Foundations of ontology-based data access under bag semantics

Charalampos Nikolaou a, *, Egor V. Kostylev a, George Konstantinidis b, *a, Mark Kaminski a, Bernardo Cuenca Grau a, Ian Horrocks a

a Department of Computer Science, University of Oxford, UK
b Department of Electronics and Computer Science, University of Southampton, UK

Abstract

Ontology-based data access (OBDA) is a popular approach for integrating and querying multiple data sources by means of a shared ontology. The ontology is linked to the sources using mappings, which assign to ontology predicates views over the data. The conventional semantics of OBDA is set-based—that is, the extension of the views defined by the mappings does not contain duplicate tuples. This treatment is, however, in disagreement with the standard semantics of database views and database management systems in general, which is based on bags and where duplicate tuples are retained by default. The distinction between set and bag semantics in databases is very significant in practice, and it influences the evaluation of aggregate queries.

In this article, we propose and study a bag semantics for OBDA which provides a solid foundation for the future study of aggregate and analytic queries. Our semantics is compatible with both the bag semantics of database views and the set-based conventional semantics of OBDA. Furthermore, it is compatible with existing bag-based semantics for data exchange recently proposed in the literature. We show that adopting a bag semantics makes conjunctive query answering in OBDA coNP-hard in data complexity. To regain tractability of query answering, we consider suitable restrictions along three dimensions, namely, the query language, the ontology language, and the adoption of the unique name assumption. Our investigation shows a complete picture of the computational properties of query answering under bag semantics over ontologies in the DL-Lite family.

© 2019 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1. Introduction

Ontology-based data access (OBDA) is an increasingly popular approach to enable uniform access to multiple data sources with diverging schemas [2–7]. An ontology in OBDA is represented using a fragment of first-order logic and provides a unifying conceptual model for the data sources. The ontology is linked to the schema of each data source by global-as-view (GAV) mappings [8], which declaratively assign views over the data to predicates in the vocabulary of the ontology. Users of OBDA systems are typically unaware of the details of the source schemas or the mappings, and access the data by means of...
(conjunctive) queries formulated using only the vocabulary of the ontology. The answers to a query are those that logically follow from the union of the ontology and the materialisation over the sources of the views defined by the mappings.

The formalism of choice for representing ontologies in OBDA is the description logic DL-Lite$_R$ [9], which underpins the Web ontology language OWL 2 QL [10]. DL-Lite$_R$ is a language designed to ensure that all input queries to the OBDA system are first-order rewritable—that is, the answers to every input query can be obtained by first reformulating the query as a set of relational queries over the source schemas, and then evaluating the reformulated queries over the source data [2]. In practice, such reformulation involves two steps known as rewriting and unfolding [8,11]. In the rewriting step, the original query is transformed into a first-order query that captures the relevant information in the ontology; in turn, in the unfolding step, the query computed in the rewriting step is further rewritten as a set of relational queries over the schemas of the sources using the relevant mappings.

OBDA has received a great deal of attention in recent years. Researchers have studied the limits of first-order rewritability in ontology languages [9,12], established bounds on the size of rewritings [13,14], developed optimisation techniques [15–17], and implemented systems well-suited for real-world applications [3,5].

Example 1. Consider a music encyclopedia that collects metadata about music records and makes it available to the public. The encyclopedia gathers data from record labels, which are maintained in separate relational tables. For instance, it contains the table Columbia(art_nm, r_title, r_year, r_date, r_loc) that provides the names of the records (r_title) artists (art_nm) have cut on the record label Columbia together with their release years (r_year), recording dates (r_date), and locations (r_loc). In addition, we consider the table Verve_Wind(name, title, year) that provides information about the names of the records (title) artists (name) playing wind instruments have cut on Verve (for the sake of illustration, we assume that records by Verve are organised in different tables according to the type of instrument played by the lead performer; in our running example, we restrict ourselves to wind instruments). The following is a relational database instance $D_{ex}$ providing information about the records that trumpeter Miles Davis and pianist Keith Jarrett have cut on these two labels.

| Columbia: | art_nm | r_title           | r_year | r_date | r_loc    |
|----------|--------|-------------------|--------|--------|----------|
| M. Davis | Kind of Blue | 1959 | 2/3/59 and 22/4/59 | New York |
| M. Davis | A Tribute to Jack Johnson | 1971 | 7/4/70 | New York |
| K. Jarrett | Expectations | 1972 | 5/4/72–27/4/72 | New York |

| Verve_Wind: | name | title | year |
|------------|------|-------|------|
| M. Davis   | Ascenseur pour l’Échafaud | 1958 |

To integrate this data, the music encyclopedia relies on a DL-Lite$_R$ ontology with TBox $T_{ex}$, which defines unary predicates, called concepts, such as Musician, WindPlayer, and Record, and binary predicates, called roles, such as hasMusician. TBox $T_{ex}$ describes the meaning of these predicates using the following axioms, called inclusions:

WindPlayer $\sqsubseteq$ Musician, saying that every player of a wind instrument is a musician,

Record $\sqsubseteq$ 3hasMusician, saying that every record is associated to some musician.

The extension of the concepts and roles in the ontology based on the data in Columbia and Verve_Wind tables is determined by a set of GAV mapping assertions of the form $\Phi(x) \leadsto A(x)$ or $\Psi(x, y) \leadsto P(x, y)$, where $\Phi(x)$ and $\Psi(x, y)$ are SQL queries, $A$ is a concept, and $P$ is a role. Such mappings can also be seen as database views with names $A$ or $P$ and definitions $\Phi(x)$ or $\Psi(x, y)$, respectively. The following set $\Sigma_{ex}$ of mappings is used to populate concepts Musician, WindPlayer, and Record, as well as role hasMusician with data from $D_{ex}$:

$\sigma_1$: SELECT art_nm AS x FROM Columbia $\leadsto$ Musician(x),
$\sigma_2$: SELECT name AS x FROM Verve_Wind $\leadsto$ WindPlayer(x),
$\sigma_3$: SELECT r_title AS x FROM Columbia $\leadsto$ Record(x),
$\sigma_4$: SELECT title AS x FROM Verve_Wind $\leadsto$ Record(x),
$\sigma_5$: SELECT r_title AS x, art_nm AS y FROM Columbia $\leadsto$ hasMusician(x, y),
$\sigma_6$: SELECT title AS x, name AS y FROM Verve_Wind $\leadsto$ hasMusician(x, y).

Finally, user queries are formulated using only the vocabulary of $T_{ex}$, and users are typically unaware of the schema of $D_{ex}$ and the definition of the mappings. $\leftarrow$
The distinction between set and bag semantics in databases is significant in practice; in particular, it influences the evaluation of aggregate queries that combine various aggregation functions such as MIN, MAX, SUM, COUNT, and AVG with the grouping functionality provided in SQL by the GROUP BY construct.

The mismatch between the set semantics of OBDA and the bag semantics of database views manifests itself already in our running example.

Example 2. Consider TBox $T_{ex}$, mappings $\Sigma_{ex}$, and database instance $D_{ex}$ specified in Example 1. Consider also the query $q_{ex}(x) = \text{Musician}(x)$, which asks for all musicians. Under the conventional OBDA semantics, the virtual ABox $A_{ex}$ corresponding to $D_{ex}$ and $\Sigma_{ex}$ comprises the following assertions:

\begin{align*}
\text{Musician}(M. Davis), \text{Musician}(K. Jarrett), \text{WindPlayer}(M. Davis), \\
\text{Record}(\text{Kind of Blue}), \text{Record}(\text{A Tribute to Jack Johnson}), \text{Record}(\text{Expectations}), \\
\text{Record}(\text{Ascenseur pour l’Échafaud}), \\
\text{hasMusician}(\text{Kind of Blue, M. Davis}), \text{hasMusician}(\text{A Tribute to Jack Johnson, M. Davis}), \\
\text{hasMusician}(\text{Expectations, K. Jarrett}), \text{hasMusician}(\text{Ascenseur pour l’Échafaud, M. Davis}).
\end{align*}

The answers to $q_{ex}(x)$ over ontology $\langle T_{ex}, A_{ex} \rangle$ are given by the set $\{M. Davis, K. Jarrett\}$ since $M. Davis$ and $K. Jarrett$ are the only instances of concept Musician entailed by the ontology $\langle T_{ex}, A_{ex} \rangle$.

To compute these answers in practice, however, an OBDA system would exploit the first-order rewritability property of OBDA. In particular, it would first rewrite $q_{ex}(x)$ into the union of queries $\text{Musician}(x)$ and $\text{WindPlayer}(x)$ using inclusion $\text{WindPlayer} \sqsubseteq \text{Musician}$ in $T_{ex}$. Then, in a second step, the system would unfold each disjunct of the rewritten query into a query over the database $D_{ex}$ using the mappings in $\Sigma_{ex}$. The result is the SQL query $\Phi_{ex}(x)$ comprising the union of the SQL queries $\Phi_{\sigma_{1}}(x)$ and $\Phi_{\sigma_{2}}(x)$ mentioned on the left-hand side of mappings $\sigma_{1}$ and $\sigma_{2}$, respectively. The query $\Phi_{ex}(x)$ is finally evaluated directly over $D_{ex}$.

According to the semantics of OBDA, the answers to $\Phi_{ex}(x)$ over ontology $\langle T_{ex}, A_{ex} \rangle$ should coincide with the evaluation of SQL query $\Phi_{ex}(x)$ over $D_{ex}$. This is, however, not the case in our example. In particular, evaluating $\Phi_{ex}(x)$ over $D_{ex}$ yields a bag containing two occurrences of $M. Davis$ and one occurrence of $K. Jarrett$. This is because duplicates in the answers to SQL queries are kept by default unless duplicate elimination is explicitly requested by using the DISTINCT operator in the SELECT clause of a query.

This discrepancy between OBDA semantics and the semantics of database views may occur even if the TBox of the ontology is empty. In particular, in such a case the evaluation of $q_{ex}(x)$ over ABox $A_{ex}$ does not coincide with the evaluation of the rewritten query $\langle \Phi_{\sigma_{1}}(x) \rangle$ in this case) over $D_{ex}$.  

Example 2 suggests that the conventional approach to OBDA can faithfully represent only a subset of GAV mapping assertions—those whose SQL query contains the DISTINCT operator in the top-level SELECT clause.

In this paper, we propose and study a bag semantics for OBDA, which provides a solid foundation for future research on aggregate and analytic queries. Our semantics is compatible with (i) the bag semantics of database views; (ii) the set-based conventional semantics of OBDA; and (iii) the bag semantics recently proposed by Hernich and Kolaitis [20] in the context of data exchange.

1.1. Contributions and organisation

The contributions and organisation of this paper are as follows. In Section 3 we introduce the bag semantics of an OBDA setting $\langle T, \Sigma, D \rangle$ consisting of a DL-Lite$^R$ TBox $T$, a set of GAV mappings $\Sigma$, and a (bag) database instance $D$. We also define the notion of certain answers to conjunctive queries as well as the associated query answering problem.

In Section 4 we define the ontology language DL-Lite$^R_{B}$ and two of its natural fragments. A distinctive feature of DL-Lite$^R_{B}$ is that ABoxes are bags of assertions rather than sets. Syntactically, this language allows for the same TBoxes as DL-Lite$^R$, but their semantics also take multiplicities into account. We show that, as in the case of conventional OBDA, the certain answers to a query over an OBDA setting $\langle T, \Sigma, D \rangle$ can be characterised as those that logically follow from the union of the TBox $T$ and a virtual bag ABox $A_{\Sigma, D}$ representing the materialisation of the views defined by the mappings $\Sigma$ over the database $D$. As a result, the data complexity of OBDA query answering under bag semantics coincides with that of query answering over DL-Lite$^R_{B}$.

In Section 5 we then establish the relationship between bag and set semantics of DL-Lite$^R_{B}$ and DL-Lite$^R$, respectively. In particular, we show that, on the one hand, satisfiability checking in DL-Lite$^R_{B}$ reduces to satisfiability checking in DL-Lite$^R$ and, on the other hand, query answers under bag and set semantics coincide if multiplicities are ignored. There are, however, key properties of the conventional semantics of DL-Lite$^R$ that are no longer satisfied by the bag semantics: unlike the set

---

1 Bag semantics offers two types of union, called maximal and arithmetic; the first computes the maximum number of occurrences for every tuple in the provided operands whereas the second adds these numbers up. The answer we provide here is based on maximal union for reasons that will be made clear later on in this article. The use of arithmetic union does not affect our motivation.
case, DL-Lite$_{R}^{b}$ ontologies may not have a universal model for conjunctive queries—that is, a single model the answers over which are precisely the certain answers to all such queries; moreover, query answers may be sensitive to the adoption of the unique name assumption (UNA).

In Section 6 we show that conjunctive query answering under bag semantics is computationally more challenging than in the set case. In particular, we establish three incomparable coNP lower bounds for the data complexity of the problem. We first show that it is coNP-hard even if we restrict the ontology language to DL-Lite$_{b}^{R}$—that is, the language where role inclusions are disallowed—regardless of whether the UNA is adopted or not. Second, we show that without the UNA the problem is hard even if the ontologies do not have existential quantification on the right-hand side of concept inclusions (i.e., in the DL-Lite$_{b}$$_{KDFS}$ ontology language) and the queries are restricted to the so-called rooted conjunctive queries [21]—that is, conjunctive queries with all their connected components containing at least one constant or answer variable; this class of queries comprises most practical OBDA queries. The third hardness result is established for the same settings as in the second case except that the UNA is adopted, but there are no restrictions on inclusion axioms.

In Section 7 we make a first step on the way to regain tractability of conjunctive query answering. In particular, we show that rooted conjunctive queries admit a universal model over DL-Lite$_{b}^{R}$$_{CORE}$ ontologies, regardless of the adoption of the UNA.

In Section 8 we employ this result to show that rooted conjunctive queries are rewritable over DL-Lite$_{b}^{R}$$_{CORE}$ to queries in a bag analogue of relational calculus, BCALC, which can be evaluated directly on the ABox of the ontology. Using known results on bag databases, we conclude that the corresponding query answering problem is tractable, in particular, in LOGSPACE.

In Section 9 we establish similar results for arbitrary conjunctive queries and DL-Lite$_{KDFS}$—that is, the language that allows for role inclusions, but does not allow for existential quantification on the right-hand side of concept inclusions; however, these positive results hold only when the UNA is adopted, while we already know that without the UNA the problem is coNP-hard and hence non-rewritable to BCALC.

In Section 10 we combine the results of the previous two sections and establish rewritability and tractability of query answering for the ontology language DL-Lite$_{b}^{R}$—capturing both DL-Lite$_{b}^{R}$$_{CORE}$ and DL-Lite$_{KDFS}$. This language allows for both role inclusions and existential quantification on the right-hand side of concept inclusions. To establish rewritability, however, the language essentially forbids interactions between these two features. Also, the setting inherits all of the restrictions imposed on the previous cases—that is, it adopts the UNA and considers only rooted conjunctive queries.

Finally, in Section 11 we provide a comprehensive discussion of related work, and in Section 12 we discuss possible extensions of our OBDA framework.

2. Preliminaries

In this section we recapitulate the basic definitions that we use in the remainder of the paper. In Section 2.1 we introduce the syntax and (set-based) model-theoretic semantics of the standard ontology and query languages for OBDA. Then, in Section 2.2, we introduce conjunctive queries and define the associated query answering problem. In Section 2.3 we review the common operations on bags. Finally, in Section 2.4 we specify a bag relational calculus that we will exploit later on to express query rewritings over ontologies; our calculus is embeddable into the bag algebra for relational databases by Grumbach and Milo [22], which we discuss in the accompanying Appendix.  

2.1. Syntax and semantics of DL-Lite$_{R}$ ontologies

We fix a vocabulary consisting of countably infinite and pairwise disjoint sets of individuals $I$ (or constants), variables $X$, atomic concepts $C$ (unary predicates) and atomic roles $R$ (binary predicates). A role is an atomic role $P$ in $R$ or its inverse $P^{-}$. A concept is an atomic concept in $C$ or an existentially quantified concept $\exists R$, where $R$ is a role. An inclusion axiom (or just inclusion) is an expression of the form $S_{1} \subseteq S_{2}$ with $S_{1}$ and $S_{2}$ either both concepts, in which case we speak of a concept inclusion, or both roles, in which case we speak of a role inclusion. A disjointness axiom is an expression of the form $\text{Disj}(S_{1}, S_{2})$ with $S_{1}$ and $S_{2}$ either both concepts or both roles. A DL-Lite$_{R}$ TBox is a finite set of inclusions and disjointness axioms. A concept assertion is of the form $A(a)$ with $a \in I$ and $A \in C$. A role assertion is of the form $P(a, b)$ with $a, b \in I$ and $P \in R$. A DL-Lite$_{R}$ ABox is a finite set of concept and role assertions. A DL-Lite$_{R}$ ontology is a pair $(T, A)$ with $T$ a DL-Lite$_{R}$ TBox and $A$ a DL-Lite$_{R}$ ABox.

An interpretation $I$ (or set interpretation, when the context matters) is a pair $(\Delta^{I}, I^{I})$, where the domain $\Delta^{I}$ is a non-empty set, and the interpretation function $^{I}$ maps each individual $a \in I$ to an element $a^{I} \in \Delta^{I}$, each atomic concept $A \in C$ to a subset $A^{I}$ of $\Delta^{I}$, and each atomic role $P \in R$ to a subset $P^{I}$ of $\Delta^{I} \times \Delta^{I}$. The interpretation function extends to other concepts and roles as follows:

$$(R^{-})^{I} = \{(u, v) \mid (v, u) \in R^{I}\},$$
$$(\exists R)^{I} = \{u \mid \exists v \text{ such that } (u, v) \in R^{I}\}.$$
Interpretation $I$ is finite if so is $\Delta^I$. An interpretation $I = (\Delta^I, \cdot)$ satisfies the unique name assumption (or UNA) whenever $I$ interprets distinct individuals from $\mathbf{I}$ with distinct elements from $\Delta^I$—that is, the inequality $a' \neq b'$ holds whenever $a, b \in \mathbf{I}$ with $a \neq b$. An interpretation $I$ satisfies an inclusion $S_1 \subseteq S_2$ if $S_1^I \subseteq S_2^I$, and it satisfies a disjunction axiom $\text{Disj}(S_1, S_2)$ if $S_1^I \cap S_2^I = \emptyset$. An interpretation $I$ satisfies a TBox $T$, written $I \models T$, if $I$ satisfies every axiom in $T$. A TBox $T$ entails an axiom $\alpha$, written $T \models \alpha$, if every interpretation satisfying $T$ also satisfies $\alpha$; TBox $T$ entails $\alpha$ under the UNA if the same holds for every interpretation satisfying the UNA. An interpretation $I$ satisfies an ABox $\mathcal{A}$ if $a' \in A^I$ for all $A(a) \in \mathcal{A}$ and $(a', b') \in P^I$ for all $P(a, b) \in \mathcal{A}$. An interpretation is a model of a DL-Lite$_R$ ontology $(T, \mathcal{A})$ if it satisfies $T$ and $\mathcal{A}$. An ontology $K = (T, \mathcal{A})$ is satisfiable if it has a model, and it is satisfiable under the UNA if it has a model satisfying the UNA.

It is well-known that a DL-Lite$_R$ ontology is satisfiable if and only if it is satisfiable under the UNA, and this can be tested in NLocSpace in general and in AC$^9$ if the TBox is fixed; similarly, a DL-Lite$_R$ TBox entails an axiom if and only if it entails the axiom under the UNA [24].

In this paper, we will also consider the following three sublanguages of DL-Lite$_R$:

- DL-Lite$_\text{CORE}$ restricts DL-Lite$_R$ by disallowing in TBoxes role inclusions and role disjunction axioms;
- DL-Lite$_\text{RDFS}$ restricts DL-Lite$_R$ by disallowing in TBoxes concept and role disjunction axioms as well as existentially quantified concepts on the right-hand side of concept inclusions;
- DL-Lite$_{R\neg}$ restricts DL-Lite$_R$ by disallowing in TBoxes $T$ inclusions of the form $C \sqsubseteq \exists R$ whenever $T$ contains inclusion $R \sqsubseteq S$ for some role $S$ different from $R$.

Note that DL-Lite$_{\text{CORE}}$ and DL-Lite$_{\text{RDFS}}$ are incomparable fragments of DL-Lite$_R$, whereas DL-Lite$_{R\neg}$ extends both DL-Lite$_{\text{CORE}}$ and DL-Lite$_{\text{RDFS}}$. The restrictions imposed in DL-Lite$_{R\neg}$ limit the interaction between concept and role inclusions in TBoxes.

2.2. Queries over ontologies

A conjunctive query (or CQ) $q(x)$ with answer variables $x$ is a formula $\exists y. \phi(x, y)$ in first-order logic with equality, where $x$ and $y$ are (possibly empty) repetition-free tuples of variables from $x$ and $\phi(x, y)$ is a conjunction of atoms of the form $A(t)$, $P(t_1, t_2)$ or $Q(t) = z$, where $A \in \mathbf{C}$, $P \in \mathbf{R}$, $z \in \mathbf{x} \cup \mathbf{y}$, and $t_1, t_2 \in \mathbf{x} \cup \mathbf{y} \cup \mathbf{I}$. If $x$ is clear from the context, then we may write $q$ instead of $q(x)$. The equality atoms of the form $(z = t)$ in a CQ $q(x) = \exists y. \phi(x, y)$ yield an equivalence relation $\sim$ on terms $x \cup y \cup I$, and we write $\bar{t}$ for the equivalence class of a term $t$. The Gaifman graph of $q(x)$ has a node $\bar{t}$ for each $t \in x \cup y \cup I$ in $\phi$, and an edge $[\bar{t}_1, \bar{t}_2]$ for each atom $P(t_1, t_2)$ in $\phi$. In what follows, we (silently) assume that all CQs $q(x) = \exists y. \phi(x, y)$ are safe—that is, such that for each $z \in x \cup y$, the class $\bar{z}$ contains either an individual from $I$ or a variable mentioned in an atom of $\phi(x, y)$ that is not an equality. A CQ is Boolean if its answer variables $x$ are the empty tuple (\). Furthermore, following Bienvenu et al. [21], a CQ $q(x)$ is rooted if each connected component of its Gaifman graph has a node with a term in $x \cup I$. A union of CQs (UCQ) is a disjunction of CQs with the same answer variables. A UCQ is Boolean (or rooted, or both) if so are all of its component CQs.

The answers $q^I$ to a (UC)Q $q(x)$ over an interpretation $I$ are the set of all tuples $a$ of individuals from $\mathbf{I}$ with $|a| = |x|$ such that the formula $q(a)$ holds in $I$ (where $|a|$ and $|x|$ are the sizes of $a$ and $x$, respectively). The certain answers to a (UC)Q $q(x)$ over a DL-Lite$_R$ ontology $K$ are the intersection of the answers to $q(x)$ over all models of $K$. The certain answers $q^K_0$ to $q(x)$ over $K$ under the UNA are the intersection of the answers to $q(x)$ over all models of $K$ satisfying the UNA. In fact, for DL-Lite$_R$, the (usual) certain answers always coincide with the certain answers under UNA, and checking whether a tuple of individuals is in the certain answers to a (UC)Q $q$ over a DL-Lite$_R$ ontology $(T, \mathcal{A})$ is an NP-complete problem with AC$^9$ data complexity (i.e., when $T$ and $q$ are fixed) [9,24]. The latter follows from the rewritability of the class of UCQs to itself over DL-Lite$_R$ [9]. Informally, the key ideas for rewritability is as follows. First, every DL-Lite$_R$ ontology possesses a so-called canonical interpretation, which is a model if the ontology is satisfiable, and which is homomorphically embeddable to every other model. Moreover, this model is universal for UCQs in the sense that the answers to every UCQ on the canonical interpretation coincide with the certain answers over the ontology. Finally, for every UCQ and every TBox it is always possible to construct another UCQ such that, for every ABox, the answers to the original UCQ over the universal model of the resulting ontology are the same as the answers to the new UCQ over just the ABox. We will formally define the notions of rewritability, canonical interpretation, and universal model later in the article.

To conclude, we note that our definition of CQs is slightly non-standard in that it allows for equality atoms, which are usually regarded as inessential. Making equalities explicit in the query will be convenient later on when computing query rewritings, where we will sometimes need to force an answer variable to become equal to another answer variable or to an individual. This is, however, just a technicality; in particular, none of our complexity lower bounds to query answering or negative rewritability results depend on the presence of equalities in the query.

2.3. Bags

A bag over a set $M$ is a function $\Omega : M \to \mathbb{N}_0^\infty$, where $\mathbb{N}_0^\infty$ is the set of non-negative integers $\mathbb{N}_0$ extended with infinity $\infty$. The value $\Omega(c)$ is the multiplicity of $c$ in $\Omega$. A bag $\Omega$ is finite if there are finitely many $c \in M$ with $\Omega(c) > 0$ and there is no $c$ with $\Omega(c) = \infty$. The empty bag $\emptyset$ over $M$ is the bag satisfying $\Omega(c) = 0$ for all $c \in M$. Given bags $\Omega_1$ and
\( \Omega_2 \) over \( M \), we say that \( \Omega_1 \) is a subbag of \( \Omega_2 \), in symbols \( \Omega_1 \subseteq \Omega_2 \), if \( \Omega_1(c) \leq \Omega_2(c) \) for each \( c \in M \). Often, especially in examples, we will use an alternative syntax for bags: for instance, we will write \( \| c : 5, d : 3 \| \) for the bag that assigns 5 to \( c \), 3 to \( d \), and 0 to all other elements.

In this paper, we use the following common set of operators encountered in the literature on algebras over bags [22,23,25–27]. The intersection \( \cap \), maximal union \( \cup \), arithmetic union \( \oplus \), and difference \( - \) are the binary operators defined for bags \( \Omega_1 \) and \( \Omega_2 \) over the same set \( M \), and for every \( c \in M \) as follows:

\[
(\Omega_1 \cap \Omega_2)(c) = \min(\Omega_1(c), \Omega_2(c)), \\
(\Omega_1 \cup \Omega_2)(c) = \max(\Omega_1(c), \Omega_2(c)), \\
(\Omega_1 \oplus \Omega_2)(c) = \Omega_1(c) + \Omega_2(c), \text{ and} \\
(\Omega_1 - \Omega_2)(c) = \max(0, \Omega_1(c) - \Omega_2(c)).
\]

Note that bag difference is well-defined only when \( \Omega_2 \) does not assign \( \infty \) to any element in \( M \). Also, the unary duplicate elimination \( \epsilon \) operator is defined for a bag \( \Omega \) over a set \( M \) and for every \( c \in M \) as follows:

\[
(\epsilon(\Omega))(c) = \begin{cases} 1, & \text{if } \Omega(c) > 0, \\ 0, & \text{otherwise}. \end{cases}
\]

Note that, for every two finite bags \( \Omega_1 \) and \( \Omega_2 \) over the same set \( M \), the following identities hold [26]:

\[
\Omega_1 \cup \Omega_2 = \Omega_1 \cup (\Omega_2 - \Omega_1) \quad \text{and} \quad \Omega_1 \cap \Omega_2 = (\Omega_1 \cup \Omega_2) - (\Omega_1 \cup \Omega_2).
\]

However, these identities may not hold if the bags are infinite.

### 2.4. A calculus for querying bag databases

A database schema is a non-empty finite set \( S \) of predicates with non-negative arities that are disjoint from \( C \) and \( R \). Given a database schema \( S \) and a set of constants (i.e., individuals) \( I \), a database fact is an expression of the form \( S(a) \), where \( a \in S \) and \( a \) is a tuple of constants from \( I \) of size equal to the arity of \( S \). Then, a bag database instance \( D \) is a finite bag over all the facts over \( S \) and \( I \).

Grumbach and Milo [22] proposed an algebraic query language for bag databases called BALG, which is sufficiently powerful to capture relational algebra over bags [19]. BALG allows for nesting of bags and can be seen as the union of the sublanguages BALG\( k \), \( k \geq 1 \), each of which allowing for up to \( k - 1 \) levels of nesting. In this article we restrict ourselves to BALG\( 1 \), which does not allow for bag nesting. Grumbach and Milo [22] studied the data complexity of answering BALG\( 1 \) queries under a unary encoding of numbers in the input and showed that it is strictly “sandwiched” between complexity classes \( \mathcal{AC}^0 \) and \( \text{LogSpace} \); it is therefore tractable, but strictly harder than the data complexity of relational queries over set databases. We defer a full treatment of Grumbach and Milo’s algebra and its associated decision problem to Appendix A.

We next introduce BCALC—a calculus for querying bag databases based on Grumbach and Milo’s algebra. Using a calculus formulation instead of an algebraic one will significantly simplify the presentation of our query rewriting algorithms later on. In Appendix B we show that our calculus can be easily embedded into BALG\( 1 \) and thus inherits its \( \text{LogSpace} \) upper bound for data complexity of query answering.

The syntax of BCALC for a database schema, formally presented in the following inductive definition, extends the syntax of \((U)CQs\) given in Section 2.2 with several new operations. Domain-dependent queries, unexpressible in algebraic query languages, are precluded by introducing restrictions on the use of variables (intuitively, a query is domain-dependent if its answers over a fixed database instance may change when the underlying set of constants is modified; see [28] for details).

**Definition 3.** Given a database schema \( S \) and a set of constants \( I \), a BCALC query \( \Phi(\mathbf{x}) \) with answer variables \( \mathbf{x} \) is any of the following, where \( \Psi, \Psi_1, \text{ and } \Psi_2 \) are BCALC queries:

- \( S(t) \), where \( S \in S \) and \( t \) is a tuple over \( X \cup I \) of size equal to the arity of \( S \) mentioning all variables in \( x \);
- \( \Psi_1(\mathbf{x}_1) \land \Psi_2(\mathbf{x}_2) \), where \( \mathbf{x} = \mathbf{x}_1 \cup \mathbf{x}_2 \);
- \( \Psi(\mathbf{x}) \land (x = t) \), where \( x \in X_0 \), \( t \in X \cup I \), and \( \mathbf{x} = \mathbf{x}_0 \cup \{t\} \setminus I \);
- \( \exists \mathbf{y} \Psi(\mathbf{x}, \mathbf{y}) \), where \( \mathbf{y} \) is a tuple of distinct variables from \( X \) that are not in \( \mathbf{x} \);
- \( \Psi_1(\mathbf{x}) \lor \Psi_2(\mathbf{x}) \), where \( \lor \in \{\lor, \lor, \land\}; \) or
- \( \delta \Psi(\mathbf{x}) \).

A BCALC query is positive if it does not mention the difference operator \( \setminus \). A positive BCALC query is a BCALC conjunctive query (CQ) if it additionally does not mention operators \( \lor, \lor, \) and \( \delta \). A BCALC maximal (or arithmetic) union of CQs is a BCALC query of the form \( \Phi_1(\mathbf{x}) \lor \cdots \lor \Phi_n(\mathbf{x}) \) (or of the form \( \Phi_1(\mathbf{x}) \lor \cdots \lor \Phi_n(\mathbf{x}) \), respectively), where each \( \Phi_i \) is a BCALC CQ.
Next we formally define the semantics of BCALC queries, which are bags of tuples of constants.

**Definition 4.** The bag answers $\Phi^D$ to a BCALC query $\Phi(x)$ over a bag database instance $D$ is the finite bag over $I^{|I|}$ defined inductively by the following equations for every tuple $a$ over $I$ with $|a| = |x|$, where $\nu : x \cup I \rightarrow I$ is the function such that $\nu(x) = a$ and $\nu(a) = a$ for all $a \in I$:

- $\Phi^D(a) = D(\nu(t(t)))$, if $\Phi(x) = S(t)$;
- $\Phi^D(a) = \Psi^D(\nu(x_1)) \times \Psi^D(\nu(x_2))$, if $\Phi(x) = \Psi_1(x_1) \land \Psi_2(x_2)$;
- $\Phi^D(a) = \Psi^D(\nu(x_0))$, if $\Phi(x) = \Psi(x_0) \land (x = t)$ and $\nu(x) = \nu(t)$;
- $\Phi^D(a) = 0$, if $\Phi(x) = \Psi(x_0) \land (x = t)$ and $\nu(x) \neq \nu(t)$;
- $\Phi^D(a) = \sum y \rightarrow \Psi^D(a, \nu(y))$, if $\Phi(x) = 3y. \Psi(x, y)$;
- $\Phi^D(a) = (\nu^D \lor \nu^D)(a)$, if $\Phi(x) = \nu^D \lor \nu^D(a)$, where $\lor$ is $\cup$, $\lor$, or $\land$, and $\lor^D$ is $\lor$, $\lor$, or $\land$, respectively;
- $\Phi^D(a) = (\nu^D)(a)$, if $\Phi(x) = \delta \Psi(x)$.

The decision problem of query answering for BCALC is defined as follows, where all numbers in the input are assumed to be represented in unary and the bag (i.e., the database instance) is explicitly defined only for a finite number of facts while the multiplicities of all other facts are assumed to be 0.

| QUERYANSWERING[BCALC] |
|------------------------|
| Input: BCALC query $\Phi(x)$, bag database instance $D$, tuple $a$ of constants over $I$, and number $k \in \mathbb{N}_0$. |
| Question: Is $\Phi^D(a) \geq k$? |

The data complexity of this problem is the complexity when query $\Phi$ is considered to be fixed and only $D$, $a$, and $k$ form the input.

The LogSPACE upper bound for the data complexity of QUERYANSWERING[BCALC] is obtained by showing that for each BCALC query $\Phi$ one can construct a BALG1 algebra expression $E_\Phi$ such that the bag answers to $\Phi$ over every bag database $D$ coincide with the bag answers to $E_\Phi$ over $D$; this serves the need, because, as we already mentioned, BALG1 can be evaluated in LogSPACE. The proof of this claim, which is rather technical but conceptually straightforward, is deferred to Appendix B.

**Proposition 5.** QUERYANSWERING[BCALC] is in LogSPACE in data complexity.

We conclude by observing that in the literature on query optimisation under bag semantics [29–31] it is common to encounter the notion of bag-set semantics for databases, where input database instances are sets—that is, do not allow for multiplicities greater than 1—while permitting query answers and views to be bags. The bag semantics we consider in this article generalises the bag-set semantics, and restricting ourselves to bag-set semantics would not change any of our results.

### 3. Ontology-based data access under bag semantics

In this section, we introduce our OBDA framework as a natural generalisation of that by Poggi et al. for OBDA under set semantics [2]. We start by defining the syntax of OBDA settings.

**Definition 6.** A bag OBDA setting is a triple $(\mathcal{T}, \Sigma, \mathcal{D})$ where

- $\mathcal{T}$ is a DL-Lite$_R$ TBox;
- $\Sigma$ is a set of global-as-view (GAV) mapping assertions (or mappings) of the form $\Phi(x) \sim A(x)$ or $\Psi(x, y) \sim P(x, y)$,

where $\Phi$ and $\Psi$ are BCALC queries, while $A$ and $P$ are an atomic concept and atomic role, respectively; and
- $\mathcal{D}$ is a bag database instance.

A couple of observations about Definition 6 are in order. First, recall that in all our motivating examples so far we have written mappings using SQL queries, which reflects the way in which mappings are defined in practice. To formally study OBDA, however, the use of a bag query language close to first-order logic, such as BCALC, is more appropriate. Second, in contrast to the definitions by Poggi et al. [2], we do not allow for function symbols on the right-hand side of mappings.
This restriction does not affect the computational properties of query answering in OBDA under set semantics [2], and it is adopted in most theoretical papers on data integration [8,32]; it is also immaterial to our technical results, and yet allows us to simplify the presentation.

**Example 7.** The mappings $\Sigma_{ex}$ in Example 1 are equivalently expressed using BCALC as follows:

$$
\begin{align*}
\sigma_1 &= \exists y_1, y_2, y_3, y_4. \text{Columbia}(x, y_1, y_2, y_3, y_4) \leadsto \text{Musician}(x), \\
\sigma_2 &= \exists y_1, y_2. \text{Verve_Wind}(x, y_1, y_2) \leadsto \text{WindPlayer}(x), \\
\sigma_3 &= \exists y_1, y_2, y_3, y_4. \text{Columbia}(y_1, x, y_2, y_3, y_4) \leadsto \text{Record}(x), \\
\sigma_4 &= \exists y_1, y_2. \text{Verve_Wind}(y_1, x, y_2) \leadsto \text{Record}(x), \\
\sigma_5 &= \exists y_1, y_2, y_3. \text{Columbia}(y, x, y_1, y_2, y_3) \leadsto \text{hasMusician}(x, y), \\
\sigma_6 &= \exists y_1. \text{Verve_Wind}(y, x, y_1) \leadsto \text{hasMusician}(x, y).
\end{align*}
$$

The semantics of bag OBDA settings is based on bag interpretations $I$, which are defined as set interpretations (see Section 2.1) with the exception that concepts and roles are now interpreted as bags rather than sets. The extension of the interpretation function to non-atomic concepts and roles is defined in a natural way: for example, the concept $\exists P$ for an atomic role $P$ is interpreted as the bag projection of the interpretation $P^I$ of $P$ to its first component, where each occurrence of a pair $(u, v)$ in $P^I$ contributes separately to the multiplicity of a domain element $u$ in $(\exists P)^I$.

**Definition 8.** A bag interpretation $I$ is a pair $(\Delta^I, \cdot^I)$ where the domain $\Delta^I$ is a non-empty set, and the interpretation function $\cdot^I$ maps each individual $a \in I$ to an element $a^I \in \Delta^I$, each atomic concept $A \in C$ to a bag $A^I$ over $\Delta^I$, and each atomic role $P \in R$ to a bag $P^I$ over $\Delta^I \times \Delta^I$. Interpretation function $\cdot^I$ extends to complex concepts $(P^-)^I$ and $(\exists R)^I$, for $P \in R$ and $R$ a role, as follows, for all $u, v \in \Delta^I$:

$$(P^-)^I(u, v) = P^I(v, u) \quad \text{and} \quad (\exists R)^I(u) = \sum_{v \in \Delta^I} R^I(u, v).$$

A bag interpretation $I$ is finite if $\Delta^I$ is a finite set and $I$ assigns a finite bag to each $A \in C$ and each $P \in R$.

We are now ready to specify the semantics of bag OBDA settings in terms of bag interpretations (note that satisfaction of axioms is defined in the same way as in the set case, but the symbols $\subseteq, \cap$, and $\emptyset$ denote the subbag relation, bag intersection, and the empty bag, respectively).

**Definition 9.** A bag interpretation $I$ satisfies the UNA if $a^I \neq b^I$ whenever $a, b \in I$ with $a \neq b$. A bag interpretation $I$ satisfies an inclusion $S_1 \subseteq S_2$ if $S_1^I \subseteq S_2^I$, and it satisfies a disjointness axiom $\text{Disj}(S_1, S_2)$ if $S_1^I \cap S_2^I = \emptyset$. A bag interpretation $I$ satisfies a TBox $\mathcal{T}$, written $I \models^b \mathcal{T}$, if $I$ satisfies every axiom in $\mathcal{T}$. A TBox $\mathcal{T}$ entails an axiom $\alpha$ under bag semantics, written $\mathcal{T} \models^b \alpha$, if every bag interpretation satisfying $\mathcal{T}$ satisfies $\alpha$; $\mathcal{T}$ entails $\alpha$ under bag semantics and the UNA if only bag interpretations satisfying the UNA are considered.

A bag interpretation $I$ satisfies a set of mappings $\Sigma$ with respect to a bag database instance $\mathcal{D}$ if the following holds, for all mappings $\Phi(x) \leadsto A(x)$ and $\Psi(x, y) \leadsto P(x, y)$ in $\Sigma$, and all individuals $a, a_1, a_2 \in I$:

$$
\sum_{b \models^I I} \Phi^D(b) \leq A^I(a^I) \quad \text{and} \quad \sum_{b_i \models^I i} \Psi^D(b_1, b_2) \leq P^I(a_1^I, a_2^I).$$

A bag interpretation $I$ is a model of a bag OBDA setting $(\mathcal{T}, \Sigma, \mathcal{D})$, denoted by $I \models^b (\mathcal{T}, \Sigma, \mathcal{D})$, if $I \models^b \mathcal{T}$ and $I$ satisfies $\Sigma$ with respect to $\mathcal{D}$. An OBDA setting $(\mathcal{T}, \Sigma, \mathcal{D})$ is satisfiable if it has a model, and it is satisfiable under the UNA if it has a model satisfying the UNA.

Note that if a bag interpretation $I$ satisfies the UNA, then the notion of satisfaction for a set of mappings $\Sigma$ as defined above becomes equivalent to requiring that inequalities $\Phi^D(a) \leq A(a^I)$ and $\Psi^D(a_1, a_2) \leq P(a_1^I, a_2^I)$ hold for all mappings $\Phi(x) \leadsto A(x)$ and $\Psi(x, y) \leadsto P(x, y)$ in $\Sigma$, and all individuals $a, a_1, a_2 \in I$.

**Example 10.** Consider the bag OBDA setting $(\mathcal{T}_{ex}, \Sigma_{ex}, \mathcal{D}_{ex})$ in Example 1. Let $I_{ex}$ be the bag interpretation with domain $\Delta_{ex} = I$ that maps all individuals to themselves, and assigns bags to concepts and roles as follows:
Musician\textsubscript{ex} = \{ M. Davis : 2, K. Jarrett : 1 \},
WindPlayer\textsubscript{ex} = \{ M. Davis : 1 \},
Record\textsubscript{ex} = \{ Kind of Blue : 1, A Tribute to Jack Johnson : 1, Expectations : 1, Ascenseur pour l’Échafaud : 1 \},
hasMusician\textsubscript{ex} = \{ (Kind of Blue, M. Davis) : 1, (A Tribute to Jack Johnson, M. Davis) : 1, (Expectations, K. Jarrett) : 1, (Ascenseur pour l’Échafaud, M. Davis) : 1 \}.

To show that \( l_{\text{ex}} \models_{b} (T_{\text{ex}}, \Sigma_{\text{ex}}, D_{\text{ex}}) \), we argue that \( l_{\text{ex}} \) satisfies \( \Sigma_{\text{ex}} \) with respect to \( D_{\text{ex}} \) and that \( l_{\text{ex}} \models_{b} T_{\text{ex}} \). For the former, we first compute the bag answers to the BCALC queries \( \Phi_{1}, \ldots, \Phi_{6} \) appearing on the left-hand side of the mappings \( \sigma_{1}, \ldots, \sigma_{6} \) over database instance \( D_{\text{ex}} \). These are specified as follows:

\begin{align*}
\Phi_{1}^{\text{ex}} & = \{ M. Davis : 2, K. Jarrett : 1 \}, \\
\Phi_{2}^{\text{ex}} & = \{ M. Davis : 1 \}, \\
\Phi_{3}^{\text{ex}} & = \{ Kind of Blue : 1, A Tribute to Jack Johnson : 1, Expectations : 1 \}, \\
\Phi_{4}^{\text{ex}} & = \{ Ascenseur pour l’Échafaud : 1 \}, \\
\Phi_{5}^{\text{ex}} & = \{ (Kind of Blue, M. Davis) : 1, (A Tribute to Jack Johnson, M. Davis) : 1, (Expectations, K. Jarrett) : 1 \}, \\
\Phi_{6}^{\text{ex}} & = \{ (Ascenseur pour l’Échafaud, M. Davis) : 1 \}.
\end{align*}

It is now immediate to verify that the inequalities stipulated by Definition 9 hold for \( \sigma_{1}, \ldots, \sigma_{6} \); thus, \( l_{\text{ex}} \) satisfies \( \Sigma_{\text{ex}} \) with respect to \( D_{\text{ex}} \). We next argue that \( l_{\text{ex}} \models_{b} T_{\text{ex}} \). Indeed, by Definition 8, the interpretation of \( \exists \text{hasMusician} \) is the bag

\[ \{ (Kind of Blue, M. Davis) : 1, (A Tribute to Jack Johnson, M. Davis) : 1, (Expectations, K. Jarrett) : 1 \} \],

so \( C_{\text{ex}} \) is a subbag of \( D_{\text{ex}} \) for each inclusion \( C \subseteq D \) in \( T_{\text{ex}} \), as required. \( \checkmark \)

We next discuss an important aspect of Definition 9 concerning the presence of different mappings defining the same view. In such cases, the extension of the view intuitively corresponds to the union of the answers to the queries specified on the left-hand side of the contributing mappings. However, bag query languages, such as BCALC, come with two versions of the union operation: maximal and arithmetic union. Moreover, in different settings one of these unions can be more intuitive and preferable than the other. On the one hand, Definition 9 tacitly commits to the maximal union by requiring that a model for \( \Sigma \) and \( D \) satisfies each contributing mapping independently. On the other hand, this is not a limitation of our OBDA framework since GAV mapping assertions can always be rewritten to reflect the alternative choice based on arithmetic union.

Example 11. Consider the bag OBDA setting \( (T_{\text{ex}}, \Sigma_{\text{ex}}, D_{\text{ex}}) \) obtained from our running example by augmenting \( \Sigma_{\text{ex}} \) to the set \( \Sigma_{\text{ex}}'' \) that additionally contains mapping

\[ \sigma_{7} = \exists y_{1}, y_{2}. \text{Verve_Wind}(x, y_{1}, y_{2}) \leadsto \text{Musician}(x). \]

Note that both mappings \( \sigma_{1} \) and \( \sigma_{7} \) define the extension of concept \text{Musician}. By Definition 9 interpretations such as \( l_{\text{ex}} \) in Example 10, which interpret \text{Musician} as the maximal union of the bags corresponding to the musicians mentioned in the Columbia and Verve_Wind tables, are valid models of \( (T_{\text{ex}}, \Sigma_{\text{ex}}, D_{\text{ex}}) \). Let us now define \( \Sigma_{\text{ex}}'' \) as the mappings obtained from \( \Sigma_{\text{ex}} \) by replacing \( \sigma_{1} \) by

\[ (\exists y_{1}, y_{2}, y_{3}, y_{4}. \text{Columbia}(x, y_{1}, y_{2}, y_{3}, y_{4}) \lor (\exists y_{1}, y_{2}. \text{Verve_Wind}(x, y_{1}, y_{2})) \leadsto \text{Musician}(x). \]

In this case, every model of \( (T_{\text{ex}}, \Sigma_{\text{ex}}'', D_{\text{ex}}) \) interprets \text{Musician} as the arithmetic union of the bags corresponding to musicians in the relevant tables. In particular, interpretation \( l_{\text{ex}} \) in Example 10 is not a model, as required. \( \checkmark \)

We are now ready to define CQ answering under bag semantics. We first define the answers \( q^{l} \) to a CQ \( q(x) \) over a bag interpretation \( l \); this is a natural extension to (possibly infinite) interpretations of the notion of bag answers to a CQ over a bag database (see Section 2.4). Specifically, \( q^{l} \) is a bag of tuples of individuals such that each valid embedding \( \lambda \) of the atoms in \( q \) into \( l \) contributes separately to the multiplicity of the tuple \( \lambda(x) \) in \( q^{l} \), and where the contribution of each specific \( \lambda \) is the product of the multiplicities of the images of the query atoms under \( \lambda \) in \( l \).

Definition 12. Let \( q(x) = \exists y. \phi(x, y) \) be a CQ and \( l = (\Delta^{\ell}, I) \) be a bag interpretation. The bag answers \( q^{l} \) to \( q \) over \( I \) are the bag over tuples of individuals from \( I \) of size \( |x| \) such that, for every such tuple \( a \),

\[ q^{l}(a) = \sum_{\lambda \in \Lambda} \prod_{S(t) \in \phi(x, y)} S^{l}((\lambda(t)), \]

where \( \Lambda \) is the set of all valuations \( \lambda : x \cup y \cup I \rightarrow \Delta^{\ell} \) such that \( \lambda(x) = a^{l}, \lambda(a) = a^{l} \) for each \( a \in I \), and \( \lambda(z) = \lambda(t) \) for each \( z = t \) in \( \phi(x, y) \).
If \( q \) is Boolean then the bag answers \( q^I \) are defined only for the empty tuple \( () \). Also, conjunction \( \phi(x, y) \) may contain repeated atoms, and hence can be seen as a bag of atoms; while repeated atoms are redundant in the set case, they are essential in the bag setting [29,33], and thus the definition of \( q^I(a) \) should be read in the way that it treats each copy of a query atom \( \exists(x) \) separately in the product.

The following definition of certain answers, which captures open-world query answering, is a natural extension of the set notion to bags: a query answer is certain with multiplicity \( k \) if it is an answer with multiplicity at least \( k \) over every model of the OBDA setting.

**Definition 13.** The bag certain answers \( q^I(T, \Sigma, D) \) to a CQ \( q \) over a bag OBDA setting \((T, \Sigma, D)\) are the bag

\[
\bigcap_{l \models^b (T, \Sigma, D)} q^I.
\]

The bag certain answers under the UNA are defined in the same way except that the intersection ranges only over the models satisfying the UNA.

**Example 14.** Consider the OBDA setting \((T_{ex}, \Sigma_{ex}, D_{ex})\) of our running example. Let \( T_{ex}' \) augment \( T_{ex} \) with the additional inclusion \( \text{\texttt{hasMusician}} \sqsubseteq \text{\texttt{Musician}} \) specifying the range of role hasMusician, and let \( q_{ex}(x) = \text{\texttt{Musician}}(x) \). We argue that \( q_{ex}(T_{ex}', \Sigma_{ex}, D_{ex}) (\text{\texttt{M. Davis}}) = 3 \). On the one hand, consider interpretation \( J_{ex} \) extending \( I_{ex} \) in Example 10 by setting \( \text{\texttt{Musician}}^{I_{ex}} = \{ \text{\texttt{M. Davis}} : 3, \text{\texttt{K. Jarrett}} : 1 \} \); it can be easily checked that \( J_{ex} \) is a model of \((T_{ex}', \Sigma_{ex}, D_{ex})\) satisfying \( q_{ex}^I (\text{\texttt{M. Davis}}) = 3 \). On the other hand, we argue that every bag model \( l \) of \((T_{ex}', \Sigma_{ex}, D_{ex})\) satisfies \( q_{ex}^I (\text{\texttt{M. Davis}}) \geq 3 \). The fact that \( l \) satisfies \( \Sigma_{ex} \) with respect to \( D_{ex} \) (and, in particular, mappings \( \sigma_5 \) and \( \sigma_6 \)) implies that \( \text{\texttt{hasMusician}} \) associates with \( \text{\texttt{M. Davis}} \) at least three elements. So, \((\text{\texttt{hasMusician}})') (\text{\texttt{M. Davis}}) \geq 3 \). Then, since \( T_{ex}' \) contains \( \text{\texttt{hasMusician}} \sqsubseteq \text{\texttt{Musician}} \) and \( l \) satisfies \( T_{ex}' \), we have that \( \text{\texttt{Musician}}' (\text{\texttt{M. Davis}}) \geq 3 \). Therefore, \( q_{ex}^I (T_{ex}', \Sigma_{ex}, D_{ex}) (\text{\texttt{M. Davis}}) = 3 \) holds as well. \( \triangleright \)

The decision problem \( \text{\texttt{BagCertOBDA}}[Q, O] \) corresponding to computing the bag certain answers to a CQ from a class \( Q \) over an OBDA setting with a TBox from an ontology language \( O \) (i.e., \( \text{\texttt{DL-Lite}}_R \) or one of its sublanguages) is defined as follows, where we again assume that all numbers in the input are represented in unary and the bag is explicitly defined only for a finite number of facts.

**BagCertOBDA[Q, O]**

- **Input:** CQ \( q \) from \( Q \), bag OBDA setting \((T, \Sigma, D)\) with \( T \) from \( O \), tuple \( a \) of individuals from \( I \), and number \( k \in \mathbb{N}_0^\infty \).
- **Question:** Is \( q^I(T, \Sigma, D)(a) \geq k \)?

The UNA version \( \text{\texttt{BagCertOBDA}}[Q, O] \) of this problem is defined in the same way as \( \text{\texttt{BagCertOBDA}}[Q, O] \) except that the certain answers are considered under the UNA. The data complexity of these problems is the complexity when the query \( q \), TBox \( T \), and mappings \( \Sigma \) are considered to be fixed, and only \( D, a, \) and \( k \) form the input.

4. The ontology language \( \text{\texttt{DL-Lite}}_R^b \)

In Section 4.1 we define the ontology language \( \text{\texttt{DL-Lite}}_R^b \) and its natural fragments, where the distinctive feature of \( \text{\texttt{DL-Lite}}_R^b \) is that ABoxes consist of bags of facts rather than sets. We then show in Section 4.2 that, analogously to the case of conventional OBDA, the materialisation of the mappings over the sources can be represented by a virtual bag ABox; as a result, the data complexity of OBDA query answering under bag semantics coincides with that of query answering over \( \text{\texttt{DL-Lite}}_R^b \).

4.1. The syntax and semantics of \( \text{\texttt{DL-Lite}}_R^b \)

We start by introducing the notion of a bag ABox and describing its semantics in terms of bag interpretations.

**Definition 15.** A bag ABox is a finite bag over the set of concept and role assertions. A bag interpretation \( I = (\Delta^I, \cdot^I) \) satisfies a bag ABox \( A \), written \( I \models^b A \), if, for each concept assertion \( A(a) \) and role assertion \( P(a_1, a_2) \), the following holds:

\[
\sum_{b \in I: b^I = a^I} \mathbb{A}(A(b)) \leq A^I(a^I) \quad \text{and} \quad \sum_{b_i \in I: b_i^I = a_i^I, i = 1, 2} \mathbb{A}(P(b_1, b_2)) \leq P^I(a_1^I, a_2^I).
\]
If a bag interpretation $I$ satisfies $A$, then ABox satisfaction amounts to checking whether the inequalities $A(A(a)) \leq A^1(a^1)$ and $A(P(a_1, a_2)) \leq P^1(a_1^1, a_2^1)$ hold for each concept and role assertion $A(a)$ and $P(a_1, a_2)$, respectively. We can now introduce the notion of a bag ontology and define the ontology language DL-Lite$_R^b$ and its fragments.

**Definition 16.** A DL-Lite$_R^b$ ontology is a pair $(T, A)$ of a DL-Lite$_R$ TBox $T$ and bag ABox $A$. The sublanguages DL-Lite$_R^b$$_{core}$, DL-Lite$_R^b$$_{dofs}$, and DL-Lite$_R^b$$_{core}$ of DL-Lite$_R^b$$_{core}$ are defined in the same way except that only DL-Lite$_R^b$$_{core}$, DL-Lite$_R^b$$_{dofs}$, and DL-Lite$_R^b$$_{core}$ TBoxes are allowed, respectively.

A bag interpretation $I$ is a model of a DL-Lite$_R^b$ ontology $(T, A)$, written $I \models^b (T, A)$, if $I \models T$ and $I \models^b A$. A DL-Lite$_R^b$ ontology is satisfiable if it has a model; it is satisfiable under the UNA if it has a model satisfying the UNA.

The following definition of certain answers, which captures open-world query answering, is a natural extension of the set notion for DL-Lite$_R$ to bags: a query answer is certain for a given multiplicity if it occurs with at least that multiplicity in the bag answers to the query over every model of the ontology.

**Definition 17.** The bag certain answers $q^k$ to a CQ $q$ over a DL-Lite$_R^b$ ontology $K$ are the bag $\bigcap_{I \models^b K} q^I$. The bag certain answers under the UNA are defined in the same way except that only models satisfying the UNA are considered in the intersection.

Similarly to the OBDA case, the decision problem corresponding to computing the bag certain answers to a CQ from a class $Q$ over an ontology in a bag ontology language $\mathcal{O}$ (e.g., DL-Lite$_R^b$$_{core}$ or one of its sublanguages) is defined as follows, where we again assume that all numbers in the input are represented in unary and the bag is explicitly defined only for a finite number of facts.

$$\text{BagCert}(Q, \mathcal{O})$$

**Input:** CQ $q$ from $Q$, ontology $\mathcal{K} = (T, A)$ from $\mathcal{O}$, tuple $a$ of individuals from $I$, and number $k \in \mathbb{N}_0^\infty$.

**Question:** Is $q^k(a) \geq k$?

The UNA version $\text{BagCert}^{\text{UNA}}(Q, \mathcal{O})$ of this problem is defined in the same way as $\text{BagCert}(Q, \mathcal{O})$ except that the certain answers are considered under the UNA. The data complexity of this problem is the complexity when the query $q$ and the TBox $T$ are considered to be fixed, and only $A, a,$ and $k$ form the input.

We conclude this section by introducing the notion of rewritability for our bag semantics. Our definition is analogous to that of Calvanese et al. [9] for the set case. Note that a bag ABox $A$ can be seen as a database instance, so we can write $\Phi^A$ for a BCALC query $\Phi$ over the unary and binary predicates for atomic concepts and roles, respectively.

**Definition 18.** A BCALC query $\Phi$ is a rewriting of a CQ $q$ with respect to a TBox $T$ if all the individuals, atomic concepts and atomic roles of $\Phi$ appear in $q$ or $T$, and $q^T = \Phi^A$ for every bag ABox $A$ with satisfiable $(T, A)$. A class of CQs $Q$ is rewritable to a class of BCALC queries $Q'$ over an ontology language $\mathcal{O}$ if, for every query in $Q$ and every TBox in $\mathcal{O}$, there exists in $Q'$ a rewriting of the query with respect to the TBox. Rewritability and rewritability under the UNA are defined in the same way except that only ontologies satisfiable under the UNA and certain answers under the UNA are considered.

Note that by restricting the signature of $\Phi$ to that of $q$ and $T$ in this definition we are considering the problem of finding a pure rewriting of $q$ with respect to $T$. In the set case, it was shown that pure rewritings may have to be of larger size than their impure counterparts [13]. However, our negative results on rewritability (i.e., Propositions 45 and 63) do not depend on this restriction, and may be shown for the more general case; we impose this restriction to facilitate exposition of some proofs.

Note also that in the set case the target query language for rewriting is typically a class of UCQs, since Calvanese et al. [9] showed that arbitrary CQs are rewritable to UCQs over DL-Lite$_R$. In the case of bags, however, we will see that the situation is markedly different, and we will focus on BCALC as the target language for rewriting.

4.2. Relationship to query answering in OBDA

In conventional OBDA, the certain answers to a query over $(T, \Sigma, D)$ can be characterised as those that logically follow from the union of the TBox $T$ and the virtual ABox $A_{\Sigma, D}$, which represents the materialisation of the views defined by the mappings $\Sigma$ over the database instance $D$. As a result, query answering in OBDA amounts to query answering over DL-Lite$_R$, and the rewritability and data complexity properties of both problems coincide [2].
In what follows, we show that an analogous correspondence holds under bag semantics. We start by introducing the notion of a virtual bag ABox, which captures the materialisation of the views specified in a bag OBDA setting.

**Definition 19.** The virtual DL-Lite$_R$ ABox of the bag OBDA setting $⟨T, Σ, D⟩$ is the bag ABox $A_{Σ, D}$ defined as follows, for all $A ∈ C, P ∈ R$, and $a, a_1, a_2 ∈ I$:

$$A_{Σ, D}(a) = \max \{ \Phi^D(a) \mid \Phi(x) \sim A(x) ∈ Σ \},$$

$$A_{Σ, D}(P(a_1, a_2)) = \max \{ \Psi^D(a_1, a_2) \mid \Psi(x, y) \sim P(x, y) ∈ Σ \}.$$

The following example illustrates the notion of virtual ABoxes.

**Example 20.** The virtual ABox $A_{ex}$ of the bag OBDA setting $⟨T_{ex}, Σ_{ex}, D_{ex}⟩$ where $T_{ex}$ is specified in Example 1 and $Σ_{ex}, D_{ex}$ in Example 7 is the following bag of assertions:

$$\| \text{Musician}(M. Davis) : 2, \text{Musician}(K. Jarrett) : 1, \text{WindPlayer}(M. Davis) : 1, \text{Record}(\text{Kind of Blue}) : 1, \text{Record}(\text{A Tribute to Jack Johnson}) : 1, \text{Record}(\text{Expectations}) : 1, \text{Record}(\text{Ascenseur pour l’Échafaud}) : 1, \text{hasMusician}(\text{Kind of Blue}, M. Davis) : 1, \text{hasMusician}(\text{A Tribute to Jack Johnson}, M. Davis) : 1, \text{hasMusician}(\text{Expectations}, K. Jarrett) : 1, \text{hasMusician}(\text{Ascenseur pour l’Échafaud}, M. Davis) : 1 \| \triangleleft$$

The following lemma shows that the models of a bag OBDA setting $⟨T, Σ, D⟩$ coincide with those models satisfying the TBox $T$ and the virtual bag ABox $A_{Σ, D}$.

**Lemma 21.** For every bag OBDA setting $⟨T, Σ, D⟩$ and every bag interpretation $I$, we have that $I \models_b ⟨T, Σ, D⟩$ if and only if $I \models_b ⟨T, A_{Σ, D}⟩$.

**Proof.** It suffices to show that $I$ is a model of $A_{Σ, D}$ if and only if $I$ satisfies $Σ$ with respect to $D$ as in Definition 9. For this, let $I$ be a model of $A_{Σ, D}$ and, for all $S ∈ C ∪ R$ and tuples $a$ of individuals, let $Ψ_{S,a}(x) \sim S(x)$ be the mapping in $Σ$ such that $Ψ^D_{S,a}(a) = \max \{ \Phi^D(a) \mid \Phi(x) \sim S(x) ∈ Σ \}$ where $x$ has the same arity as $a$. By the definition of $I \models_b A_{Σ, D}$ and Definition 19, the following inequality holds for every $S ∈ C ∪ R$, tuples $a$ of individuals, and mapping $Φ(x) \sim S(x)$ in $Σ$:

$$S^I(a^I) ≥ \sum_{b \text{ over } I, b^I = a^I} A_{Σ, D}(S(b)) = \sum_{b \text{ over } I, b^I = a^I} Ψ^D_{S,b}(b) ≥ \sum_{b \text{ over } I, b^I = a^I} Φ^D(b).$$

By Definition 9, this is equivalent to requiring that $I$ satisfies $Σ$ with respect to $D$, as desired. \qed

Having Lemma 21 at our disposal, we can relate the problems of satisfiability checking and query answering for bag OBDA settings to the corresponding problems for DL-Lite$_R$.

**Theorem 22.** The following statements hold:

1. a bag OBDA setting $⟨T, Σ, D⟩$ is satisfiable if and only if $⟨T, A_{Σ, D}⟩$ is satisfiable;
2. for every bag OBDA setting $⟨T, Σ, D⟩$ and every CQ $q$ we have $q(⟨T, Σ, D⟩) = q(⟨T, A_{Σ, D}⟩)$; and
3. BagCertObda$[Q, O]$ and BagCert$[Q, O]$ are mutually reducible in LogSpace with respect to data complexity, for $Q$ a class of CQs and for $O$ a sublanguage of DL-Lite$_R^b$ or DL-Lite$_R^c$ itself.

All three statements also hold when the problems are considered under the UNA.

**Proof.** We concentrate on the general case; the case of the UNA is analogous. The first two statements are direct consequences of Lemma 21 and the definitions of satisfiability and certain answers.

To show the third statement, we start by reducing BagCertObda$[Q, O]$ to BagCert$[Q, O]$. For this, fix an instance of BagCertObda$[Q, O]$ consisting of a CQ $q ∈ Q$, an OBDA setting $⟨T, Σ, D⟩$ with TBox $T$ in $O$, a tuple $a$ of individuals, and a number $k$. By Statement 2, BagCertObda$[Q, O]$ is true on the aforementioned instance if and only if BagCert$[Q, O]$ is true on an instance consisting of $q, ⟨T, A_{Σ, D}⟩, a$, and $k$, where $A_{Σ, D}$ is the virtual ABox corresponding to $Σ$ and $D$. Because $T$ and $Σ$ are of fixed size, it is clear by Definition 19 and Proposition 5 on the data complexity of BCALC that the construction of $A_{Σ, D}$ is feasible in LogSpace.

The reduction of BagCert$[Q, O]$ to BagCertObda$[Q, O]$ is straightforward: for an instance of BagCert$[Q, O]$ consisting of a CQ $q ∈ Q$, an ontology $⟨T, A⟩$ in $O$, a tuple $a$ of individuals, and a number $k$, we just consider an instance of BagCertObda$[Q, O]$ consisting of $q, ⟨T, Σ, A⟩, a$, and $k$, where $A$ is considered as a bag database instance and $Σ$ is a set of identity mappings $S(x) \sim S(x)$ for all atomic concepts and roles $S$ appearing in $A$. \qed
This theorem allows us to talk only about \( DL-Lite^b_R \) ontologies in the rest of the paper, silently assuming that all the results apply to OBDA settings as well.

5. Relationship of bag and set semantics in the context of \( DL-Lite^b_R \)

In this section we discuss the main similarities and differences between our bag semantics of \( DL-Lite^b_R \) and the conventional set semantics of \( DL-Lite_R \). First, in Section 5.1, we argue that our bag semantics can be seen as a generalisation of the set semantics, in the sense that, on the one hand, satisfiability checking in \( DL-Lite^b_R \) reduces to satisfiability checking in \( DL-Lite_R \) and, on the other hand, query answers under bag and set semantics coincide if multiplicities are ignored. There are, however, key properties of the conventional semantics for \( DL-Lite_R \) that are no longer satisfied by the bag semantics. In particular, in Section 5.2, we discuss the influence of the UNA on query answering and show fundamental differences with the set case. Furthermore, in Section 5.3, we show that a universal model—a representative model of a satisfiable ontology over which each CQ can be correctly evaluated—may not exist under bag semantics; this is in contrast to the set case, where the fact that a universal model always exists is key to ensuring favourable computational properties of query answering.

5.1. Satisfiability, entailment of axioms, and query answering

The following theorem shows that our bag semantics is compatible with the conventional set semantics of \( DL-Lite_R \).

Theorem 23. Let \( K = \langle T, A \rangle \) be a \( DL-Lite_R \) ontology and let \( K’ = \langle T’, A’ \rangle \) be a \( DL-Lite^b_R \) ontology with the same TBox such that \( A = \{ S(t) | A(S(t)) \geq 1 \} \). Then, the following statements hold:

1. \( K \) is satisfiable if and only if \( K’ \) is satisfiable;
2. \( T \models \alpha \) if and only if \( T’ \models^b \alpha \), for each \( DL-Lite_R \) axiom \( \alpha \); and
3. \( a \in q^K \) if and only if \( q^{K’}(a) \geq 1 \), for each CQ \( q \) and each tuple \( a \) over \( I’ \).

All three statements also hold when the problems are considered under the UNA.

Proof. We concentrate on the general case; the case of the UNA is analogous.

Statement 1) Assume that \( K \) has a model \( I = \langle \Delta^I, .^I \rangle \). Let \( I’ = \langle \Delta^I, .^I \rangle \) be the bag interpretation defined as follows, for each \( a \in I, A \in C, P \in R \), and \( u, v \in \Delta^I \):

\[
a^I = a^I, \quad A^I(u) = \begin{cases} \infty, & \text{if } u \in A^I, \\ 0, & \text{otherwise}, \end{cases} \quad P^I(u, v) = \begin{cases} \infty, & \text{if } (u, v) \in P^I, \\ 0, & \text{otherwise}. \end{cases}
\]

Bag interpretation \( I’ \) satisfies \( A’ \) and all axioms in \( T \). Thus, \( I’ \) is a model of \( K’ \) and, therefore, \( K’ \) is satisfiable, as required. Conversely, suppose that \( K’ \) has a model \( I’ = \langle \Delta^I, .^I \rangle \). We construct an interpretation \( I = \langle \Delta^I, .^I \rangle \) as follows, for each \( a \in I, A \in C, P \in R \), and \( u, v \in \Delta^I \):

\[
a^I = a^I, \\
u \in A^I \text{ if and only if } A^I(u) > 0, \\
(u, v) \in P^I \text{ if and only if } P^I(u, v) > 0.
\]

Interpretation \( I \) is a model of \( K \) by construction, which completes the proof of Statement 1.

Statement 2) We show the claim by considering a case for each kind of axiom in \( T \).

Let first \( \alpha \) be \( S_1 \sqsubseteq S_2 \) where \( S_1 \) and \( S_2 \) are either both concepts or both roles. To show that \( T \models^b S_1 \sqsubseteq S_2 \) implies \( T \models S_1 \sqsubseteq S_2 \), assume that \( T \models^b S_1 \sqsubseteq S_2 \) but \( T \not\models S_1 \sqsubseteq S_2 \). Then, the following \( DL-Lite_R \) ontology must be satisfiable [9]: \( \langle T \cup \{ S \subseteq S_1 \} \cup \{ \text{Disj}(S, S_2) \} \cup \{ S(a) \} \rangle \), where \( S \) is fresh. By Statement 1, the \( DL-Lite^b_R \) ontology consisting of the same TBox and a bag ABox \( A’ = \{ S(a) : 1 \} \) is satisfiable. Thus, there exists a bag interpretation \( I \) such that \( I \models^b T \), \( S_1 \subseteq S_2 \), \( S_1 \cap S_2 = \emptyset \), and \( S_1(u) > 0 \) for some tuple \( u \) of domain elements. From this we derive that \( S_1^I(u) > 0 \) and \( S_2^I(u) = 0 \) and hence \( S_1^I \not\subseteq S_2^I \), which then implies \( T \not\models S_1 \sqsubseteq S_2 \), contradicting our assumption.
We now show that $\mathcal{T} \models S_1 \subseteq S_2$ implies $\mathcal{T} \models b^S S_1 \subseteq S_2$. Let $\mathcal{T}'$ be the TBox extending $\mathcal{T}$ with the following inclusions for each role inclusion $R_1 \subseteq R_2$ in $\mathcal{T}$: $\exists R_1 \subseteq \exists R_2$, $\exists R_1 \subseteq \exists R_1^-$, and $\exists R_1^+ \subseteq \exists R_2^+$. for $R_1$ and $R_2$, respectively. Following [34], $\mathcal{T} \models S_1 \subseteq S_2$ implies that either $\mathcal{T} \models \text{Disj}(S_1, S_1)$, or there exists a chain of inclusions $T_0 \subseteq T_1, \ldots, T_{n-1} \subseteq T_n$ in $\mathcal{T}'$ such that $T_0 = S_1$ and $T_n = S_2$. In the first case, we have that $\mathcal{T} \models b^S \text{Disj}(S_1, S_1)$ by definition. In the second case, the chain of inclusions in $\mathcal{T}'$ implies that $\mathcal{T}' \models b^S S_1 \subseteq S_2$ since $T_0 \subseteq T_1 \subseteq \cdots \subseteq T_n$ should hold for every bag interpretation $I$ satisfying $\mathcal{T}'$. Then, $\mathcal{T} \models b^S S_1 \subseteq S_2$ follows from the fact that every bag interpretation satisfying $\mathcal{T}'$ satisfies also the additional inclusions in $\mathcal{T}'$ by construction.

Let now $\alpha$ be $\text{Disj}(S_1, S_2)$ where $S_1, S_2$ are either both concepts or both roles. If $\mathcal{T} \not\models b^S \text{Disj}(S_1, S_2)$, then there exists a bag interpretation $I$ such that $\mathcal{T} \models b^S$ and $S_1^I \cap S_2^I \neq \emptyset$. Let $I'$ be the set interpretation constructed in the proof of Statement 1 on the basis of $I$. By construction we have that $I' \models \mathcal{T}$ and $S_1^I \cap S_2^I \neq \emptyset$; thus $\mathcal{T} \not\models \text{Disj}(S_1, S_2)$, as required. The other direction can be shown in exactly the same way.

\begin{itemize}
\item [(Statement 3)] First, let $a \in q^K$—that is, let $a$ belong to the certain answers to $q$ over $K$—and assume for the sake of contradiction that $q^K(a) = 0$. The latter means that there exists a model $I'$ of $K'$ such that $q^I(a) = 0$. Consider the interpretation $I$ constructed on the basis of $I'$, as in the proof of Statement 1. Interpretation $I$ is a model of $K$ such that $I \not\models q(a)$, which yields a contradiction. Thus, $q^K(a) \geq 1$, as required. The other direction can be shown in exactly the same way. \hfill \Box
\end{itemize}

5.2. Unique name assumption

As we mentioned before, in the set case general satisfiability and satisfiability under the UNA coincide for $DL-Lite_R$, and the same holds for the axioms of entailment. The following corollary, which states similar claims for bag semantics, is an immediate consequence of Statements 1 and 2 of Theorem 23 and this property.

**Corollary 24.** A $DL-Lite^b_R$ ontology is satisfiable if and only if it is satisfiable under the UNA. A $DL-Lite^b_R$ TBox entails an axiom under bag semantics if and only if it entails this axiom under bag semantics and the UNA.

Artale et al. [24] showed that query answering over satisfiable $DL-Lite^b_R$ ontologies is also independent of whether the UNA is adopted or not; indeed the UNA may influence query answers under set semantics only if the ontology language allows for some form of equality (e.g., functionality constraints). We next argue that the situation is markedly different under bag semantics. The following proposition shows that the UNA can influence query answering under bag semantics as soon as role inclusions are allowed in the ontology language.

**Proposition 25.** There exists a satisfiable $DL-Lite^b_R$ ontology $K$ and a rooted CQ $q$ such that the (general) certain answers to $q$ over $K$ differ from the certain answers under the UNA.

\begin{proof}
Let $K = \langle T, A \rangle$ be a $DL-Lite^b_R$ ontology with $T = \{P' \subseteq P\}$ and $A = \{P(a, b_1) : 1, P'(a, b_2) : 1\}$. Let also $q = \exists y. P(a, y)$. Under the UNA, we have that $q^K(1) = 2$. Indeed, in all models $I$ satisfying the UNA, individuals $b_1$ and $b_2$ are interpreted as different elements; as a result, in each such $I$, the element $a'$ is associated with at least two elements in $P'$. In contrast, if the UNA is not adopted, then the certain answer is 1, which is witnessed by an interpretation that maps $b_1$ and $b_2$ to the same element of the domain. \hfill \Box
\end{proof}

As we will see later on (in Corollary 43), the presence of role inclusions is crucial for this mismatch, and the certain answers to rooted CQs over $DL-Lite^b_CORE$ ontologies do not depend on whether the UNA is adopted.

5.3. Universal models

An important property of each satisfiable $DL-Lite_R$ ontology $K$ is the existence of so-called universal models for CQs—that is, models $I$ such that the certain answers to every CQ $q$ over $K$ can be obtained by evaluating $q$ over $I$ [9]. Existence of such universal models is critical to the favourable computational properties of $DL-Lite_R$. The notion of a universal model for bags is the same as for sets.

**Definition 26.** A model $I$ of an ontology $K$ is universal for a class of queries $Q$ if $q^K = q^I$ for all $q \in Q$. It is universal under the UNA if the certain answers under the UNA are considered.

In the set case, it is well-known that universal models for the class of CQs always exist for satisfiable ontologies $K$. In fact, they are canonical interpretations—that is, interpretations that can be obtained by a restricted chase procedure applied to $K$ [35]. It is also well-known that a model of $K$ is universal for the class of CQs and only if it can be homomorphically embedded into every other model of $K$ [9]. Unfortunately, in contrast to the set case, even $DL-Lite^b_CORE$ ontologies may not admit universal models for all CQs.
Proposition 27. There exists a satisfiable DL-Lite\textsuperscript{b} \textsubscript{CORE} ontology \( K \) that has neither a universal model nor a universal model under the UNA for the class of all CQs.

Proof. Let \( K = (T, A) \) be the DL-Lite\textsuperscript{b} \textsubscript{CORE} ontology with \( T = \{ A \sqsubseteq \exists P \cdot \exists P^- \sqsubseteq B \} \) and \( A = \{ A(a) : 1, B(b) : 1 \} \). Consider the bag interpretation \( I_1 \) with domain \( \{ a, b \} \) that interprets the individuals by themselves and interprets the concepts and roles as follows:

\[
A^I_1 = \{ a : 1 \}, \quad B^I_1 = \{ b : 1 \}, \quad P^I_1 = \{ (a, b) : 1 \}.
\]

Similarly, consider the bag interpretation \( I_2 \) with domain \( \{ a, b, u \} \) that interprets the individuals by themselves and interprets concepts and roles as follows:

\[
A^I_2 = \{ a : 1 \}, \quad B^I_2 = \{ b : 1, u : 1 \}, \quad P^I_2 = \{ (a, u) : 1 \}.
\]

It is immediate to verify that both \( I_1 \) and \( I_2 \) are models of \( K \). Moreover, for the Boolean CQs

\[
q_r = P(a, b) \quad \text{and} \quad q_{nr} = \exists y. B(y),
\]

we have that

\[
q^I_r(\langle \rangle) = 1, \quad q^I_{nr}(\langle \rangle) = 0,
\]

\[
q^I_1(\langle \rangle) = 1, \quad q^I_{nr}(\langle \rangle) = 2;
\]

thus, neither model is universal for \( \{ q_r, q_{nr} \} \).

Suppose now there is a universal model \( I \) for \( \{ q_r, q_{nr} \} \). Then, since \( q^I_r(\langle \rangle) \) must be 0, we have that \( P^I(\langle d, b \rangle) = 0 \). Since assertion \( A(a) \) occurs in \( A \) with multiplicity 1 and inclusion \( A \sqsubseteq \exists P \) belongs to \( T \), we have that \( P^I(\langle a, v \rangle) \geq 1 \) for some \( v \in \Delta^* \) distinct from element \( b \). Since inclusion \( \exists P^- \sqsubseteq B \) is in \( T \), it follows that \( B^I(\langle v \rangle) \geq 1 \), and, hence, \( q^I_{nr}(\langle \rangle) \geq 2 \), contradicting universality of \( I \).

Finally, note that the proof also works under the UNA. \( \square \)

6. Lower bounds for the data complexity of query answering under bag semantics

The lack of universal models illustrated in Section 5.3 suggests that CQ answering under bag semantics is computationally more challenging than in the set case. In this section, we show that this is indeed the case and establish three incomparable coNP lower bounds in data complexity. These are in stark contrast to the well-known \( \text{AC}^0 \) upper bound in the set case for CQ answering over DL-Lite\textsubscript{R}. We also establish two new lower bounds in this setting.

The first lower bound is given in Theorem 28, where we show that CQ answering is coNP-hard even if we restrict the ontology language to DL-Lite\textsubscript{b} \textsubscript{CORE} regardless of the adoption of the UNA. The second and third lower bounds are established in Theorem 29, where we show similar coNP-hardness results for the cases where the query language is restricted to the class of rooted CQs and the ontology language is allowed to contain role inclusions.

Theorem 28. Both \( \text{BagCert}^\text{CQs. DL-Lite}^\text{b} \text{CORE} \) and \( \text{BagCert}^\text{UNA} \text{CQs. DL-Lite}^\text{b} \text{CORE} \) are coNP-hard in data complexity.

Proof. We prove that there exists a DL-Lite\textsuperscript{b} \textsubscript{CORE} TBox \( T \) and a Boolean CQ \( q \) such that checking whether \( q(T, A) \geq k \) for an input bag ABox \( A \) and \( k \in \mathbb{N}_{\geq 0} \) is coNP-hard regardless of whether the UNA is adopted or not. Following ideas of Kostylev and Reutter [36], we provide a reduction of the complement of the 3-colourability problem for directed graphs, a well-known coNP-complete problem, to query answering.

We first address the case of BagCert\textsubscript{UNA}. Let \( G = (V, E) \) be a directed graph with vertices \( V \) and edges \( E \). We construct a DL-Lite\textsubscript{CORE} TBox \( T \) and a Boolean CQ \( q \), neither of which depends on \( G \), as well as a bag ABox \( A_G \) based on \( G \), such that \( G \) is not 3-colourable if and only if \( q(T, A_G) \geq 3 \times |V| + 2 \).

First, let \( T \) consist of the inclusions

\[
\text{Vertex} \sqsubseteq \exists \text{hasColour} \quad \text{and} \quad \exists \text{hasColour}^- \sqsubseteq \text{Colour},
\]

where \( \text{Vertex} \) and \( \text{Colour} \) are atomic concepts, and \( \text{hasColour} \) is an atomic role, and let \( q \) be the Boolean CQ

\[
\exists x, y, z, w. \text{Edge}(x, y) \land \text{hasColour}(x, z) \land \text{hasColour}(y, z) \land \text{Colour}(w).
\]

Then, let \( A_G \) be the bag ABox defined as given next, where we use an individual \( a_v \) for each vertex \( v \in V \), an individual \( a \) representing an auxiliary “vertex”, individuals \( r, g \), and \( b \) representing three colours, atomic role \( \text{Edge} \), as well as the concepts and roles introduced before:

- \( \text{Vertex}(a_v) \) has multiplicity 1, for each vertex \( v \in V \);
- \( \text{Edge}(a_v, a_{v'}) \) has multiplicity 1, for each edge \( (v_1, v_2) \in E \);
– Colour(r) has multiplicity |V| + 1 for colour r;
– Colour(g) and Colour(b) each has multiplicity |V| for colours g and b;
– Vertex(a), Edge(a, r) and hasColour(a, r) each has multiplicity 1 for the auxiliary “vertex” a; and
– all other assertions have multiplicity 0.

Concept Vertex and role Edge are used to encode G. The role hasColour represents a colour assignment to the vertices of G, where inclusions Vertex ⊑ hasColour ⊑ Colour necessitate the association of each vertex to a colour. Concept Colour provides a sufficient number of pre-defined copies of the three colours; every proper colour assignment of G shall use at most |V| times each of these colours. We next exploit these properties for showing that

\[ G \text{ is not 3-colourable if and only if } q^{(T, A_C)}(\langle \rangle) \geq 3 \times |V| + 2. \]

First, let G be not 3-colourable. Consider an arbitrary model I of \( \langle T, A_C \rangle \). We next show that \( q^I(\langle \rangle) \geq 3 \times |V| + 2. \) Since I is a model of bag ABox \( A_C \), the valuation \( \lambda \) defined as \( \lambda(x) = \lambda(y) = a^I \) and \( \lambda(z) = \lambda(w) = r^I \) contributes to \( q^I(\langle \rangle) \) a multiplicity of at least \(|V| + 1\); this is because \( A_C \) contains assertion Colour(r) with multiplicity \(|V| + 1\) and assertions Vertex(a), Edge(a, a), and hasColour(a, r) with multiplicity 1. Similarly, each valuation that differs from \( \lambda \) by sending \( u \) to either \( g^I \) or \( b^I \) contributes to \( q^I(\langle \rangle) \) a multiplicity of at least \(|V| \). We have two possibilities for \( I \): either there exists an element \( u \in \Delta^I \) different from \( r^I, g^I \) and \( b^I \) such that Colour \( ^I(u) \geq 1 \) or not. In the first case the valuation that differs from \( \lambda \) by sending \( u \) to \( r^I, g^I \) and \( b^I \) contributes to \( q^I(\langle \rangle) \) a multiplicity of at least 1, so overall we have \( q(\langle \rangle) \geq 3 \times |V| + 2 \), as required. In the second case we can consider a colour assignment \( \gamma \) to \( V \) such that, for every \( v \in V \), there exists \( I \) such that \( \gamma(v_1) = \gamma(v_2) \). Consider the valuation \( \gamma' \) such that \( \gamma'(x) = a^I \), \( \gamma'(y) = a^I \), \( \gamma'(z) = b^I \), \( \gamma'(w) = r^I \). By construction, \( \gamma' \) contributes to \( q^I(\langle \rangle) \) a multiplicity of at least \(|V| + 1\). Therefore, overall we have that \( q^I(\langle \rangle) \geq 3 \times |V| + 2 \), as required.

Assume now that \( G \) is 3-colourable. It suffices to show that there exists a model I of \( \langle T, A_C \rangle \) for which \( q^I(\langle \rangle) < 3 \times |V| + 2. \) Since G is 3-colourable, there is an assignment \( \gamma : V \to \{r, g, b\} \) such that, for every \( (v_1, v_2) \in E \), \( \gamma(v_1) \neq \gamma(v_2) \). Consider a bag interpretation I with the domain consisting of all the individuals (i.e., \( \Delta^I = \{a \mid v \in V \} \cup \{a, r, g, b\} \) that interprets all the individuals by themselves, and such that \( \text{Vertex}, \text{Edge}\) and Colour\( ^I \) are defined precisely according to \( A_C \) (e.g., Colour\( ^I(a) = A_C(\text{Vertex}(a)) \) for every individual \( a \)), while

\[
\text{hasColour}^I(u_1, u_2) = \begin{cases} 
1, & \text{if } u_1 = a \text{ and } u_2 = \gamma(v) \text{ for } v \in V, \\
1, & \text{if } u_1 = a \text{ and } u_2 = r, \\
0, & \text{otherwise}.
\end{cases}
\]

In other words, interpretation I is defined on the basis of the 3-colouring of G. By construction, I is a model of \( \langle T, A_C \rangle \). Next, we show that \( q^I(\langle \rangle) = 3 \times |V| + 1. \) First, we observe that the first three atoms \( \text{Edge}(x, y), \text{hasColour}(x, z) \), and \( \text{Colour}(y, z) \) of \( q \) match exactly once (i.e., under the valuation sending \( x \) to \( a \) and \( z \) to \( r \)). Next, there are precisely three possibilities for variable \( w \), namely \( r, g, \) and \( b \), contributing multiplicity \( 3 \times |V| + 1 \) in total. Consequently, \( q^I(\langle \rangle) = 3 \times |V| + 1 \), as desired.

We now address the case of BagCert0 by discussing the required modifications in the aforementioned reduction. For this, it is enough to ensure that, first, the auxiliary “vertex” a is not interpreted by the same element as any of the vertices of \( G \); and, second, that the colour individuals \( r, g, b \) are interpreted by pairwise different elements. To ensure this, we use atomic concepts \( V_a, V_g, \text{Red}, \text{Green}, \) and \( \text{Blue} \). We add the following disjointness axioms to TBox \( T \): Disj\( (V_a, V_g) \), Disj\( (\text{Red}, \text{Green}) \), Disj\( (\text{Red}, \text{Blue}) \), and Disj\( (\text{Green}, \text{Blue}) \). We also modify bag ABox \( A_C \) by setting the multiplicity of \( V_a(a), \text{Red}(r), \text{Green}(g), \text{Blue}(b), \) and \( V_g(a) \), for every vertex \( v \in V \), to 1 (and the multiplicity of all other assertions over the new concepts to 0). Following the same argumentation as for the case of BagCert0UNA, we can show that the above reduction works when the UNA is dropped. \( \square \)

Note that the query constructed in the proof of Theorem 28 is not rooted; furthermore, the use of the disconnected atom Colour\( ^I(\langle \rangle) \) in the query is instrumental to the correctness of the reduction. In Section 8 we show that rooted CQs are rewritable to BCALC over DL-Lite\( _{rcore} \) regardless of the adoption of the UNA—that is, the problems are in LogSPACE in data complexity.

Unfortunately, the restriction to rooted CQs alone is not sufficient to ensure tractability of query answering for bag ontology languages allowing for role inclusions. In the first part of Theorem 29 we show that answering rooted CQs is intractable (coNP-hard) even if we restrict ourselves to DL-Lite\( _{rcore} \) ontologies, which allow for role inclusions while at the same time disallowing existential quantification on the right-hand side of concept inclusions. This lower bound, however, critically depends on the fact that the UNA is not adopted; indeed, in Section 9 we will show that all CQs (and not just rooted ones) are rewritable to BCALC over DL-Lite\( _{ epis } \) under the UNA. On the other hand, even if adopting the UNA can make rooted CQ answering easier, in the second part of Theorem 29 we show that it remains intractable in general: answering rooted CQs over DL-Lite\( _{rip} \) is coNP-hard under the UNA.
Theorem 29. BagCert$^b_{\text{rooted CQs}, \ DL-Lite^b_{\text{RDLP}}}$ and BagCert$^{\text{UNA}}_{\text{rooted CQs}, \ DL-Lite^b_{\text{RDLP}}}$ are coNP-hard in data complexity.

Proof. We first prove the claim for the case of BagCert, and then show how to adapt the proof to the case of BagCert$^{\text{UNA}}$. The proof is again by reduction of the complement of the 3-colouring problem for directed graphs; however, the reduction is more involved. Let $G = (V, E)$ be a directed graph with vertices $V$ and edges $E$. Next, we define a $DL-Lite^b_{\text{RDLP}}$ TBox $\mathcal{T}$ and Boolean rooted CQ $q$, neither of which depends on $G$, as well as a bag ABox $\mathcal{A}_G$, based on $G$, such that $G$ is not 3-colourable if and only if $q^{(\mathcal{T}, \mathcal{A}_G)}() \geq 3 \times |V| + 2$.

First, let $\mathcal{T}$ consist of a single role inclusion

\[
\text{hasColour} \sqsubseteq \text{Colour},
\]

where hasColour and Colour are atomic roles. Let also $q$ be the following Boolean rooted CQ, where $a_0$ is a "root" individual, and Edge, Beg, End and Vertex are atomic roles:

\[
\exists x_v, x_c, y_e, y_1^v, y_2^v, y_c. \text{Vertex}(a_0, x_v) \land \text{Colour}(x_v, x_c) \land \text{Edge}(a_0, y_e) \land \text{Beg}(y_e, y_1^v) \land \text{hasColour}(y_1^v, y_c) \land \text{End}(y_e, y_2^v) \land \text{hasColour}(y_2^v, y_c).
\]

Finally, let the bag ABox $\mathcal{A}_G$ mention the "root" individual $a_0$ of $q$, individuals $a_v$ and $c_v$ associated to vertices $v \in V$, individuals $r, g, b$ corresponding to the three colours, individuals $a_e$ associated to edges $e \in E$, as well as an auxiliary "vertex" individual $a$ and "edge" individual $a_e$: let also $\mathcal{A}_G$ assign 1 to the following assertions (and 0 to all others):

- $\text{Vertex}(a_0, a_e)$, $\text{Colour}(a_e, r)$, $\text{Colour}(a_e, g)$, $\text{Colour}(a_e, b)$, $\text{hasColour}(a_v, c_v)$, for each vertex $v \in V$,
- $\text{Edge}(a_0, a_e)$, $\text{Beg}(a_e, a_v)$, $\text{End}(a_e, a_v)$, for each $e = (v_1, v_2) \in E$,
- $\text{Vertex}(a_0, a)$, $\text{Colour}(a, r)$, $\text{hasColour}(a, r)$, and
- $\text{Edge}(a_0, a_e)$, $\text{Beg}(a_e, a)$, $\text{End}(a_e, a)$.

Having the reduction complete, next we show that it is correct—that is, that

$G$ is not 3-colourable if and only if $q^{(\mathcal{T}, \mathcal{A}_G)}() \geq 3 \times |V| + 2$.

Intuitively, every model of $(\mathcal{T}, \mathcal{A}_G)$ has $3 \times |V| + 1$ contributing valuations for $q$ that send the subquery of $q$ over the $y$ variables to the interpretations of the assertions of the subquery over the auxiliary $a_v$, $a$, and $r$, while the subquery over the $x$ variables to the assertions over one of $a_v$ and $a$, and one of $r$, $g$, and $b$. Then, if some $c_v$ is interpreted as neither $r$, nor $g$, nor $b$, we can construct one more valuation sending $x_v$ to the interpretation of $c_v$ (here we make use of the TBox $\mathcal{T}$). Otherwise, the identifications of $c_v$ can be seen as a colouring of the vertices (represented by $a_v$ individuals), and every valid colouring corresponds to the model possessing exactly $3 \times |V| + 1$ valuations. Next, we make this intuition formal. In fact, in the both directions of the correctness proof we make use of the following fact.

Claim 30. For every bag interpretation $I$ satisfying all assertions in $\mathcal{A}_G$, $q^I() \geq 3 \times |V| + 1$.

Proof. Let $I$ be a bag interpretation satisfying ABox $\mathcal{A}_G$. Consider all the valuations $\lambda$ such that $\lambda(y_e) = a_e^r$, $\lambda(y_1^v) = \lambda(y_2^v) = a^1$, $\lambda(y_c) = r^1$, as well as $\lambda(x_v)$ is one of $a_v$ for $v \in V$, and $\lambda(x_c)$ is one of $r^1$, $g^1$, and $b^1$. Since $I$ satisfies $\mathcal{A}_G$, each of these valuations contribute at least 1 to $q^I()$, and there are overall $3 \times |V|$ of them. Note that we rely only on the cardinality of the ABox here, so even if the interpretations of the individuals are not pairwise distinct—that is, if the UNA is violated—and some of these valuations may coincide, the total contribution of these valuations is still at least $3 \times |V|$ by Definition 15. Consider now the valuation $\lambda$ that is the same as before on $y_e$, $y_1^v$, $y_2^v$ and $y_c$, but such that $\lambda(x_v) = a^1$ and $\lambda(x_c) = r^1$. This valuation also contributes a multiplicity of at least 1. Moreover, for the same reason as before, the contribution of each of the considered valuations is separate—that is, the total contribution is at least $3 \times |V| + 1$. □

Having this claim at hand, we are ready to show correctness of the reduction. Let first $G$ be not 3-colourable. Consider an arbitrary model $I$ of $(\mathcal{T}, \mathcal{A}_G)$. Since $I$ satisfies all assertions in $\mathcal{A}_G$, by Claim 30 we know that there are valuations that contribute $3 \times |V| + 1$ to $q^I()$. So, it is enough to show that there is a valuation with a non-zero and different contribution. We have two possibilities: either there is a vertex $v \in V$ such that $c_v^1$ is distinct from $r^1$, $g^1$, and $b^1$, or not.

In the first case, consider such a vertex $v$ and the valuation $\lambda$ that is the same as in Claim 30 on $y_e$, $y_1^v$, $y_2^v$ and $y_c$, but such that $\lambda(x_v) = a^1$ and $\lambda(x_c) = c_v^1$. On the one hand, $\mathcal{A}(\text{hasColour}(a_v, c_v)) = 1$ and $\mathcal{T}$ has the inclusion $\text{hasColour} \sqsubseteq \text{Colour}$, so $\text{Colour}^I(a_v, c_v^1) \geq 1$ and, therefore, the contribution of this valuation is at least 1. On the other hand, we have not considered this valuation yet, because $c_v^1$ is different from $r^1$, $g^1$, and $b^1$ by assumption.

Consider now the second case—that is, the case when $c_v^1$ is among $r^1$, $g^1$, and $b^1$ for each $v$. Construct a colour assignment $\gamma$ to vertices such that, for each such vertex $v$, $\gamma(v)$ is red if $c_v^1 = r^1$, it is green if $c_v^1 = g^1$, and it is blue if $c_v^1 = b^1$ (if some of $r^1$, $g^1$, and $b^1$ coincide, then there are many of such assignments and $\gamma$ can be any of them). We know that $G$ does not have a valid colouring, so there is an edge $e = (v_1, v_2)$ in $E$ with both $v_1$ and $v_2$ assigned to the same colour. For brevity,
consider only the case when this colour is red; the other two cases are symmetric. Let \( \lambda \) be the valuation that agrees with a valuation in Claim 30 on the variables \( x_a \) and \( x_b \), and follows the following assignment for the rest of the variables: \( \lambda(y_e) = a_e^r, \lambda(y_1^v) = a_1^r, \lambda(y_2^v) = a_2^v, \lambda(y_e) = r^f \). On the one hand, the contribution of this valuation to \( q^l(\overline{v}) \) is at least 1 by construction. On the other hand, this valuation is different from the ones considered in Claim 30.

Therefore, in both cases we have that \( q^l(\overline{v}) \geq 3 \times |V| + 2 \), as required.

Let now \( G \) be 3-colourable—that is, there is a colour assignment \( \gamma \) to \( V \) such that the vertices of each edge are coloured differently. We next show that \( q^{(T, A^C)}(\overline{v}) < 3 \times |V| + 2 \). To this end, consider the bag interpretation \( I \) defined as follows:

- for each vertex \( v \in V \), \( c^I_v \) is \( r \) if \( \gamma(v) \) is red, it is \( g \) if \( \gamma(v) \) is green, and it is \( b \) if \( \gamma(v) \) is blue;
- all other individuals are interpreted by themselves; and
- all the atomic roles are interpreted as dictated by the ABox (i.e., \( S^I(u_1^1, u_2^1) = A^C(S(u_1, u_2)) \)) for every atomic role \( S \) and every pair of individuals \( u_1, u_2 \).

On the other hand, interpretation \( I \) is a model of \( (T, A^C) \) by construction, because the identification of all \( c^I_v \) with one of \( r, g \) and \( b \) makes \( I \) satisfy the inclusion of \( T \) for the only relevant assertions \( hasColour(a_v, c_v) \). On the other hand, \( q^l(\overline{v}) = 3 \times |V| + 1 \): indeed, it is at least \( 3 \times |V| + 1 \) by Claim 30, and it is immediate to check that there are no more valuations with a non-zero contribution to \( q^l(\overline{v}) \). So, \( I \) is a witness for the fact that \( q^{(T, A^C)}(\overline{v}) < 3 \times |V| + 2 \).

We are left to show the second part of the theorem—that is, coNP-hardness of BagCERT for the case when the UNA is adopted, but arbitrary DL-Lite\textsubscript{R} TBoxes are allowed. In fact, we can essentially repurpose the same reduction as in the first part. The only modifications in the reduction are that the ABox \( A^C \) does not have assertions \( hasColour(a_v, c_v) \), for \( v \in V \) (i.e., does not use individuals \( c_v \) at all), while the TBox \( T \) additionally has the inclusion \( \exists v \overline{vertex} \subseteq \exists hasColour \). Then, the correctness proof goes along the same lines as in the first case, except that anonymous domain elements \( u_v \), which are enforced by the new inclusion for each \( v \in V \), are used instead of the \( c^I_v \).

Since the data complexity of BCALC is strictly contained in LogSpace, the lower bounds in Theorems 28 and 29 imply non-rewritability to BCALC for the relevant query and ontology languages.

**Corollary 31.** The class of all CQs is not rewritable to BCALC over DL-Lite\textsubscript{B} over DL-Lite\textsubscript{R}, both in general and under the UNA. The class of rooted CQs is not rewritable to BCALC over DL-Lite\textsubscript{B} over DL-Lite\textsubscript{R} under the UNA.

In the following sections, we investigate how to regain tractability of query answering and rewritability to BCALC by considering suitable restrictions on the query and ontology languages that allow us to circumvent the bounds in Theorems 28 and 29. In Sections 7 and 8 we focus on DL-Lite\textsubscript{CORE} and show that the class of rooted CQs is rewritable to BCALC in general and under the UNA. In Section 9 we focus on the ontology language DL-Lite\textsubscript{R} and show that all CQs (and not just rooted ones) are rewritable to BCALC under the UNA. Finally, in Section 10 we show rewritability of rooted CQs over DL-Lite\textsubscript{R}, which extends both DL-Lite\textsubscript{CORE} and DL-Lite\textsubscript{R} under the UNA. For the convenience of the reader, we summarise in Table 1 all the data complexity results proved in this paper.

### 7. Universal models for rooted conjunctive queries over DL-Lite\textsubscript{B} ontologies

Our next main goal is to show tractability of answering rooted CQs over DL-Lite\textsubscript{B} in data complexity and their BCALC rewritability, both regardless of the adoption of the UNA. Towards this goal, in this section we show that every satisfiable DL-Lite\textsubscript{B} ontology admits a universal model for rooted CQs, both in general and under the UNA. To this end, we proceed as in the set case: we first define a special bag interpretation for each DL-Lite\textsubscript{B} ontology, which we call *canonical*, and then, after developing dedicated machinery, prove that it is indeed universal for the class of rooted CQs when the ontology is satisfiable. However, in contrast to the set case, the requirement for CQs to be rooted is crucial here: recall Proposition 27, where we constructed a DL-Lite\textsubscript{B} ontology that does not have a universal model for all CQs.

To formalise canonical bag interpretations, we need two auxiliary notions. First, the concept closure \( \text{cl}_T[u, l] \) of an element \( u \in \Delta^I \) in a bag interpretation \( I = (\Delta^I, I) \) over a TBox \( T \) is the bag of concepts such that, for every concept \( C \),...
ccl\(\mathcal{T}[u, I](C) = \max\{C^I_0(u) \mid \mathcal{T} \models C_0 \subseteq C\}\).

In other words, ccl\(\mathcal{T}[u, I](C)\) is the maximum value of \(C^I_0(u)\) amongst all concepts \(C_0\) satisfying \(\mathcal{T} \models C_0 \subseteq C\)—that is, ccl\(\mathcal{T}[u, I](C)\) is the minimal multiplicity of \(C^I(u)\) required for an extension \(J\) of \(I\) to satisfy TBox \(\mathcal{T}\) locally in \(u\).

Second, the union \(I \cup J\) of two bag interpretations \(I = \langle \Delta^I \cdot J_1 \rangle\) and \(J = \langle \Delta^J \cdot J_2 \rangle\) interpreting all the individuals in the same way—that is, such that \(a^I = a^J\) for all \(a \in I\)—is the bag interpretation \(\langle \Delta^I \cup \Delta^J \cdot J_1 \cup J_2 \rangle\) with \(a^I \cup J_1 = a^J\) for all individuals \(a \in I\) and \(S^I \cup J_1 = S^I \cup J_2\) for all atomic concepts and roles \(S \in C \cup R\) (recall that \(S^I\) and \(S^J\) are bags, so \(S^I \cup S^J\) is the bag maximal union).

Definition 32. The canonical bag interpretation \(Can(K)\) of a DL-Lite\(^b\)\_CORE ontology \(K = \langle \mathcal{T}, \mathcal{A} \rangle\) is the bag interpretation that is the union \(\bigcup_{i \geq a} Can_i(K)\) of the bag interpretations \(Can_i(K)\) defined as follows:

\(- \ Can_0(K) = \langle \Delta^{Can_0(K)} \cdot Can_0(K) \rangle\) is the bag interpretation corresponding to bag ABox \(\mathcal{A}\)—that is, such that \(\Delta^{Can_0(K)} = I\), \(a^{Can_0(K)} = a\) for each \(a \in I\), and \(S^{Can_0(K)}(a) = \mathcal{A}(a)\) for each \(S \in C \cup R\) and individuals \(a\);

\(- \) for each \(i > 0\), \(Can_i(K) = \langle \Delta^{Can_i(K)} \cdot Can_i(K) \rangle\) extends \(Can_{i-1}(K)\) by satisfying all the inclusions that are not satisfied in \(Can_{i-1}(K)\)—that is,

\[\Delta^{Can_i(K)} = \Delta^{Can_{i-1}(K)} \cup \{w_{u \cdot R}^1, \ldots, w_{u \cdot R}^\delta \mid u \in \Delta^{Can_{i-1}(K)}\}\]

and \(R\) is a role such that

\[\delta = \text{ccl}_{\mathcal{T}}[u, Can_{i-1}(K)](3R) \cdot (3R)\text{Can}_{i-1}(K)(u)\]

where \(w_{u \cdot R}^j\) are fresh domain elements, called anonymous, and, for all \(a \in I, A \in C, R \in R,\) and domain elements \(u\) and \(v\),

\(a^{Can_i(K)} = a\),

\(A^{Can_i(K)}(u) = \begin{cases} 
\text{ccl}_{\mathcal{T}}[u, Can_{i-1}(K)](A, \cdot) & \text{if } u \in \Delta^{Can_{i-1}(K)} \\
0 & \text{otherwise}
\end{cases}\)

\(p^{Can_i(K)}(u, v) = \begin{cases} 
P^{Can_{i-1}(K)}(u, v) & \text{if } u, v \in \Delta^{Can_{i-1}(K)} \\
1 & \text{if } u = w_{u \cdot R}^j \text{ or } v = w_{v \cdot R}^j \\
0 & \text{otherwise}
\end{cases}\)

We have just defined canonical bag interpretations in a declarative way. Note, however, that they can also be obtained by applying a variant of the restricted chase procedure [35] extended to bags—a procedure where, starting from the ABox, violations of the inclusions in the TBox are successively “repaired” by extending the interpretation of concepts and roles in a minimal way. We now illustrate Definition 32 with an example.

Example 33. Consider the DL-Lite\(^b\)\_CORE ontology \(K = \langle \mathcal{T}, \mathcal{A} \rangle\) specified as

\[\mathcal{T} = \{ A \sqsubseteq B, B \sqsubseteq \exists P, \exists P^\neg \sqsubseteq \exists R \} \text{ and } \mathcal{A} = \{ A(a) : 2, A(b) : 4, B(b) : 3, P(b, d) : 2, P(a, d) : 1, R(d, c) : 1 \}\]

The canonical bag interpretation of \(K\) is depicted in Fig. 1. Let us now follow the steps involved in its construction according to Definition 32. Interpretation \(Can_0(K)\), enclosed in the shaded area, reflects the bag ABox \(\mathcal{A}\): interpretations of concepts are shown by labels of the individuals, which include the multiplicities (e.g., label \(A : 2\) of \(a\) indicates that \(A^{Can_0(K)}(a) = 2\)), while interpretations of roles are shown by solid lines (e.g., two lines between \(b\) and \(d\) labelled by \(P\) indicate that \(P^{Can_0(K)}(b, d) = 2\)). Interpretation \(Can_1(K)\) increases the multiplicities of the elements \(a\) and \(b\) in the interpretation of \(B\) (e.g., \(b\) has a label \(B : 3\) inside the shaded area indicating that \(B^{Can_1(K)}(b) = 3\) and a label \(B : 4\) outside of it indicating...
that $B_{\text{Can}_1(K)}(b) = 4$, introduces the anonymous domain elements $w^{1}_{b,p}$, $w^{2}_{b,p}$, $w^{1}_{a,p}$, $w^{1}_{d,R}$ and $w^{2}_{d,R}$, and extends the interpretations of the roles accordingly, which is shown by dashed lines. Interpretation $\text{Can}_2(K)$ further introduces three anonymous elements and extends the interpretations of the roles. No further changes occur for $i > 2$, and hence $\text{Can}(K) = \text{Can}_0(K) \cup \text{Can}_1(K) \cup \text{Can}_2(K)$. - \\

Inspecting Definition 32 and Example 33, we observe that every canonical bag interpretation interprets each concept with a bag of individuals but only with a set of anonymous elements; similarly, multiplicities greater than 1 in the interpretations of roles are possible only for pairs of elements that are both individuals. This is an important property of canonical bag interpretations, which we are going to use in this section.

Note that the canonical bag interpretation satisfies the UNA. The following is another simple and intuitive observation, which holds regardless of the UNA and can be checked by the construction.

**Proposition 34.** If a DL-Lite$^b_\text{Core}$ ontology is satisfiable then its canonical bag interpretation is its model.

In the rest of this section, we show that the canonical bag interpretations of satisfiable DL-Lite$^b_\text{Core}$ ontologies are universal models for the class of rooted CQs regardless of the adoption of the UNA. There are two key ideas here, which are similar to the set case, but more subtle.

First, the canonical bag interpretation of a satisfiable DL-Lite$^b_\text{Core}$ ontology admits a homomorphism of a special type to every model of the ontology. Such homomorphisms, which we call multiplicity-preserving on the individuals, have a hybrid nature: for the concept interpretations, they preserve multiplicities of the interpretations of individuals, but are not required to do so for anonymous elements; similarly, for role interpretations, they preserve multiplicities of the pairs having at least one element being the interpretation of an individual, but are not required to do so for pairs of anonymous elements.

Second, each valuation of a rooted CQ over the canonical bag interpretation sends at least one term of each connected component to the interpretation of an individual. So, since DL-Lite$^b_\text{Core}$ does not allow for role inclusions, if two such valuations are different, then they are different on the non-anonymous part of the canonical interpretation. Moreover, the canonical bag interpretation is essentially set-based on the anonymous elements. Therefore, a valuation contributing to the answers and its multiplicity are determined solely by the non-anonymous part of the image of the valuation.

Putting these two ideas together, we can conclude that the composition of a valuation of a rooted CQ over the canonical bag interpretation and a homomorphism from the canonical interpretation to another model that is multiplicity-preserving on the individuals is also a valuation, and its contribution to the certain answers over the latter model is at least as large as the contribution of the former valuation over the canonical interpretation; moreover, different valuations over the canonical interpretation contribute independently to the certain answers over the latter model. This means that the canonical bag interpretation has the smallest possible certain answers to every rooted CQ—that is, by definition, it is the universal model for the class of such queries.

Even though these ideas may seem quite intuitive, their formalisation requires additional machinery, which we do not have yet. The problem is that with the current terminology we cannot unambiguously refer to each particular occurrence of an element in a bag, which is highly desirable for the formalisation. Therefore, we start by introducing new terminology for bags and other bag-based notions.

**Definition 35.** An enumerated bag (or e-bag) $\Theta$ over a set $M$ is a set of pairs $[c:m]$ with $c \in M$ and positive integer $m \in \mathbb{N}$, such that if $[c:m] \in \Theta$ then $[c:m-1] \notin \Theta$ for all $m \in \mathbb{N}$.

There is a straightforward one-to-one correspondence between bags and e-bags, and we call the e-bag corresponding to a bag $\Omega$ the enumerated version of $\Omega$ and denote it $\Omega^e$. We can extend this notation to bag interpretations and consider the enumerated version $I^e$ of a bag interpretation $I = (\Delta^I, J^I)$ defined as the pair $(\Delta^I, J^e)$ such that $a^e = a^I$ for each individual $a$ and $S^e = (S^I)^e$ for each concept or role $S$.

The use of enumerated versions of interpretations allows us to refer, in an unambiguous way, to the different occurrences of elements and pairs of elements in the bags corresponding to concepts and roles, respectively. An enumerated homomorphism between two interpretations is then defined as a standard homomorphism that additionally establishes a correspondence for each enumerated tuple of elements in each bag of the relevant bag interpretations.

**Definition 36.** Given two bag interpretations $I$ and $J$, an enumerated homomorphism (or e-homomorphism) from $I^e = (\Delta^I, J^e)$ to $J^e = (\Delta^J, J^e)$ is a family $(h, h_S, \ldots)$, with $S \in C \cup R$, of functions

$$
\begin{align*}
  h : \Delta^I & \to \Delta^J, \\
  h_S : S^I_e & \to S^J_e, \quad \text{for all } S \in C \cup R,
\end{align*}
$$

such that

- $h(a^e) = a^I$ for each $a \in I$,
- $h_A([u:m]) = [h(u):\ell]$ for all $A \in C$ and $[u:m] \in A^I$, where $\ell$ is a number in $\mathbb{N}$,
− \( h_P([(u, v):m]) = [(h(u), h(v)): \ell] \) for all \( P \in \mathcal{R} \) and \( [(u, v):m] \in P^{\ell} \), where \( \ell \) is a number in \( \mathbb{N} \).

To handle some cases uniformly, we sometimes write \( h_P-([(u, v):m]) \) instead of \( h_P([(u, v):m]) \), for \( P \in \mathcal{R} \).

E-homomorphisms have no essential differences with usual homomorphisms because they can send several enumerated tuples to just one without any restrictions. In contrast, the next definition formalises the aforementioned idea of multiplicity preservation: e-homomorphisms that are multiplicity-preserving on the individuals preserve multiplicities on the non-anonymous part of the source interpretation.

**Definition 37.** An e-homomorphism \((h, h_5, \ldots)\) from \( I^n = (\Delta^I, I^e) \) to \( J^n = (\Delta^J, J^e) \) is multiplicity-preserving on individuals \( I \) if, for each \( a \in I \), the following holds, where \( u = a^{\Delta^I} \):

− \( h_A([(u:m)]) \neq h_A([(u:\ell)]) \) for all \( A \in \mathcal{C} \) and all \( [(u:m)], [(u:\ell)] \in A^{\Delta^I} \) with \( m \neq \ell \),
− \( h_R([(u, v_1):m]) \neq h_R([(u, v_2):\ell]) \) for all \( R \) and all \( [(u, v_1):m], [(u, v_2):\ell] \in R^{\Delta^I} \) with \( v_1 \neq v_2 \) or \( m \neq \ell \).

The following lemma then formalises the first key idea about the canonical bag interpretation in terms of e-homomorphisms that are multiplicity-preserving on the individuals.

**Lemma 38.** For every DL-Lite\(_{\mathcal{C}_\text{core}}\) ontology \( K \) and every model \( I \) of \( K \) there exists an e-homomorphism from \( \text{Can}^e(K) \) to \( I^n \) that is multiplicity-preserving on \( I \).

**Proof.** Consider a DL-Lite\(_{\mathcal{C}_\text{core}}\) ontology \( K = (\mathcal{T}, \mathcal{A}) \) with the canonical bag interpretation \( \text{Can}(K) = \bigcup_{i \geq 0} \text{Can}(K)_i \) such that \( \text{Can}(K)_0 = (\Delta^\mathcal{T}(K), I^e(K)) \) and a model \( I = (\Delta^I, I^e) \) of \( K \). We first define a witnessing e-homomorphism \((h, h_5, \ldots)\) from \( \text{Can}^e(K) \) to \( I^e \) for the elements in \( \Delta^\mathcal{T}(K) \)—that is, for the (interpretations of the) individuals \( I \)—that is multiplicity-preserving on \( I \), then extend it to the elements introduced in \( \text{Can}(K)_1 \)—that is, to the anonymous elements on the first level of the canonical bag interpretation—and, finally, inductively define it on all other elements.

To begin, recall that \( a = a^{\Delta^\mathcal{T}(K)} \) for every \( a \in I \) and let \( h : \Delta^\mathcal{T}(K) \rightarrow \Delta^I \) be such that \( h(a^{\Delta^\mathcal{T}(K)}) = a \) for every \( a \in \Delta^\mathcal{T}(K) \)—that is, for every \( a \in I \). We now define function \( h_A : \text{Can}^e(K)_0 \rightarrow I^e \) for every \( S \in \mathcal{C} \cup \mathcal{R} \). For this, consider a tuple of individuals \( a \) such that \( S^{\text{Can}(K)}(a) = k \), for \( k \in \mathbb{N} \)—that is, such that \( [a:m] \in S^{\text{Can}(K)}_m \) for all \( m \in \mathbb{N} \) with \( m \leq k \). By the definition of \( \text{Can}(K)_0 \), we have that \( A(S(a)) = k \). Since \( I \) is a model of \( K \), it satisfies \( A \), and, in particular,

\[
S^I(a^I) \geq \sum_{b \text{ tuple over } I} A(S(b)) \geq A(S(a)) = k.
\]

As a result, since \( h(a) = a^I \), we have that \( [h(a):m] \in \text{Can}^e(K) \) for all \( m \leq k \); thus we can set \( h_5([(a:m)]) = [h(a):m] \) for all such \( m \). By construction, \((h, h_5, \ldots)\) satisfies all conditions stipulated by Definitions 36 and 37, thus it is an e-homomorphism from \( \text{Can}^e(K)_0 \) to \( I^e \) that is multiplicity-preserving on \( I \).

We now show the claim for \( \text{Can}^e(K)_1 \). To this end, we extend \((h, h_5, \ldots)\) from the case of \( \text{Can}^e(K)_0 \). By construction of \( \text{Can}^e(K)_1 \), it is enough to extend \( h_A \) on \( \Delta^\mathcal{T}(K) \) to \( I \) for \( A \in \mathcal{C} \), to define \( h \) on the anonymous elements introduced to the domain of \( \text{Can}^e(K)_1 \), and to define \( h_A \) on the role links between these anonymous elements and corresponding individuals.

For the extension for atomic concepts, consider an arbitrary concept name \( A \in \mathcal{C} \) and an element \( u \in \Delta^\mathcal{T}(K) \) to \( I \). By definition, we have \( A^{\text{Can}(K)}(u) = ccl_T\{u, \text{Can}(K)(u)\}(A) \). Therefore, to show that it is possible to extend \( h_A \) in such a way that different enumerated elements are sent to different enumerated elements, it suffices to prove that \( A^I(u^I) \geq ccl_T\{u, \text{Can}(K)(u)\}(A) \). By the definition of concept closure, there must exist a concept \( C \) such that \( T \models C \subseteq A \) and \( C^{\text{Can}(K)}(u) = ccl_T\{u, \text{Can}(K)(u)\}(A) \). If \( C \) is an atomic concept, then \( C^{\text{Can}(K)}(u) = A(u) \) by construction, and \( A^I(u^I) \geq A(C(u)) \) follows from the fact that \( I \) is a model of both \( A \) and \( T \), and the fact that \( T \models C \subseteq A \) is equivalent to \( T \models b \subseteq C \subseteq A \) (see Statement 2 of Theorem 23).

For the extension for introduced anonymous elements and corresponding role links, it is enough to consider arbitrary \( P \in \mathcal{R} \) and \( u \in \Delta^\mathcal{T}(K) \) to \( I \). By definition, \( \Delta^\mathcal{T}(K) \) extends \( \text{Can}^e(K) \) with fresh anonymous elements \( w_{I,u,P}^j \), for \( j = 1, \ldots, ccl_T\{u, \text{Can}(K)(u)\}(\exists P) \), \( \text{Can}^e(K)(u) \), and \( p^{\text{Can}(K)}(u, w_{I,u,P}^j) \) is set to 1 for each of these elements; the same is done for \( P^{-} \) instead of \( P \). We consider only the first case, since the second can be handled in the same way. To extend \( h \) to the new elements \( w_{I,u,P}^j \) and \( h_P \) to the pairs \( (u, w_{I,u,P}^j) \), it suffices to show that \( (\exists P)^{\text{Can}(K)}(u) \geq ccl_T\{u, \text{Can}(K)(u)\}(\exists P) \). This can be done in exactly the same way as the proof of \( A^I(u^I) \geq ccl_T\{u, \text{Can}(K)(u)\}(A) \) in the atomic concept case, by taking \( \exists P \) instead of \( A \).

To complete the proof we argue that having an e-homomorphism \((h, h_5, \ldots)\) from \( \text{Can}^e(K) \) to \( I^n \) that is multiplicity-preserving on \( I \) for \( i \geq 1 \), we can extend it to \( \text{Can}^e_{i+1}(K) \). This can be shown analogously to the case of \( \text{Can}^e(K) \). The only additional observation is that for every \( i \geq 1 \) and every \( S \in \mathcal{C} \cup \mathcal{R} \), bag \( S^{\text{Can}(K)} \) extends \( S^{\text{Can}(K)}_i \) with tuples \( u \) of elements that are all anonymous, for which \( S^{\text{Can}(K)}_i(u) = 1 \) by definition.
We next move to the formalisation of the second intuitive idea. Recall that a Boolean CQ can be seen as a bag of atoms. Therefore, in the following definition we can consider the enumerated version $q^e$ of a Boolean CQ $q$, which is the e-bag of its atoms; this formulation allows us to distinguish different occurrences of atoms in $q$. Then, an enumerated valuation from a CQ to an interpretation is essentially an e-homomorphism where the CQ is seen as a bag interpretation.

**Definition 39.** An enumerated valuation (e-valuation) of a Boolean CQ $q$ over a bag interpretation $I = (\Delta^I, \iota)$ is a family $(\nu, \nu_S, \nu_R)$ for $S \in \mathcal{C} \cup \mathcal{R}$, of the following functions, where $q_S$ is the subquery of $q$ consisting of all its atoms over $S$:

$$\nu : y \cup I \rightarrow \Delta^I,$$

$$\nu_S : q_S^e \rightarrow \mathcal{S}^a, \quad \text{for all } S \in \mathcal{C} \cup \mathcal{R},$$

such that

- $\nu(a) = a^I$ for each $a \in I$,
- $\nu(y) = \nu(t)$ for all equality atoms $y = t$ in $q$,
- $\nu_A([A(t):m]) = [\nu(t) : \ell]$ for all $A \in \mathcal{C}$ and $[A(t):m] \in q_S^e$, where $\ell$ is a number in $\mathbb{N}$, and
- $\nu_P([P(t_1, t_2):m]) = ([\nu(t_1), \nu(t_2)]: \ell)$ for all $P \in \mathcal{R}$ and $[P(t_1, t_2):m] \in q_S^e$, where $\ell$ is a number in $\mathbb{N}$.

We sometimes write $\nu_{P^-}([P^-(t_2, t_1):m])$ instead of $\nu_P([P(t_1, t_2):m])$, for $P \in \mathcal{R}$.

It is straightforward to check that the number of e-valuations of a Boolean CQ $q$ over a bag interpretation $I$ is precisely the multiplicity of the empty tuple in the certain answers to $q$ over $I$.

**Proposition 40.** The number of e-valuations of a Boolean CQ $q$ over a bag interpretation $I$ is $q^I(\iota)$.

The following lemma formalises the second idea: if two e-valuations over the canonical bag interpretation coincide on all the (enumerated occurrences of the) atoms of a rooted CQ that involve terms evaluating to (the interpretations of) individuals, then they are the same e-valuation.

**Lemma 41.** Let $q$ be a rooted Boolean CQ and $\mathcal{K}$ be a DL-Lite$^b_{\text{core}}$ ontology. If two e-valuations $(\nu^1, \nu^2, \ldots)$ and $(\nu^2, \nu^2, \ldots)$ of $q$ over $\text{Can}^e(\mathcal{K})$ are different, then there exists an individual $a \in I$, an atom $S(t)$ in $q$, a number $m \in \mathbb{N}$, and $i \in \{1, 2\}$ such that $\nu^i(a)$ appears in the tuple $\nu^i(t)$ and $\nu^i(S(t):m) \neq \nu^i(S(t):m)$.

**Proof.** Let e-valuations $(\nu^1, \nu^2, \ldots)$ and $(\nu^2, \nu^2, \ldots)$ of $q^e$ over $\text{Can}^e(\mathcal{K})$ be different, but, for the sake of contradiction, $\nu^1(S(t):m) = \nu^2(S(t):m)$ for all $a \in I$, $[S(t):m] \in q^e$ and $i \in \{1, 2\}$ such that $\nu^i(a)$ is in $\nu^i(t)$. Since the e-valuations are different, there exists $[S(t):m] \in q^e$ such that $\nu^1(S(t):m) \neq \nu^2(S(t):m)$. Moreover, by assumption $t$ consists of only variables. We consider only the case when $S(t) = P(x_1, x_2)$, where $P \in \mathcal{R}$ (and the case when $S(t)$ is $A(x)$ for $A \in \mathcal{C}$ can be handled in the same way).

Boolean CQ $q$ is rooted, so there exists a sequence

$$[R_1(t_0, t_1): m_1], [R_2(t_1, t_2): m_2], \ldots, [R_k(t_{k-1}, t_k): m_k]$$

such that $t_0 \in I$, $[R_k(t_{k-1}, t_k): m_k]$ is either $[P(x_1, x_2): m]$ or $[P^-(x_2, x_1): m]$ and, for each $j = 1, \ldots, k$, $t_j \sim t_j$ and either $[R_j(t'_j, t_j): m_j]$ is in $q^e$, if $R_j$ is an atomic role, or $[P_j(t'_j, t_j): m_j]$ is in $q^e$, if $R_j = P_j$ for an atomic role $P_j$.

We claim that

$$\nu^1_{R_j}(R_j(t'_j, t_j): m_j) = \nu^2_{R_j}(R_j(t'_j, t_j): m_j) \quad \text{for all } j = 1, \ldots, k$$

(2)

for all $j = 1, \ldots, k$ (which, in particular, contradicts our assumption on $[P(x_1, x_2): m]$). To prove this claim, suppose for the sake of contradiction that it is not the case, and let $i \in \{1, \ldots, k\}$ be the smallest number such that (2) does not hold. By assumption, we know that $\nu^i(t'_i) \neq \nu^i(a)$ for both $i = 1, 2$ and every $a \in I$ (therefore, $j \neq 1$, because $t_0 \in I$). However, since $j$ is the smallest number, $\nu^1(t'_i) = \nu^2(t'_i)$.

So, the element $u = v^i(t'_i)$ in the canonical bag interpretation $\text{Can}^e(\mathcal{K})$ was not introduced in $\text{Can}^0(\mathcal{K})$, which implies, by construction, that $\exists R_j(\text{Can}^e(\mathcal{K}))(u) \leq 1$. In fact, since $(\nu^1, \nu^2, \ldots)$ is an e-valuation, $(\exists R_j)^{\text{Can}^e(\mathcal{K})}(u) = 1$—that is, there exists just one $v \in \Delta^\text{Can}^e(\mathcal{K})$ such that $R_j^\text{Can}^e(\mathcal{K})(u, v) \geq 1$, and, moreover, $R_j^\text{Can}^e(\mathcal{K})(u, v) = 1$. In other words, $[(u, v)] \in R_j^\text{Can}^e(\mathcal{K})$, but $[(u, v)] \notin R_j^\text{Can}^0(\mathcal{K})$. Since $(\nu^1, \nu^2, \ldots)$ and $(\nu^2, \nu^2, \ldots)$ are e-valuations, $\nu^1_{R_j}$ and $\nu^2_{R_j}$ send $[R_j(t'_j, t_j): m_j]$ to some enumerated pairs in $R_j^\text{Can}^e(\mathcal{K})$, which, by assumption, are different. However, we also know that $\nu^1(t_{j-1}) = \nu^2(t_{j-1})$, so the only possibility for both $\nu^1_{R_j}(R_j(t'_j, t_j): m_j)$ and $\nu^2_{R_j}(R_j(t'_j, t_j): m_j)$ is $[(u, v)]$. Therefore, our assumption on the existence of $j$ was wrong and (2) indeed holds for all $j$. In particular, it holds for $j = k$, which contradicts the fact that $\nu^1_{P}(P(x_1, x_2): m) \neq \nu^2_{P}(P(x_1, x_2): m)$. \(\square\)
Lemma 41 relies both on the fact that there are no role inclusions in the TBox and on the fact that the CQ is rooted. It is easy to construct counter-examples to this lemma if any one of these requirements is violated.

Having Lemmas 38 and 41 at hand, we are ready to prove that, for satisfiable DL-Lite\textsuperscript{b\textsubscript{CORE}} ontologies, the canonical bag interpretation is the universal model for the class of rooted CQs. The idea is that the composition of an e-valuation of a rooted CQ and an e-homomorphism that is multiplicity-preserving on the individuals is also an e-valuation; moreover, the mapping between e-valuations defined by the e-homomorphism in this way is injective and therefore preserving the size of the domain of the mapping.

**Theorem 42.** The canonical bag interpretation Can(K) of a satisfiable DL-Lite\textsuperscript{b\textsubscript{CORE}} ontology K is a universal model for the class of rooted CQs. The same holds if universality is considered under the UNA.

**Proof.** First, note that it is enough to consider only Boolean rooted CQs, because the required property for a non-Boolean rooted CQ q(x) follows from the property for all Boolean CQs obtained from q(x) by replacing variables x by individuals from I. For a Boolean rooted CQ q it is enough to show that for every DL-Lite\textsuperscript{b\textsubscript{CORE}} ontology K, every model I of K, and every e-valuation (v, v₁, ..., vₙ) of q over Can(K) there exists a unique e-valuation (v', v₁', ..., vₙ') of q over I. By Lemma 38 we know that there exists an e-homomorphism (h, h₁, ..., hₙ) from Can\textsuperscript{e}(K) to I\textsuperscript{e} that is multiplicity-preserving on I. Therefore, we can take the composition (v, v₁, ..., vₙ) \circ (h, h₁, ..., hₙ) = (v \circ h, v₁ \circ h₁, ..., vₙ \circ hₙ) as (v', v₁', ..., vₙ'); indeed, the result of this composition is an e-valuation of q over I and, by Lemma 41, this result is unique among all e-valuations of q over Can(K).

Given that Can(K) satisfies the UNA, this result implies also that Can(K) is universal under the UNA for the class of rooted CQs and satisfiable DL-Lite\textsuperscript{b\textsubscript{CORE}} ontologies K. □

The following is an important corollary of Theorem 42 and Corollary 24, which allows us to forget the UNA in the rest of the paper when talking about rooted CQ answering over DL-Lite\textsuperscript{b\textsubscript{CORE}}.

**Corollary 43.** The certain answers to rooted CQs over DL-Lite\textsuperscript{b\textsubscript{CORE}} ontologies do not depend on the adoption of the UNA.

Another important corollary of Theorem 42, the structural properties of rooted CQs, and the definition of the canonical interpretation is that, similarly to the set case, the bag certain answers q\textsuperscript{K} to a rooted CQ q over a satisfiable DL-Lite\textsuperscript{b\textsubscript{CORE}} ontology K can be computed over the sub-interpretation Can\textsubscript{n}(K) of Can(K) with n depending only on q.

**Corollary 44.** If K is a satisfiable DL-Lite\textsuperscript{b\textsubscript{CORE}} ontology with Can(K) = \bigcup\textsubscript{i≥0} Can\textsubscript{i}(K) and q is a rooted CQ having n atoms, then q\textsuperscript{K} = q\textsuperscript{Can\textsubscript{n}(K)}.

8. **Rewritability of rooted conjunctive queries over DL-Lite\textsuperscript{b\textsubscript{CORE}}**

First-order rewritability of CQs is a key property of DL-Lite query answering under set semantics. In this section, we show rewritability of rooted CQs over DL-Lite\textsuperscript{b\textsubscript{CORE}} to BCALC (recall that by Corollary 43 this result is agnostic to the adoption of the UNA).

8.1 **Non-rewritability to BCALC unions of conjunctive queries**

In the case of set semantics, the target language for rewritings is that of unions of conjunctive queries (UCQs). There are two natural counterparts to UCQs in the bag setting: BCALC maximal union of CQs and BCALC arithmetic union of CQs. Our first result is negative and in stark contrast to the set case: in general, rewriting to either of these classes of BCALC queries is not possible, even over DL-Lite\textsuperscript{b\textsubscript{CORE}}.

**Proposition 45.** Rooted CQs are rewritable neither to BCALC maximal nor to BCALC arithmetic unions of CQs over DL-Lite\textsuperscript{b\textsubscript{CORE}}.

**Proof.** We first prove the claim for BCALC maximal unions of CQs. Consider the satisfiable DL-Lite\textsuperscript{b\textsubscript{CORE}} ontology K = (T, A) over atomic concepts A and B, and atomic role P with TBox T = {A \sqsubseteq \exists P. \exists P \sqsubseteq B}, and ABox A = \{ A(a) : 3. P(a, b) : 2, B(b) : 3 \}. Let also q(x) = \exists y. P(x, y) \land B(y). Then, Can(K) is the interpretation with domain Δ\textsuperscript{Can(K)} = I \cup \{ w\textsubscript{b, p} \} that interprets individuals by themselves and predicates as follows:

\[ A\textsuperscript{Can(K)} = \{ a : 3 \}, \quad P\textsuperscript{Can(K)} = \{ (a, b) : 2, (a, w\textsubscript{a, p}) : 1 \}, \quad B\textsuperscript{Can(K)} = \{ b : 3, w\textsubscript{b, p} : 1 \}. \]

Evaluating q over Can(K), we get q\textsuperscript{Can(K)}(a) = 7 for individual a. Assume for the sake of contradiction that there exists a rewriting of q to a BCALC maximal union \Phi(x) of CQs with respect to T. By the semantics of the BCALC maximal union, there exists a BCALC CQ q₀ in \Phi with q₀\textsuperscript{A}(a) = q\textsuperscript{Can(K)}(a). Observe that A contains three distinct assertions with multiplicities 3, 2, and 3. Hence, whenever there is a valuation of the terms of q₀ that maps an atom of q₀ to one of these assertions, the multiplicity is either 2 or 3. Because q₀ is a CQ, every valuation of q₀ contributes to q₀\textsuperscript{A}(a) a multiplicity that
is a multiple of 2 or 3. Since 7 is prime, there can be no valuation contributing multiplicity 7. Moreover, there are only two ways to get 7 as an instance of a polynomial with coefficients 2 and 3, namely, $2 + 2 + 3$ and $2 \times 2 + 3$. For the former sum, this means that there exist three distinct valuations contributing to $q_0^A(a)$ with multiplicities 2, 2, and 3, respectively, which is impossible given the fact that, to get 2, query $q_0$ must be set equal to $3y. P(x, y)$, which excludes the possibility of getting the multiplicity 3. For the latter sum, there must exist two distinct valuations contributing to $q_0^A(a)$ with multiplicities 4 and 3, respectively, which is again impossible given the fact that, to get 4, query $q_0$ must be set to $3y. P(x, y) \land P(x, z)$ or to $3y. P(x, y) \land P(x, y)$, which in either case excludes the possibility of getting the multiplicity 3.

We now prove the claim for BCALC arithmetic unions of CQs. Consider the $DL-Lite^b_{\text{CORE}}$ ontology $K = \langle T, A \rangle$ over atomic concepts $A$ and $B$ with TBox $T = \{ A \subseteq B \}$ and ABox $A = \{ A(a) : 3, A(b) : 2, B(a) : 2, B(b) : 3 \}$, and the rooted query $q(x) = B(x)$. Observe that $q^C = B^{\text{Can}(K)} = \{ a : 3, b : 3 \}$. We have the following possible cases for a BCALC arithmetic union $\Phi(x)$ of CQs:

- if $\Phi(x) = A(x)$, then $\Phi^A = \{ a : 3, b : 2 \}$;
- if $\Phi(x) = B(x)$, then $\Phi^A = \{ a : 2, b : 3 \}$;
- if $\Phi(x) = A(x) \land B(x)$, then $\Phi^A = \{ a : 6, b : 6 \}$;
- if $\Phi(x) = A(x) \lor B(x)$, then $\Phi^A = \{ a : 5, b : 5 \}$;
- if $\Phi(x)$ contains $A(x)$ at least twice, then $\Phi^A(a) \geq 9$; and
- if $\Phi(x)$ contains $B(x)$ at least twice, then $\Phi^A(b) \geq 9$.

So, none of these cases satisfies $\Phi^A = B^{\text{Can}(K)}$, which is required for a rewriting of $q$ with respect to $T$.

### 8.2. General ideas for rewriteability to BCALC queries

Next, we show that rooted CQs are rewriteable to a richer fragment of BCALC over $DL-Lite^b_{\text{CORE}}$, which also features the operation of difference. This ensures LocSpace membership in data complexity of query answering. Our rewriting algorithm is inspired by that of Kikot et al. [37] for $DL-Lite_{\text{Q}}$. The key observation behind our approach is that, for a $DL-Lite^b_{\text{CORE}}$ ontology $K$ and a rooted CQ $q(x) = \exists y. \phi(x, y)$, the bag answers to $q$ over the canonical bag interpretation $\text{Can}(K)$, which, by Theorem 42, coincide with the bag certain answers to $q$ over $K$, can be partitioned as

$$q^{\text{Can}(K)} = \bigcup_{z \subseteq y} [q, z]^{\text{Can}(K)},$$

where each $[q, z]^{\text{Can}(K)}$ is the bag of answers to $q$ over $\text{Can}(K)$ supported by valuations of $q(x)$ over $\text{Can}(K)$ that send all the variables in a subset $z$ of the variables $y$ to anonymous elements and all other variables to individuals. Next, we define such partitions formally.

**Definition 46.** Let $q(x) = \exists y. \phi(x, y)$ be a rooted CQ and let $K$ be a $DL-Lite^b_{\text{CORE}}$ ontology. Given a subset $z$ of variables $y$, let $[q, z]^{\text{Can}(K)}$ be the bag of tuples over $I$ such that, for each tuple $a$ of individuals,

$$[q, z]^{\text{Can}(K)}(a) = \sum_{\lambda \in \Lambda_z} \prod_{s(t) \in \phi(x, y)} S^{\text{Can}(K)}(\lambda(t)),$$

where $\Lambda_z$ is the set of valuations $\lambda : x \cup y \cup I \rightarrow \Delta^{\text{Can}(K)}$ such that $\lambda(x) = a$, $\lambda(a) = a$ for each $a \in I$, $\lambda(x) = \lambda(t)$ for each $x = t$ in $\phi(x, y)$, $\lambda(z)$ is an anonymous element for each $z \in z$, and $\lambda(y) \in I$ for each $y \in y \setminus z$.

Following this key observation, given a rooted CQ, the rewriting algorithm first constructs a set of BCALC queries each one accounting for the bag $[q, z]^{\text{Can}(K)}$ for a subset $z$ and then adjoins them using the arithmetic union to produce the actual rewriting, which is still a BCALC query. In particular, every subset $z$ is processed along the following three steps:

1. $z$ is checked for $T$-realisability—that is, whether the corresponding subquery can be folded to the anonymous part of a canonical bag interpretation—and disregarded from consideration in $\text{(3)}$ if the check fails;
2. each connected component of the subquery corresponding to $z$ is replaced in the query of $[q, z]^{\text{Can}(K)}$ by a single representative role atom; and
3. each concept atom and each representative atom is rewritten to a BCALC query that takes into account the TBox and the fact that $z$ should be sent to anonymous elements.

In the following three sections we formalise each of these steps and prove their correctness.

### 8.3. Step 1: checking for realizability

In the first step, every subset $z$ of existentially quantified variables $y$ in a rooted CQ $q(x)$ is checked for $T$-realisability. Intuitively, $z$ is $T$-realisable if the subquery of $q(x)$ induced by $z$ can be “folded” into the anonymous forest-shaped part of
Can(K) for some ontology K having TBox T. Therefore, non-realisable z cannot contribute to partitioning (3) and their associated subqueries can be disregarded for the purpose of query rewriting. To provide the formal definition of T-realisability, we need to introduce some preliminary definitions and notations.

Let q(x) = ∃y. φ(x, y) be a rooted CQ and T be a DL-LiteCORE TBox.

First, recall that the Gaifman graph G of q has the equivalence class t for each term t ∈ x ∪ y ∪ I in φ as a node, and an edge [t1, t2] for each atom P(t1, t2) in φ. A subset z of y is equality-consistent if z ⊆ z for every z ∈ z.

Second, each equality-consistent subset z of y has a corresponding subgraph G Iz of Gaifman graph G—that is, the subgraph on the set of nodes [z | z ∈ z]. This subgraph may have several connected components; a subset v of z is maximally connected if it is also equality-consistent and the subgraph of G Iz corresponding to v is a connected component of G Iz. Therefore, z can be partitioned to its maximal connected subsets.

Third, for every maximally connected subset v of an equality-consistent z ⊆ y and for each individual a, we define the query

qv = ∃v, v′. ϕv ∧ \( \bigwedge_{t \in t_v} (t = a) \land \bigwedge_{v \in v} (v \neq a) \),

where ϕv is the conjunction of all atoms in φ mentioning at least one variable in v, t_v is the set of all terms appearing in ϕ_v but not in v, and v = (x ∪ y) ∩ t_v. This query is a Boolean CQ, except that it may have equalities of two individuals and inequalities of terms. The semantics of CQs in Definition 12 can be extended to such queries in a straightforward way: the additional requirement on each valuation λ contributing to the sum is that λ should satisfy λ(x) ≠ λ(t) for each inequality atom (x ≠ t) in the query (and the requirement for equalities of individuals is the same as for usual equalities).

Example 47. Consider the rooted CQ

q(x) = ∃y. P(x, y1) ∧ P(x, y2) ∧ P(x, y3) ∧ P(x, y4) ∧ P(d, y4) ∧ R(y3, y5) ∧ (y1 = c) ∧ (y3 = y4)

with y = y1, …, y5 over atomic roles P and R, and its Gaifman graph, depicted in Fig. 2a. Observe that no subset of y containing y1 is equality-consistent because q contains equality y1 = c and c is not in y. Furthermore, every subset of y containing y3 or y4 but not both is also not equality-consistent. However, z = {y2, y3, y5} is equality-consistent. The corresponding subgraph G Iz is depicted in Fig. 2b. It has two connected components, and therefore z partitions into two maximally connected subsets, v1 = {y2} and v2 = {y3, y4, y5}. For the first, we have ϕv1 = P(x, y2), t_v1 = {x}, and, for an individual a,

qv1 = ∃v2, x. P(x, y2) ∧ (x = a) ∧ (y2 ≠ a).

(4)

For the second, we have ϕv2 = P(x, y3) ∧ P(x, y4) ∧ P(d, y4) ∧ R(y3, y5) ∧ (y3 = y4), t_v2 = {x, d}, and

qv2 = ∃v2, x. ϕv2 ∧ (x = d) ∧ (d = d) ∧ (y3 ≠ d) ∧ (y4 ≠ d) ∧ (y5 ≠ d).

(5)

We are now ready to define the notion of realisability, which is inspired by the notion of tree witnesses proposed by Kikot et al. [37] in the context of query rewriting over DL-LiteR.

Definition 48. Let q(x) = ∃y. φ(x, y) be a rooted CQ and T be a DL-LiteCORE TBox. A subset z of variables y is T-realisable if it is equality-consistent and every maximally connected subset v of z satisfies the following conditions, where, as before, ϕ_v is the conjunction of all atoms in φ mentioning at least one variable in v and t_v is the set of all terms appearing in ϕ_v but not in v:

1. there is at most one individual in t_v;
2. all the atoms in ϕ_v mentioning terms in t_v are over the same atomic role P_v and have these terms at the same position P_v ∈ \{1, 2\};
3. \( |\phi_v^{\mathrm{CQ}(K_v)}(v)| \geq 1 \), where a is the individual in t_v if it exists or a fresh individual otherwise, and, for a fresh individual b, \( K_v \) is \( \langle T, \| P_v(a, b) : 1 \| \rangle \) if \( P_v = 1 \) or \( \langle T, \| P_v(b, a) : 1 \| \rangle \) if \( P_v = 2 \).
Note that realisability checking is clearly decidable. In particular, by Corollary 44, Condition 3 can be checked over a bounded fragment of the relevant canonical bag interpretation.

Example 49. Consider the DL-Lite core TBox $\mathcal{T} = \{A \sqsubseteq \exists P, \exists P^\neg \sqsubseteq \exists R\}$ over atomic concept $A$ and atomic roles $P$ and $R$, the rooted CQ $q(x)$ specified in Example 47, and the equality-consistent subset $z = \{y_1, \ldots, y_5\}$ with two maximally connected subsets, $v_1 = \{y_1, y_2\}$ and $v_2 = \{y_3, y_4, y_5\}$. Subset $z$ is $\mathcal{T}$-realisable. To see this, note first that Conditions 1 and 2 are immediately satisfied by both $v_1$ and $v_2$, with no individuals in $v_1$, one individual $d$ in $v_2$, $P_{v_1} = P_{v_2} = P$ and $P_{v_1} = P_{v_2} = 1$. To verify Condition 3 for $v_1$, note that $K_{v_1} = (\mathcal{T}, \{P(a, b) : 1\})$ for fresh individuals $a$ and $b$, and therefore

$$A^{Can(K_{v_1})} = \emptyset, \quad P^{Can(K_{v_1})} = \{(a, b) : 1\}, \quad R^{Can(K_{v_1})} = \{(b, w_1^2) : 1\};$$

so $q_{v_1}$, defined in (4), evaluates to 1 for $\emptyset$ over $Can(K_{v_1})$. Condition 3 for $v_2$ can be verified in exactly the same way, except that $K_{v_2}$ and $Can(K_{v_2})$ have $d$ instead of $a$, and $q_{v_2}$ is defined in (5).

The next lemma establishes the key property of realisability: for any ABox, a non-realisable $z$ cannot contribute to the partitioning (3) of the bag query answers over the canonical bag interpretation.

Lemma 50. Let $q(x) = \exists y. \phi(x, y)$ be a rooted CQ and let $\mathcal{T}$ be a DL-Lite core TBox. If a subset $z$ of $y$ is not $\mathcal{T}$-realisable, then $[q, z]^{Can((\mathcal{T}, A))} = \emptyset$ for every bag ABox $A$.

Proof. Let $K = (\mathcal{T}, A)$ where $A$ is an arbitrary bag ABox, and $z$ be a subset of $y$ that is not $\mathcal{T}$-realisable. By Definition 48, either $z$ is not equality-consistent, or there exists a maximally connected subset of $z$ for which one of the three conditions does not hold.

It is straightforward to check that $[q, z]^{Can(K)} = \emptyset$ if $z$ is not equality-consistent: indeed, in this case the CQ contains an equality atom $(y = t)$ with $y \in z$ and $t \notin z$, which cannot be satisfied over the canonical bag interpretation $Can(K)$ by any valuation contributing to $[q, z]^{Can((\mathcal{T}, A))}$ because $y$ is mapped to anonymous elements of $Can(K)$ by such valuations, whereas $t$ is mapped to individuals.

Similarly, if there is a maximally connected $v \subseteq z$ for which Condition 1 does not hold, then $[q, z]^{Can(K)} = \emptyset$: indeed, if $t_v$ contains two different individuals, then no contributing valuation can send $v$ to anonymous elements, because $v$ is connected, while $Can(K)$ has independent tree-shaped anonymous parts connected to these two individuals.

Next, if there is a maximally connected $v \subseteq z$ for which Condition 2 does not hold, then again $[q, z]^{Can(K)} = \emptyset$: indeed, if $\phi_v$ contains two atoms over different atomic roles with terms in $t_v$ or two atoms over the same atomic role but with the terms in $t_v$ in different positions, then no contributing valuation can send $v$ to anonymous elements by the same reasons as in the case of Condition 1 and the fact that no anonymous element is connected to another element in two different ways in $Can(K)$ by construction.

Finally, consider the case when there is a maximally connected $v \subseteq z$ for which Conditions 1 and 2 hold but Condition 3 does not. Reasoning as in the previous two cases, we conclude that all variables in $v$ are sent by every contributing valuation to anonymous elements generated by the same individual in the canonical bag interpretation, and all terms in $t_v$ are sent to this individual. By Condition 2, every atom in $\phi_v$ that has a term in $t_v$ is over atomic role $P_v$ and has the term in position $P_v$. Therefore, $(q_{v_v}^{Can(K_v)}(\emptyset))$ is the factor corresponding to $\phi_v$ in the multiplicity of every valuation satisfying the conditions of $[q, z]^{Can(K)}$, and $(q_{v_v}^{Can(K_v)}(\emptyset)) = 0$ means that there are no valuations with non-zero contribution.

8.4. Step 2: replacing subqueries with representatives

Consider a $\mathcal{T}$-realisable subset $z$ of the variables $y$ in a rooted CQ $q(x) = \exists y. \phi(x, y)$ for a DL-Lite core TBox $\mathcal{T}$. As established in the previous section, maximally connected subsets $v$ of $z$ are always disjoint, so the corresponding conjunctions of atoms $\phi_v$ from $\phi$ that mention at least one variable in $v$ do not share atoms either. So, $q(x)$ can be put into the following form for the given $z$ in a unique way, where $\psi_z$ consists of all atoms in $\phi(x, y)$ not contained in any $\phi_v$:

$$q(x) = \exists y. \psi_z \land_{\text{maximally connected } v \subseteq z} \phi_v.$$ (6)
that all terms of $q_v$ that are outside $z$ are sent to one of the individuals of the assertion. Therefore, when computing $[q, z]^{\text{Can}(T, \mathcal{A})}$, the whole $q_v$ can be replaced in $q$ by just a single role atom mentioning a representative variable and a term not in $v$ as well as several equalities identifying all the terms not in $v$.

Next, we formalise this idea and prove its correctness.

**Definition 51.** Let $q(x) = \exists y, \phi(x, y)$ be a rooted CQ and let $T$ be a DL-Lite$_{\text{core}}$ $T$Box. For every $T$-realisable subset $z$ of $y$, let

$$q_z(x) = \exists y_z, z' \cdot \psi_z \land \left( \bigwedge_{\text{maximally connected } v \subseteq z} \left( \alpha_v \land \bigwedge_{y \in (x \cup y') \cap t_v} (y = t) \right) \right),$$

where $\psi_z$ is defined as in (6), $y_z$ is the set of all variables in $y$ appearing in $\psi_z$, atom $\alpha_v$ is defined as follows, for every maximally connected $v \subseteq z$ with terms $t_v$, role $P_v$, and position $p_v$ defined as in Definition 48, as well as for a term $t \in t_v$ and a fresh variable $y_v$:

$$\alpha_v = \left\{ \begin{array}{ll} P_v(y, t), & \text{if } p_v = 1, \\
P_v(t, y), & \text{if } p_v = 2, \end{array} \right.$$ 

and $z'$ is the set of all variables $y_v$ introduced for the atoms $\alpha_v$.

Formally speaking, this definition is non-deterministic, because the terms $t$ in the atoms $\alpha_v$ are chosen arbitrarily from $t_v$. However, this choice does not influence the semantics of $q_z(x)$ because of the equalities introduced in (7). Therefore, we assume that $q_z(x)$ is well-defined.

**Example 52.** Consider the rooted CQ $q(x)$, DL-Lite$_{\text{core}}$ $T$Box $T$, and $T$-realisable subset $z$ with maximally connected subsets $v_1$ and $v_2$ introduced in Examples 47 and 49. We know that $t_{v_1} = \{x\}$, $t_{v_2} = \{x, d\}$, $P_{v_1} = P_{v_2} = P$, and $p_{v_1} = p_{v_2} = 1$, so $\alpha_{v_1} = P(x, y_{v_1})$, $\alpha_{v_2} = P(x, y_{v_2})$, and

$$q_z(x) = \exists y_1, y_{v_1}, y_{v_2}, P(x, y_1) \land (y_1 = c) \land \left( P(x, y_{v_1}) \land (x = c) \right) \land \left( P(x, y_{v_2}) \land (x = d) \land (x = c) \right).$$

In contrast to $q$, CQ $q_z$ does not contain any $R$ atoms. Indeed, such atoms are not needed: sending $y_3$ and $y_4$ to the same anonymous element $w$ and given that $w$ must be a $P$-successor of $d$ in the canonical bag interpretation, it follows that $w$ must have an $R$-successor due to inclusion $\exists P^- \subseteq \exists R$ in $T$. Thus, we only need a representative $y_{v_1}$ for $y_3$ and $y_4$. $\langle$

In the second step, our algorithm generates the query $q_z(x)$ for each $T$-realisable subset $z$ of $y$. The next lemma justifies this step by showing that we can consider in (3) only $T$-realisable $z$ and replace $q$ with $q_z$ for each such $z$.

**Lemma 53.** Let $q(x) = \exists y, \phi(x, y)$ be a rooted CQ and let $T$ be a DL-Lite$_{\text{core}}$ $T$Box. If a subset $z$ of $y$ is $T$-realisable then $[q, z]^{\text{Can}(T, \mathcal{A})} = [q_z, z']^{\text{Can}(T, \mathcal{A})}$ for every bag ABox $\mathcal{A}$, where $z'$ as in Definition 51.

**Proof.** Let $K = (T, \mathcal{A})$ where $\mathcal{A}$ is an arbitrary bag ABox, and let $z$ be a $T$-realisable subset of $y$. First, for every valuation $\lambda \in \Lambda_{z}$ as in Definition 46 of $[q, z]^{\text{Can}(K)}$, let $\lambda_z$ be the valuation of $x \cup y_z \cup z' \cup I$ to $\Delta^{\text{Can}(K)}$ that is the same as $\lambda$ on $I$ and all the terms of $\psi_z$, and, for each maximally connected subset $v$ of $z$, $\lambda_z(y_v) = \lambda(y_v)$, where $y_v$ is a variable in $v$ such that $\phi_v$ contains an atom $P_v(t, y)$ or an atom $P_v(y, t)$, for some $t \in t_v$; in other words, $\lambda_z$ is the same as $\lambda$ except that each $v$ is replaced by its representative with the corresponding value. It suffices to show that every valuation $\lambda \in \Lambda_{z}$ contributes to $[q, z]^{\text{Can}(K)}$ the same multiplicity as $\lambda_z$ contributes to $[q_z, z']^{\text{Can}(K)}$.

Consider first a valuation $\lambda \in \Lambda_{z}$ contributing to $[q, z]^{\text{Can}(K)}$ a non-zero multiplicity. By construction, $\text{Can}(K)$ interprets all concepts with multiplicity at most 1 on the anonymous elements and all roles with multiplicity at most 1 on all pairs with at least one anonymous element. Thus, the contribution of $\lambda$ is equal to the contribution of (the relevant part of) $\lambda$ to the evaluation of $\psi_z$. So, it is enough to show that, for each maximally connected $v \subseteq z$, $\lambda_z(\alpha_v)$ has multiplicity 1 in $\text{Can}(K)$, and all the equalities on $t_v$ introduced in (7) to $q_z$ hold for $\lambda_z$ (i.e., for $\lambda$, because they coincide on $t_v$). The former holds immediately by construction of $\lambda_z$, while the latter follows from the fact that $v$ are connected in $q$ and sent to anonymous elements by $\lambda$, while $\text{Can}(K)$ is tree-shaped on the anonymous elements.

Consider now a valuation $\lambda \in \Lambda_{z}$ such that $\lambda_z$ contributes to $[q, z]^{\text{Can}(K)}$ a non-zero multiplicity. Similarly to the previous case, the contribution of $\lambda_z$ is equal to the contribution of (the relevant part of) $\lambda_z$ to the evaluation of $\psi_z$. So, it is enough to show that, for each maximally connected $v \subseteq z$, $\lambda_z(\phi_v)$ has multiplicity 1 in $\text{Can}(K)$. However, this follows from the fact that $z$ is $T$-realisable (in particular, Condition 3) and the fact that $\lambda_z(\alpha_v)$ has multiplicity 1 in $\text{Can}(K)$. $\square$
8.5. Step 3: rewriting atoms to BCALC queries

In the last step, our algorithm first transforms each CQ \( q_z(x) \) computed in Step 2 for a \( T \)-realisable \( z \) to a BCALC query \( \Phi_z(x) \) satisfying equality \( [q_z, z]^\text{Can}(\langle T, A \rangle) = \Phi_z^A \) and then constructs the final rewriting of the input CQ \( q(x) = \exists y. \phi(x, y) \) with respect to the input DL-Lite\textsubscript{core} TBox \( T \) to the BCALC query

\[
\Phi_q(x) = \bigvee_{T\text{-realisable } z \subseteq y} \Phi_z(x).
\]

The intuition behind the construction of \( \Phi_z(x) \) from \( q_z(x) \) and \( T \) hinges on the observation that, for every bag ABox \( A \), the multiplicities of individuals in the interpretations of concepts and roles in \( \text{Can}(\langle T, A \rangle) \) are determined by the multiplicities of the assertions in \( A \) as follows:

- for a concept \( C \), the multiplicity of an individual \( a \) in \( C^\text{Can}(\langle T, A \rangle) \) is the maximum multiplicity of \( a \) in the interpretation of the concepts subsumed by \( C \) with respect to \( T \) in the bag interpretation corresponding to \( A \);
- for a role \( P \), the multiplicity of a pair of individuals \( (a, b) \) in \( P^\text{Can}(\langle T, A \rangle) \) coincides with the multiplicity of assertion \( P(a, b) \) in \( A \) (which is justified by the fact that DL-Lite\textsubscript{core} TBoxes do not allow for role inclusions).

Therefore, for a role \( R \), the number of anonymous \( R \)-successors of an individual \( a \) in the canonical bag interpretation is precisely the multiplicity of \( a \) in the interpretation of the concept \( \exists R \) minus the number of individual \( R \)-successors of \( a \) in the ABox. We can then exploit these observations to construct a BCALC query \( \Phi_z(x) \) such that the bag answers to \( \Phi_z(x) \) over every ABox \( A \) coincide with the bag \( [q_z, z]^\text{Can}(\langle T, A \rangle) \) (where all valuations contributing to the latter bag map \( z \) to anonymous elements and the rest to individuals). For this, it suffices to apply to \( q_z \), which has the form (7), the following replacements:

- each atom over an atomic concept \( A \) in the conjunction \( \psi_z \) of \( q_z \) with a BCALC query retrieving the maximum multiplicity over all concepts subsumed by \( A \) in \( T \);
- each atom \( \alpha_v = P_v(t, y_v) \), for a maximally connected \( v \subseteq z \), with a BCALC query that subtracts the number of \( P_v \)-successors in the ABox from the maximum multiplicity over all concepts subsumed by \( \exists P_v \) in \( T \); and
- each atom \( \alpha_v = P_v(y_v, t) \), for a maximally connected \( v \subseteq z \), with a BCALC query that subtracts the number of \( P_v \)-predesigners in the ABox from the maximum multiplicity over all concepts subsumed by \( \exists P_v \) in \( T \).

Note that atoms over atomic roles in \( \psi_z \) are left intact because \( T \) does not allow for role inclusions.

Next, we formalise this intuition and define the BCALC query \( \Phi_z(x) \) in terms of \( q_z(x) \) and \( T \).

**Definition 54.** Let \( q(x) = \exists y. \phi(x, y) \) be a rooted CQ, \( T \) be a DL-Lite\textsubscript{core} TBox, and let \( z \) be a \( T \)-realisable subset of \( y \). The BCALC query \( \Phi_z(x) \) is obtained from \( q_z(x) \) by replacing

- each occurrence of an atom \( A(t) \) in \( \psi_z \) with

\[
\bigvee_{T\models C \subseteq A} \zeta_C(t),
\]

- for each maximally connected \( v \subseteq z \), the atom \( P_v(t, y_v) \) and the atom \( P_v(y_v, t) \) with the following BCALC query, where \( R_v \) is \( P_v \) and \( P_v^\neg \), respectively:

\[
\left( \bigvee_{T\models C \subseteq \exists R_v} \zeta_C(t) \right) \setminus \zeta_{\exists R_v}(t).
\]

where, for every concept \( C \) and term \( t \), and for a fresh variable \( y \),

\[
\zeta_C(t) = \begin{cases} 
C(t), & \text{if } C \text{ is an atomic concept}, \\
\exists y. \xi_R(t, y), & \text{if } C \text{ is of the form } \exists R, 
\end{cases}
\]

and, for every atomic role \( P \) and terms \( t_1 \) and \( t_2 \),

\[
\xi_P(t_1, t_2) = P(t_1, t_2) \quad \text{and} \quad \xi_{P^\neg}(t_1, t_2) = P(t_2, t_1).
\]

Before proving correctness of the last step of the rewriting, we illustrate the definitions on our running example.

**Example 55.** Consider the DL-Lite\textsubscript{core} TBox \( T = \{ \text{Record} \sqsubseteq \exists \text{hasMusician}, \exists \text{hasMusician}^- \sqsubseteq \text{Musician} \} \) and the rooted CQ

\[
q(x) = \exists y. \text{hasMusician}(x, y) \land \text{Musician}(y).
\]
There are two subsets of $y$, namely $\emptyset$ and $y$, and it is immediate to check that both of them are $T$-realisable. Moreover,

$$q_{\emptyset}(x) = q(x) \quad \text{and} \quad q_y(x) = \exists y'. \text{hasMusician}(x, y').$$

For the first of these CQs, all valuations contributing to the bag $[q_{\emptyset}, \emptyset]^{\text{Can}(K)}$ map all variables of $q_{\emptyset}$ to individuals, for every $K$ with TBox $T$. Therefore, the atom hasMusician($x, y$) remains intact in $\Phi_{q_{\emptyset}}$ whereas the atom Musician($y$) is rewritten to a BCALC query retrieving the multiplicity of an individual $a$ in Musician$^{\text{Can}(K)}$, which is equal to the maximum multiplicity of $a$ amongst the concepts Musician and $\exists$hasMusician$^{-}$ over the ABox of $K$. As a result,

$$\Phi_{q_{\emptyset}}(x) = \exists y. \text{hasMusician}(x, y) \land (\text{Musician}(y) \lor \exists z. \text{hasMusician}(z, y)).$$

For the second CQ, $q_y$, all valuations contributing to the bag $[q_y, y']^{\text{Can}(K)}$ map $y'$ to an anonymous element and $x$ to an individual, for every $K$ with $T$. Hence, each such valuation contributes to $[q_y, y']^{\text{Can}(K)}$ multiplicity 1, while all valuations $\lambda$ agreeing on $x$ contribute to $[q_y, y']^{\text{Can}(K)}$ an overall multiplicity equal to the number of anonymous hasMusician-successors of $\lambda(x)$. This number is the multiplicity of $\lambda(x)$ in the interpretation of $\exists$hasMusician under $\text{Can}(K)$ minus the number of individual hasMusician-successors of $\lambda(x)$. Inspecting $T$, we finally derive

$$\Phi_y(x) = (\text{Record}(x) \lor \exists y'. \text{hasMusician}(x, y')) \setminus \exists y''. \text{hasMusician}(x, y'').$$

Consider now the ABox

$$A = [\text{Record(Expectations)} : 2, \text{hasMusician(Expectations, K. Jarrett)} : 1].$$

Evaluating the two rewritings on this ABox, we get $\Phi^{A}_{q_{\emptyset}} = \Phi^{A}_y = [\text{Expectations} : 1]$, respectively, we have $\Phi^{A}_{q} = (\Phi_{q_{\emptyset}} \lor \Phi_{y})^{A} = [\text{Expectations} : 2]$, which is equal to $q^{\text{Can}((T, A))} = q^{(T, A)}$ as expected. \hfill \blacksquare

The following lemma establishes correctness of Step 3 of our rewriting approach as formalised in Definition 54.

**Lemma 56.** Let $q(x) = \exists y. \phi(x, y)$ be a rooted CQ and $T$ a DL-Lite$\text{core}$ TBox. Then $[q_A, z]^{\text{Can}((T, A))} = \Phi^A_q$ for every bag $A$Box $A$ and every $T$-realisable subset $z$ of $y$ with $z$ as in Definition 51.

**Proof.** We first claim that it is enough to show that, for every $T$-realisable subset $z$ of $y$ and every maximally consistent $v \subseteq z$, the bag answers to the rewritings (9) and (10) of atoms $A(t)$ and $\exists \xi_r(t, y_v)$, respectively, in $q_z$ over every bag $A$Box $A$ are equal to the bag answers $[A(t), \emptyset]^{\text{Can}(K)}$ and $[\exists y_v. \xi_r(t, y_v), y_v]^{\text{Can}(K)}$, respectively, for $K = (T, A)$. Indeed, on the one hand, $\Phi_{z}$ is obtained from $\Phi_{z}$ by applying only these replacements; on the other hand, the atoms that are not rewritten in $\Phi_{z}$ are of the form $P(t_1, t_2)$ or $(t_1 = t_2)$, for terms $t_1$ and $t_2$ that are mapped to individuals by each valuation $\lambda \in A_\lambda$, so such individuals do not need to be rewritten by the fact that atoms $P(t_1, t_2)$ satisfy $p^{\text{Can}(K)}(\lambda(t_1), \lambda(t_2)) = A(P(\lambda(t_1), \lambda(t_2)))$ for every $\lambda \in A_\lambda$, whereas equalities do not contribute to multiplicities.

We now argue the correctness of replacements (9) and (10). By the definitions of canonical bag interpretation and concept closure, for all individuals $a$ and concepts $C$,

$$C^{\text{Can}(K)}(a) = [C^{\text{Can}(K)}(a) = \text{ccl}_T[a, C_{00}(K)](C) = \max(\max\{A(A_0(a)) \mid A_0 \in C, T \models A_0 \subseteq C\},$$

$$\max\{\sum_{b \in 1} A(P_0(b, a)) \mid P_0 \in R, T \models \exists P_0 \subseteq C\}, \max\{\sum_{b \in 1} A(P_0(b, a)) \mid P_0 \in R, T \models \exists P_0 \subseteq C\}).$$

(11)

First, by substituting $A$ for $C$ in (11) and by the semantics of BCALC queries, we immediately derive the following for every $a \in 1$ and $A \in C$:

$$[A(a), \emptyset]^{\text{Can}(K)} = \left(\bigvee_{T \models C_\lambda \subseteq A} \xi_C(a)\right)^A,$$

which proves the claim for concept atoms. For $R_v$ atoms, note that, for each individual $a$, $[\exists y_v. \xi_r(a, y_v), y_v]^{\text{Can}(K)}$ is the number of anonymous $R_v$-successors of $a$ in the bag canonical interpretation, which is $(\exists R_v)C^{\text{Can}(K)}(a) - (\exists R_v)^A(a)$. Therefore, by substituting $\exists R_v$ for $C$ in (11) and by the semantics of BCALC queries, we get the following for every $a \in 1$ and $A \in C$:

$$[\exists y_v. \xi_r(a, y_v), y_v]^{\text{Can}(K)} = \left(\bigvee_{T \models C_\lambda \subseteq \exists R_v} \xi_{C_\lambda}(a)\right)^A \setminus \left(\bigvee_{T \models C_\lambda \subseteq \exists R_v} \xi_{R_v}(a)\right)^A,$$

which proves the claim for $R_v$ atoms. \hfill \blacksquare
8.6. Rewriting and complexity

Putting the results of the previous three sections together, we obtain the following theorem, which establishes the correctness of our rewriting approach.

**Theorem 57.** For every rooted CQ $q$ and every $DL$-Lite$^b_{\text{CORE}}$ ontology $K = \langle T, A \rangle$ we have that $q^{\text{Can}(K)} = \Phi_q^{A}$.

**Proof.** Consider a rooted CQ $q(x) = \exists y. \phi(x, y)$ and a $DL$-Lite$^b_{\text{CORE}}$ ontology $K = \langle T, A \rangle$. By (3), by Lemmas 50, 53, and 56, and by (8) we have the sequence of equalities

$$q^{\text{Can}(K)} = \bigcup_{z \subseteq y} \bigcup_{T\text{-realisable } z \subseteq y} [q, z]^{\text{Can}(K)} = \bigcup_{T\text{-realisable } z \subseteq y} \Phi_z^A = \bigcup_{T\text{-realisable } z \subseteq y} (\bigvee \Phi_z)^A = \Phi_q^A,$$

which proves the statement of the theorem. □

Theorems 42 and 57, together with the fact that $\Phi_q$ does not depend on $A$, imply our main rewritability result.

**Corollary 58.** The class of rooted CQs is rewritable to BCALC over $DL$-Lite$^b_{\text{CORE}}$.

Recall that by Corollary 43 the rewritability result applies regardless of the adoption of the UNA. Note also that we need to manipulate only finite bags when evaluating the rewriting $\Phi$. Since BCALC maximal union is expressible via BCALC arithmetic union and difference for such bags according to equation (1), we can strengthen Corollary 58 and claim that there always exists a rewriting that uses only $\wedge$, $\vee$, $\\setminus$, equalities, and existential quantification.

We conclude with the LogSpace upper bound on the data complexity of the query answering for rooted CQs over $DL$-Lite$^b_{\text{CORE}}$. We can decide this problem by checking the ontology for non-satisfiability as in the usual set setting and, if the check fails, evaluating the BCALC rewriting of the input query on the ABox. The algorithm is correct by Statement 1 of Theorem 23 on the equivalence of satisfiability under bag and set semantics, and by Corollaries 43 and 58. Both steps can be done in LogSpace by Proposition 5 on the LogSpace membership of the query answering problem for BCALC and the results obtained by Calvanese et al. [9].

**Corollary 59.** $\text{BAGCERT}[\text{rooted CQs, } DL$-Lite$^b_{\text{CORE}}]$ and $\text{BAGCERT}^{\text{UNA}}[\text{rooted CQs, } DL$-Lite$^b_{\text{CORE}}]$ are in LogSpace in data complexity.

9. Rewritability of conjunctive queries over $DL$-Lite$^b_{\text{RDFS}}$ under UNA

In this section we show BCALC rewritability of CQs over $DL$-Lite$^b_{\text{RDFS}}$ under the UNA as well as tractability of the corresponding query answering problem in data complexity. This result holds only under the UNA since, as shown in Theorem 29, even rooted CQ answering over $DL$-Lite$^b_{\text{RDFS}}$ is coNP-hard in data complexity if the UNA is dropped. However, that the results of this section hold for arbitrary CQs and not just rooted ones.

We proceed analogously to the case of $DL$-Lite$^b_{\text{CORE}}$ described in the previous two sections; however, the absence of existential quantification on the right-hand side of concept inclusions considerably simplifies the exposition.

First, to formalise canonical bag interpretations for $DL$-Lite$^b_{\text{RDFS}}$, we introduce the notion of role closure, which is analogous to that of concept closure in Section 7. The role closure $\text{rcI}(\langle u, v \rangle, I)$ of a pair $(u, v)$ of elements $u, v \in \Delta^l$ in a bag interpretation $I = \langle \Delta^l, l \rangle$ over a TBox $\mathcal{T}$ is the bag of roles such that, for every role $R$,

$$\text{rcI}(\langle u, v \rangle, I)(R) = \max\{R_0^l(u, v) \mid \mathcal{T} \models R_0 \subseteq R \}.$$

In other words, $\text{rcI}(\langle u, v \rangle, I)(R)$ is the maximum value of $R_0^l(u, v)$ amongst all roles $R_0$ satisfying $\mathcal{T} \models R_0 \subseteq R$.

We are now ready to define canonical bag interpretations for $DL$-Lite$^b_{\text{RDFS}}$ ontologies.

**Definition 60.** Let $K = \langle T, A \rangle$ be a $DL$-Lite$^b_{\text{RDFS}}$ ontology and $\text{Can}_0(K)$ be the bag interpretation corresponding to $A$—that is, such that $\Delta^l \text{Can}_0(K) = \mathcal{I}$, $a \text{Can}_0(K) = a$ for each $a \in \mathcal{I}$, and $s \text{Can}_0(K)(a) = A(S(a))$ for each $S \in C \cup R$ and tuple of individuals $a$. Let also $\text{Can}_R(K)$ be the bag interpretation with domain $\mathcal{I}$ that interprets individuals and atomic concepts as $\text{Can}_0(K)$ and, for each atomic role $P$ and individuals $a, b$,

$$p \text{Can}_R(K)(a, b) = \text{rcI}(\langle a, b \rangle, \text{Can}_0(K))(P).$$

The canonical bag interpretation $\text{Can}(K)$ of $K$ is the bag interpretation with domain $\mathcal{I}$ that interprets individuals and atomic roles as $\text{Can}_R(K)$, and, for each atomic concept $A$ and individual $a$, satisfies

$$A \text{Can}(K)(a) = \text{cclI}(a, \text{Can}(K))(A).$$
There are two main differences between the canonical bag interpretations for DL-Lite$_{RDFS}^b$ and DL-Lite$_{CORE}^b$ ontologies. On the one hand, canonical bag interpretations for DL-Lite$_{RDFS}^b$ do not involve anonymous domain elements; this is because the logic does not support existentially quantified concepts on the right-hand side of inclusions. On the other hand, canonical bag interpretations for DL-Lite$_{RDFS}^b$ need to satisfy role inclusions, which is ensured using the notion of role closure. As expected, the aforementioned definitions of a canonical bag interpretation coincide for ontologies that are both in DL-Lite$_{RDFS}^b$ and DL-Lite$_{CORE}^b$.

We next argue that the canonical bag interpretation in Definition 60 is a universal model for CQs under the assumptions in this section.

**Proposition 61.** The canonical bag interpretation of a DL-Lite$_{RDFS}^b$ ontology is a universal model for the class of all CQs under the assumptions in this section.

**Proof.** First, note that every DL-Lite$_{RDFS}^b$ ontology is satisfiable under the UNA, because it does not have any disjointness axioms (or any other axioms that may cause an inconsistency). Moreover, the canonical bag interpretation Can$(K)$ is a model of a DL-Lite$_{RDFS}^b$ ontology $K$ by construction. Second, every bag interpretation $I$ satisfying the UNA has the corresponding bag interpretation $I_s$ satisfying the standard name assumption—that is, $I_s$ is the same as $I$ except that, for every $a \in I$, the individual $a$ itself is used in $I_s$ as an element instead of $a^I$; moreover, $q^I = q^{I_s}$ for every query $q$. Therefore, it is enough to show that, for every model $I$ of $K$, $I_s$ contains Can$(K)$, in the sense that $\Delta^\text{Can}(K) \subseteq \Delta^{I_s}$ and $\Sigma^\text{Can}(K) \subseteq \Sigma^{I_s}$ for every $S \in \mathbb{C} \cup \mathbb{R}$. However, this follows from the definitions of a bag interpretation and the canonical bag interpretation, because concept and role closures essentially encode satisfaction of the inclusions. $\square$

**Example 62.** Consider the DL-Lite$_{RDFS}^b$ ontology $K = (\mathcal{T}, \mathcal{A})$ where $\mathcal{T}$ is the TBox defined as

$$\mathcal{T} = \{ \text{hasMusician} \sqsubseteq \text{Record}, \text{playsOn} \sqsubseteq \text{hasMusician} \}$$

and $\mathcal{A}$ is the bag ABox defined as

$$\mathcal{A} = \{ \text{Record(Expectations)} : 1, \text{hasMusician(Expectations, K. Jarrett)} : 1, \text{playsOn(P. Motian, Expectations)} : 1 \}.$$

The canonical bag interpretation Can$(K)$ is the bag interpretation having domain $\Delta^\text{Can}(K) = I$, interpreting all individuals by themselves, and assigning bags to atomic concepts and atomic roles as follows:

$$\text{Record}^\text{Can}(K) = \| \text{Expectations} : 2 \|,$$

$$\text{playsOn}^\text{Can}(K) = \| \{ \text{P. Motian, Expectations} \} : 1 \|,$$

$$\text{hasMusician}^\text{Can}(K) = \| \{ \text{Expectations, K. Jarrett} \} : 1, \{ \text{Expectations, P. Motian} \} : 1 \|.$$

By Definition 60, the bag interpretations Can$_0(K)$ and Can$_R(K)$ are involved in the construction of Can$(K)$. These are specified as follows. Bag interpretation Can$_0(K)$ reflects the bag ABox $\mathcal{A}$ and maps Record to $\| \text{Expectations} : 1 \|$, hasMusician to $\| \{ \text{Expectations, K. Jarrett} \} : 1 \|$, and playsOn to $\| \{ \text{P. Motian, Expectations} \} : 1 \|$. Bag interpretation Can$_R(K)$ extends Can$_0(K)$ by setting the multiplicity of tuple (Expectations, P. Motian) in role hasMusician to 1, thus, satisfying the role inclusion axiom in $\mathcal{T}$. Then, interpretation Can$(K)$ is defined on the basis of Can$_R(K)$ by setting the multiplicity of element Expectations in Record to 2, thus, satisfying the concept inclusion in $\mathcal{T}$.

Observe that satisfying concept inclusions only after role inclusions have been satisfied is crucial for deriving a model of $K$. This is because satisfaction of role inclusions may result in the increase of an individual’s multiplicity in the interpretation of a concept, which clearly affects satisfaction of concept inclusions. Indeed, it is easy to verify that the interpretation resulting from the parallel satisfaction of the inclusions in $\mathcal{T}$ is not a model of $K$, because it assigns to role Record the bag $\| \text{Expectations} : 1 \|$, hence, violating the concept inclusion in $\mathcal{T}$. $<$

Next, we investigate BCALC rewritability of CQs over DL-Lite$_{RDFS}^b$ under the UNA. We start by arguing that rooted CQs can be rewritten to neither BCALC maximal nor BCALC arithmetic unions of CQs.

**Proposition 63.** The class of rooted CQs is rewriteable neither to BCALC maximal nor to BCALC arithmetic unions of CQs over DL-Lite$_{RDFS}^b$ under the UNA.

**Proof.** For BCALC maximal unions of CQs, consider the Boolean rooted CQ $q = A(a)$ for an atomic concept $A$ and individual $a$, and the DL-Lite$_{RDFS}^b$ ontology $K = (\mathcal{T}, \mathcal{A})$ with $\mathcal{T} = \{ P \subseteq R, \exists R \subseteq A \}$ and $\mathcal{A} = \{ P(a, b) : 1, R(a, c) : 1 \}$, for atomic roles $P$ and $R$ as well as individuals $b$ and $c$. By Proposition 61, we have $q^\text{Can}(K) = q^\text{Can}(K)$; moreover, $\Delta^\text{Can}(K) = I$, while

$$A^\text{Can}(K) = \{ a : 2 \}, \quad \text{p}^\text{Can}(K) = \{ (a, b) : 1 \}, \quad R^\text{Can}(K) = \{ (a, b) : 1, (a, c) : 1 \}.$$

As a result, we have $q^\text{Can}(K) = 2$. Assume now for the sake of contradiction that there exists a BCALC maximal union of CQs $\Phi$ such that $\Phi^\text{-A}(() \leq 2$. By the semantics of maximal union, this means that $\Phi$ contains a CQ $q_0 = \exists y. \Phi(y)$ satisfying $q_0^\text{-A}(() = 2$. There are only two ways in which this is possible:
– if there is only one valuation \( \lambda : y \cup 1 \rightarrow I \) for \( q_0 \) contributing multiplicity 2 to \( q_0^A(\langle \rangle) \), then \( \phi(y) \) must contain an atom \( S(t) \) such that \( S^A(\lambda(t)) = 2 \); however, this is not possible because \( A \) does not have any assertions with multiplicity 2;
– there exist two distinct valuations for \( q_0 \), each contributing multiplicity 1; however, this is not possible because \( A \) does not have two assertions with multiplicity 1 over the same atomic concept or role.

As a result, we conclude that \( \Phi \) cannot contain any such \( CQ \) \( q_0 \), and hence \( \Phi \) cannot be a rewriting of \( q \) with respect to \( T \).

Finally, non-rewritability to BCALC arithmetic unions of \( CQs \) follows already from the proof of the second part of Proposition 45, where the relevant TBox consists only of an inclusion between two atomic concepts. \( \square \)

In what follows, we show how to construct a BCALC rewriting of an arbitrary \( CQ \) with respect to a \( DL-Lite_{RDFS} \) TBox. The construction is much simpler than that for \( DL-Lite_{RDFS}^{core} \) in Section 8 since the canonical bag interpretation of a \( DL-Lite_{RDFS}^{b} \) ontology does not contain anonymous elements. In particular, there is no need to identify \( T \)-realisable subsets of existentially quantified variables and to transform the input \( CQ \) accordingly, which implies that Steps 1 and 2 in Section 8 are no longer required. As a result, a BCALC rewriting of the \( CQ \) can be directly constructed following an analogous approach to that of Step 3 in Section 8, and yet the operations of arithmetic union and difference are no longer required. Note, however, that the BCALC query resulting from the rewriting is not a BCALC maximal union of \( CQs \) since it interleaves maximal unions with existential quantification.

**Definition 64.** Let \( q(x) = \exists y. \phi(x, y) \) be a \( CQ \) and \( T \) be a \( DL-Lite_{RDFS} \) TBox. The BCALC query \( \Phi_q(x) \) is obtained from \( q(x) \) by replacing each occurrence of an atom \( A(t) \) and an atom \( P(t_1, t_2) \) with

\[
\bigvee_{T \models C \subseteq A} \zeta_C(t) \quad \text{and} \quad \bigvee_{T \models R \subseteq P} \xi_R(t_1, t_2),
\]

respectively, where, for every concept \( C \) and term \( t \), and for a fresh variable \( y \),

\[
\zeta_C(t) = \begin{cases} C(t), & \text{if } C \text{ is an atomic concept,} \\ \exists y. \left( \bigvee_{T \models R_0 \subseteq R} \xi_{R_0}(t, y) \right), & \text{if } C \text{ is of the form } \exists R, \end{cases}
\]

and, for every atomic role \( P \) and terms \( t_1 \) and \( t_2 \), \( \xi_P(t_1, t_2) = P(t_1, t_2) \) and \( \xi_{P^2}(t_1, t_2) = P(t_2, t_1) \).

So, analogously to Definition 54 in the \( DL-Lite_{RDFS}^{core} \) case, we replace each concept atom \( A(t) \) with a BCALC maximal union of all of them over \( t \) corresponding to concepts subsumed by \( A \) in \( T \); in contrast to Definition 54, however, we also need to take into account role inclusions, which results in the inclusion of further disjuncts in the second part of the definition of \( \zeta_C(t) \). Then, each role atom \( P(t_1, t_2) \) is replaced by the BCALC maximal union of all binary atoms over \( t_1 \) and \( t_2 \) corresponding to roles subsumed by \( P \) in \( T \).

We now provide an example illustrating the construction of a rewriting.

**Example 65.** Consider TBox \( T \) from Example 62 and the Boolean \( CQ \) \( q = \exists y_1, y_2. \text{Record}(y_1) \land \text{hasMusician}(y_1, y_2) \). Following Definition 64, the relevant concepts for rewriting the atom \( \text{Record}(y_1) \) are \( \text{Record} \) itself, \( \exists \text{hasMusician} \), and \( \exists \text{playsOn} \) as these are all the concepts subsumed by \( \text{Record} \) in \( T \). Similarly, the relevant roles for rewriting the atom \( \text{hasMusician}(y_1, y_2) \) are \( \text{hasMusician} \) itself and \( \text{playsOn} \) as these are all the roles subsumed by \( \text{hasMusician} \) in \( T \). Thus atoms \( \text{Record}(y_1) \) and \( \text{hasMusician}(y_1, y_2) \) are replaced in \( q \) with expressions

\[
\zeta_{\text{Record}}(y_1) \lor \zeta_{\exists \text{hasMusician}}(y_1) \lor \zeta_{\exists \text{playsOn}}(y_1, y_2) = \text{Record}(y_1) \lor \exists y. (\text{hasMusician}(y_1, y) \lor \text{playsOn}(y, y_1)) \lor \exists y. \text{playsOn}(y_1, y),
\]

and

\[
\zeta_{\exists \text{hasMusician}}(y_1, y_2) \lor \zeta_{\exists \text{playsOn}}(y_1, y_2) = \text{hasMusician}(y_1, y_2) \lor \text{playsOn}(y_2, y_1),
\]

respectively, resulting in the rewriting

\[
\Phi_q = \exists y_1, y_2. (\text{Record}(y_1) \lor \exists y. (\text{hasMusician}(y_1, y) \lor \text{playsOn}(y, y_1)) \lor \exists y. \text{playsOn}(y_1, y)) \land \text{hasMusician}(y_1, y_2) \lor \text{playsOn}(y_2, y_1)).
\]

Consider now ABox

\[
\mathcal{A} = \{ \text{Record}(\text{Expectations}) : 1, \text{hasMusician}(\text{Expectations.K.Jarrett}) : 1, \text{playsOn}(P.\text{Motion.\text{Expectations}}) : 1 \},
\]

from Example 62 and the canonical bag interpretation \( \text{Can}(K) \) of the \( DL-Lite_{RDFS}^{b} \) ontology \( K = \langle T, \mathcal{A} \rangle \), specified there. On the one hand, evaluating \( q \) over \( \text{Can}(K) \) results in the bag \( q^{\text{Can}(K)} = \{ (\langle \rangle : 4) \} \). On the other hand, it is immediate to check that \( \Phi_q \) also evaluates to \( \{ (\langle \rangle : 4) \} \) over \( \mathcal{A} \). \( \Diamond \)
We next prove correctness of the rewriting formalised in Definition 64.

**Theorem 66.** For every CQ $q$ and every DL-Lite$^b_{\text{RDFS}}$ ontology $K = (\mathcal{T}, \mathcal{A})$ we have that $q_{\text{Can}}(K) = \Phi^A_q$.

**Proof.** Let $q(x) = \exists y. \phi(x, y)$ be a CQ and $K = (\mathcal{T}, \mathcal{A})$ be DL-Lite$^b_{\text{RDFS}}$ ontology. To prove the claim, it suffices to show that for each atom $A(t)$ and $P(t_1, t_2)$ in $\phi(x, y)$, and for each valuation $\lambda : x \cup y \cup I \rightarrow \Delta_{\text{Can}}(K)$, the multiplicities $A_{\text{Can}}(K)(\lambda(t))$ and $P_{\text{Can}}(K)(\lambda(t_1), \lambda(t_2))$ are equal to the multiplicities of $\lambda(t)$ and $(\lambda(t_1), \lambda(t_2))$ in the bag answers to the BCALC queries of (12) over ABox $\mathcal{A}$, respectively.

Canonical interpretation $\text{Can}(K)$ assigns to every atomic concept $A$ and every atomic role $P$ the bag defined as $A_{\text{Can}}(K)(u) = \text{cel}_T[u, \text{Can}(K)(A)]$ and $P_{\text{Can}}(K)(u, v) = \text{rel}_T[(u, v), \text{Can}(K)(P)]$ for all elements $u, v \in \Delta_{\text{Can}}(K)$, respectively. By unfolding the definition of closures, we get the following, for every atom $A(t)$ and $P(t_1, t_2)$ in $\phi(x, y)$, and every valuation $\lambda$:

$$A_{\text{Can}}(K)(\lambda(t)) = \max(\max\{A(A_0(\lambda(t))) | A_0 \in C, \mathcal{T} \models A_0 \subseteq A\}, \max(\exists R)_{\text{Can}(K)}(\lambda(t)) | R \text{ a role, } \mathcal{T} \models \exists R \subseteq A\}),$$

$$P_{\text{Can}}(K)(\lambda(t_1), \lambda(t_2)) = \max(\max\{A(P_0(\lambda(t_1), \lambda(t_2))) | P_0 \in R, \mathcal{T} \models P_0 \subseteq P\}, \max\{A(P_0(\lambda(t_2), \lambda(t_1))) | P_0 \in R, \mathcal{T} \models P_0 \subseteq P\}).$$

Consider first the right-hand side of (13). Observe that the first subexpression involving the max function can be equivalently written as $\left(\bigvee_{A_0 \in C, \mathcal{T} \models A_0 \subseteq A} A_0(t)\right)^A(\lambda(t))$. As for the second subexpression, it can be equivalently written as $(\bigcup_{R \models \exists R} \max\{A(P)(\lambda(t)) | P \in R, \mathcal{T} \models P \subseteq P\})^A$, where bag $(\exists R)_{\text{Can}(K)}(\lambda(t))$ can be further written as $(\exists y'. (\bigvee_{T' = R_{\subseteq R} \lambda \mathcal{R}_0(t, y')})^A$. Thus, by substituting $\bigvee$ for $\cup$ in the aforementioned expression and for the outer max function in (13), and by the definition of $\zeta_c(t)$ in Definition 64, we derive

$$A_{\text{Can}}(K)(\lambda(t)) = \left(\bigvee_{T' = C_{\subseteq A}} \zeta_c(t)^A\right)(\lambda(t)),$$

as required. To complete the proof of the theorem, observe that from (14) and the semantics of BCALC queries we immediately obtain $P_{\text{Can}}(K)(\lambda(t_1), \lambda(t_2)) = \left(\bigvee_{T' = B_{\subseteq P}} \xi_t(t_1, t_2)^A\right)(\lambda(t_1), \lambda(t_2))$, as required. □

From Proposition 61 and Theorem 66 we obtain the rewrirtability result of this section.

**Corollary 67.** The class of all CQs is rewritable to positive BCALC over DL-Lite$^b_{\text{RDFS}}$ under the UNA.

All DL-Lite$^b_{\text{RDFS}}$ ontologies are satisfiable, so the data complexity upper bound for the corresponding query answering problem can be established as a straightforward consequence of Proposition 5 and Corollary 67.

**Corollary 68.** BagCert$^{\text{UNA}}[\text{CQ5, DL-Lite}^b_{\text{RDFS}}]$ is in LogSpace in data complexity.

### 10. Rewrirtability of rooted conjunctive queries over DL-Lite$^b_{\text{R}}$ under UNA

In this section we consider BCALC rewritings of rooted CQs over the ontology language DL-Lite$^b_{\text{R}}$, which extends both DL-Lite$^b_{\text{RDFS}}$ and DL-Lite$^b_{\text{CORE}}$. As defined in Section 2.1, this language provides all the constructs available in DL-Lite$^b_{\text{R}}$, but allows concept inclusions of the form $C \subseteq R$ only for roles $R$ that do not have more general roles in the ontology.

Similarly to the case of DL-Lite$^b_{\text{RDFS}}$ considered in the previous section, we focus on the semantics that adopts the UNA; this is justified by Theorem 29, where we showed that (rooted) CQ answering over ontology languages allowing for role inclusions is coNP-hard in data complexity if the UNA is not adopted. However, in contrast to the previous section, we focus on rooted CQs rather than general CQs; this choice is justified by Theorem 28, where we established coNP-hardness of CQ answering over DL-Lite$^b_{\text{RDFS}}$ under the UNA.

The strategy behind our proof of BCALC rewritability of rooted CQs over DL-Lite$^b_{\text{R}}$ is to seamlessly combine the techniques developed in the previous three sections for DL-Lite$^b_{\text{RDFS}}$ and DL-Lite$^b_{\text{CORE}}$. We start with the definition of the canonical bag interpretation. The construction is divided into two steps: first, we build the canonical bag interpretation of the DL-Lite$^b_{\text{RDFS}}$-sub-ontology consisting only of role inclusions, and then use the resulting interpretation as the starting point for the recursive model construction done for DL-Lite$^b_{\text{CORE}}$ where we now ignore the role inclusions.

**Definition 69.** The canonical bag interpretation $\text{Can}(K)$ of a DL-Lite$^b_{\text{R}}$ ontology $K = (\mathcal{T}, \mathcal{A})$ is the bag interpretation $\bigcup_{i \geq 0} \text{Can}_i(K)$, where $\text{Can}_i(K)$ is the canonical bag interpretation of the DL-Lite$^b_{\text{RDFS}}$ ontology obtained from $K$ by keeping only the role inclusions in $\mathcal{T}$ (as defined in Definition 60), and, for each $i \geq 1$, interpretation $\text{Can}_i(K)$ is defined on the basis of $\text{Can}_{i-1}(K)$ as in Definition 32 by considering only the concept inclusions in $\mathcal{T}$. 
It is immediate to check that this definition generalises the definitions of canonical bag interpretations for $\text{DL-Lite}^\text{b-core}$ and $\text{DL-Lite}^\text{b-DFS}$. We next illustrate the aforementioned definition with an example.

**Example 70.** Consider the $\text{DL-Lite}^\text{b-}\mathcal{R}_{\land}$ ontology $\mathcal{K} = (\mathcal{T}, \mathcal{A})$, where
\[
\mathcal{T} = \{ \text{Record} \sqsubseteq \exists \text{hasMusician}, \exists \text{hasMusician} \sqsubseteq \text{Musician}, \text{playsOn} \sqsubseteq \text{hasMusician} \},
\mathcal{A} = \{ \text{Record} (\text{Expectations}) : 2, \text{playsOn} (\text{K. Jarrett, Expectations}) : 1 \}.
\]
The canonical bag interpretation $\text{Can}(\mathcal{K})$ has domain $\Delta^\text{Can}(\mathcal{K}) = \mathcal{I} \cup \{ w^1_{\text{Expectations}, \exists \text{hasMusician}} \}$, interprets all the individuals by themselves, and assigns bags to atomic concepts and atomic roles as follows:
\[
\begin{align*}
\text{Record}^\text{Can}(\mathcal{K}) &= \{ \text{Expectations} : 2 \}, \\
\text{Musician}^\text{Can}(\mathcal{K}) &= \{ \text{K. Jarrett} : 1, w^1_{\text{Expectations}, \exists \text{hasMusician}} : 1 \}, \\
\text{playsOn}^\text{Can}(\mathcal{K}) &= \{ (\text{K. Jarrett, Expectations}) : 1 \}, \\
\text{hasMusician}^\text{Can}(\mathcal{K}) &= \{ (\text{Expectations, K. Jarrett}) : 1, (\text{Expectations, w}^1_{\text{Expectations, \exists \text{hasMusician}}}) : 1 \}. \quad \text{□}
\end{align*}
\]

Note that the canonical bag interpretation of a satisfiable $\text{DL-Lite}^\text{b-}\mathcal{R}_{\land}$ ontology $\mathcal{K} = (\mathcal{T}, \mathcal{A})$ is always a model of $\mathcal{K}$ by construction and the fact that $\mathcal{T}$ does not have concept inclusions of the form $C \sqsubseteq \exists R$ whenever $R$ has a more general role in $\mathcal{T}$. This fact guarantees that the subinterpretation of $\text{Can}(\mathcal{K})$ corresponding to the set of concept inclusions in $\mathcal{T}$, namely $\bigcup_{i=1}^n \text{Can}_i(\mathcal{K})$, does not violate the role inclusions in $\mathcal{T}$ that have been already satisfied in $\text{Can}_0(\mathcal{K})$. Given this property, the following is a generalisation of Proposition 34 for the canonical bag interpretations of $\text{DL-Lite}^\text{b-core}$ ontologies, which holds again regardless of whether the UNA is adopted or not (recall that this was automatic for the case of $\text{DL-Lite}^\text{b-DFS}$ because this language does not allow for any inconsistencies).

**Proposition 71.** If a $\text{DL-Lite}^\text{b-}\mathcal{R}_{\land}$ ontology $\mathcal{K}$ is satisfiable then its canonical bag interpretation is its model.

We next show that the canonical bag interpretation is universal for the class of rooted CQs. For this, we establish the counterpart of Lemma 38 for the case of $\text{DL-Lite}^\text{b-}\mathcal{R}_{\land}$ ontologies.

**Lemma 72.** For every $\text{DL-Lite}^\text{b-}\mathcal{R}_{\land}$ ontology $\mathcal{K}$ and every model $\mathbf{I}$ of $\mathcal{K}$ there exists an $e$-homomorphism from $\text{Can}^\mathbf{I}(\mathcal{K})$ to $\mathbf{I}^e$ that is multiplicity-preserving on $\mathbf{I}$.

**Proof.** The proof goes along the same lines as the proof of Lemma 38. The only difference is the way in which we define a witnessing multiplicity-preserving e-homomorphism $(h, h_1, \ldots)$ for atomic roles and the elements in $\Delta^\text{Can}(\mathcal{K})$—that is, on the (interpretations of the) individuals $\mathbf{I}$. However, the existence of such an e-homomorphism follows from the proof of Proposition 61, which shows universality of the canonical bag interpretation for $\text{DL-Lite}^\text{b-DFS}$. Indeed, the fact that for every model $\mathbf{I}$ of $\mathcal{K}$ the corresponding model $\mathbf{I}_\Phi$ (i.e., the model satisfying the standard name assumption) is such that $\Delta^\text{Can}_0(\mathcal{K}) \subseteq \Delta^\mathbf{I}_\Phi$ and $\Delta^\text{Can}_0(\mathcal{K}) \subseteq \Delta^\Phi$ for every $\Phi \in \mathcal{K} \cup \mathcal{R}$ implies the existence of an $e$-homomorphism. \text{□}

Using Lemma 72 and the fact that Lemma 41 transfers trivially to $\text{DL-Lite}^\text{b-}\mathcal{R}_{\land}$ ontologies, we can conclude that Theorem 42 (as well as Proposition 61) generalises to the case of $\text{DL-Lite}^\text{b-}\mathcal{R}_{\land}$.

**Theorem 73.** The canonical bag interpretation $\text{Can}(\mathcal{K})$ of a satisfiable $\text{DL-Lite}^\text{b-}\mathcal{R}_{\land}$ ontology $\mathcal{K}$ is a universal model for the class of rooted CQs under the UNA.

We next move to the BCALC rewritability of rooted CQs over $\text{DL-Lite}^\text{b-}\mathcal{R}_{\land}$. We start by pointing out that this class of queries is rewritable to neither BCALC maximal nor to BCALC arithmetic unions of CQs over $\text{DL-Lite}^\text{b-}\mathcal{R}_{\land}$ under the UNA, which is a direct consequence of Propositions 45 and 63, as well as the fact that $\text{DL-Lite}^\text{b-}\mathcal{R}_{\land}$ extends both $\text{DL-Lite}^\text{b-core}$ and $\text{DL-Lite}^\text{b-DFS}$. In fact, the rewriting algorithm can be seen as a combination of the corresponding algorithms for $\text{DL-Lite}^\text{b-core}$ and $\text{DL-Lite}^\text{b-DFS}$, described in Sections 8 and 9, respectively. When given as input a $\text{DL-Lite}^\mathcal{R}_{\land}$ TBox $\mathcal{T}$ and a rooted CQ $q(\mathbf{x}) = \exists \mathbf{y}. \phi(\mathbf{x}, \mathbf{y})$, the algorithm considers each subset $\mathbf{z}$ of $\mathbf{y}$ independently and then takes the BCALC arithmetic union of the resulting rewritings. In particular, for each such $\mathbf{z}$ we proceed according to the following three steps:

1. as specified in Definition 48, $\mathbf{z}$ is checked for $\mathcal{T}_C$-realisability, where $\mathcal{T}_C$ is the set of concept inclusions in $\mathcal{T}$, and disregarded from consideration if the check fails,
2. as specified in Definition 51, each maximally connected component of the subquery corresponding to $\mathbf{z}$ is replaced by a single representative role atom, resulting in a CQ; and
3. all atoms are rewritten to a BCALC query that takes into account the TBox and the fact that $z$ should be mapped to anonymous elements.

The only essential difference to the algorithm for DL-Lite$^b_{\text{core}}$ is in the last step, where we now take into account also the presence of role inclusions, as in the case of DL-Lite$^b_{\text{RDFS}}$.

The construction of the rewriting is formalised by the following definition.

**Definition 74.** Let $q(x) = \exists y. \phi(x, y)$ be a rooted CQ and $T$ a DL-Lite$_{R^{-}}$ TBox. For a $T_C$-realisable subset $z$ of $y$, let $\Phi_z(x)$ be the query obtained from the CQ $q_z(x)$, given in Definition 51, by replacing

- each occurrence of an atom $A(t)$ and an atom $P(t_1, t_2)$ with
  
  \[
  \bigvee_{T_C=C \subseteq A} \zeta_C(t) \quad \text{and} \quad \bigvee_{T_P=R \subseteq P} \xi_R(t_1, t_2),
  \]

- the atom $P_V(t, y_V)$ and the atom $P_V(y_V, t)$ for each maximally connected $v \subseteq z$ with the following BCALC query, where $R_V$ is $P_V$ and $P^\prime_V$, respectively:
  
  \[
  \left(\bigvee_{T_C=C \subseteq R_V} \zeta_C(t) \right) \setminus \zeta_{R^\prime_V}(t),
  \]

where, for every concept $C$ and term $t$, and for a fresh variable $y$,

\[
\zeta_C(t) = \begin{cases} 
C(t), & \text{if } C \text{ is an atomic concept,} \\
\exists y. \left(\bigvee_{T_P=R \subseteq R} \xi_R(t, y)\right), & \text{if } C \text{ is of the form } \exists R,
\end{cases}
\]

and, for every atomic role $P$ and terms $t_1, t_2$, $\xi_P(t_1, t_2) = P(t_1, t_2)$ and $\xi_{P^\prime}(t_1, t_2) = P(t_2, t_1)$.

Finally, let

\[
\Phi_q(x) = \bigvee_{T_C=\text{realisable } y \subseteq z} \Phi_z(x).
\]

The structure of the rewriting in Definition 74 is similar to that for DL-Lite$^b_{\text{core}}$ in Definition 54. The main differences are as follows:

- when rewriting atoms not having variables in $z$, we take into account the role inclusions as in the DL-Lite$^b_{\text{RDFS}}$ case; and

- when rewriting role atoms $P_V(t, y_V)$ and $P_V(y_V, t)$ for maximally connected subsets $v$ of variables, we distinguish between different types of subconcepts of $\exists P_V$ and $\exists P^\prime_V$, respectively; for atomic concepts we follow the rewriting for DL-Lite$^b_{\text{core}}$, while for existentially quantified subconcepts we take into account the role inclusions as in the DL-Lite$^b_{\text{RDFS}}$ case.

**Example 75.** Consider TBox $T$ from Example 70 and the rooted CQ $q(x) = \exists y. \text{hasMusician}(x, y) \land \text{Musician}(y)$, same as in Example 55. As before, there are two $T_C$-realisable subsets of $y$, namely $\emptyset$ and $y$, and

\[
q_\emptyset(x) = q(x) \quad \text{and} \quad q_y(x) = \exists y'. \text{hasMusician}(x, y').
\]

The rewritings of these CQs are

\[
\Phi_q(x) = \exists y. (\text{hasMusician}(x, y) \lor \text{playsOn}(y, x)) \land (\text{Musician}(y) \lor \exists y'. (\text{hasMusician}(y', y) \lor \text{playsOn}(y, y'))) ,
\]

\[
\Phi_y(x) = (\text{Record}(x) \lor \exists y'. (\text{hasMusician}(x, y') \lor \text{playsOn}(y', x)) \setminus \exists y'. (\text{hasMusician}(x, y') \lor \text{playsOn}(y', x))).
\]

Evaluating the two rewritings on ABox $A$ from Example 70, we get $\Phi_q^A = \Phi_y^A = \{\text{Expectations : 1}\}$. Therefore, $\Phi_q^A = (\Phi_\emptyset \lor \Phi_y)^A = \{\text{Expectations : 2}\}$, which is equal to $q^\text{can}(T^-, A) = q(T^-, A)$ as expected. 

We show the correctness of the approach by means of the following generalisation of Lemma 56.

**Lemma 76.** Let $q(x)$ be a rooted CQ and $T$ a DL-Lite$_{R^{-}}$ TBox. Then $[q_z, z]^\text{can}(T^-, A) = \Phi_z^A$ for every bag ABox $A$ and every $T_C$-realisable subset $z$ of $y$. 


Proof. The proof is similar to the proofs of Lemma 56 and Theorem 66. Consider an arbitrary bag ABox $\mathcal{A}$ and let $\mathcal{K} = \langle T, \mathcal{A} \rangle$ be the resulting DL-Lite$_b^{\mathcal{R}}_{\mathcal{K}}$ ontology. We need to show that the bag answers over each bag ABox $\mathcal{A}$ to the rewritings of atoms $A(t), P(t_1, t_2)$ and $\xi_y \sqrt{P(t, y)}$ in $q_2$ as these appear on the left- and right-hand side of (15), as well as in (16), respectively, are equal to the bag answers $[A(t), \emptyset]^{\text{can}}(\mathcal{K}), [P(t_1, t_2), \emptyset]^{\text{can}}(\mathcal{K})$, and $[\exists y. \xi_y \sqrt{P(t, y)}, y]^{\text{can}}(\mathcal{K})$, respectively.

We now argue the correctness of replacements (15) and (16) in Definition 74. To begin, let $\mathcal{T}_R$ be the set of role inclusions in $T$, let $I_A = \langle \Delta^{I_A}, \Lambda^{I_A} \rangle$ be the bag interpretation corresponding to ABox $\mathcal{A}$—that is, the interpretation defined as $\Delta^{I_A} = I$, $\Lambda^{I_A} = a$ for all $a \in I$, and $S^{I_A}(a) = A(S(a))$ for all $S \in C \cup R$ and tuples of individuals $a$, and let $\text{Can}_{\text{bag}}(\langle \mathcal{T}_R, \mathcal{A} \rangle)$ be the bag interpretation for the DL-Lite$_b^{\mathcal{R}}_{\mathcal{K}}$ ontology $\langle \mathcal{T}_R, \mathcal{A} \rangle$ defined on the basis of $I_A$ according to Definition 60 (note that we use $I_A$ instead of $\text{Can}_{\text{bag}}(\mathcal{K})$ since the definition of the latter in DL-Lite$_b^{\mathcal{R}}_{\mathcal{K}}$ differs from the one in DL-Lite$_b^{\mathcal{R}}_{\mathcal{K}}$). By the definitions of canonical bag interpretations for DL-Lite$_b^{\mathcal{R}}_{\mathcal{K}}$ ontologies as well as concept and role closures, for all individuals $a$ and $b$, concepts $C$, and atomic roles $P$, we have the following equalities:

$$c^{\text{Can}_{\mathcal{K}}}(a) = c^{\text{Can}_{\mathcal{K}_1}}(a) = \text{cc}(T_\mathcal{A}[a, \text{Can}_{\mathcal{K}_1}(\mathcal{C})] = \max(\max\{\mathcal{A}(A_0(a)) \mid A_0 \subseteq \mathcal{C}, T \models A_0 \subseteq \mathcal{C}\}, \max\{\exists R^{\text{Can}_{\mathcal{K}_1}(\langle \mathcal{T}_R, \mathcal{A} \rangle)}(a) \mid R \in \mathcal{A}\}, T \models \exists R \subseteq \mathcal{C}) \).$$

$$p^{\text{Can}_{\mathcal{K}}}(a, b) = \text{cc}(T_\mathcal{A}[a, b, I_A](P) = \text{cc}(T_\mathcal{A}[a, (b), I_A](P) = \max(\max\{\mathcal{A}(P_0(a, b)) \mid P_0 \in \mathcal{R}, T \models P_0 \subseteq P\}, \max\{\mathcal{A}(P_0(a, b)) \mid P_0 \in \mathcal{R}, T \models P_0 \subseteq P\}).$$

From (17), (18), and the semantics of BCALC queries, we immediately derive the following, for every $a, b \in I, A \subseteq \mathcal{C},$ and $P \in \mathcal{R}$ (see also the derivations for (13) and (14) in the proof of Theorem 66):

$$[A(a), \emptyset]^{\text{Can}_{\mathcal{K}}}(a) = \left( \bigcup_{T \models \exists R \subseteq P} \xi_{R \subseteq P}(a) \right)^{\mathcal{A}} \text{ and } [P(a, b), \emptyset]^{\text{Can}_{\mathcal{K}}}(a) = \left( \bigcup_{T \models \exists R \subseteq P} \xi_{R \subseteq P}(a, b) \right)^{\mathcal{A}},$$

which proves the claim for atoms of the form $A(t)$ and $P(t_1, t_2)$. The claim for $P_y$ atoms is proved using (17) and similarly to the proof of Lemma 56. □

Using this lemma, the following theorem can be proved in exactly the same way as Theorem 57.

Theorem 77. For every rooted CQ $q(x)$ and every DL-Lite$_b^{\mathcal{R}}_{\mathcal{K}}$ ontology $\mathcal{K} = \langle T, \mathcal{A} \rangle$ we have that $q^{\text{Can}_{\mathcal{K}}}(a) = \Phi_q^{\mathcal{A}}$.

Combining Theorem 73 and Theorem 77, we derive the following corollary.

Corollary 78. The class of rooted CQs is rewritable to BCALC over DL-Lite$_b^{\mathcal{R}}_{\mathcal{K}}$ under the UNA.

Similarly to Corollary 58 in the DL-Lite$_b^{\mathcal{R}}_{\mathcal{K}}$ case, we can strengthen Corollary 78 and claim that there always exists a rewriting that uses only $\land, \lor, \neg$, equalities, and existential quantification.

We conclude this section by establishing the data complexity of rooted CQ answering over DL-Lite$_b^{\mathcal{R}}_{\mathcal{K}}$. Problem BagCert$_{\text{in}}$[rooted CQs, DL-Lite$_b^{\mathcal{R}}_{\mathcal{K}}$] can be decided in the same way as BagCert[rooted CQs, DL-Lite$_b^{\mathcal{R}}_{\mathcal{K}}$] (see Section 8.6) with the only difference that the rewriting $\Phi_q$ is now computed as in Definition 74. Correctness of the algorithm follows from Theorems 73 and 77, and yields the following data complexity bound.

Corollary 79. BagCert$_{\text{in}}$[rooted CQs, DL-Lite$_b^{\mathcal{R}}_{\mathcal{K}}$] is in LOGSPACE in data complexity.

11. Related work

In this section, we establish bridges between our bag semantics for OBDA and existing work in the literature on data exchange and query answering over DL-Lite.

11.1. Bag semantics in data exchange

Hernich and Kolaitis [20] have recently studied CQ answering under bag semantics in the context of data exchange. As is customary in conventional treatments of data exchange [32], their setting considers disjoint source and target database schemas, which are related via source-to-target GLAV mappings. In the data exchange literature it is common to also consider dependencies over the target schema (which can equivalently be seen as ontological axioms); however, the semantics by Hernich and Kolaitis is defined only under the assumption that no such dependencies exist.
Definition 80. A bag data exchange setting is a tuple \( \langle S, T, \Sigma, \mathcal{D} \rangle \) where

- \( S \) is a source database schema;
- \( T \) is a target database schema disjoint from \( S \);
- \( \Sigma \) is a finite set of global-and-local-as-view (GLAV) mapping assertions (or mappings) of the form
  \[ q(x) \sim p(x) \]
  where \( q(x) \) is a CQ over \( S \) and \( p(x) \) is a CQ over \( T \); and
- \( \mathcal{D} \) is a bag database instance over \( S \).

In the conventional set-based case, a solution of a data exchange setting is a finite database over the target schema that satisfies all the mappings together with the source database. Hernich and Kolaitis [20] defined two possible generalisations of this notion to bag semantics.

Definition 81. Let \( \langle S, T, \Sigma, \mathcal{D} \rangle \) be a bag data exchange setting. A bag database instance \( B \) over \( T \) is

- an incognizant solution (or i-solution), if \( q^D \subseteq p^B \) for every mapping \( q(x) \sim p(x) \) in \( \Sigma \);
- a cognizant solution (or c-solution), if, for each mapping \( \sigma = q(x) \sim p(x) \) in \( \Sigma \), there exists a bag database instance \( B_\sigma \) over \( T \) with \( q^D \subseteq p^{B_\sigma} \) such that
  \[ \bigcup_{\sigma \in \Sigma} B_\sigma \subseteq B. \]

In other words, the incognizant semantics adopts the maximal union approach for interpreting mappings defining the same view—that is, when applied to GAV mappings \( q_1(x) \sim S(x) \) and \( q_2(x) \sim S(x) \) for a predicate \( S \), and a database \( \mathcal{D} \), an incognizant solution \( B \) must satisfy \( q_1^D \cup q_2^D \subseteq S^B \); in contrast, the cognizant semantics adopts the arithmetic union approach and requires that a solution \( B \) satisfies \( q_1^D \cup q_2^D \subseteq S^B \).

Query answering in bag data exchange settings is defined as the problem of computing the bag certain answers to a query over the target schema with respect to the set of solutions.

Definition 82. Let \( \langle S, T, \Sigma, \mathcal{D} \rangle \) be a bag data exchange setting and let \( q \) be a CQ over \( T \). The incognizant certain answers i-certain\( \langle S, T, \Sigma, \mathcal{D} \rangle \)(\( q \)) to \( q \) with respect to \( \langle S, T, \Sigma, \mathcal{D} \rangle \) are defined as follows:

\[ \text{i-certain}_\langle S, T, \Sigma, \mathcal{D} \rangle(q) = \bigcap_{B \text{ is an i-solution for } \langle S, T, \Sigma, \mathcal{D} \rangle} q^B. \]

The cognizant certain answers c-certain\( \langle S, T, \Sigma, \mathcal{D} \rangle \)(\( q \)) to \( q \) with respect to \( \langle S, T, \Sigma, \mathcal{D} \rangle \) are defined in the same way except that c-solutions are considered in the intersection instead of i-solutions.

We are now ready to establish the connection between our OBDA framework and that of Hernich and Kolaitis for data exchange. Note that there are several differences between our OBDA settings, introduced in Definition 6, and data exchange settings. First, target schemas in OBDA are restricted to predicates of arity one and two, whereas in data exchange the arity of target predicates is unbounded. Second, data exchange mappings have CQs on both sides, whereas OBDA mappings have arbitrary BCalC queries on the source side, but only atomic queries of the form \( A(x) \) and \( P(x, y) \) on the target side. Third, an OBDA setting comes with a TBox, whereas the data exchange setting in [20] does not allow for any dependencies over the target schema. Finally, from a semantics point of view, certain answers in OBDA are defined in terms of (possibly infinite) models, whereas certain answers in data exchange are defined in terms of (finite) database solutions.

We next show that, despite these mismatches, our semantics is compatible with that of Hernich and Kolaitis. For this we note that OBDA and data exchange settings are syntactically compatible if we assume target predicates of arity only one and two, absence of an ontology, and restrict ourselves to GAV mappings with CQs on the source side and atoms \( A(x) \) and \( P(x, y) \) on the target side. Under these assumptions, we argue that our semantics coincides with the incognizant semantics of Hernich and Kolaitis; furthermore, their cognizant semantics can be simulated in our setting by means of a suitable rewriting of their mappings. Note also that the following proposition is stated for the OBDA semantics without the UNA; however, adopting the UNA would not affect this result, because in case of an empty ontology the two semantics coincide.

Proposition 83. Let \( \langle S, T, \Sigma, \mathcal{D} \rangle \) be a data exchange setting such that \( T \) consists only of unary and binary predicates and all mappings in \( \Sigma \) have atoms of the form \( A(x) \) or \( P(x, y) \) on the target side. Then, for every CQ \( q \) over \( T \),

---

Note that Hernich and Kolaitis do not allow for equality atoms in CQs, but this does not affect expressivity.
1. \( \text{i-certain}_{\mathbf{S}, \mathbf{T}, \Sigma, \mathcal{D}}(q) = q^{(\emptyset, \Sigma, \mathcal{D})} \), and

2. \( \text{c-certain}_{\mathbf{S}, \mathbf{T}, \Sigma, \mathcal{D}}(q) = q^{(\emptyset, \Sigma', \mathcal{D})} \), where \( \Sigma' \) is obtained from \( \Sigma \) by replacing, for each atom \( S(x) \), all the mappings with \( S(x) \) on the target side with

\[
\bigvee_{q(x) \sim S(x)} q(x).
\]

**Proof.** Given a bag database \( B \) over \( T \), let \( I_B \) be the bag interpretation corresponding to \( B \)—that is, the interpretation having the set of individuals (i.e., constants) appearing in \( B \) as the domain and satisfying \( S^{\text{Bi}} = S^B \) for every predicate \( S \in T \). We claim that, for every bag database \( B \) over \( T \),

1. \( B \) is an i-solution for \( (\mathbf{S}, \mathbf{T}, \Sigma, \mathcal{D}) \) if and only if \( I_B \) is a model of \( (\emptyset, \Sigma, \mathcal{D}) \), and
2. \( B \) is a c-solution for \( (\mathbf{S}, \mathbf{T}, \Sigma, \mathcal{D}) \) if and only if \( I_B \) is a model of \( (\emptyset, \Sigma', \mathcal{D}) \).

The first claim follows from Definitions 81 and 9 of the semantics of data exchange and OBDA, respectively.

Consider now the second claim. The fact that \( B \) is a c-solution for \( (\mathbf{S}, \mathbf{T}, \Sigma, \mathcal{D}) \) means that there are bag database instances \( B_\sigma \), for \( \sigma \in \Sigma \), such that \( q^D \subseteq (S(x))^{\text{Bi}} \) for each \( \sigma = q(x) \sim S(x) \) and \( \bigcup_{\sigma \in \Sigma} B_\sigma \subseteq B \). This is equivalent to the fact that, for each atom \( S(x) \),

\[
\bigcup_{q(x) \sim S(x)} q^D \subseteq S^{\text{Bi}},
\]

which means that \( I_B \) is a model of \( (\emptyset, \Sigma', \mathcal{D}) \) by the definition of \( \Sigma' \). To conclude the proof, note that, since an ontology with an empty TBox has a unique finite model that is minimal with respect to bag inclusion, by Lemma 21, it is enough to consider only finite interpretations when computing the bag certain answers to \( q \). \( \square \)

Note that the second statement in this proposition is essentially illustrated in Example 11.

We conclude this section by observing that, at least in principle, we could have defined a “cognizant” bag semantics for DL-Lite\( _R^R \), which is based on arithmetic union rather than maximal union. Under such a semantics, inclusions \( A \sqsubseteq C \) and \( \emptyset \sqsubseteq C \) would be satisfied by a bag interpretation \( I \) only if \( A^I \sqsubseteq \emptyset \sqsubseteq C^I \); this is in contrast to our current “incognizant” semantics where we require that \( A^I \sqcup \emptyset \sqsubseteq C^I \). Such a semantics would, however, come with clear disadvantages. For example, every model \( I \) of the ontology with TBox \( \{ A \sqsubseteq B, B \sqsubseteq A, C \sqsubseteq B \} \) and ABox \( \| A(a) : 1. B(a) : 1. C(a) : 1 \| \) would need to satisfy

\[
A^I \sqsubseteq B^I, \quad B^I \sqsubseteq A^I, \quad C^I \sqsubseteq B^I, \quad A^I \sqcup C^I \sqsubseteq B^I,
\]

and, hence, would have to include infinitely many occurrences of element \( a^I \) in both \( A^I \) and \( B^I \).

11.2. Count aggregate queries over ontologies

Our approach to OBDA query answering under bag semantics is closely related to existing approaches for answering conjunctive counting aggregate queries over DL-Lite\( _R \) [36,38]. Indeed, under bag semantics, CQs are intrinsically equipped with counting power: the result of evaluating a CQ over a (set or bag) interpretation under bag semantics is a bag of answer tuples, where each tuple comes with a multiplicity (i.e., a “count”).

Calvanese et al. [38] proposed an epistemic semantics, where query answers are obtained by evaluating the query over the (finite) set of all ABox facts entailed by the ontology. Although this approach is well-suited for practical implementations, it can easily lead to counter-intuitive answers. To remedy this, Kostylev and Reutter [36] proposed a certain answers semantics that requires the query to be evaluated over all models of the ontology, which yields more intuitive answers at the expense of increased computational cost.

In what follows we take a closer look at Kostylev and Reutter’s framework, which is formalised in the following definition. The main difference between their setting and ours is that they consider set ABoxes and conventional set-based semantics of ontologies.

**Definition 84.** A count aggregate query is the expression \( q_c(x, \text{Count}(i)) = \exists y. \phi(x, y) \), where \( \phi(x, y) \) is a conjunction of atoms over atomic concepts and roles. A (set) interpretation \( I \) satisfies \( q_c(a, m) \), for a tuple \( a \) over \( I \) with \( |a| = |x| \) and a number \( m \in \mathbb{N}_0^\infty \), if there are exactly \( m \) valuations \( \lambda : x \cup y \cup I \rightarrow \Delta^I \) with \( \lambda(x) = a^I \) and \( \lambda(a) = \emptyset^I \) for each \( a \in I \) that make \( \phi(x, y) \) true in \( I \). A number \( n \in \mathbb{N}_0^\infty \) is in the count aggregate certain answers \( \text{Cert}(q_c, a, K) \) to \( q_c \) for a tuple of individuals \( a \) and DL-Lite\( _R \) ontology \( K \) if \( n \leq \min \{ m \in \mathbb{N}_0^\infty \mid \text{there is } I \models K \text{ satisfying } q_c(a, m) \} \). Count aggregate certain answers \( \text{Cert}^{\text{UNA}}(q_c, a, K) \) under the UNA are defined in the same way except that only interpretations satisfying the UNA are considered.
Intuitively, the count associated to \( a \) in the count aggregate certain answers to \( q_c \) is the minimum number of matching valuations over all models of the ontology. The following proposition establishes a correspondence between our setting and theirs: under suitable restrictions on the TBox, assuming only set ABoxes, and adopting the UNA, the certain answers to CQs under our bag semantics coincide with the answers to the corresponding count query as given in Definition 84.

**Proposition 85.** For every DL-Lite\(_R\) ontology \( K^b = \langle T, \mathcal{A}^s \rangle \) such that \( T \) does not contain any inclusions of the form \( \exists R \subseteq C \), every count aggregate query \( q_c(x, \text{Count}(i)) = \exists y. \phi(x, y) \), every tuple of individuals \( a \), and every \( n \in \mathbb{N}_0^\infty \).

\[
n \in \text{Cert}_{UNA}(q_c, a, K^b) \text{ if and only if } n \leq q^{K^b,UNA}(a),
\]

where \( q \) is the CQ \( \exists y. \phi(x, y) \), \( K^b \) is the DL-Lite\(_R\) ontology obtained from \( K^c \) by considering \( \mathcal{A}^b \) as a bag ABox (i.e., as the bag ABox assigning 1 to all assertions in \( \mathcal{A}^a \) and 0 to all others), and \( q^{K^b,UNA} \) is the certain answers to \( q \) over \( K^b \) under the UNA.

**Proof.** First, we claim that for every set interpretation \( I^b \) that is a model of \( K^c \) there exists a bag interpretation \( I^b \) that is a model of \( K^b \) such that, for every tuple of individuals \( a \), \( q^{b}(a) = m \) where \( m \) is the number with \( I^b \) satisfying \( q_c(a, m) \). Indeed, given such an \( I^b \) we can take as \( I^b \) the bag version of \( I^b \)—that is, the bag interpretation such that \( S^{-b}(b) = 1 \) for an atomic concept or role \( S \) and individuals \( b \) if \( b \in S^{-b} \) and \( S^{-b}(b) = 0 \) otherwise. Bag interpretation \( I^b \) is a model of \( K^c \) because the restriction on the TBox rules out any forced increase of multiplicities, while \( q^{b}(a) = m \) holds for every \( a \) because each valuation contributes to \( q^{b}(a) \) with multiplicity 0 and 1, and the ones with 1 are precisely those that turn \( \phi \) to true.

Second, we claim that for every bag interpretation \( I^b \) that is a model of \( K^b \) there exists a set interpretation \( I^b \) that is a model of \( K^c \) such that \( I^b \) satisfies \( q_c(a, m) \) for every tuple of individuals \( a \) and for a number \( m \leq q^{b}(a) \). Indeed, given such an \( I^b \) we can take as \( I^b \) the “characteristic” version of \( I^b \)—that is, the set interpretation such that \( b \in S^{I^b} \) for an atomic concept or role \( S \) and individuals \( b \) if and only if \( S^{b}(b) \geq 1 \). Set interpretation \( I^b \) is a model of \( K^c \) just by definition, while \( I^b \) additionally satisfies \( q_c(a, m) \) for a tuple \( a \) and a number \( m \leq q^{b}(a) \) because, by construction, a valuation contributes to \( q^{b}(a) \) a multiplicity greater than 0 if and only if it turns \( \phi \) to true.

These two claims immediately imply the statement of the proposition. \( \square \)

Note, however, that both the UNA and the restriction on TBoxes are necessary for Proposition 85 to hold, and dropping any of these makes the two frameworks incomparable. Indeed, if the UNA is not adopted, then, for the simple ontology \( \langle \emptyset, \{A(a), A(b)\} \rangle \), the aggregate query counting \( A(y) \) has certain answer 1, while the corresponding CQ has the empty tuple with multiplicity 2 in the answer. Similarly, for the ontology \( \langle \{B \subseteq \exists P, \exists P^\prime \subseteq A\}, \{B(a), B(b)\} \rangle \), we have the same situation if we drop the restriction on the TBox.

We conclude by pointing out that the work by Kostylev and Reutter is also strongly related to existing approaches in the database literature for answering counting aggregate queries in the presence of incomplete information [39–41]. As observed by Kostylev and Reutter, however, these approaches are not directly applicable to answering counting queries in the presence of an ontology, and we refer to [36] for a detailed discussion.

### 11.3. Other related work

Jiang [42] proposed a bag semantics for the description logic \( ALC \). The author focuses on satisfiability checking and provides a tableaux-based decision procedure. Their semantics is, however, incomparable to ours. For example, concepts of the form \( \exists R.T \) are not interpreted as the bag projection on the first argument of role \( R \), which makes the semantics incompatible with SQL.

Query answering and optimisation under bag semantics have received significant attention in the database literature [22,23,27,29–31,33,43]. These works study the relative expressive power of bag algebra primitives, the relationship with set-based algebras, and establish the data complexity of query answering, query containment, and query equivalence. More recently, Console et al. [44] studied query answering under bag semantics in incomplete relational databases. Query answering and its data complexity under bag semantics have been recently studied as well in the setting of the Semantic Web query languages [45,46].

### 12. Conclusion and future work

In this article, we have proposed a novel bag semantics for OBDA and studied the computational properties of its associated query answering problems. The key advantage of our semantics is that it allows us to faithfully represent arbitrary GAV mappings (and not just those whose source query involves duplicate elimination) in a way that is compatible with SQL. Furthermore, our semantics is compatible with existing bag-based semantics for databases with incomplete information and data exchange.

We see many interesting directions for future work. First, we are planning to extend the query language to allow for database-style aggregate functions and to study suitable restrictions ensuring re-writability of such queries. Second, it would
be interesting to try to push the rewritability boundaries for CQs to include some constant-free Boolean queries. For this, an interesting starting point could be the notion of local concepts and queries proposed by Gutiérrez-Basulto et al. [47] as a means of regaining decidability of query answering over DL-Lite\(g\) for the class of CQs with inequalities. Finally, our rewriting algorithms are not designed with an efficient implementation in mind. We plan to develop a practically applicable rewriting algorithm for our bag semantics.

Acknowledgements

This work was supported by the Royal Society under a University Research Fellowship, the EPSRC projects ED3 (EP/N014359/1) and DBOnto (EP/L012138/1), and the Research Council of Norway via the Sirius SFI (237889).

Appendix A. BALG\(^1\): algebraic query language for bag databases

In this appendix we introduce the algebraic query language for bag databases of Grumbach and Milo [22], called BALG\(^1\). Queries in BALG\(^1\) are algebra expressions built upon predicates in a database schema by the composition of the operations defined in Section 2.3 for bags, as well as the operations of tupling \(\tau\), attribute projection \(\alpha\), bagging \(\beta\), Cartesian product \(\times\), selection \(\sigma\), and projection \(\pi\).\(^4\) The semantics of these algebra expressions is defined formally below, where the operations are grouped into those that operate on tuples over the database domain \(I\) and those that operate on bags of tuples.

1. Operations on tuples:
   - tupling \(\tau(a_{1},...,a_{k})\) for \(a_{1},...,a_{k} \in I\) is the \(k\)-ary tuple \((a_{1},...,a_{k})\);
   - attribute projection \(\alpha_i(a)\) for a tuple \(a\) over \(I\) of size \(k\) and \(i \in [1,k]\) is the \(i\)-th component of \(a\);
   - bagging \(\beta(a)\) for a tuple \(a\) over \(I\) is the bag of tuples over \(I\) of size \(|a|\) assigning 1 to \(a\) and 0 to other tuples.

2. Operations on bags of tuples:
   - intersection \(\cap\), maximal union \(\cup\), arithmetic union \(\omega\), difference \(\setminus\), and duplicate elimination \(\varepsilon\) on bags of tuples over \(I\) of the same size are defined as in Section 2.3;
   - Cartesian product \(\Omega_{1} \times \Omega_{2}\) of bags \(\Omega_{1}\) and \(\Omega_{2}\) of tuples of size \(k\) and \(\ell\), respectively, over \(I\) is the bag of tuples of size \(k+\ell\) assigning \(\Omega_{1}(a) \times \Omega_{2}(b)\) to the concatenation \(a \times b\) of each two tuples \(a\) and \(b\) of size \(k\) and \(\ell\), respectively;
   - selection \(\sigma_{E_{1}(X_{1})=E_{2}(X_{2})}(\Omega)\) of a bag \(\Omega\) of tuples over \(I\) of size \(k\) for BALG\(^1\) algebra expressions \(E_{1}\) and \(E_{2}\) with a tuple variable \(X\) is the bag of tuples over \(I\) of size \(k\) that assigns \(E_{1}(a) = E_{2}(a)\) to each \(a\) with \(E_{1}(a) = E_{2}(a)\) and 0 to all other tuples;
   - projection \(\pi_{i_{1},...,i_{n}}(\Omega)\) of a bag \(\Omega\) of tuples over \(I\) of size \(k\) for \(i_{1},...,i_{n} \in [1,k]\) and \(n \geq 1\) is the bag of tuples of size \(n\) over \(I\) that assigns to each tuple \(a\) of size \(n\) the sum of \(\Omega(a)\) over all \(a\) that have \(a\) as components \(i_{1},...,i_{n}\).

The bag answers \(E^{D}\) to a BALG\(^1\) algebra expression \(E\) (operating on bags) over a database instance \(D\) over the same schema is the bag of tuples resulting in the evaluation of \(E\) on \(D\).

We now define the decision problem corresponding to computing the bag answers to BALG\(^1\) algebra expressions as it was introduced in [22], where we again assume that all numbers in the input are represented in unary and the bag database instance is explicitly defined only for a finite number of facts.

\[
\text{QUERY}\text{Answering}^{=}[\text{BALG}\text{\(^1\)}]
\]

\begin{align*}
\text{Input:} & \quad \text{BALG}\text{\(^1\)} \text{ algebra expression } E, \text{ bag database instance } D, \\
\text{tuple } a \text{ of constants over } I, \text{ and number } k \in \mathbb{N}_{0}.
\end{align*}

\begin{align*}
\text{Question:} & \quad E^{D}(a) = k?
\end{align*}

The data complexity of this problem is the complexity when the expression \(E\) is considered to be fixed and only \(D, a,\) and \(k\) form the input.

We stress here that in this problem the question is whether the equality \(\Phi^{D}(a) = k\) holds. However, we are more interested in the threshold version QUERY\text{Answering}^{=}[\text{BALG}\text{\(^1\)}] where \(\Phi^{D}(a) \geq k\) is checked instead.

The following proposition shows that, in data complexity, our problem is not any harder than the one of Grumbach and Milo.

**Proposition 86.** QUERY\text{Answering}^{=}[\text{BALG}\text{\(^1\)}] is \(\text{AC}^{0}\) reducible to QUERY\text{Answering}^{=}[\text{BALG}\text{\(^1\)}] in data complexity.

\(^4\) Note that the language of Grumbach and Milo [22] includes a restructuring operation instead of projection. Our convenient deviation does not change the expressivity of the language since these two operations are expressible via each other in the presence of the other operations [22].
Proof. Let $E$ be a fixed BALG$^1$ algebra expression, $D$ be a bag database instance, $a$ be a tuple of constants over $I$, and $k$ be a number in $\mathbb{N}_0$. Consider the bag database instance $D_0$ over $I$ and schema extended with a fresh predicate $T$ of arity $|a|$ in which $T$ assigns $k$ to $a$ and 0 to all other tuples of size $|a|$, while every other predicate is as in $D$.

Let $E_D = T - E$. We claim that $E^D(a) \geq k$ if and only if $E_D^{D_0}(a) = 0$. Indeed, assuming $E^D(a) \geq k$, we derive $E_D^{D_0}(a) = 0$. Similarly, assuming $E^D(a) < k$, we derive $E_D^{D_0}(a) > 0$.

The above many-one reduction is computable, for each $D$, $a$, and $k$, by a Boolean circuit with arbitrary fan-in AND and OR gates whose depth depends only on $E$. We conclude that language $\{(D, a, k) \mid E^D(a) \geq k\}$ is contained in $\{(D, a, k) \mid E^D(a) = k\}$ under LocSpace-uniform AC$^0$ reductions, as required. \hfill $\Box$

Grumbach and Milo [22] studied the expressive power of QueryAnswering$^+$[BALG$^1$] and showed that it is strictly in between AC$^0$ and LocSpace. Therefore, we can conclude the following fact.

Corollary 87. QueryAnswering$^+$[BALG$^1$] is in LocSpace in data complexity.

Appendix B. Relationship between BCALC and BALG$^1$

Theorem 88. For each BCALC query $\Phi$ there is a BALG$^1$ algebra expression $E_\Phi$ such that $E^{D_\Phi} = E_D^{D_0}$ for every bag database instance $D$.

Proof. In this proof we assume that all the variables $X$ are globally ordered, and each tuple of repetition-free variables has its variables according to this order; moreover, for a BCALC query $\exists y. \Phi(x, y)$ we assume that all $y$ are after all $x$ in the order, which is done without loss of generality, because BCALC is agnostic to renaming of variables.

We define BALG$^1$ algebra expression $E_\Phi$ for each BCALC query $\Phi$ by induction on the structure of $\Phi$ as follows:

- if $\Phi(x_1, \ldots, x_n) = S(t_1, \ldots, t_k)$ for a predicate $S$ and a tuple of terms $t_1, \ldots, t_k$ with the first occurrence of each $x_i$ in a position $p_i$, then, for $c_1, \ldots, c_k \in I$ such that $c_j$ is fresh if $t_j$ is a variable and $c_j$ is $t_j$ otherwise for each $j$.

\[
E_\Phi = \prod_{i=1}^k \pi_{p_i} (\sigma_{a_{\ell}}(x_i) = a_{m_{\ell}}(x_i) (S \times \beta(\tau(c_1, \ldots, c_k))) \ldots)),
\]

where, for each $j$, $m_j = p_i$ if $t_j$ is $x_i$ and $m_j = j + k$ otherwise;

- if $\Phi(x) = \Psi(x) \land (x_i = x_2)$ with $x_1$ and $x_2$ in positions $i$ and $j$ in $x$, then $E_\Phi = \sigma_{a_i}(x_i) = a_j(x_j)(E_\Psi)$;

- if $\Phi(x_1, \ldots, x_k) = \Psi(x_1, x_j, \ldots, x_k) \land (x_i = x_j)$ for $i$, $j \in [1, k]$, $i \neq j$, then $E_\Phi = \sigma_{a_1}(x_1) = a_{j+1}(x_j)(E_\Psi)$;

- if $\Phi(x_1, \ldots, x_k) = \Psi(x_1, x_j, x_k) \land (x_i = a)$ for $i \in [1, k]$, then $E_\Phi = \prod_{i=1}^k (\sigma_{a_i}(x_j) = a_{j+1}(x_i)(E_{\Psi(x_1)} \times \beta(\tau(a))))$;

- if $\Phi(x_1, \ldots, x_k) = \Psi_1(x_1, \ldots, x_m) \land \Psi_2(x_j, \ldots, x_m)$ with $[i_1, \ldots, i_m, j_1, \ldots, j_n] = [1, k]$, then

\[
E_\Phi = \prod_{i=1}^k \pi_{p_i} \left(\sigma_{a_{m_{\ell}}}(x_i) = a_{m_{\ell}}(x_i) \left(\Psi_1(x_1) \times \Psi_2(x_j, \ldots, x_m)\right)\right),
\]

where, for each $\ell \in [1, k]$, $p_\ell$ is the position of the first occurrence of $\ell$ in $i_1, \ldots, i_m, j_1, \ldots, j_n$ and, for each $\ell \in [1, n]$, $s_\ell$ is the position of the first occurrence of $j_\ell$ in $i_1, \ldots, i_m, j_1, \ldots, j_n$;

- if $\Phi(x) = \exists y. \Psi(x, y)$, then $E_\Phi = \prod_{i=1}^k (E_\Psi)$;

- if $\Phi(x) = \Psi_1(x) \lor \Psi_2(x)$, then $E_\Phi = E_{\Psi_1} \lor E_{\Psi_2}$;

- if $\Phi(x) = \Psi_1(x) \lor \Psi_2(x)$, then $E_\Phi = E_{\Psi_1} \lor E_{\Psi_2}$;

- if $\Phi(x) = \delta \Psi(x)$, then $E_\Phi = E(\delta \Psi)$.

It is now straightforward to check that $E^{D_\Phi} = E_D^{D_0}$ for every bag database $D$. \hfill $\Box$

The following fact is a direct consequence of Corollary 87 and Proposition 88.

Proposition 88. QueryAnswering$^+$[BCALC] is in LocSpace in data complexity.

References

[1] C. Nikolaou, E.V. Kostylev, G. Konstantinidis, M. Kaminski, B. Cuenca Grau, I. Horrocks, The bag semantics of ontology-based data access, in: Proceedings of the 26th International Joint Conference on Artificial Intelligence, 2017, pp. 1224–1230.

[2] A. Poggi, D. Lembo, D. Calvanese, G. De Giacomo, M. Lenzerini, R. Rosati, Linking data to ontologies, J. Data Semant. 10 (2008) 133–173.

[3] D. Calvanese, G. De Giacomo, D. Lembo, M. Lenzerini, A. Poggi, M. Rodriguez-Muro, R. Rosati, M. Ruzzi, D.F. Savo, The MASTRO system for ontology-based data access, Semant. Web 2 (1) (2011) 43–53.

[4] D. Calvanese, G. De Giacomo, D. Lembo, M. Lenzerini, A. Poggi, M. Rodriguez-Muro, R. Rosati, Ontologies and databases: the DL-Lite approach, in: Proceedings of 5th International Summer School on Reasoning Web: Semantic Technologies for Information Systems, Tutorial Lectures, 2009, pp. 255–356.

[5] D. Calvanese, B. Cogrel, S. Komla-Ebri, R. Kontchakov, D. Lanti, M. Rezk, M. Rodriguez-Muro, G. Xiao, Ontop: answering SPARQL queries over relational databases, Semant. Web 8 (3) (2017) 471–487.
