerCount}(\mathbb{C})$ is $\mathrm{W}[1]$-equivalent.*
3. *If the contract treewidths of CQs in $\mathsf{cl}^{\exists}_{\mathsf{CM}}(\mathbb{C})$ are unbounded and the dominating starsizes of CQs in $\mathsf{cl}^{\exists}_{\mathsf{CM}}(\mathbb{C})$ are bounded, then $\mathsf{AnswerCount}(\mathbb{C})$ is $\#\mathrm{W}[1]$-equivalent.*
4. *If the dominating starsizes of CQs in $\mathsf{cl}^{\exists}_{\mathsf{CM}}(\mathbb{C})$ are unbounded, then $\mathsf{AnswerCount}(\mathbb{C})$ is $\#\mathrm{W}[2]$-hard.*

The upper bounds are easy to obtain by replacing the given OMQ with its existential rewriting, constructing the (then finite) chase, and then applying the results for UCQs discussed above. In the lower bounds, we also first transition to the existential rewriting. We then give parameterized Turing-reductions from CQ-based OMQs to UCQ-based ones and, in a second step, from CQs without ontologies to CQ-based OMQs. Our constructions are strongly inspired by those from [10,11] and partially reuse results from there.

It is interesting to note that the ontology interacts with all three measures from Theorem 1, that is, there are classes $\mathbb{C} \subseteq (\mathbb{G}, \mathbb{UCQ})$ such that the treewidths of CQs in the OMQs in $\mathbb{C}$ are unbounded while the treewidths of CQs in $\mathsf{cl}^{\exists}_{\mathsf{CM}}(\mathbb{C})$ are bounded, and likewise for contract treewidth and dominating starsize.

Regarding ramifications for description logic, we recall that every ontology formulated in $\mathcal{ELIH}$ can be transformed into a well-known normal form that avoids syntactic nesting of concept constructors, and that $\mathcal{ELIH}$-ontologies in normal form fall within $\mathbb{G}$. Consequently, Theorem 1 still holds when $\mathbb{G}$ is replaced with $\mathcal{ELIH}$, assuming normal form. The case where $\mathcal{ELIH}$-ontologies are not in normal form can be covered by an easy modification of our proofs. In fact, we could extend Theorem 1 to ontologies that are sets of frontier-guarded TGDs in which the treewidth of the rule bodies is bounded by a constant. This also captures $\mathcal{ELIH}$-ontologies that are not in normal form.

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

References

1. Barceló, P., Dalmau, V., Feier, C., Lutz, C., Pieris, A.: The limits of efficiency for open- and closed-world query evaluation under guarded TGDs. In: Proc. of PODS (2020)
2. Barceló, P., Feier, C., Lutz, C., Pieris, A.: When is ontology-mediated querying efficient? In: LICS. pp. 1–13 (2019)

3. 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)
4. Bienvenu, M., Manière, Q., Thomazo, M.: Counting query answers over dl-lite knowledge base. In: Proc. of IJCAI (2020)
5. Bienvenu, M., Ortiz, M.: Ontology-mediated query answering with data-tractable description logics. In: Proc. of Reasoning Web. LNCS, vol. 9203, pp. 218–307. Springer (2015)
6. Calì, A., Gottlob, G., Kifer, M.: Taming the infinite chase: Query answering under expressive relational constraints. J. Artif. Intell. Res. **48**, 115–174 (2013)
7. Calvanese, D., Corman, J., Lanti, D., Razniewski, S.: Counting query answers over dl-lite knowledge base. In: Proc. of IJCAI (2020)
8. Calvanese, D., Giacomo, G.D., Lembo, D., Lenzerini, M., Rosati, R.: Data complexity of query answering in description logics. Artif. Intell. **195**, 335–360 (2013)
9. Calvanese, D., Giacomo, G.D., Lenzerini, M.: On the decidability of query containment under constraints. In: Proc. of PODS. pp. 149–158. ACM Press (1998)
10. Chen, H., Mengel, S.: A trichotomy in the complexity of counting answers to conjunctive queries. In: Proc. of ICDT. pp. 110–126 (2015)
11. Chen, H., Mengel, S.: Counting answers to existential positive queries: A complexity classification. In: Proc. of PODS. pp. 315–326 (2016)
12. Dalmau, V., Jonsson, P.: The complexity of counting homomorphisms seen from the other side. J. Theor. Comput. Sci. **329**(1-3), 315–323 (2004)
13. Dell, H., Roth, M., Wellnitz, P.: Counting answers to existential questions. In: Proc. of ICALP. pp. 113:1–113:15 (2019)
14. Durand, A., Mengel, S.: Structural tractability of counting of solutions to conjunctive queries. J. Theory Comput. Syst. **57**(4), 1202–1249 (2015)
15. Eiter, T., Gottlob, G., Ortiz, M., Simkus, M.: Query answering in the description logic horn-. In: Proc. of JELIA. LNCS, vol. 5293, pp. 166–179. Springer (2008)
16. Glimm, B., Lutz, C., Horrocks, I., Sattler, U.: Conjunctive query answering for the description logic SHIQ. J. Artif. Intell. Res. **31**, 157–204 (2008)
17. Kostov, B., Kremen, P.: Count distinct semantic queries over multiple linked datasets. OJSW **5**(1), 1–11 (2018)
18. Kostylev, E.V., Reutter, J.L.: Complexity of answering counting aggregate queries over dl-lite. J. Web Semant. **33**, 94–111 (2015)
19. Krisnadhi, A., Lutz, C.: Data complexity in the $\mathcal{EL}$ family of dls. In: Proc. of DL. CEUR Workshop Proceedings, vol. 250. CEUR-WS.org (2007)
20. Krötzsch, M., Rudolph, S., Hitzler, P.: Complexities of horn description logics. ACM Trans. Comput. Log. **14**(1), 2:1–2:36 (2013)
21. Lutz, C.: The complexity of conjunctive query answering in expressive description logics. In: Proc. of IJCAR. LNCS, vol. 5195, pp. 179–193. Springer (2008)
22. Lutz, C., Sabellek, L.: Ontology-mediated querying with the description logic $\mathcal{EL}$: Trichotomy and linear datalog rewritability. In: Proc. of IJCAI. pp. 1181–1187. ijcai.org (2017)
23. Lutz, C., Wolter, F.: The data complexity of description logic ontologies. Logical Methods in Computer Science **13**(4) (2017)
24. Rosati, R.: The limits of querying ontologies. In: Proc. of ICDT. LNCS, vol. 4353, pp. 164–178. Springer (2007)