Relaxed Regular Path Queries in Lightweight DLs (Extended Abstract)*

Oliver Fernández Gil and Anni-Yasmin Turhan

Theoretical Computer Science, TU Dresden, Germany
firstname.lastname@tu-dresden.de

Regular path queries (RPQs) is a well-investigated query language that dates back to the early 90's, where its capabilities to navigate graph-structured data attracted much attention in research on semistructured data and graph databases [10,6]. This interest was revived in recent years, since in many application areas data is graph-structured and represented in graph database models. Notable examples of applications of RPQs are querying biological networks, the semantic web and social networks. Moreover, RPQs and its extensions are part of SPARQL, which is the standard language recommended by the W3C to query RDF data. Formally, a graph database consists of a labeled directed graph, where edge labels correspond to binary predicates stating relations between data items. A RPQ consists of a regular language over these labels, and retrieves pairs of data items (a, b) that are connected by paths complying to the specified regular language. The extension of *two-way* RPQs (2RPQ) allows to traverse edges backwards, and the more expressive language of *conjunctive* 2RPQs (C2RPQ) allows conjunctions of 2RPQs that can share variables.

In scenarios where a RPQ yields no answers over a particular database, it can be useful to relax the query to retrieve more than the classical answers, i.e., pairs that are connected by paths that are "similar enough" to the paths required by the query. This can be practical to provide feasible alternatives in applications where data is gathered automatically from heterogeneous data sources and exact semantics need not yield the expected results. Similarly, in applications where the data is irregular and evolves in structure and content, it can be hard for users to have full knowledge of its vocabulary and structure and queries that approximate/relax the set of answers may be helpful.

Several approaches have been considered to address this problem, as for instance, [8,9,7,12]. In particular, [7] proposes an elegant and tractable solution that uses a weighted finite-state transducers to define the approximation semantics. Roughly speaking, such a transducer is a mechanism that transforms input words into corresponding output words, and computes a weight quantifying the cost of the transformation. The idea is to use a transducer as a means to specify which paths are allowed to be considered approximations/distortions of the "ideal" paths specified by the query, and to specify their distortion costs.

* Supported by the DFG in grant 389792660 as part of TRR 248 (https://perspicuous-computing.science) and grant BA 1122/20-1.

Approximate answers are then tuples (a, b, η), where η is the minimal cost of distorting a path that complies with the query, into a path leading from a to b.

Path queries have also been investigated for ontology-mediated query answering (OMQA), in which semantic knowledge provided in a background ontology is used to enrich the data. Ontologies are often formulated in a description logic (DL) which can be used to represent the conceptual knowledge of an application domain in a structured and formally well-understood way. In contrast to query answering over a (graph) database, OMQA usually adopts the open world assumption where all possible models of the ontology and the data are considered when computing the answers. Answering conjunctive 2-way regular path queries (C2RPQs) has been studied for very expressive DLs [4], and for the $\mathcal{EL}$ and DL-Lite families of lightweight DLs [3,13,2]. However, approaches for query answering under approximate semantics in the OMQA setting are scarce. There is prior work on the simple case of instance queries [5] and on C2RPQs in the restricted setting of acyclic ontologies using RDFs schema [12] or non-gradual variants of conjunctive queries [11]. The goal of this paper is to define approximate semantics for answering C2RPQs in DLs and to devise computation algorithms for answering them in the DLs $\mathcal{ELH}$ and DL-Lite$_\mathcal{R}$.

Our contributions are threefold. First, we extend the known transducer-based approximate semantics from RPQs to the more general query language of C2RPQs in the graph database setting. In contrast to RPQs, C2RPQs may contain quantified variables and more than one query atom. This requires to regard several matchings of the quantified variables and to combine the costs of the distortions of each query atom. Second, we consider the setting of OMQA and define approximate semantics for answering C2RPQs over DL ontologies. We define the notion of *certain approximate answers* as a generalization of the classical *certain answers*. Third, we investigate two reasoning problems for certain approximate answers. We start with the decision problem that asks, given a threshold value τ and a tuple $\bar{a}$, is $\bar{a}$ a certain approximate answer with approximation cost of at most τ? Then we consider the problem of computing the exact approximation cost of $\bar{a}$. For 2RPQs, we devise a polynomial time algorithm that can be used to solve both problems. Regarding C2RPQs, we prove that a) both problems can be solved in polynomial time in *data complexity*, b) the decision problem is in NExpTime in combined complexity, and c) we provide a double exponential time algorithm (in combined complexity) to compute the approximation cost.

Several of our upper bounds are not tight, in the sense that they do not match the hardness results inherited from the classical case. The exact data complexity for DL-Lite$_\mathcal{R}$ and the combined complexity of C2RPQs in both DLs remain open: NL$-$PTime and PSpace$-$NexpTime (2ExpTime), respectively. Trying to close (or reduce) these gaps constitutes ongoing work, where the case concerning the combined complexity appears to represent the biggest challenge. On the one hand, we are analysing whether the PSpace rewriting procedure proposed in [2] can really be adapted to answer C2RPQs under approximate semantics. On the other hand, we are investigating if the approximation mechanism allows to

simulate more difficult query answering problems, like for instance, answering nested 2RPQs which is ExpTime-hard in both DLs [1].

Regarding future research, we plan to consider other semi-rings and study which approximation patterns they allow to express, as well as the impact on the computational complexity of the considered problems.

References

1. Bienvenu, M., Calvanese, D., Ortiz, M., Simkus, M.: Nested regular path queries in description logics. In: Baral, C., Giacomo, G.D., Eiter, T. (eds.) Principles of Knowledge Representation and Reasoning: Proceedings of the Fourteenth International Conference, KR 2014, Vienna, Austria, July 20-24, 2014. AAAI Press (2014)
2. Bienvenu, M., Ortiz, M., Simkus, M.: Regular path queries in lightweight description logics: Complexity and algorithms. J. Artif. Intell. Res. 53, 315–374 (2015)
3. 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. Autom. Reasoning 39(3), 385–429 (2007)
4. Calvanese, D., Eiter, T., Ortiz, M.: Answering regular path queries in expressive description logics via alternating tree-automata. Inf. Comput. 237, 12–55 (2014)
5. Ecke, A., Peñaloza, R., Turhan, A.Y.: Similarity-based relaxed instance queries. Journal of Applied Logic 13(4, Part 1), 480–508 (2015), special Issue for the Workshop on Weighted Logics for AI 2013
6. Florescu, D., Levy, A.Y., Suciu, D.: Query containment for conjunctive queries with regular expressions. In: Mendelzon, A.O., Paredaens, J. (eds.) Proceedings of the Seventeenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, June 1-3, 1998, Seattle, Washington, USA. pp. 139–148. ACM Press (1998)
7. Grahne, G., Thomo, A.: Regular path queries under approximate semantics. Ann. Math. Artif. Intell. 46(1-2), 165–190 (2006)
8. Jagadish, H.V., Mendelzon, A.O., Milo, T.: Similarity-based queries. In: Yannakakis, M. (ed.) Proceedings of the Fourteenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, May 22-25, 1995, San Jose, California, USA. pp. 36–45. ACM Press (1995)
9. Kanza, Y., Sagiv, Y.: Flexible queries over semistructured data. In: PODS. ACM (2001)
10. Mendelzon, A.O., Wood, P.T.: Finding regular simple paths in graph databases. SIAM J. Comput. 24(6), 1235–1258 (1995)
11. Peñaloza, R., Thost, V., Turhan, A.Y.: Query answering for rough $\mathcal{EL}$ ontologies. In: Thielscher, M., Toni, F. (eds.) Proceedings of 16. International Conference on Principles of Knowledge Representation and Reasoning (KR 2018). pp. 399–408. AAAI (2018)
12. Poulovassilis, A., Selmer, P., Wood, P.T.: Approximation and relaxation of semantic web path queries. J. Web Semant. 40, 1–21 (2016)
13. Stefanoni, G., Motik, B., Krötzsch, M., Rudolph, S.: The complexity of answering conjunctive and navigational queries over OWL 2 EL knowledge bases. J. Artif. Intell. Res. 51, 645–705 (2014)