Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI-23)
Sister Conferences Best Papers Track

Finite Entailment of UCRPQs over $\mathcal{ALC}$ Ontologies (Extended Abstract)*

Víctor Gutiérrez-Basulto[1] , **Albert Gutowski**[2] , **Yazmín Ibáñez-García**[1] , **Filip Murlak**[2]
[1]Cardiff University, UK
[2]University of Warsaw, Poland
{a.gutowski, f.murlak}@mimuw.edu.pl, {gutierrezbasultov, ibanezgarciay}@cardiff.ac.uk

Abstract

We investigate the problem of *finite* entailment of ontology-mediated queries. We consider the expressive query language, unions of conjunctive regular path queries (UCRPQs), extending the well-known class of unions of conjunctive queries, with regular expressions over roles. We look at ontologies formulated using the description logic $\mathcal{ALC}$, and show a tight 2EXPTIME upper bound for finite entailment of UCRPQs.

1 Introduction

At the intersection of knowledge representation and database theory lies the fundamental problem of *ontology-mediated query entailment (OMQE)*, where the background knowledge provided by an ontology is used to enrich the answers to queries posed to databases. In this context, description logics (DLs) are a widely accepted family of logics used to formulate ontologies. Description logics provide the logical basis of the web ontology language OWL 2, the medical ontology SNOMED CT, and the NCI thesaurus. Furthermore, DLs provide a good trade-off between expressivity and computational complexity. By now, the OMQE problem under the unrestricted semantics (reasoning over arbitrary models) is well understood for various query languages and DLs [Schneider and Simkus, 2020]. In contrast, for the *finite* OMQE problem, where one is interested in reasoning over finite models only, the overall landscape is rather incomplete. However, in recent years, the study of finite OMQE has been gaining traction, considering both lightweight and expressive DLs and (mostly) unions of conjunctive queries [Rosati, 2008; Ibáñez-García *et al.*, 2014; Rudolph, 2016; Gogacz *et al.*, 2018; Gogacz *et al.*, 2019; Danielski and Kieronski, 2019; Gogacz *et al.*, 2020; Bednarczyk and Kieroński, 2022].

With this in mind, in our KR2022 paper [Gutiérrez-Basulto *et al.*, 2022] we considered the problem of finite OMQE with unions of conjunctive regular path queries (UCRPQs) as the query language. UCRPQs [Florescu *et al.*, 1998; Calvanese *et al.*, 2000] are a powerful navigational query language for graph databases in which one can express that two entities are related by a path of edges that can be specified by a regular language over binary relations. For example, using UCRPQs, from a genealogical graph database one can retrieve all the ancestors of a person. So, UCRPQs extend unions of conjunctive queries (UCQs) with atoms that might contain regular expressions that traverse the edges of the database. Indeed, path navigation is included in the query language XPath 2.0 for XML data, and it is also present in the SPARQL 1.1 query language for RDF data through the property path feature. Given the resemblance of instance data stored in so-called ABoxes in DLs to graph-like data, several investigations on unrestricted entailment of various types of navigational query languages mediated by DL ontologies have been carried out [Stefanoni *et al.*, 2014; Calvanese *et al.*, 2014; Bienvenu *et al.*, 2015; Gutiérrez-Basulto *et al.*, 2018; Gogacz *et al.*, 2019; Bednarczyk and Rudolph, 2019], yielding algorithmic approaches and optimal complexity bounds. For finite entailment of regular path queries mediated by DL ontologies, there are only undecidability results available [Rudolph, 2016]. The most relevant positive news are the decidability and computational complexity results by Danielski and Kieronski (2019) and Gogacz *et al.* (2020) on finite entailment of conjunctive queries with transitive closure over roles mediated by expressive DL ontologies.

We focus on ontologies formulated using the description logic $\mathcal{ALC}$, which is the most basic Boolean-complete DL [Baader *et al.*, 2017]. Note that entailment of UCRPQs over $\mathcal{ALC}$ ontologies is not *finitely controllable*, i.e. finite and unrestricted entailment do not coincide as it is *not* the case that for any $\mathcal{ALC}$ knowledge base $\mathcal{K}$ and any UCRPQ φ, it holds that $\mathcal{K}$ entails φ over all (unrestricted) models iff $\mathcal{K}$ entails φ over all finite models.

Example 1. *Consider the ABox* $\mathcal{A} = \emptyset$*, and TBox* $\mathcal{T} = \{\top \sqsubseteq \exists r.\top\}$*, together with the RPQ* $\varphi = \exists x.r^+(x,x)$ *for some role* r*. We have that* $(\mathcal{T},\mathcal{A}) \not\models \varphi$*, but for every finite model* $\mathcal{I}$ *of* $(\mathcal{T},\mathcal{A})$*, we have* $\mathcal{I} \models \varphi$*.*

Therefore, when the represented world is assumed to be finite, we cannot reuse existing complexity bounds or algorithmic approaches to UCRPQ entailment. It has been acknowledged that developing specialised algorithms for finite reasoning brings additional challenges as certain regularities present in infinite models do not occur in finite ones [Hod-

*The full version of this extended abstract was published in the proceedings of the 19th International Conference on Principles of Knowledge Representation and Reasoning (KR 2022), where it received the *Best Student Paper Award*.

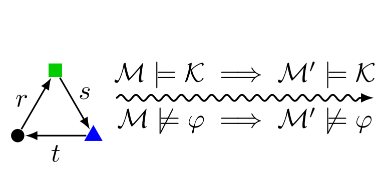

Figure 1: Colored Blocking

kinson and Otto, 2003]. From a usability perspective, the suitability of the finiteness assumption depends on the potential applications. A particular interest for navigational queries comes from bioinformatics and cheminformatics [Lysenko *et al.*, 2016; Cook *et al.*, 2016; Galgonek *et al.*, 2016; Hu *et al.*, 2015; Rajabi and Sanchez-Alonso, 2021; Chen *et al.*, 2020]. For instance, experts often need to find associations between entities in protein, cellular, drug, and disease networks (represented as graph databases), so that e.g. gene-disease-drug associations (corresponding to paths in the database) can be discovered for developing new treatment methods. In this type of applications, databases and the models they represent are clearly meant to be finite. Importantly, biochemical networks contain complex motifs involving e.g. *cycles* or cliques. This type of patterns can be described using UCRPQs, however, without the finiteness assumption these patterns could be disregarded as the associated query might not be entailed when reasoning over all models (including infinite ones).

2 Contribution

The main technical contribution of our KR 2022 paper is a dedicated automata-based method for finite entailment of UCRPQs over $\mathcal{ALC}$ ontologies, providing an optimal complexity upper bound. More precisely, we obtain the following result, where the matching lower bound is inherited from [Ortiz and Simkus, 2014].

Theorem 1. *Finite entailment of UCRPQs over* $\mathcal{ALC}$ *ontologies is* 2EXPTIME*-complete.*

Theorem 1 (and its proof) is important in two ways. First, it provides a key step towards delimiting the decidability boundary of finite OMQE with navigational queries. It substantially reduces the gap between what is known to be decidable and what is known to be undecidable: prior work showed that finite entailment of 2RPQs in $\mathcal{ALCIOF}$ (extending $\mathcal{ALC}$with inverse relations and nominals) is undecidable [Rudolph, 2016]. Second, it introduces a novel approach to finite OMQE that amalgamates extensions of existing model theoretical and automata-based techniques.

Previous works on finite query entailment over DL ontologies [Gogacz *et al.*, 2018; Gogacz *et al.*, 2019; Gogacz *et al.*, 2020] have already made substantial progress on the development of a variety of techniques. Two techniques that will play a key role in our framework are the following. The *colored blocking principle* [Gogacz *et al.*, 2018] allows us to construct finite counter-models for an ontology $\mathcal{K}$ and a query φ by expanding cycles in the model, cf. Figure 1. In a nutshell, this technique looks for sufficiently large neighborhoods of nodes (seeing models as graphs) and uses coloring of nodes to distinguish them. Another key technique is a decomposition of arbitrarily complex finite models into more regular structures, which guarantees that it is enough to solve the finite entailment problem on these simpler structures. More precisely, Gogacz *et al.* (2020) introduce a decomposition of finite models into tree-like structures whose nodes are labelled with strongly connected components (SCC)[1], see Figure 2. Importantly, considering such decompositions allows us to use standard machinery, such as type elimination or tree automata, to decide finite entailment.

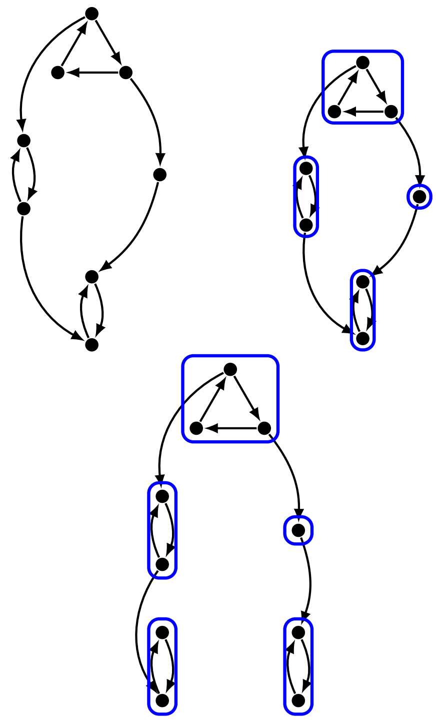

Figure 2: Tree-like decomposition with nodes labelled with strongly connected components

Given all these advancements, one might wonder why they are not enough for solving the finite entailment problem of UCRPQs over $\mathcal{ALC}$ ontologies. The answer is a simple one. In all previous works, the query language is weaker than UCRPQs. In particular, full regular expressions are disallowed in queries. This restriction makes the problem easier to tackle: to solve the entailment problem, it is enough to reduce to the case where DL roles are treated separately. In other words, one does not have to reason over graphs (models) in which edges might be labelled with multiple roles, but instead it is enough to separately consider graphs in which all edges are labelled with the same role. Note that this is not possible for UCRPQs as in this case atoms contain arbitrary regular expressions e.g. union of various roles under Kleene star. So, one unavoidably has to consider graphs labelled with multiple roles. In particular, in a path one has to consider different valid sequences of roles labelling edges.

[1]Recall that a directed graph is said to be strongly connected if every node is reachable from every other node

3 Approach

To overcome these challenges, we developed a novel model decomposition that only depends on the input query and model, which allows to devise a recursive algorithm for finite entailment of UCRPQs over $\mathcal{ALC}$ ontologies. This algorithm effectively makes use of tree-shaped SCC decompositions and colored blocking. At the heart of our decomposition there is a stratification of interpretations induced by the deterministic finite automaton underlying the UCRPQ. This stratification builds upon the so-called *tape construction*, previously used to efficiently evaluate queries in the extension of XPath 1.0 where arbitrary regular expressions may appear as path expressions [Bojańczyk and Parys, 2011]. To realize the tape construction, we work with a representation of UCRPQs by means of a semiautomaton $\mathcal{B}$ [Ginzburg, 1968]. A semiautomaton essentially is a deterministic finite automaton without initial and final states; a run of a semiautomaton over a word is defined just like for a finite automaton, except that it can begin in any state and there is no notion of accepting runs. We then rely on an expansion of $\mathcal{B}$, that allows us to trace runs of $\mathcal{B}$ that begin in all possible states, on all infixes of the input word. For a graphic example of a semiautomaton expansion, see Figure 3 where the upper part is the query represented by an automaton and the part below is its expansion on the depicted input word. Notice that two threads that begin at different levels can meet at the same level somewhere along the run; if this happens they remain equal until the end of the run. Also, threads can be born in the middle of a run of the expansion, but they never disappear. A crucial property of threads is that they are non-decreasing.

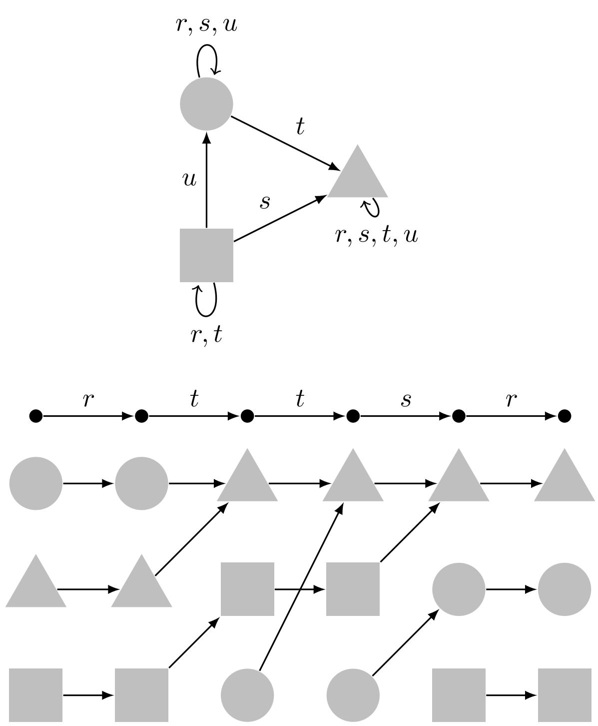

Figure 3: Tape construction

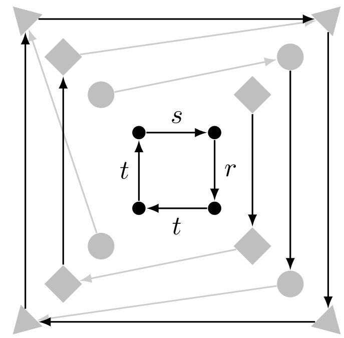

Figure 4: Runs of the automaton on all paths of a model (black square in the center)

We also make interpretations aware of the expansion by enriching their paths with possible runs of the expansion; see Figure 4 for an example, where the black square in the middle is the given interpretation. We can then obtain a decomposition of an interpretation by associating its edges with levels induced by the transitions of the expansion. For example, Figure 5 shows a decomposition of the interpretation in Figure 4 into three different levels, where each level is determined using the black transitions depicted in Figure 4.

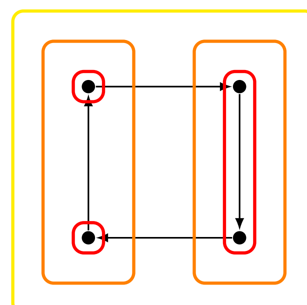

Figure 5: Decomposition of an interpretation in levels

In a similar fashion, we make CRPQs aware of levels. With this at hand, we develop a recursive algorithm that makes use of existing techniques, such a decomposition of models into trees of strongly connected components and colored blocking. In particular, we tackle finite entailment by eliminating the lowest level from a query and from an interpretation, and then recursively solving the simpler problem. At each step of this process, we should be able to arrange solutions to simpler problems in a hierarchical way so that we can reason over them. To this aim, we consider a variant of entailment that includes an *environment*, which provides the necessary information to position the arranged solutions to simpler problems in the context of larger interpretations. To better keep track of the complexity of our recursive method, we introduce a modification of the entailment problem modulo environment in which we look at a particular type of finite models: (ℓ, ℓ')-*models*, which are models with edges of levels ℓ or higher that are 'consistent' w.r.t. queries referring to edges of level ℓ' or higher. We solve the problem of finding (ℓ, ℓ')-models recursively by increasing ℓ and ℓ' in an alternating way, until both reach the maximum level $n + 1$, with n the number of states of $\mathcal{B}$. This provides a solution to finite entailment modulo environment, and thus to standard finite entailment as well.

4 Outlook

We provided first positive results on finite entailment of CRPQs over DLs ontologies. Along the way, we extend the existing toolkit for finite model reasoning in the presence of DL ontologies by introducing a novel decomposition, which supports recursive algorithms for finite entailment.

This work leaves open several interesting questions. It would be interesting to consider other ontology languages besides $\mathcal{ALC}$. We believe that our method can be adapted to allow inverses, nominals or counting. However, some combinations of these constructs are problematic, e.g. inverses and counting (already functionality). Regarding more expressive query languages, the natural next step is to consider *two-way* CRPQs. Our current approach relies on the fact that information only flows forward, and it is not clear whether it can be adapted to deal with queries that can go back. Besides this, it would also be interesting to develop techniques that are better suited for implementation, since our current approach relies on automata- and type-based techniques which are always worst-case complexity.

Acknowledgements

Albert Gutowski and Filip Murlak were supported by Poland's NCN grant 2018/30/E/ST6/00042.

References

[Baader *et al.*, 2017] Franz Baader, Ian Horrocks, Carsten Lutz, and Ulrike Sattler. *An Introduction to Description Logic*. Cambridge University Press, 2017.

[Bednarczyk and Kieroński, 2022] Bartosz Bednarczyk and Emanuel Kieroński. Finite entailment of local queries in the $\mathcal{Z}$ family of description logics. In *Proc. of AAAI-2022*, 2022.

[Bednarczyk and Rudolph, 2019] Bartosz Bednarczyk and Sebastian Rudolph. Worst-case optimal querying of very expressive description logics with path expressions and succinct counting. In *IJCAI*, pages 1530–1536. ijcai.org, 2019.

[Bienvenu *et al.*, 2015] Meghyn Bienvenu, Magdalena Ortiz, and Mantas Simkus. Regular path queries in lightweight description logics: Complexity and algorithms. *J. Artif. Intell. Res.*, 53:315–374, 2015.

[Bojańczyk and Parys, 2011] Mikolaj Bojańczyk and Pawel Parys. XPath evaluation in linear time. *J. ACM*, 58(4):17:1–17:33, 2011.

[Calvanese *et al.*, 2000] Diego Calvanese, Giuseppe De Giacomo, Maurizio Lenzerini, and Moshe Y. Vardi. Containment of conjunctive regular path queries with inverse. In *KR*, pages 176–185. Morgan Kaufmann, 2000.

[Calvanese *et al.*, 2014] Diego Calvanese, Thomas Eiter, and Magdalena Ortiz. Answering regular path queries in expressive description logics via alternating tree-automata. *Inf. Comput.*, 237:12–55, 2014.

[Chen *et al.*, 2020] Chuming Chen, Hongzhan Huang, Karen E. Ross, Julie E. Cowart, Cecilia N. Arighi, Cathy H. Wu, and Darren A. Natale. Protein ontology on the semantic web for knowledge discovery. *Scientific Data*, 7(1):337, 2020.

[Cook *et al.*, 2016] Charles E Cook, Mary Todd Bergman, Robert D Finn, Guy Cochrane, Ewan Birney, and Rolf Apweiler. The European Bioinformatics Institute in 2016: Data growth and integration. *Nucleic acids research*, 44:D20–D26, 2016.

[Danielski and Kieronski, 2019] Daniel Danielski and Emanuel Kieronski. Finite satisfiability of unary negation fragment with transitivity. In *MFCS*, volume 138 of *LIPIcs*, pages 17:1–17:15, 2019.

[Florescu *et al.*, 1998] Daniela Florescu, Alon Y. Levy, and Dan Suciu. Query containment for conjunctive queries with regular expressions. In *PODS*, pages 139–148. ACM Press, 1998.

[Galgonek *et al.*, 2016] Jakub Galgonek, Tomáš Hurt, Vendula Michlíková, Petr Onderka, Jan Schwarz, and Jiří Vondrášek. Advanced SPARQL querying in small molecule databases. *Journal of Cheminformatics*, 8(1):31, 2016.

[Ginzburg, 1968] Abraham Ginzburg. *Algebraic Theory of Automata*. ACM Monograph Series. Academic Press, 1968.

[Gogacz *et al.*, 2018] Tomasz Gogacz, Yazmin Ibáñez-García, and Filip Murlak. Finite query answering in expressive description logics with transitive roles. In *Proc. of KR-18*, 2018.

[Gogacz *et al.*, 2019] Tomasz Gogacz, Víctor Gutiérrez-Basulto, Yazmín Ibáñez-García, Jean Christoph Jung, and Filip Murlak. On finite and unrestricted query entailment beyond SQ with number restrictions on transitive roles. In *IJCAI*, pages 1719–1725. ijcai.org, 2019.

[Gogacz *et al.*, 2020] Tomasz Gogacz, Víctor Gutiérrez-Basulto, Albert Gutowski, Yazmín Ibáñez-García, and Filip Murlak. On finite entailment of non-local queries in description logics. In *KR*, pages 424–433, 2020.

[Gutiérrez-Basulto *et al.*, 2018] Víctor Gutiérrez-Basulto, Yazmín Angélica Ibáñez-García, and Jean Christoph Jung. Answering regular path queries over SQ ontologies. In *AAAI*, pages 1845–1852. AAAI Press, 2018.

[Gutiérrez-Basulto *et al.*, 2022] Víctor Gutiérrez-Basulto, Albert Gutowski, Yazmín Ibáñez-García, and Filip Murlak. Finite entailment of ucrpqs over ALC ontologies. In *KR*, 2022.

[Hodkinson and Otto, 2003] Ian M. Hodkinson and Martin Otto. Finite conformal hypergraph covers and gaifman cliques in finite structures. *Bull. Symb. Log.*, 9(3):387–407, 2003.

[Hu *et al.*, 2015] Wei Hu, Honglei Qiu, and Michel Dumontier. Link analysis of life science linked data. In *ISWC (2)*, 2015.

[Ibáñez-García *et al.*, 2014] Yazmín Angélica Ibáñez-García, Carsten Lutz, and Thomas Schneider. Finite model reasoning in Horn description logics. In *KR*, 2014.

[Lysenko *et al.*, 2016] Artem Lysenko, Irina A. Roznovăţ, Mansoor Saqi, Alexander Mazein, Christopher J. Rawlings, and Charles Auffray. Representing and querying disease networks using graph databases. *BioData Mining*, 9(1):23, 2016.

[Ortiz and Simkus, 2014] Magdalena Ortiz and Mantas Simkus. Revisiting the hardness of query answering in expressive description logics. In *RR*, volume 8741 of *Lecture Notes in Computer Science*, pages 216–223. Springer, 2014.

[Rajabi and Sanchez-Alonso, 2021] Enayat Rajabi and Salvador Sanchez-Alonso. Knowledge discovery using SPARQL property path: The case of disease data set. *Journal of Information Science*, 47(5):677–687, 2021.

[Rosati, 2008] Riccardo Rosati. Finite model reasoning in DL-Lite. In *ESWC*, 2008.

[Rudolph, 2016] Sebastian Rudolph. Undecidability results for database-inspired reasoning problems in very expressive description logics. In *KR*, pages 247–257. AAAI Press, 2016.

[Schneider and Simkus, 2020] Thomas Schneider and Mantas Simkus. Ontologies and data management: A brief survey. *Künstliche Intell.*, 34(3):329–353, 2020.

[Stefanoni *et al.*, 2014] Giorgio Stefanoni, Boris Motik, Markus Krötzsch, and Sebastian Rudolph. The complexity of answering conjunctive and navigational queries over OWL 2 EL knowledge bases. *J. Artif. Intell. Res.*, 51:645–705, 2014.