Assessing the Exposure to Public Knowledge in Policy-Protected Description Logic Ontologies

Gianluca Cima[1] , **Domenico Lembo**[1] , **Lorenzo Marconi**[1] , **Riccardo Rosati**[1] and **Domenico Fabio Savo**[2]
[1]Sapienza University of Rome
[2]University of Bergamo
{lastname}@diag.uniroma1.it, domenicofabio.savo@unibg.it

Abstract

We propose a general framework for assessing the exposure of sensitive knowledge in policy-protected knowledge bases (KBs), where knowledge is represented as logical theories and data protection policies are defined declaratively using epistemic dependencies. The framework models scenarios in which confidential parts of the KB may be publicly known due to security breaches. We study two fundamental decision problems: determining whether the exposed knowledge violates the data protection policy (leakage), and whether there exists a secure view of the KB that complies with the policy. We analyze the computational complexity (specifically, data complexity) of these problems, focusing on the DL-Lite$_\mathcal{R}$ and $\mathcal{EL}_\bot$ Description Logics. Our findings show that, for DL-Lite$_\mathcal{R}$ with restricted forms of policy, both the problems can be efficiently solved through query rewriting methods. For $\mathcal{EL}_\bot$, we establish conditions for tractable computational bounds. Our results highlight the potential of this framework for practical applications in confidentiality-preserving knowledge management.

1 Introduction

Protecting sensitive knowledge in information systems is a critical and challenging task, often complicated by the existence of information that a potential adversary may already possess or acquire from external sources, also referred to as background knowledge (e.g. in [Biskup and Bonatti, 2004; Bonatti and Sauro, 2013]). Such knowledge can be exploited to compromise individual privacy or infer sensitive data from otherwise protected datasets. In this paper, we focus on scenarios where background knowledge consists of publicly available information, such as that obtained from social networks, websites, or public records; in addition, public knowledge may originate from the system itself, due to security breaches, resulting in previously private information being made public. This may compromise the protection of sensitive data, thus constituting an information leakage.

To formally study this problem, we propose a general framework in which a (first-order logic) knowledge base (KB) $\mathcal{K}$ is coupled with a policy $\mathcal{P}$ that protects confidential information in $\mathcal{K}$. Part of $\mathcal{K}$, denoted $\mathcal{K}_{pub}$, is publicly known, together with $\mathcal{P}$. The policy is given in terms of epistemic dependencies (EDs), which are formulas of the form $\forall\vec{x}\,(K\phi_b(\vec{x}) \rightarrow K\phi_h(\vec{y}))$, where $\phi_b(\vec{x})$ and $\phi_h(\vec{y})$ are queries (i.e. open formulas) over $\mathcal{K}$, $\vec{y} \subseteq \vec{x}$, and K is a modal operator. Intuitively, an ED stipulates that if the answers to the query $\phi_b(\vec{x})$ are known, then also the answers to the query $\phi_h(\vec{y})$ must be known (note that if $\phi_h(\vec{y})$ is a contradiction, the answers to $\phi_b(\vec{x})$ cannot be disclosed at all). EDs were introduced in [Console and Lenzerini, 2020] and employed as constraints within an Ontology-based Data Management setting. Recently, they have been adopted in [Cima *et al.*, 2024a] in the context of Controlled Query Evaluation (CQE), where they serve the same purpose as in our framework. Notably, using EDs in policy design enables highly expressive forms of protection rules that go beyond the policies commonly considered in many studies in privacy-preserving query answering [Bonatti and Sauro, 2013; Cuenca Grau *et al.*, 2015; Cima *et al.*, 2024b].

The example below illustrates our framework and highlights the confidentiality issues that may arise in this context.

Example 1. *Consider a KB of patients' genetic data. It is public knowledge that the* $\mathsf{\Delta F508}$ *variant of the* CFTR *gene may indicate the presence of cystic fibrosis and is thus classified as pathogenic. In contrast, the* $\mathsf{R577X}$ *variant of the* $\mathsf{ACTN3}$ *gene, which implies improved muscle endurance in individuals, is generally not considered pathogenic. We denote the former variant with* $\mathsf{v_1}$ *and the latter with* $\mathsf{v_2}$*. The KB associates patients to the gene variants they possess (*$\mathsf{hasVariant}$ *predicate). Patients may consent to disclose these associations (*$\mathsf{consentToShare}$ *predicate). Pathogenic variants are classified as sensitive (*$\forall x\,(\mathsf{pathogenic}(x) \rightarrow \mathsf{sensitive}(x))$ *axiom). The policy stipulates that patient-variant pairs may be disclosed only with patient consent, but pairs involving sensitive variants can never be disclosed.*

The KB $\mathcal{K}$*, the policy* $\mathcal{P}$ *and the public portion* $\mathcal{K}_{pub}$ *of* $\mathcal{K}$ *are as follows (where* $\bot$ *denotes a contradiction).*

$$\mathcal{P} = \{\forall p, v\,(K\,\mathsf{hasVariant}(p, v) \rightarrow K\,\mathsf{consentToShare}(p, v)),$$
$$\forall p\,(K\,\exists v\,(\mathsf{hasVariant}(p, v) \wedge \mathsf{sensitive}(v)) \rightarrow K\bot)\},$$
$$\mathcal{K}_{pub} = \{\forall x\,(\mathsf{pathogenic}(x) \rightarrow \mathsf{sensitive}(x)), \mathsf{pathogenic}(\mathsf{v_1})\},$$
$$\mathcal{K} = \mathcal{K}_{pub} \cup \{\mathsf{hasVariant}(\mathsf{ann}, \mathsf{v_1}), \mathsf{consentToShare}(\mathsf{ann}, \mathsf{v_1}),$$
$$\mathsf{hasVariant}(\mathsf{sam}, \mathsf{v_2}), \mathsf{consentToShare}(\mathsf{sam}, \mathsf{v_2}),$$
$$\mathsf{hasVariant}(\mathsf{bob}, \mathsf{v_2})\}$$

It is easy to see that $\mathcal{K}_{pub}$ is compliant with the policy. $\mathcal{K}_{pub}$ might even contain the fact $\mathsf{hasVariant}(\mathsf{sam}, \mathsf{v}_2)$*, because* sam *provided his consent for the publication of this data, and* v_2 *is not classified as sensitive. In contrast,* $\mathsf{hasVariant}(\mathsf{bob}, \mathsf{v}_2)$ *cannot be in $\mathcal{K}_{pub}$, since* bob *did not provide his consent. Finally, the fact* $\mathsf{hasVariant}(\mathsf{ann}, \mathsf{v}_1)$ *cannot be added to $\mathcal{K}_{pub}$, even though* ann *gave her consent, because together with the other formulas in $\mathcal{K}_{pub}$, this addition causes a violation of the second rule in $\mathcal{P}$.* ∎

Let us now introduce the two fundamental decision problems that we study within the framework:

- verifying whether there is a *leakage*, that is, if the public knowledge implies some confidential information protected by the policy;
- verifying whether there exists a *secure view*, i.e. a set of sentences inferred by the KB, expressed in a language $\mathcal{L}$, that complies with the policy and the public knowledge.

We adopt a very general approach, according to which the only action that the system can do to protect itself is to provide the user with a view of the KB that complies with the policy. Thus a leakage occurs only if such a secure view does not exist (in this case, $\mathcal{K}_{pub}$ implies some confidential data). The second problem may appear to be the complement of the first, but it has a distinguishing characteristic: it requires views to be expressed in a given language $\mathcal{L}$. Consequently, solving the first problem does not guarantee a solution to the second. The language is a crucial parameter for the practical usage of the notion of secure view, and has an impact on the computational complexity of the problem, as we will show.

We study the data complexity [Vardi, 1982] of both problems, focusing on certain instantiations of the framework. As for the policy, we consider EDs in which both $\phi_b(\vec{x})$ and $\phi_h(\vec{y})$ are *conjunctive queries* (CQs), considering also the special cases in which $\phi_h(\vec{y})$ does not contain existentially quantified variables (full CQ-EDs), or the EDs respect an acyclicity condition (acyclic CQ-EDs), or both the previous conditions hold (acyclic full CQ-EDs). As for the KB, we concentrate on two fragments of first-order (FO) logic that are commonly used in contexts involving the management of large amounts of data through ontologies: DL-Lite$_\mathcal{R}$ [Calvanese *et al.*, 2007b] and $\mathcal{EL}_\bot$ [Baader *et al.*, 1999]. These are the logical counterparts of the OWL 2 profiles OWL 2 QL and OWL 2 EL, respectively [Motik *et al.*, 2009]. Finally, regarding the view language $\mathcal{L}$, we analyze the case where $\mathcal{L}$ is either the language of Boolean CQs or the language of ground atoms (GA), i.e. when the view is simply a set of facts. Both languages are particularly relevant for the practical usage of the notion of secure view.

All our complexity results are summarized in Table 1. The table also includes the case where both $\mathcal{K}$ and $\mathcal{K}_{pub}$ consist of ABoxes (i.e. sets of facts). This scenario corresponds to KBs without an intensional component (the so-called TBox), as in plain databases. We emphasize that our complexity results encompass all possible combinations of the languages mentioned above for the various components of the framework. Notably, we identified many tractable cases, several of which are even in AC^0. This latter result indicates that in these cases, the problem can be efficiently solved by evaluating an FO query over an ABox, a task achievable using standard relational database technology. These findings highlight the potential of our framework for practical applications in confidentiality-preserving knowledge management.

Due to space limitations, complete proofs are deferred to the supplemental material accompanying this submission.

2 Related Work

Our work is inspired by research on controlled query evaluation (CQE), a declarative framework for privacy-preserving query answering, where confidential data is protected by a policy, as is the case in our study. Unlike work on CQE, we do not focus here on query answering but instead address the leakage and secure view existence problems in the presence of public knowledge. Public knowledge is often not explicitly considered in CQE research on ontologies (e.g. in [Cuenca Grau *et al.*, 2013; Cuenca Grau *et al.*, 2015; Cima *et al.*, 2024b; Cima *et al.*, 2024a]), although a common assumption is that the TBox of the KB is known to the user, whereas the ABox is treated as private. A distinguishing feature of our approach is that public knowledge may also include extensional knowledge, and the private portion of $\mathcal{K}$ may contain parts of the TBox. Two previous papers on CQE over ontologies provide similiar abilities, i.e. [Bonatti and Sauro, 2013; Studer and Werner, 2014]. The settings considered in these papers are incomparable to ours. In particular, their frameworks cannot replicate the expressiveness and flexibility given by the EDs we use for the policy. Other works on CQE have addressed attacks in the presence of users' background knowledge [Biskup and Bonatti, 2001; Biskup and Bonatti, 2004; Biskup and Weibert, 2008], but they consider the context of propositional databases.

Privacy issues in $\mathcal{EL}$ ontology publishing are studied in [Baader and Nuradiansyah, 2019], where the knowledge about individuals to be published is an $\mathcal{EL}$ instance store, i.e. an ABox instantiating $\mathcal{EL}$ concepts only, without a TBox. Both the privacy policy and the potential background knowledge of an attacker are expressed as $\mathcal{EL}$ concepts. In [Baader *et al.*, 2019], the background knowledge is represented by concepts in more expressive DLs. The focus of these papers is on verifying compliance and safety of the attackers' background concepts with respect to the policy, problems that differ from those examined here.

Various studies in data and knowledge bases adopt the idea of protecting confidential information through views for users' access control [Calvanese *et al.*, 2012; Stoffel and Studer, 2005; Chirkova and Yu, 2017], an approach that can be somehow considered as complementary to ours. In these papers, the views and their extensions are given, and there is no explicit protection policy. Forms of leakage are also studied in [Benedikt *et al.*, 2018], in the context of ontology-based data integration, where the focus is on establishing whether the system discloses a source query, considered as secret. The same problem is studied in [Benedikt *et al.*, 2019], where an attacker may exploit additional knowledge on source constraints. Note that these studies do not address the case of attackers with extensional public knowledge.

We finally remark that our approach is deterministic, i.e. we study the assessment of exact leakages. Examples of probabilistic approaches to privacy-preserving information management are instead [Miklau and Suciu, 2007; Sweeney, 2002; Dwork, 2011].

3 Preliminaries

We adopt standard notions of function-free first-order (FO) logic and consider pairwise disjoint sets of *predicate* names Σ_p, *constant* (a.k.a. *individual*) names Σ_i and *variable* names Σ_v. We define Σ as $\Sigma_p \cup \Sigma_i \cup \Sigma_v$. Throughout the paper, an FO formula ϕ is sometimes denoted as $\phi(\vec{x})$, where $\vec{x}$ represents the tuple of free variables in ϕ. Closed FO formulas (namely, the ones without free variables) are also called *sentences*, while variable-free formulas are said to be *ground*. In particular, ground atoms are called *facts*. A *knowledge base (KB)* Φ is defined as a set of FO sentences. Its semantics is provided in terms of FO interpretations over the signature $\Sigma_p \cup \Sigma_i$. We focus on interpretations that share a common infinite countable domain $\Delta = \Sigma_i$, where each element of Σ_i is mapped to itself. In other words, we assume the use of *standard names*, a common practice when working with epistemic operators [Calvanese *et al.*, 2007a].

We use $\mathit{eval}(\phi, \mathcal{I})$ to denote the outcome (i.e. true or false) of the evaluation of an FO sentence ϕ over an FO interpretation $\mathcal{I}$. An FO interpretation $\mathcal{I}$ is a *model* of a KB Φ if every sentence in Φ evaluates to true over $\mathcal{I}$. A KB Φ *entails* an FO sentence ϕ, denoted by $\Phi \models \phi$, if $\mathit{eval}(\phi, \mathcal{I})$ is true, for every model $\mathcal{I}$ of Φ. Given a KB Φ', we say that Φ *entails* Φ', denoted by $\Phi \models \Phi'$, if Φ entails every sentence in Φ'.

In this work, we also consider the language of conjunctive queries and their variants. A *conjunctive query* (CQ) is an FO formula of the form $\exists \vec{y}\, \phi(\vec{x}, \vec{y})$, where $\vec{x} \cup \vec{y} \subseteq \Sigma_v$, and $\phi(\vec{x}, \vec{y})$ is a finite, non-empty conjunction of atoms $p(\vec{t})$ (with $p \in \Sigma_p$ and $\vec{t} \subseteq \Sigma_i \cup \vec{x} \cup \vec{y}$). We also include the special CQ $\bot$, assuming that $\mathit{eval}(\bot, \mathcal{I})$ is false for any FO interpretation $\mathcal{I}$. A *union of conjunctive queries* (*UCQ*) is a disjunction of CQs sharing the same free variables. For convenience, we sometimes treat UCQs as sets of CQs. We call *Boolean CQs* (*BCQs*) the CQs without free variables.

Given any language $\mathcal{L}$ of formulas, we denote by $\mathcal{L}_b$ the set of sentences belonging to $\mathcal{L}$. Throughout the paper we refer to the languages $\mathbf{FO} \supseteq \mathbf{CQ} \supseteq \mathbf{CQ}_b \supseteq \mathbf{GA}$, which are defined as, respectively, the set of FO formulas over Σ, the set of CQs (possibly containing free variables) over Σ, the set of BCQs over Σ, and the set of ground atoms over Σ. Given a language $\mathcal{L} \subseteq \mathbf{FO}_b$ and a KB Φ, we denote by $Cons_{\mathcal{L}}(\Phi)$ the set of sentences $\{\phi \in \mathcal{L} \mid \Phi \models \phi\}$.

A *ground substitution* for a sequence $\vec{x} = x_1, \ldots, x_k$ of variables is a sequence of constants $\vec{c} = c_1, \ldots, c_k$. Furthermore, if $\vec{x}$ are the free variables of an FO formula $\phi(\vec{x})$, we indicate as $\phi(\vec{c})$ the FO sentence obtained from $\phi(\vec{x})$ by replacing each x_i with c_i, for $1 \leq i \leq k$.

In this paper, we will specifically focus on Description Logics (DLs), i.e. decidable fragments of FO [Baader *et al.*, 2007]. Typically, in such logics, the predicate set Σ_p is partitioned into two sets: Σ_c, containing unary predicates (called *concepts*), and Σ_r, containing binary predicates (called *roles*).

A *DL ontology* $\mathcal{K} = \mathcal{T} \cup \mathcal{A}$ comprises a TBox $\mathcal{T}$ and an ABox $\mathcal{A}$, which are finite sets of assertions that capture intensional and extensional knowledge, respectively. In this paper, ABoxes are defined as sets of ground atoms. As already said, we focus on the DLs DL-Lite$_{\mathcal{R}}$ [Calvanese *et al.*, 2007b] and $\mathcal{EL}_\bot$, which is an extension $\mathcal{EL}$ [Baader *et al.*, 1999] that allows for the use of the empty concept $\bot_c$.

A DL-Lite$_{\mathcal{R}}$ TBox $\mathcal{T}$ is a set of axioms, which can either be inclusions of the form $C \sqsubseteq C'$ and $R \sqsubseteq R'$ (concept and role inclusions) or disjointness assertions of the form $C \sqsubseteq \neg C'$ and $R \sqsubseteq \neg R'$ (concept and role disjointnesses). In these axioms, the concepts C and C' (respectively, the roles R and R') can take the form A, $\exists P$, or $\exists P^-$ (respectively, P or P^-), where $A \in \Sigma_c$, $P \in \Sigma_r$, and P^- denotes the inverse of the role P. Terms $\exists P$ and $\exists P^-$, called *unqualified existential restrictions*, represent the sets of objects appearing as the first and second argument of P, respectively.

An $\mathcal{EL}_\bot$ TBox $\mathcal{T}$ is a set of concept inclusions of the form $C \sqsubseteq C'$, where the C and C' can take the form A, $\exists P.C$, $C_1 \sqcap C_2$, $\bot_c$, and $\top_c$. Here, C and C' are called *general concepts*, which can be an atomic concept $A \in \Sigma_c$, a concept of the form $\exists P.C$, with $P \in \Sigma_r$, called *qualified existential restriction* and denoting the set of objects that the atomic role P relates to some instance of the general concept C, a concept of the form $C_1 \sqcap C_2$, i.e. a conjunction of two general concepts, $\bot_c$, i.e. the empty concept, or $\top_c$, i.e. the top concept.

For DL-Lite$_{\mathcal{R}}$ ontologies, we make use of the query rewriting algorithm *PerfectRef* [Calvanese *et al.*, 2007b], which enjoys the following property.

Proposition 1. *Let $\mathcal{T}$ be a DL-Lite$_{\mathcal{R}}$ TBox and let $q(\vec{x})$ be a CQ. For every ABox $\mathcal{A}$ and every ground substitution $\vec{c}$ for $\vec{x}$, we have that $\mathcal{T} \cup \mathcal{A} \models q(\vec{c})$ iff $\mathcal{A} \models q_r(\vec{c})$, where $q_r(\vec{x})$ is the UCQ returned by PerfectRef$(q(\vec{x}), \mathcal{T})$.*

All our complexity results pertain to *data complexity* [Vardi, 1982], which in our context is the complexity computed with respect to the size of the ABox.

4 Framework

We first formalize the policy $\mathcal{P}$ of our framework as a finite set of *epistemic dependencies*, each of which can be seen as an *EQL-Lite(FO)* sentence [Calvanese *et al.*, 2007a].

Definition 1 (Epistemic dependency, policy)**.** *Given a language $\mathcal{L} \subseteq \mathbf{FO}$, an* $\mathcal{L}$-epistemic dependency (ED) *is a sentence τ of the following form:*

$$\forall \vec{x}\, (K\phi_b(\vec{x}) \rightarrow K\phi_h(\vec{x_h})) \tag{1}$$

where $\vec{x_h} \subseteq \vec{x}$, $\phi_b(\vec{x})$ (called the body *of τ) is an $\mathcal{L}$-formula with free variables $\vec{x}$, $\phi_h(\vec{x_h})$ (called the* head *of τ) is an $\mathcal{L}$-formula with free variables $\vec{x_h}$, and K is a modal operator. An* $\mathcal{L}$-policy *is a finite set of $\mathcal{L}$-EDs.*

To simplify our terminology, we call *ED* an $\mathbf{FO}$-ED and call *policy* an $\mathbf{FO}$-policy. Also, we use the notation $\tau(\vec{x})$ to indicate an ED τ whose universally quantified variables are $\vec{x}$. In addition, given a ground substitution $\vec{c}$ for $\vec{x}$, we denote by $\tau(\vec{c})$ the ED obtained from τ by removing the universal quantification of the variables $\vec{x}$ and replacing all their occurrences with the corresponding constants in $\vec{c}$. Finally, given

an ED τ of the form (1), we denote by *body*(τ) and *head*(τ) the formulas $\phi_b(\vec{x})$ and $\phi_h(\vec{x_h})$, respectively.

Intuitively, an ED of form (1) should be read as follows: if the sentence $\phi_b(\vec{c})$ is *known* to hold, then the sentence $\phi_h(\vec{c_h})$ is *known* to hold, for every ground substitution $\vec{c}$ for $\vec{x}$, where $\vec{c_h}$ is the corresponding ground substitution of $\vec{x_h}$.

Formally, we define when a KB Φ *satisfies* an ED τ, denoted $\Phi \models_{\mathsf{EQL}} \tau$. To this aim, we consider the set E of all models of Φ, and say that $\Phi \models_{\mathsf{EQL}} \tau$ if, for every ground substitution $\vec{c}$ for $\vec{x}$, the fact that *eval*($\phi_b(\vec{c}), \mathcal{I}$) is true for every $\mathcal{I} \in E$ implies that *eval*($\phi_h(\vec{c_h}), \mathcal{I}$) is true for every $\mathcal{I} \in E$. We say that Φ *satisfies* a policy $\mathcal{P}$ (denoted $\Phi \models_{\mathsf{EQL}} \mathcal{P}$) if Φ satisfies τ, for each $\tau \in \mathcal{P}$. We remark that, as already said, EDs of the form (1) have been originally introduced in [Console and Lenzerini, 2020], although in a slightly differnt form.

Following [Calvanese *et al.*, 2007a], given an FO formula $\phi(\vec{x})$ and a KB Φ, we say that ϕ is Φ-*range restricted* [Calvanese *et al.*, 2007a] if the set of ground substitutions $\vec{c}$ for $\vec{x}$ such that $\Phi \models \phi(\vec{c})$ is finite. Given a policy $\mathcal{P}$ and a KB Φ, $\mathcal{P}$ is said Φ-*range restricted* if, for every $\tau \in \mathcal{P}$ of the form (1), the formulas $\phi_b(\vec{x})$ and $\phi_h(\vec{x_h})$ are Φ-range restricted.

Let us now introduce the new notion of KB that we adopt in the paper, in which the information to be protected is specified by the policy as a set of EDs, and the public knowledge is explicitly defined.

Definition 2 (Policy-protected KB with public knowledge)**.** *A* policy-protected KB with public knowledge *(PPKB for short) is a triple $\mathcal{E} = \langle \mathcal{K}, \mathcal{K}_{pub}, \mathcal{P} \rangle$, where $\mathcal{K}$ and $\mathcal{K}_{pub}$ are KBs such that $\mathcal{K}$ has at least one model, $\mathcal{K} \models \mathcal{K}_{pub}$, and $\mathcal{P}$ is a $\mathcal{K}$-range restricted policy.*

Notice that, in the above definition, we assume that the information in $\mathcal{K}_{pub}$ is either contained within or a consequence of the information in $\mathcal{K}$. In other words, we embrace the idea that $\mathcal{K}_{pub}$ represents a portion of the information in $\mathcal{K}$ that, in some way, has been publicly exposed. An example of PPKB is the triple $\langle \mathcal{K}, \mathcal{K}_{pub}, \mathcal{P} \rangle$ provided in Example 1.

Our aim is to determine whether the sensitive information in $\mathcal{K}$, i.e. the knowledge safeguarded by $\mathcal{P}$, remains secure when the knowledge in $\mathcal{K}_{pub}$ is exposed. To this end, we first provide the following notion of policy violation.

Definition 3 (Policy violation)**.** *Given a KB $\mathcal{K}$ and a policy $\mathcal{P}$, a* policy violation (PV) for $\langle \mathcal{K}, \mathcal{P} \rangle$ *is a KB Φ for which there exists no KB Φ' such that $\mathcal{K} \models \Phi'$ and $\Phi \cup \Phi' \models_{\mathsf{EQL}} \mathcal{P}$.*

Informally, a PV for $\langle \mathcal{K}, \mathcal{P} \rangle$ is a fragment of knowledge that conflicts with the policy $\mathcal{P}$ and cannot be reconciled by adding further knowledge from $\mathcal{K}$.

Example 2. *Consider Example 1. Both the KB $\Phi_1 = \{\mathsf{hasVariant}(\mathsf{ann}, \mathsf{v}_1), \mathsf{sensitive}(\mathsf{v}_1)\}$ and the KB $\Phi_2 = \{\mathsf{hasVariant}(\mathsf{bob}, \mathsf{v}_2)\}$ are PVs for $\langle \mathcal{K}, \mathcal{P} \rangle$.* ∎

When a PV follows from $\mathcal{K}_{pub}$, it reveals that certain data in $\mathcal{K}$ safeguarded by $\mathcal{P}$ has been exposed. In such a case, we say that the PV is a *leakage* in $\mathcal{E}$.

Definition 4 (Leakage)**.** *Given a PPKB $\mathcal{E} = \langle \mathcal{K}, \mathcal{K}_{pub}, \mathcal{P} \rangle$ and a PV Φ for $\langle \mathcal{K}, \mathcal{P} \rangle$, we say that Φ is a* leakage *in $\mathcal{E}$ if $\mathcal{K}_{pub} \models \Phi$.*

Even if there is no leakage, $\mathcal{K}_{pub}$ alone can still not satisfy the policy. However, this can be overcome by exposing additional knowledge from the system (see Φ' in Definition 3). In other words, it is still possible to expose a view of $\mathcal{K}$ that is secure with respect to $\mathcal{P}$. In practical applications, it is useful to parameterize the view notion to the language chosen to specify it. All of this is formalized in the following definition.

Definition 5 (Secure $\mathcal{L}$-view)**.** *Given a PPKB $\mathcal{E} = \langle \mathcal{K}, \mathcal{K}_{pub}, \mathcal{P} \rangle$ and a language $\mathcal{L} \subseteq \mathbf{FO}$, a* secure $\mathcal{L}$-view *of $\mathcal{E}$ is a KB $\Phi \subseteq \mathcal{L}$ such that $\mathcal{K} \models \Phi$ and $\mathcal{K}_{pub} \cup \Phi \models_{\mathsf{EQL}} \mathcal{P}$.*

For the sake of brevity, we may refer to a secure **FO**-view of $\mathcal{E}$ simply as a *secure view* of $\mathcal{E}$.

Example 3. *Recall Example 2. Note that Φ_1 is not a leakage in $\mathcal{E}$ as $\mathcal{K}_{pub} \models \Phi_1$ does not hold. Suppose now that, after a security breach, the data $\mathsf{hasVariant}(\mathsf{ann}, \mathsf{v}_1)$ is made public. We thus consider the new PPKB $\mathcal{E}' = \langle \mathcal{K}, \mathcal{K}'_{pub}, \mathcal{P} \rangle$, where $\mathcal{K}'_{pub} = \mathcal{K}_{pub} \cup \{\mathsf{hasVariant}(\mathsf{ann}, \mathsf{v}_1)\}$. We have that the PV Φ_1 of Example 2 is a leakage in $\mathcal{E}'$.*

Moreover, an example of secure **GA***-view of $\mathcal{E}$ is the set $\{\mathsf{hasVariant}(\mathsf{sam}, \mathsf{v}_2), \mathsf{consentToShare}(\mathsf{sam}, \mathsf{v}_2)\}$, while $\mathcal{E}'$ has no secure view.* ∎

We close this section by outlining the two decision problems that will be the focus of the remainder of the paper.

Leakage-existence (LE) problem: given a PPKB $\mathcal{E}$, deciding whether there exists a leakage in $\mathcal{E}$.

$\mathcal{L}$-view-existence ($\mathcal{L}$-VE) problem: given a PPKB $\mathcal{E}$, deciding whether there exists a secure $\mathcal{L}$-view of $\mathcal{E}$.

5 General Results

In this section, we study the problems introduced earlier, focusing on PPKBs whose components are expressed in a generic language $\mathcal{L} \subseteq \mathbf{FO}$.

First, given an ED τ and a PPKB $\mathcal{E} = \langle \mathcal{K}, \mathcal{K}_{pub}, \mathcal{P} \rangle$, we denote by *Ground*($\tau, \mathcal{E}$) the set of EDs obtained through all possible ground substitutions of the universally quantified variables of τ with constants occurring in $\mathcal{E}$. We also denote by *GroundPol*($\mathcal{E}$) the set $\{\textit{Ground}(\tau, \mathcal{E}) \mid \tau \in \mathcal{P}\}$.

The following property immediately follows from the assumption that $\mathcal{P}$ is $\mathcal{K}$-range restricted.

Lemma 1. *Let $\mathcal{E} = \langle \mathcal{K}, \mathcal{K}_{pub}, \mathcal{P} \rangle$ be a PPKB. For every KB Φ such that $\mathcal{K} \models \Phi$, $\Phi \models_{\mathsf{EQL}} \mathcal{P}$ iff $\Phi \models_{\mathsf{EQL}}$ GroundPol($\mathcal{E}$).*

We now provide a notion of consequences of a KB Φ with respect to (the EDs of) a PPKB $\mathcal{E}$. Such a notion will be used throughout the rest of the paper.

Definition 6. *Given a PPKB $\mathcal{E} = \langle \mathcal{K}, \mathcal{K}_{pub}, \mathcal{P} \rangle$ and a KB Φ such that $\mathcal{K} \models \Phi$, we define ED-Cons($\Phi, \mathcal{E}$) as the minimal set Φ' of FO sentences such that, for every $\tau \in$ GroundPol($\mathcal{E}$), if $\Phi \cup \Phi' \models$ body(τ) then head(τ) $\in \Phi'$.*

Example 4. *Recall Example 1. Given the KB $\Phi_3 = \{\mathsf{hasVariant}(\mathsf{sam}, \mathsf{v}_2)\}$, the set ED-Cons($\Phi_3, \mathcal{E}$) is equal to $\{\mathsf{consentToShare}(\mathsf{sam}, \mathsf{v}_2)\}$. Now, let $\Phi_4 = \{\mathsf{hasVariant}(\mathsf{ann}, \mathsf{v}_1)\} \cup \mathcal{K}_{pub}$. In this case, ED-Cons($\Phi_4, \mathcal{E}$) contains $\mathsf{consentToShare}(\mathsf{ann}, \mathsf{v}_1)$ and $\bot$, due to the first and the second ED in $\mathcal{P}$, respectively.* ∎

Observe that, if $\mathcal{P}$ is a set of $\mathcal{L}$-EDs for some $\mathcal{L} \subseteq \mathbf{FO}$, then *ED-Cons*$(\Phi, \mathcal{E}) \subseteq \mathcal{L}_b$ for every $\Phi \subseteq \mathbf{FO}_b$.

We now exploit Lemma 1 for proving the following important property of the *ED-Cons* function.

Lemma 2. *Let $\mathcal{E} = \langle \mathcal{K}, \mathcal{K}_{pub}, \mathcal{P} \rangle$ be a PPKB. For every KB Φ such that $\mathcal{K} \models \Phi$, $\Phi \models_{\mathsf{EQL}} \mathcal{P}$ iff $\Phi \models$ ED-Cons$(\Phi, \mathcal{E})$.*

The above result allows us to derive the next property, which relates the notion of *ED-Cons*$(\Phi, \mathcal{E})$ to the one of PV.

Lemma 3. *Let $\mathcal{E} = \langle \mathcal{K}, \mathcal{K}_{pub}, \mathcal{P} \rangle$ be a PPKB. For every KB Φ such that $\mathcal{K} \models \Phi$, Φ is a PV for $\langle \mathcal{K}, \mathcal{P} \rangle$ iff $\mathcal{K} \not\models$ ED-Cons$(\Phi, \mathcal{E})$.*

In turn, as an immediate consequence of Lemma 3, we obtain the following key property.

Theorem 1. *Given a PPKB $\mathcal{E} = \langle \mathcal{K}, \mathcal{K}_{pub}, \mathcal{P} \rangle$, there exists a leakage in $\mathcal{E}$ iff $\mathcal{K} \not\models$ ED-Cons$(\mathcal{K}_{pub}, \mathcal{E})$.*

The previous theorem is crucial for establishing a general decidability result for the LE problem.

Theorem 2. *If $\mathcal{E} = \langle \mathcal{K}, \mathcal{K}_{pub}, \mathcal{P} \rangle$ is such that $\mathcal{P}$ is a set of $\mathcal{L}$-EDs and deciding entailment of $\mathcal{L}_b$-formulas with respect to theories in $\mathcal{K} \cup \mathcal{L}_b$ is decidable, then the* LE *problem is decidable for $\mathcal{E}$.*

The next two properties establish an important connection between the LE and $\mathcal{L}$-VE problems.

Theorem 3. *Let $\mathcal{E} = \langle \mathcal{K}, \mathcal{K}_{pub}, \mathcal{P} \rangle$ be a PPKB, let $\mathcal{L} \subseteq \mathbf{FO}$ and let $\mathcal{P}$ be a set of $\mathcal{L}$-EDs. Then, there exists a leakage in $\mathcal{E}$ iff there exists no secure $\mathcal{L}_b$-view of $\mathcal{E}$.*

For $\mathcal{L} = \mathbf{FO}$, the above theorem specializes as follows.

Corollary 1. *Let $\mathcal{E} = \langle \mathcal{K}, \mathcal{K}_{pub}, \mathcal{P} \rangle$ be a PPKB. Then, there exists a leakage in $\mathcal{E}$ iff there exists no secure view of $\mathcal{E}$.*

Finally, as a corollary of Theorem 2 and Theorem 3, we obtain the following property.

Corollary 2. *If $\mathcal{E} = \langle \mathcal{K}, \mathcal{K}_{pub}, \mathcal{P} \rangle$ is such that $\mathcal{P}$ is a set of $\mathcal{L}$-EDs and deciding entailment of $\mathcal{L}_b$-formulas with respect to theories in $\mathcal{K} \cup \mathcal{L}_b$ is decidable, then the $\mathcal{L}$-*VE *problem is decidable for $\mathcal{E}$.*

6 Results for DL Ontologies

In this section, we study the computational properties of the LE and $\mathcal{L}$-VE problems, focusing on PPKBs where $\mathcal{K}$ and $\mathcal{K}_{pub}$ are DL ontologies and $\mathcal{P}$ is a finite set of **CQ**-EDs. As is customary in the context of DLs, we distinguish between intensional and extensional knowledge. A *DL PPKB* is thus defined as a triple $\mathcal{E} = \langle \mathcal{T} \cup \mathcal{A}, \mathcal{T}_{pub} \cup \mathcal{A}_{pub}, \mathcal{P} \rangle$, where $\mathcal{T}$ and $\mathcal{T}_{pub}$ are TBoxes, $\mathcal{A}$ and $\mathcal{A}_{pub}$ are ABoxes, such that $\mathcal{T} \models \mathcal{T}_{pub}$, and $\mathcal{T} \cup \mathcal{A} \models \mathcal{T}_{pub} \cup \mathcal{A}_{pub}$. The triple $\mathcal{S} = \langle \mathcal{T}, \mathcal{T}_{pub}, \mathcal{P} \rangle$ is named *PPKB specification*. Specifically, we focus on the DLs DL-Lite$_\mathcal{R}$ and $\mathcal{EL}_\bot$, referring to a PPKB $\langle \mathcal{K}, \mathcal{K}_{pub}, \mathcal{P} \rangle$ as a *DL-Lite$_\mathcal{R}$ PPKB* or *$\mathcal{EL}_\bot$ PPKB* when both $\mathcal{K}$ and $\mathcal{K}_{pub}$ are in DL-Lite$_\mathcal{R}$ or in $\mathcal{EL}_\bot$, respectively.

We call *full $\mathcal{L}$-policy* a finite set of full $\mathcal{L}$-EDs, where a *full $\mathcal{L}$-ED* is an $\mathcal{L}$-ED without existential variables in its head.

Also, for a DL-Lite$_\mathcal{R}$ TBox $\mathcal{T}$, we recall the subclass of the policy language consisting in EDs that are acyclic for $\mathcal{T}$, originally introduced in [Cima *et al.*, 2024a].

Given a DL-Lite$_\mathcal{R}$ TBox $\mathcal{T}$ and a policy $\mathcal{P}$, the *dependency graph* of $\mathcal{T}$ and $\mathcal{P}$, denoted by $G(\mathcal{T}, \mathcal{P})$, is the directed graph defined as follows: (i) the set of nodes of $G(\mathcal{T}, \mathcal{P})$ is the set of predicates occurring in $\mathcal{T} \cup \mathcal{P}$; (ii) there is a P-edge from node p_1 to node p_2 in $G(\mathcal{T}, \mathcal{P})$ if and only if there exists an epistemic dependency of the form (1) in $\mathcal{P}$ such that p_1 occurs in ϕ_b and p_2 occurs in ϕ_h; (iii) there is a T-edge from node p_1 to node p_2 in $G(\mathcal{T}, \mathcal{P})$ if and only if there is a concept or role inclusion in $\mathcal{T}$ such that p_1 occurs in the left-hand side and p_2 occurs in the right-hand side of the inclusion.

Definition 7. *Given a DL-Lite$_\mathcal{R}$ TBox $\mathcal{T}$ and a policy $\mathcal{P}$, we say that $\mathcal{P}$ is* acyclic for $\mathcal{T}$ *if $G(\mathcal{T}, \mathcal{P})$ contains no cycle involving a P-edge. When $\mathcal{P}$ is acyclic for the empty TBox, we simply say that $\mathcal{P}$ is acyclic.*

Informally, the graph $G(\mathcal{T}, \mathcal{P})$ represents the logical dependencies between the predicates in $\mathcal{T}$ and $\mathcal{P}$: a P-edge (resp., a T-edge) from p_1 to p_2 means that predicate p_1 may have a direct implication on p_2 through $\mathcal{P}$ (resp., through $\mathcal{T}$). The notion of acyclicity given above ensures that, if (p_1, p_2) is a P-edge in $G(\mathcal{T}, \mathcal{P})$, then there is no path from p_2 to p_1, i.e. p_2 has no (direct or indirect) implication on p_1.

Then, let us define the following classes of DL PPKBs:

- *ABox PPKB*: a DL PPKB $\langle \mathcal{K}, \mathcal{K}_{pub}, \mathcal{P} \rangle$ where both $\mathcal{K}$ and $\mathcal{K}_{pub}$ are ABoxes, i.e. it has the form $\langle \mathcal{A}, \mathcal{A}_{pub}, \mathcal{P} \rangle$;
- *policy-full PPKB*: a DL-Lite$_\mathcal{R}$ or $\mathcal{EL}_\bot$ PPKB $\langle \mathcal{K}, \mathcal{K}_{pub}, \mathcal{P} \rangle$ where $\mathcal{P}$ is a full policy;
- *policy-acyclic PPKB*: a DL-Lite$_\mathcal{R}$ PPKB $\langle \mathcal{T} \cup \mathcal{A}, \mathcal{T}_{pub} \cup \mathcal{A}_{pub}, \mathcal{P} \rangle$ where $\mathcal{P}$ is acyclic for $\mathcal{T}_{pub}$.

We are now ready to present our initial complexity results for the LE problem, proving that it is tractable in general.

Theorem 4. LE *is (i) PTIME-hard in data complexity for ABox policy-full PPKBs, and (ii) in PTIME in data complexity for both DL-Lite$_\mathcal{R}$ PPKBs and $\mathcal{EL}_\bot$ PPKBs.*

Next, we examine policy-acyclic PPKBs. The main computational result that we will present for these PPKBs (Theorem 5) is obtained through the definition of an FO sentence (Definition 10) whose evaluation on the ABox of the PPKB decides the LE problem in the PPKB. To arrive at such a definition, we resort to two transformations of the policy (Definition 8 and Definition 9), which augment the policy with EDs that are logically implied by the PPKB.

First, we make use of known results for *tuple-generating dependencies* (TGDs) [Abiteboul *et al.*, 1995] (a.k.a. existential rules). We are able to translate EDs into TGDs (and vice versa), in a way such that we can use known reasoning methods for TGDs in the context of PPKBs. In the following, we denote by *UCQRewrite*$(q, \mathcal{D})$ the set of CQs returned by the algorithm presented in [König *et al.*, 2015] for rewriting a conjunctive query q with respect to a set of TGDs $\mathcal{D}$.

For technical reasons, we hereafter assume the existence of a predicate *Ind* $\in \Sigma_p$ that never occurs in a policy or a PPKB, although it may occur in CQs. Given a CQ $\phi(\vec{x})$, we denote by $\phi^+(\vec{x})$ the CQ $\phi(\vec{x}) \wedge \bigwedge_{x \in \vec{x}} \mathit{Ind}(x)$, and denote by $\phi^-(\vec{x})$ the CQ obtained from $\phi(\vec{x})$ deleting all the *Ind* atoms occurring in it. Moreover, for a policy $\mathcal{P}$, we denote by $\mathcal{P}_{tgd}$ the following set of TGDs:

$$\{\forall \vec{x}\, (\phi_b^+(\vec{x}) \rightarrow \phi_h(\vec{x_h})) \mid \forall \vec{x}\, (K\phi_b(\vec{x}) \rightarrow K\phi_h(\vec{x_h})) \in \mathcal{P}\}$$

For such a set $\mathcal{P}_{tgd}$, we define *TGDClosure*$(\mathcal{P}_{tgd})$ as the following set of TGDs:

$$\{\forall \vec{x}\,(q(\vec{x}) \rightarrow \phi_h(\vec{x_h})) \mid \forall \vec{x}\,(\phi_b(\vec{x}) \rightarrow \phi_h(\vec{x_h})) \in \mathcal{P}_{tgd} \text{ and } q(\vec{x}) \in \textit{UCQRewrite}(\phi_b(\vec{x}), \mathcal{P}_{tgd})\}$$

Definition 8. *Given an acyclic policy $\mathcal{P}$, we define PExp$(\mathcal{P})$ as the following policy:*

$$\{\forall \vec{x}\,(K\phi_b^-(\vec{x}) \rightarrow K\phi_h(\vec{x_h})) \mid \forall \vec{x}\,(\phi_b(\vec{x}) \rightarrow \phi_h(\vec{x_h})) \in \textit{TGDClosure}(\mathcal{P}_{tgd})\}$$

Intuitively, *PExp*$(\mathcal{P})$ expands $\mathcal{P}$ using its own EDs as inference rules. As a simple example, consider the policy $\mathcal{P} = \{\forall x\,(K(B(x) \wedge D(x)) \rightarrow KA(x)), \forall x\,(KC(x) \rightarrow KB(x))\}$. Then, *PExp*$(\mathcal{P})$ results in the policy $\mathcal{P} \cup \{\forall x\,(K(C(x) \wedge D(x)) \rightarrow KA(x))\}$.

The next lemma formalizes the key property of *PExp*$(\mathcal{P})$ for our purposes.

Lemma 4. *Given a KB $\mathcal{K}$, an acyclic policy $\mathcal{P}$, and a KB Φ such that $\mathcal{K} \models \Phi$, Φ is a PV in $\langle \mathcal{K}, \mathcal{P}\rangle$ iff there exists an ED $\tau(\vec{x}) \in$ PExp$(\mathcal{P})$ and a ground substitution $\vec{c}$ of $\vec{x}$ such that $\Phi \models$ body$(\tau(\vec{c}))$ and $\mathcal{K} \not\models$ head$(\tau(\vec{c}))$.*

Proof (sketch). Suppose that Φ is a PV in $\langle \mathcal{K}, \mathcal{P}\rangle$. Then, by Lemma 3, $\mathcal{K} \not\models$ *ED-Cons*$(\Phi, \mathcal{E})$. Consequently, there exist a sequence of EDs $\tau_1(\vec{x_1}), \ldots, \tau_j(\vec{x_j})$ in $\mathcal{P}$ and ground instantiations $\vec{c_1}, \ldots, \vec{c_j}$ of $\vec{x_1}, \ldots, \vec{x_j}$ and a sequence of sets of BCQs $\Psi_0 = \emptyset.\Psi_1, \ldots, \Psi_{j-1}$ such that: (i) for every i such that $1 \leq i \leq j-1$, $\Phi \cup \Psi_{i-1} \models$ body$(\tau_i(\vec{c_i}))$ and $\mathcal{K} \models$ head$(\tau_i(\vec{c_i}))$ and $\Psi_i = \Psi_{i-1} \cup \{$head$(\tau_i(\vec{c_i}))\}$; (ii) $\Phi \cup \Psi_{j-1} \models$ body$(\tau_j(\vec{c_j}))$ and $\mathcal{K} \not\models$ head$(\tau_j(\vec{c_j}))$.

Now. using the known properties of *UCQRewrite* [König *et al.*, 2015], we can prove that there exists $\tau'(\vec{x_j}) \in$ *TGDClosure*$(\mathcal{P}_{tgd})$ such that body$(\tau'(\vec{x_j})) = q(\vec{x_j})$ and head$(\tau'(\vec{x_j})) =$ head$(\tau_j(\vec{x_j}))$, therefore $\Phi \models$ body$(\tau'(\vec{c_j}))$ and $\mathcal{K} \not\models$ head$(\tau'(\vec{c_j}))$, thus proving the thesis.

Then, it is easy to verify that, for every KB Φ', $\Phi \models_{\mathsf{EQL}}$ *PExp*$(\mathcal{P})$ iff $\Phi' \models_{\mathsf{EQL}} \mathcal{P}$. This immediately implies that, if Φ is such that there exists an ED $\tau(\vec{x}) \in$ *PExp*$(\mathcal{P})$ and a ground substitution $\vec{c}$ of $\vec{x}$ such that $\Phi \models$ body$(\tau(\vec{c}))$ and $\mathcal{K} \not\models$ head$(\tau(\vec{c}))$, then Φ is a PV in $\langle \mathcal{K}, \mathcal{P}\rangle$. □

We now define a transformation that "compiles" the effect of a TBox $\mathcal{T}$ in a policy.

Definition 9. *Given a policy $\mathcal{P}$ and a DL-Lite$_\mathcal{R}$ TBox $\mathcal{T}$, we define TExp$(\mathcal{P}, \mathcal{T})$ as the following policy:*[1]

$$\{\forall \vec{x}\,(Kq(\vec{x}) \rightarrow K\textit{head}(\tau)) \mid \tau(\vec{x}) \in \mathcal{P} \wedge q(\vec{x}) \in \textit{PerfectRef}(\textit{body}(\tau), \mathcal{T})\}$$

Example 5. *Consider the TBox $\mathcal{T} = \{D \sqsubseteq B,\ E \sqsubseteq A\}$ and the policy $\mathcal{P} = \{\forall x\,(KB(x) \rightarrow KA(x)), \forall x\,(KC(x) \rightarrow KB(x))\}$. Now, observe that PerfectRef$(B(x), \mathcal{T}) = B(x) \vee D(x)$ and PerfectRef$(C(x), \mathcal{T}) = C(x)$. Thus, TExp$(\mathcal{P}, \mathcal{T})$ equals $\mathcal{P} \cup \forall x\,(KD(x) \rightarrow KA(x))$.* ■

From now on, given a DL-Lite$_\mathcal{R}$ PPKB $\mathcal{E} = \langle \mathcal{K}, \mathcal{K}_{pub}, \mathcal{P}\rangle$, where $\mathcal{K}_{pub} = \mathcal{T}_{pub} \cup \mathcal{A}_{pub}$, we denote by $\mathcal{E}_{exp}$ the PPKB $\langle \mathcal{K}, \mathcal{A}_{pub}, \textit{TExp}(\mathcal{P}, \mathcal{T}_{pub})\rangle$.

Based on the above definition of *TExp*, we are able to prove the following property.

[1] Recall that *PerfectRef*$(\textit{body}(\tau), \mathcal{T})$ is a UCQ, i.e. a set of CQs.

Lemma 5. *Let $\mathcal{E} = \langle \mathcal{K}, \mathcal{K}_{pub}, \mathcal{P}\rangle$ be a policy-acyclic PPKB. There exists a leakage in $\mathcal{E}$ iff there exists a leakage in $\mathcal{E}_{exp}$.*

Proof (sketch). First, it is possible to prove that *ED-Cons*$(\mathcal{K}_{pub}, \mathcal{E})$ = *ED-Cons*$(\mathcal{A}_{pub}, \mathcal{E}_{exp})$. Then, recall that there exists a leakage in $\mathcal{E}$ iff $\mathcal{K}_{pub}$ is a leakage in $\mathcal{E}$, and there exists a leakage in $\mathcal{E}_{exp}$ iff $\mathcal{A}_{pub}$ is a leakage in $\mathcal{E}_{exp}$. But, since *ED-Cons*$(\mathcal{K}_{pub}, \mathcal{E})$ = *ED-Cons*$(\mathcal{A}_{pub}, \mathcal{E}_{exp})$, by Lemma 3 it follows that $\mathcal{K}_{pub}$ is a leakage in $\mathcal{E}$ iff $\mathcal{A}_{pub}$ is a leakage in $\mathcal{E}_{exp}$, which proves the thesis. □

Hereinafter, we use the notation $\mathcal{I}(\mathcal{A})$ to refer to the Herbrand model of a given ABox $\mathcal{A}$.

Lemma 4 and Lemma 5 imply the following property.

Lemma 6. *Let $\mathcal{S} = \langle \mathcal{T}, \mathcal{T}_{pub}, \mathcal{P}\rangle$ be a policy-acyclic PPKB specification. For every PPKB $\mathcal{E}$ of the form $\langle \mathcal{T} \cup \mathcal{A}, \mathcal{T}_{pub} \cup \mathcal{A}_{pub}, \mathcal{P}\rangle$, there exists a leakage in $\mathcal{E}$ iff there exist $\tau(\vec{x}) \in$ PExp(TExp$(\mathcal{P}, \mathcal{T}_{pub})$) and a ground substitution $\vec{c}$ of $\vec{x}$ such that eval(body$(\tau(\vec{c})), \mathcal{I}(\mathcal{A}_{pub}))$ is true and eval$(\psi(\vec{c_h}), \mathcal{I}(\mathcal{A}))$ is false, where $\psi(\vec{x_h})$ is the UCQ returned by PerfectRef(head$(\tau), \mathcal{T})$.*

The above property allows us to define an FO sentence whose evaluation on $\mathcal{A}$ and $\mathcal{A}_{pub}$ is able to decide the LE problem for a PPKB $\langle \mathcal{K}, \mathcal{K}_{pub}, \mathcal{P}\rangle$. To this aim, given a formula ϕ, we denote by *Pr*(ϕ) the formula obtained from ϕ replacing every predicate (concept or role) symbol p with its primed version p' (we assume that the auxiliary set of primed predicates is disjoint from Σ_p), and for a set of formulas Φ we denote by *Pr*(Φ) the set $\{$*Pr*$(\phi) \mid \phi \in \Phi\}$.

Definition 10. *Given a PPKB specification $\mathcal{S} = \langle \mathcal{T}, \mathcal{T}_{pub}, \mathcal{P}\rangle$, we define the following FO sentence $\phi_{leak}(\mathcal{S})$:*

$$\bigvee_{\tau(\vec{x}) \in \textit{PExp}(\textit{TExp}(\mathcal{P}, \mathcal{T}_{pub}))} \big(\exists \vec{x}\,(\textit{body}(\tau(\vec{x})) \wedge \neg \textit{PerfectRef}(\textit{head}(\textit{Pr}(\tau)), \mathcal{T}))\big)$$

The next property, whose proof relies on Lemma 6, establishes the FO-rewritability of the LE problem for policy-acyclic PPKBs.

Lemma 7. *Let $\mathcal{S} = \langle \mathcal{T}, \mathcal{T}_{pub}, \mathcal{P}\rangle$ be a PPKB specification such that $\mathcal{T}$ and $\mathcal{T}_{pub}$ are DL-Lite$_\mathcal{R}$ TBoxes and $\mathcal{P}$ is acyclic for $\mathcal{T}_{pub}$. For every pair of ABoxes $\mathcal{A}, \mathcal{A}_{pub}$ such that $\mathcal{T} \cup \mathcal{A} \models \mathcal{T}_{pub} \cup \mathcal{A}_{pub}$, there exists a leakage in the PPKB $\langle \mathcal{T} \cup \mathcal{A}, \mathcal{T}_{pub} \cup \mathcal{A}_{pub}, \mathcal{P}\rangle$ iff eval$(\phi_{leak}(\mathcal{S}), \mathcal{I}(\mathcal{A}_{pub} \cup Pr(\mathcal{A})))$ is true.*

The following result is an immediate consequence of the FO-rewritability property provided by the above lemma.

Theorem 5. LE *is in* AC^0 *in data complexity for policy-acyclic PPKBs.*

Example 6. *Consider Example 3, and recall that Φ_1 is a leakage in the PPKB $\mathcal{E}' = \langle \mathcal{K}', \mathcal{K}_{pub}, \mathcal{P}\rangle$. In the following, we call $\mathcal{A}$, $\mathcal{A}_{pub}$, $\mathcal{T}$ and $\mathcal{T}_{pub}$ the ABoxes and DL-Lite$_\mathcal{R}$ TBoxes such that $\mathcal{K}' = \mathcal{T} \cup \mathcal{A}$, $\mathcal{K}_{pub} = \mathcal{T}_{pub} \cup \mathcal{A}_{pub}$, and $\mathcal{T} = \mathcal{T}_{pub}$ only contains the concept inclusion* $\mathsf{pathogenic} \sqsubseteq \mathsf{sensitive}$ *introduced (in FOL syntax) in Example 1. Now, we have that PExp(TExp$(\mathcal{P}, \mathcal{T}_{pub})$) = TExp$(\mathcal{P}, \mathcal{T}_{pub}) = \mathcal{P} \cup \{\tau\}$, where*

$$\tau = \forall p\,(K\exists v\,(\mathsf{hasVariant}(p, v) \wedge \mathsf{pathogenic}(v)) \rightarrow K\bot)$$

For such a τ, we have that PerfectRef(head$(Pr(\tau)), \mathcal{T}) = \bot$. Then, one of the disjuncts of $\phi_{leak}(\langle \mathcal{T}, \mathcal{T}_{pub}, \mathcal{P}\rangle)$ is:

$$\psi = \exists p, v\,(\mathsf{hasVariant}(p, v) \wedge \mathsf{pathogenic}(v)) \wedge \neg\bot$$

Algorithm 1: ExistsGAView

input: A DL-Lite$_\mathcal{R}$ or $\mathcal{EL}_\bot$ PPKB $\mathcal{E} = \langle \mathcal{K}, \mathcal{K}_{pub}, \mathcal{P} \rangle$
if *there exists* $\mathcal{A}' \subseteq Cons_{\mathbf{GA}}(\mathcal{K})$ *such that*
  $\mathcal{K}_{pub} \cup \mathcal{A}' \models$ *ED-Cons*$(\mathcal{K}_{pub} \cup \mathcal{A}', \mathcal{E})$
then return *true*;
return *false*

which is such that eval$(\psi, \mathcal{I}(\mathcal{A}_{pub} \cup Pr(\mathcal{A})))$ *is true.* ∎

As a corollary of Theorem 3, we get the following property.

Corollary 3. *Let $\mathcal{E} = \langle \mathcal{K}, \mathcal{K}_{pub}, \mathcal{P} \rangle$ be either a DL-Lite$_\mathcal{R}$ or an $\mathcal{EL}_\bot$ PPKB. Then, there exists a leakage in $\mathcal{E}$ iff there exists no secure $\mathbf{CQ}_b$-view of $\mathcal{E}$.*

A notable implication of this finding is that all the complexity results established earlier for the LE problem also apply to (the complement of) the $\mathbf{CQ}_b$-VE problem.

In the rest of this section, we focus on verifying the existence of a secure view that can be expressed in a language suitable for materialization. Specifically, we address the $\mathbf{GA}$-VE problem. We begin our discussion by providing an example that highlights the differences between secure $\mathbf{CQ}_b$-views and secure $\mathbf{GA}$-views, showing that the latter may not exist even in scenarios where the former is definable.

Example 7. *Consider the following simple PPKB $\mathcal{E} = \langle \mathcal{K}, \mathcal{K}_{pub}, \mathcal{P} \rangle$ concerning patients' medical records and their participation in medical research trials.*

$\mathcal{P} = \{\forall r, t\,(K\,\mathsf{inResTrial}(r,t) \rightarrow K\,\bot),$
$\quad \forall r\,(K\,\mathsf{medRecord}(r) \rightarrow K\,\exists t\,\mathsf{inResTrial}(r,t))\},$
$\mathcal{K}_{pub} = \{\mathsf{medRecord}(r_1)\},$
$\mathcal{K} = \mathcal{K}_{pub} \cup \{\mathsf{inResTrial}(r_1, t_1)\}.$

One can verify that no secure $\mathbf{GA}$-view of $\mathcal{E}$ exists, while $\mathcal{E}$ has the secure $\mathbf{CQ}_b$-view $\{\exists y\,\mathsf{inResTrial}(r_1, y)\}$. ∎

Now, for DL-Lite$_\mathcal{R}$ or $\mathcal{EL}_\bot$ PPKBs that are policy-full, we derive the following property as a corollary of Theorem 3.

Corollary 4. *Let $\mathcal{E} = \langle \mathcal{K}, \mathcal{K}_{pub}, \mathcal{P} \rangle$ be either a policy-full DL-Lite$_\mathcal{R}$ PPKB or a policy-full $\mathcal{EL}_\bot$ PPKB. Then, there exists a leakage in $\mathcal{E}$ iff there exists a secure $\mathbf{GA}$-view of $\mathcal{E}$.*

From Theorem 4 and Corollary 4 it immediately follows that the $\mathbf{GA}$-VE problem is PTIME-complete in data complexity in the case the given PPKB is either a policy-full DL-Lite$_\mathcal{R}$ PPKB or a policy-full $\mathcal{EL}_\bot$ PPKB. More precisely, the PTIME-hardness already holds for ABox policy-full PPKBs. It is worth noting, however, that Theorem 5 and Corollary 4 imply that in all such cases the problem is in AC^0 if the given PPKB is also policy-acyclic.

Then, we provide a general upper bound for the $\mathbf{GA}$-VE problem, which can be demonstrated using of Algorithm 1.

Theorem 6. $\mathbf{GA}$-VE *is in NP in data complexity for DL-Lite$_\mathcal{R}$ PPKBs and for $\mathcal{EL}_\bot$ PPKBs.*

We then show that $\mathbf{GA}$-VE is harder than $\mathbf{CQ}_b$-VE (or, equivalently, the complement of LE problem) in many cases.

Theorem 7. $\mathbf{GA}$-VE *is NP-hard in data complexity for policy-acyclic ABox PPKBs.*

$\mathcal{K}, \mathcal{K}_{pub}$	$\mathcal{P}$ ($\mathbf{CQ}$-EDs)	LE, $\mathbf{CQ}_b$-VE	$\mathbf{GA}$-VE
ABox	all	PTIME [T4] [T4]	NP [T6] [T7]
ABox	full	PTIME [T4] [T4]	PTIME [T4+C4] [T4+C4]
ABox	acyclic	in AC^0 [T5]	NP [T6] [T7]
ABox	acyclic full	in AC^0 [T5]	in AC^0 [T5+C4]
DL-Lite$_\mathcal{R}$	all	PTIME [T4] [T4]	NP [T6] [T7]
DL-Lite$_\mathcal{R}$	full	PTIME [T4] [T4]	PTIME [T4+C4] [T4+C4]
DL-Lite$_\mathcal{R}$	acyclic	in AC^0 [T5]	NP [T6] [T7]
DL-Lite$_\mathcal{R}$	acyclic full	in AC^0 [T5]	in AC^0 [T5+C4]
$\mathcal{EL}_\bot$	all	PTIME [T4] [T4]	NP [T6] [T7]
$\mathcal{EL}_\bot$	full	PTIME [T4] [T4]	PTIME [T4+C4] [T4+C4]

Table 1: Data complexity of the LE and $\mathcal{L}$-VE problems (with $\mathcal{L} \in \{\mathbf{CQ}_b, \mathbf{GA}\}$). All entries refer to completeness results unless stated otherwise. T and C stand for theorem and corollary, respectively, where lower and upper bounds have been established.

Proof (sketch). We can prove the NP-hardness via a reduction from 3SAT. Specifically, given a 3CNF ψ, one can construct an ABox $\mathcal{A}$ that contains two facts representing the assignment of both values 1 and 0 for every propositional variable of ψ, and such that $\mathcal{A}_{pub}$ contains all the facts for representing the structure of ψ. Then, $\mathcal{P}$ can be used for stating that every propositional variable of ψ must be assigned to a value (among the ones in $\mathcal{A}$), distinct values can not be assigned to the same variable, and for every clause c there must exist a variable whose value is 1 iff it does not occur negated in c. Intuitively, an interpretation $\mathcal{I}$ satisfying ψ exists iff one can safely disclose a set of facts that is isomorphic to $\mathcal{I}$. □

7 Conclusions

In this paper, we introduced the PPKB framework for assessing the exposure of sensitive information to public knowledge. In studying the framework, we focused on two fundamental decision problems, namely the data leakage existence problem and the secure view existence problem, providing results for DL-Lite$_\mathcal{R}$ and $\mathcal{EL}_\bot$ PPKBs. Our findings, summarized in Table 1, provide a detailed analysis of the data complexity of these problems across various scenarios. Notably, we proved that the problems are tractable in many cases, and even in AC^0 for acyclic policies in the case of DL-Lite$_\mathcal{R}$.

The results we presented open several directions for future research. A natural extension involves studying these problems within the context of different ontology languages and/or richer policy languages that go beyond $\mathbf{CQ}$-EDs to offer more flexible protection mechanisms. Furthermore, there is an opportunity to investigate approaches to privacy-preserving query answering (such as CQE) in the presence of public knowledge. Specifically, it would be worth investigating forms of privacy-preserving query answering that are tolerant to leakages. That is, in situations where public knowledge "irrecoverably" violates the protection policy, computing views that, at the same time, are maximal in terms of disclosed information and do not expose further sensitive data.

Acknowledgements

This work was partially supported by: projects FAIR (PE0000013) and SERICS (PE00000014) under the MUR National Recovery and Resilience Plan funded by the EU - NextGenerationEU; GLACIATION project funded by the EU (N. 101070141); ANTHEM (AdvaNced Technologies for Human-centrEd Medicine) project (CUP B53C22006700001) funded by the National Plan for NRRP Complementary Investments; by the MUR PRIN 2022LA8XBH project Polar (POLicy specificAtion and enfoRcement for privacy-enhanced data management); and by the EU under the HORIZON.2.1.5 project dAIbetes (grant id. 101136305).

References

[Abiteboul *et al.*, 1995] Serge Abiteboul, Richard Hull, and Victor Vianu. *Foundations of Databases*. Addison Wesley Publ. Co., 1995.

[Baader and Nuradiansyah, 2019] Franz Baader and Adrian Nuradiansyah. Mixing description logics in privacy-preserving ontology publishing. In *KI 2019: Advances in Artificial Intelligence*, pages 87–100, Cham, 2019. Springer.

[Baader *et al.*, 1999] Franz Baader, Ralf Küsters, and Ralf Molitor. Computing least common subsumers in description logics with existential restrictions. In *Proc. of the 16th Int. Joint Conf. on Artificial Intelligence (IJCAI)*, pages 96–101, 1999.

[Baader *et al.*, 2007] Franz Baader, Diego Calvanese, Deborah McGuinness, Daniele Nardi, and Peter F. Patel-Schneider, editors. *The Description Logic Handbook: Theory, Implementation and Applications*. Cambridge University Press, 2nd edition, 2007.

[Baader *et al.*, 2019] Franz Baader, Francesco Kriegel, and Adrian Nuradiansyah. Privacy-preserving ontology publishing for $\mathcal{EL}$ instance stores. In *Logics in Artificial Intelligence*, pages 323–338, Cham, 2019. Springer.

[Benedikt *et al.*, 2018] Michael Benedikt, Bernardo Cuenca Grau, and Egor V. Kostylev. Logical foundations of information disclosure in ontology-based data integration. *Artificial Intelligence*, 262:52–95, 2018.

[Benedikt *et al.*, 2019] Michael Benedikt, Pierre Bourhis, Louis Jachiet, and Michaël Thomazo. Reasoning about disclosure in data integration in the presence of source constraints. In *Proc. of the 28th Int. Joint Conf. on Artificial Intelligence (IJCAI)*, pages 1551–1557, 2019.

[Biskup and Bonatti, 2001] Joachim Biskup and Piero A. Bonatti. Lying versus refusal for known potential secrets. *Data and Knowledge Engineering*, 38(2):199–222, 2001.

[Biskup and Bonatti, 2004] Joachim Biskup and Piero A. Bonatti. Controlled query evaluation for known policies by combining lying and refusal. *Ann. of Mathematics and Artificial Intelligence*, 40(1-2):37–62, 2004.

[Biskup and Weibert, 2008] Joachim Biskup and Torben Weibert. Keeping secrets in incomplete databases. *Int. J. Inf. Sec.*, 7(3):199–217, 2008.

[Bonatti and Sauro, 2013] Piero A. Bonatti and Luigi Sauro. A confidentiality model for ontologies. In *Proc. of the 12th Int. Semantic Web Conf. (ISWC)*, pages 17–32, 2013.

[Calvanese *et al.*, 2007a] Diego Calvanese, Giuseppe De Giacomo, Domenico Lembo, Maurizio Lenzerini, and Riccardo Rosati. EQL-Lite: Effective first-order query processing in description logics. In *Proc. of the 20th Int. Joint Conf. on Artificial Intelligence (IJCAI)*, pages 274–279, 2007.

[Calvanese *et al.*, 2007b] Diego Calvanese, Giuseppe De Giacomo, Domenico Lembo, Maurizio Lenzerini, and Riccardo Rosati. Tractable reasoning and efficient query answering in description logics: The *DL-Lite* family. *J. of Automated Reasoning*, 39(3):385–429, 2007.

[Calvanese *et al.*, 2012] Diego Calvanese, Giuseppe De Giacomo, Maurizio Lenzerini, and Riccardo Rosati. View-based query answering in description logics: Semantics and complexity. *J. of Computer and System Sciences*, 78(1):26–46, 2012.

[Chirkova and Yu, 2017] Rada Chirkova and Ting Yu. Exact detection of information leakage: Decidability and complexity. *Trans. Large Scale Data Knowl. Centered Syst.*, 32:1–23, 2017.

[Cima *et al.*, 2024a] Gianluca Cima, Domenico Lembo, Lorenzo Marconi, Riccardo Rosati, and Domenico Fabio Savo. Enhancing controlled query evaluation through epistemic policies. In *Proc. of the 33th Int. Joint Conf. on Artificial Intelligence (IJCAI)*, pages 3307–3314, 2024.

[Cima *et al.*, 2024b] Gianluca Cima, Domenico Lembo, Riccardo Rosati, and Domenico Fabio Savo. Controlled query evaluation in description logics through consistent query answering. *Artificial Intelligence*, 334:104176, 2024.

[Console and Lenzerini, 2020] Marco Console and Maurizio Lenzerini. Epistemic integrity constraints for ontology-based data management. In *Proc. of the 37th AAAI Conf. on Artificial Intelligence (AAAI)*, pages 2790–2797, 2020.

[Cuenca Grau *et al.*, 2013] Bernardo Cuenca Grau, Evgeny Kharlamov, Egor V. Kostylev, and Dmitriy Zheleznyakov. Controlled query evaluation over OWL 2 RL ontologies. In *Proc. of the 12th Int. Semantic Web Conf. (ISWC)*, pages 49–65, 2013.

[Cuenca Grau *et al.*, 2015] Bernardo Cuenca Grau, Evgeny Kharlamov, Egor V. Kostylev, and Dmitriy Zheleznyakov. Controlled query evaluation for datalog and OWL 2 profile ontologies. In *Proc. of the 24th Int. Joint Conf. on Artificial Intelligence (IJCAI)*, pages 2883–2889, 2015.

[Dwork, 2011] Cynthia Dwork. Differential privacy. In Henk C. A. van Tilborg and Sushil Jajodia, editors, *Encyclopedia of Cryptography and Security, 2nd Ed*, pages 338–340. Springer, 2011.

[König *et al.*, 2015] Mélanie König, Michel Leclère, Marie-Laure Mugnier, and Michaël Thomazo. Sound, complete and minimal UCQ-rewriting for existential rules. *Semantic Web J.*, 6(5):451–475, 2015.

[Miklau and Suciu, 2007] Gerome Miklau and Dan Suciu. A formal analysis of information disclosure in data exchange. *J. of Computer and System Sciences*, 73(3):507–534, 2007.

[Motik *et al.*, 2009] Boris Motik, Achille Fokoue, Ian Horrocks, Zhe Wu, Carsten Lutz, and Bernardo Cuenca Grau. OWL Web Ontology Language profiles. W3C Recommendation, World Wide Web Consortium, October 2009. Available at http://www.w3.org/TR/owl-profiles/.

[Stoffel and Studer, 2005] Kilian Stoffel and Thomas Studer. Provable data privacy. In Kim Viborg Andersen, John K. Debenham, and Roland R. Wagner, editors, *Proc. of the 16th Int. Workshop on Database and Expert Systems Applications (DEXA)*, volume 3588 of *Lecture Notes in Computer Science*, pages 324–332, 2005.

[Studer and Werner, 2014] Thomas Studer and Johannes Werner. Censors for boolean description logic. *Trans. Data Privacy*, 7(3):223–252, 2014.

[Sweeney, 2002] Latanya Sweeney. k-anonymity: A model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl. Based Syst.*, 10(5):557–570, 2002.

[Vardi, 1982] Moshe Y. Vardi. The complexity of relational query languages. In *Proc. of the 14th ACM SIGACT Symp. on Theory of Computing (STOC)*, pages 137–146, 1982.