Controlled Query Evaluation in DL-Lite through Epistemic Protection Policies (Extended Abstract)

Gianluca Cima[1], Domenico Lembo[1], Lorenzo Marconi[1], Riccardo Rosati[1] and Domenico Fabio Savo[2]

[1]*Sapienza Università di Roma*
[2]*Università degli Studi di Bergamo*

Abstract

This extended abstract summarizes our work recently accepted to the 33rd International Joint Conference on Artificial Intelligence, in which we present a new language for data protection policies in the Controlled Query Evaluation framework, based on the use of an epistemic operator.

Keywords

Description Logics, Data Protection, First-Order Rewritability, Epistemic Modal Logic

Controlled Query Evaluation (CQE) is an approach to safeguard sensitive information by filtering query responses to prevent users from inferring data declared confidential by a data protection policy [1, 2, 3, 4]. A critical aspect of CQE revolves around the expressiveness of the policy language, which allows to determine the information that must remain undisclosed.

Previous literature has mainly considered only policies consisting of sentences, i.e. closed logical formulas, as in [5, 2, 3, 6]. Through this approach, it is only possible to impose that the truth value of a sentence in the policy cannot be inferred by asking queries to the system. For example, in [6] policy statements take the form $\forall\vec{x}(cq(\vec{x}) \rightarrow \bot)$, referred to as denial. Enforcing one such denial over the data means refraining from disclosing the inference of the *Boolean conjunctive query (BCQ)* $\exists\vec{x}\ cq(\vec{x})$ by the system, even if the inference has occurred. For instance, the rule $\delta_1 = \forall x, y(\mathsf{Patient}(x) \wedge \mathsf{admitted}(x, y) \rightarrow \bot)$ aims to keep undisclosed the fact that a patient is admitted to a hospital department. One may argue that the above dependency imposes an excessively stringent level of protection, since it denies the presence of any patient in the hospital to protect their identity, which is implausible in practice. A more effective approach could require the system not to answer the *open* query $q(x) : \exists y(\mathsf{Patient}(x) \wedge \mathsf{admitted}(x, y))$. Intuitively, a rule of this kind expresses that the set of patients that the system *knows* to be admitted to the hospital must not be disclosed.

To properly capture this behaviour, we propose to use an epistemic operator K in policy formulas, in the spirit of [7, 8]. This allows us to formalize the epistemic state of the user, i.e., which information can be safely undisclosed for her. In our example, by expressing a rule like

$$\delta_2 = \forall x\big(K\,\exists y(\mathsf{Patient}(x) \wedge \mathsf{admitted}(x, y)) \rightarrow K\bot\big)$$

one could still allow the end user to infer that some patient has been admitted, while protecting the identities of all such patients.

In fact, our proposal enables us to accomplish more than this. In a more advanced scenario, concealing the relationship between a patient and a hospital department should apply only if the patient has not signed a consensus form. This can be expressed by the following formula:

$$\delta_3 = \forall x\big(K\,\exists y(\mathsf{Patient}(x) \wedge \mathsf{admitted}(x, y)) \rightarrow K\mathsf{Consent}(x)\big)$$

DL 2024: 37th International Workshop on Description Logics, June 18–21, 2024, Bergen, Norway

cima@diag.uniroma1.it (G. Cima); lembo@diag.uniroma1.it (D. Lembo); marconi@diag.uniroma1.it (L. Marconi); rosati@diag.uniroma1.it (R. Rosati); domenicofabio.savo@unibg.it (D. F. Savo)
0000-0003-1783-5605 (G. Cima); 0000-0002-0628-242X (D. Lembo); 0000-0001-9633-8476 (L. Marconi); 0000-0002-7697-4958 (R. Rosati); 0000-0002-8391-8049 (D. F. Savo)

In essence, this formula stipulates that if a user knows a patient has been admitted, she must also know the patient has signed a consent form. Consequently, if a patient has not signed a consent form, the system cannot disclose her admission status.

Hereafter, we employ standard notions of function-free first-order (FO) logics with unary and binary predicates and, besides the alphabets of predicate names, constants and variables, we assume the existence of an infinite set of symbols called *labeled nulls*. As customary when dealing with epistemic operators, we adopt the *standard name assumption* [9]. Moreover, we write *eval*$(\phi, \mathcal{I})$ to indicate the evaluation of an FO sentence ϕ over an FO interpretation $\mathcal{I}$. A Description Logic ontology is a FO theory $\mathcal{T} \cup \mathcal{A}$, where $\mathcal{T}$ is a TBox and $\mathcal{A}$ is a *quantified ABox* [10] (i.e. a set of atomic FO formulas predicating over constants and labeled nulls), and all the predicates occurring in $\mathcal{A}$ also occur in $\mathcal{T}$. As for user queries, we focus on BCQs and on *Boolean Union of Conjunctive Queries (BUCQs)*.

Formulas as δ_2 and δ_3 belong to the language of *epistemic dependencies* (EDs) which, in general, take the following form:

$$\forall \vec{x}_1, \vec{x}_2 (K q_b(\vec{x}_1, \vec{x}_2) \rightarrow K q_h(\vec{x}_2)) \tag{1}$$

where $q_b(\vec{x}_1, \vec{x}_2)$ is a CQ with free variables $\vec{x}_1 \cup \vec{x}_2$, $q_h(\vec{x}_2)$ is a CQ with free variables $\vec{x}_2$, and K is a modal operator. EDs have been originally introduced in [8] to express integrity constraints in ontology-based data management and are indeed a special case of *domain-independent EQL-Lite(CQ)* sentences [9]. A policy consisting of EDs is called *epistemic policy*. In the present paper, we use EDs as policy rules that must be satisfied to preserve data confidentiality over Descprition Logic ontologies, similarly as integrity constraints must be satisfied to guarantee data consistency. However, our aim is totally different from that of [8], whose focus is on consistency checking.

Intuitively, an ED of form (1) should be read as follows: if the sentence $q_b(\vec{c}_1, \vec{c}_2)$ is *known* to hold, then the sentence $q_h(\vec{c}_2)$ is *known* to hold, for any sequences of constants $\vec{c}_1$ and $\vec{c}_2$ of the same length of $\vec{x}_1$ and $\vec{x}_2$, respectively. More formally, we say that a FO theory Φ *satisfies* an ED δ (denoted $\Phi \models_{\mathsf{EQL}} \delta$) if, for every sequences of constants $\vec{c}_1$ and $\vec{c}_2$ of the same length of $\vec{x}_1$ and $\vec{x}_2$, respectively, the fact that *eval*$(q_b(\vec{c}_1, \vec{c}_2), \mathcal{I})$ is true for every $\mathcal{I} \in \mathcal{M}$ implies that *eval*$(q_h(\vec{c}_2), \mathcal{I})$ is true for every $\mathcal{I} \in \mathcal{M}$, where $\mathcal{M}$ is the set of all FO models of Φ. We say that Φ *satisfies* a policy $\mathcal{P}$ (denoted $\Phi \models_{\mathsf{EQL}} \mathcal{P}$) if Φ satisfies every $\delta \in \mathcal{P}$.

Now, to completely characterize our CQE framework, we need to specify its semantics. This issue is addressed in CQE through *censors*. Here, we employ *CQ-censors*, introduced in [6]. To this aim, we first define the notion of *CQE instance* as a triple $\mathcal{E} = \langle \mathcal{T}, \mathcal{A}, \mathcal{P} \rangle$, where $\mathcal{T}$ is a TBox, $\mathcal{A}$ is a quantified ABox such that $\mathcal{T} \cup \mathcal{A}$ has at least one model, and $\mathcal{P}$ is an epistemic policy such that $\mathcal{T} \models_{\mathsf{EQL}} \mathcal{P}$. The notion of (optimal) CQ-censors is then as follows.

Definition 1 (CQ-censor)**.** *A* CQ-censor *of a CQE instance $\mathcal{E} = \langle \mathcal{T}, \mathcal{A}, \mathcal{P} \rangle$ is a set $\mathcal{C}$ of BCQs such that (i) $\mathcal{T} \cup \mathcal{C} \models_{\mathsf{EQL}} \mathcal{P}$ and (ii) $\mathcal{T} \cup \mathcal{A} \models q$ for every $q \in \mathcal{C}$. Moreover, $\mathcal{C}$ is an* optimal CQ-censor *of $\mathcal{E}$ if there does not exist any CQ-censor $\mathcal{C}'$ of $\mathcal{E}$ such that $\mathcal{C}' \supset \mathcal{C}$.*

As in [6], we define CQE as the problem of checking whether an FO sentence is entailed by all optimal censors. We call this problem *SC-entailment* (Skeptical entailment under CQ-censors). This form of CQE does not suffer from the problem of having to arbitrarily select an optimal censor among several incomparable ones [3, 11]. We also consider a sound approximation of SC-entailment, that we call *IC-entailment*, consisting in checking whether an FO sentence is entailed by the intersection of all optimal censors. We use the notation $\mathcal{E} \models_{\text{SC}} q$ (resp., $\mathcal{E} \models_{\text{IC}} q$) for indicating that an FO sentence q is SC-entailed (resp., IC-entailed) by a CQE instance $\mathcal{E}$.

Example 1. *Suppose that company C_A wants to share certain user-profiling data with a company C_B for targeted advertising. This is not allowed in general, but only in some countries with a special regulation that enables sharing based on the users' consent. C_A may use the following ED in the policy to enable C_B to access only data compliant with the above requirements:*

$$\delta_4 = \forall x, y \big(K \mathsf{profileActivity}(x, y) \rightarrow K \exists z (\mathsf{citOf}(x, z) \wedge \mathsf{SR}(z) \wedge \mathsf{Consent}(x)) \big).$$

In the rule, profileActivity *links a user with her profiling-data,* citOf *relates a user to her country of citizenship,* SR *denotes countries with special regulation, and* Consent *denotes users who have given their consent.*

Suppose that C_A also wants C_B not to be able to associate to a user with her real identity, and that this is possible by collecting the person's name and her date of birth at the same time. To this aim, C_A also specifies the ED:

$$\delta_5 = \forall x, y, z\big(K(\mathsf{name}(x, y) \wedge \mathsf{dateB}(x, z)) \rightarrow K\bot\big).$$

Now, let $\mathcal{E} = \langle \mathcal{T}, \mathcal{A}, \mathcal{P} \rangle$ be a CQE instance, where $\mathcal{T} = \emptyset$, $\mathcal{P} = \{\delta_4, \delta_5\}$, and $\mathcal{A}$ is as follows:

$$\begin{aligned}\mathcal{A} = \{\ &\mathsf{profileActivity}(\mathsf{u}_1, \mathsf{a}_1), \mathsf{Consent}(\mathsf{u}_1), \mathsf{citOf}(\mathsf{u}_1, n_1), \mathsf{SR}(n_1), \mathsf{name}(\mathsf{u}_1, \mathsf{ann}),\\ &\mathsf{dateB}(\mathsf{u}_1, \mathsf{date}_1), \mathsf{profileActivity}(\mathsf{u}_2, \mathsf{a}_2), \mathsf{citOf}(\mathsf{u}_2, n_1)\ \},\end{aligned}$$

where n_1 is a labeled null and all the other terms used in $\mathcal{A}$ are constants. Let us also consider the following four BCQs:

$$\begin{aligned}q_1 &= \exists y\,(\mathsf{profileActivity}(\mathsf{u}_1, \mathsf{a}_1) \wedge \mathsf{citOf}(\mathsf{u}_1, y) \wedge \mathsf{SR}(y)),\\ q_2 &= \mathsf{profileActivity}(\mathsf{u}_2, \mathsf{a}_2),\\ q_3 &= \exists y\, \mathsf{profileActivity}(y, \mathsf{a}_2),\\ q_4 &= \mathsf{profileActivity}(\mathsf{u}_1, \mathsf{a}_1) \wedge \mathsf{name}(\mathsf{u}_1, \mathsf{ann}).\end{aligned}$$

For $X \in \{SC,\ IC\}$, one can verify that $\mathcal{E} \models_X q_1$ since user u_1 gave the consent and she is a citizen of some country (n_1) with special regulation. Conversely, $\mathcal{E} \not\models_X q_2$ since u_2 did not give the consent. Nevertheless, one can verify that q_3 belongs to each optimal CQ-censor of $\mathcal{E}$, and so $\mathcal{E} \models_X q_3$. Finally, we have that $\mathcal{E} \not\models_X q_4$ since there is an optimal CQ-censor $\mathcal{C}$ of $\mathcal{E}$ such that $\mathsf{dateB}(\mathsf{u}_1, \mathsf{date}_1) \in \mathcal{C}$, implying that $\mathsf{name}(\mathsf{u}_1, \mathsf{ann}) \notin \mathcal{C}$ (otherwise δ_5 would be violated). □

We focus on TBoxes specified in DL-Lite$_\mathcal{R}$ [7], for which SC-entailment of BCQs is known to be tractable in data complexity when policies are expressed as denials [6]. Unfortunately, when the policy consists of epistemic dependencies, such a problem turns out to be intractable in general, and the same holds for IC-entailment.

Theorem 1. *In the case of CQE instances with DL-Lite$_\mathcal{R}$ TBoxes, both SC-entailment and IC-entailment of BUCQs are coNP-complete in data complexity.*

We remark that, in the above findings, the hardness for coNP holds already when the TBox is empty and the query is a ground atom. To regain tractability, we then impose an acyclicity condition on the policy $\mathcal{P}$ w.r.t. a DL-Lite$_\mathcal{R}$ TBox $\mathcal{T}$, defined as follows.

Definition 2 (Acyclicity)**.** *We call* dependency graph *of $(\mathcal{T}, \mathcal{P})$ the direct graph whose nodes are the predicates occurring in $\mathcal{T} \cup \mathcal{P}$, and whose edges (p_1, p_2) correspond to the presence of p_1 in the left-hand side and p_2 in right-hand side of either an ED in $\mathcal{P}$ (we call such edges* P-edge*) or a positive inclusion axiom in $\mathcal{T}$. We say that $\mathcal{P}$ is* acyclic for $\mathcal{T}$ *if there exists no cycle involving a P-edge in the dependency graph of $(\mathcal{T}, \mathcal{P})$.*

For this class of epistemic policies, we are able to provide first-order rewriting algorithms for deciding if a BUCQ is SC- or IC-entailed by a CQE instance, thus proving the following result.

Theorem 2. *In the case of CQE instances with DL-Lite$_\mathcal{R}$ TBoxes $\mathcal{T}$ and epistemic policies $\mathcal{P}$ acyclic for $\mathcal{T}$, both SC-entailment and IC-entailment of BUCQs are in AC^0 in data complexity.*

Besides the computational complexity study, we also carry out an analysis of the robustness of the above-defined entailment semantics w.r.t. confidentiality-preservation. In [12, 3, 13], it is shown that censoring mechanisms based on an indistinguishability criterion are indeed more secure than others. In this regard, we adopt a similar approach to the one described in [12] for relational databases. Intuitively, under such an approach an entailment semantics preserves confidentiality if, for every CQE instance $\langle \mathcal{T}, \mathcal{A}, \mathcal{P} \rangle$ and every finite set $\mathcal{Q}$ of queries, the answers to such queries are the same as if they were

obtained from a (possibly different) CQE instance $\langle \mathcal{T}, \mathcal{A}', \mathcal{P} \rangle$ such that $\mathcal{T} \cup \mathcal{A}' \models_{\mathsf{EQL}} \mathcal{P}$. We here describe this property formally. First, for a TBox $\mathcal{T}$, a policy $\mathcal{P}$, two ABoxes $\mathcal{A}$ and $\mathcal{A}'$, a set $\mathcal{Q}$ of BUCQs, and $X \in \{\text{SC}, \text{IC}\}$, we say that $\mathcal{A}$ and $\mathcal{A}'$ are *$\mathcal{Q}$-indistinguishable for X-entailment w.r.t.* $(\mathcal{T}, \mathcal{P})$ if, for every $q \in \mathcal{Q}$, $\langle \mathcal{T}, \mathcal{A}, \mathcal{P} \rangle \models_X q$ iff $\langle \mathcal{T}, \mathcal{A}', \mathcal{P} \rangle \models_X q$.

Definition 3 (Confidentiality preservation)**.** *For $X \in \{SC,\ IC\}$, we say that X-*entailment preserves confidentiality for BCQs *(resp., BUCQs) if, for every CQE instance $\langle \mathcal{T}, \mathcal{A}, \mathcal{P} \rangle$, and for every finite set $\mathcal{Q}$ of BCQs (resp., BUCQs), there exists an ABox $\mathcal{A}'$ such that the following holds:*

(i) $\mathcal{T} \cup \mathcal{A}' \models_{\mathsf{EQL}} \mathcal{P}$*;*
(ii) $\mathcal{A}$ *and* $\mathcal{A}'$ *are $\mathcal{Q}$-indistinguishable for X-entailment with respect to* $(\mathcal{T}, \mathcal{P})$.

In the case of DL-Lite$_{\mathcal{R}}$ ontologies, we can show that SC-entailment preserves confidentiality for BCQs, but this result does not carry over to BUCQs. We however prove that the property is enjoyed even for BUCQs in the case of IC-entailment.

Acknowledgments

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; the MUR PRIN 2022LA8XBH project Polar (POLicy specificAtion and enfoRcement for privacy-enhanced data management); and by the EU under the H2020-EU.2.1.1 project TAILOR (grant id. 952215).

References

[1] J. Biskup, For unknown secrecies refusal is better than lying, Data and Knowledge Engineering 33 (2000) 1–23.
[2] J. Biskup, P. A. Bonatti, Controlled query evaluation for enforcing confidentiality in complete information systems, Int. J. Inf. Sec. 3 (2004) 14–27.
[3] P. A. Bonatti, L. Sauro, A confidentiality model for ontologies, in: Proc. of the 12th Int. Semantic Web Conf. (ISWC), 2013, pp. 17–32.
[4] B. Cuenca Grau, E. Kharlamov, E. V. Kostylev, D. Zheleznyakov, Controlled query evaluation over OWL 2 RL ontologies, in: Proc. of the 12th Int. Semantic Web Conf. (ISWC), 2013, pp. 49–65.
[5] G. L. Sicherman, W. de Jonge, R. P. van de Riet, Answering queries without revealing secrets, ACM Trans. on Database Systems 8 (1983) 41–59.
[6] D. Lembo, R. Rosati, D. F. Savo, Revisiting controlled query evaluation in description logics, in: Proc. of the 28th Int. Joint Conf. on Artificial Intelligence (IJCAI), 2019, pp. 1786–1792.
[7] D. Calvanese, G. De Giacomo, D. Lembo, M. Lenzerini, R. Rosati, Tractable reasoning and efficient query answering in description logics: The *DL-Lite* family, J. of Automated Reasoning 39 (2007) 385–429.
[8] M. Console, M. Lenzerini, Epistemic integrity constraints for ontology-based data management, in: Proc. of the 37th AAAI Conf. on Artificial Intelligence (AAAI), AAAI Press, 2020, pp. 2790–2797. URL: https://doi.org/10.1609/aaai.v34i03.5667. doi:10.1609/AAAI.V34I03.5667.
[9] D. Calvanese, G. De Giacomo, D. Lembo, M. Lenzerini, R. Rosati, EQL-Lite: Effective first-order query processing in description logics, in: Proc. of the 20th Int. Joint Conf. on Artificial Intelligence (IJCAI), 2007, pp. 274–279.
[10] F. Baader, F. Kriegel, A. Nuradiansyah, R. Peñaloza, Computing compliant anonymisations of quantified ABoxes w.r.t. $\mathcal{EL}$ policies, in: Proc. of the 19th Int. Semantic Web Conf. (ISWC), volume 12506 of *Lecture Notes in Computer Science*, Springer, 2020, pp. 3–20.

[11] B. Cuenca Grau, E. Kharlamov, E. V. Kostylev, D. Zheleznyakov, Controlled query evaluation for datalog and OWL 2 profile ontologies, in: Proc. of the 24th Int. Joint Conf. on Artificial Intelligence (IJCAI), 2015, pp. 2883–2889.
[12] J. Biskup, T. Weibert, Keeping secrets in incomplete databases, Int. J. Inf. Sec. 7 (2008) 199–217.
[13] P. A. Bonatti, A false sense of security, Artificial Intelligence 310 (2022).