Preliminary Results on the Identity Problem in Description Logic Ontologies

Franz Baader, Daniel Borchmann, Adrian Nuradiansyah*
firstname.lastname@tu-dresden.de

Theoretical Computer Science, TU Dresden

Abstract. The work in this paper is motivated by a privacy scenario in which the identity of certain persons (represented as anonymous individuals) should be hidden. We assume that factual information about known individuals (i.e., individuals whose identity is known) and anonymous individuals is stored in an ABox and general background information is expressed in a TBox, where both the TBox and the ABox are publicly accessible. The identity problem then asks whether one can deduce from the TBox and the ABox that a given anonymous individual is equal to a known one. Since this would reveal the identity of the anonymous individual, such a situation needs to be avoided. We first observe that not all Description Logics (DLs) are able to derive any such equalities between individuals, and thus the identity problem is trivial in these DLs. We then consider DLs with nominals, number restrictions, or function dependencies, in which the identity problem is non-trivial. We show that in these DLs the identity problem has the same complexity as the instance problem. Finally, we consider an extended scenario in which users with different rôles can access different parts of the TBox and ABox, and we want to check whether, by a sequence of rôle changes and queries asked in each rôle, one can deduce the identity of an anonymous individual.

1 Introduction

In order to illustrate the privacy scenario sketched in the abstract, assume that you are asked to perform a survey regarding the satisfaction of employees with the management of a company. Since the boss of the company is known not to respond well to criticism, the employees insist that you perform the survey such that the identity of persons voicing criticism cannot be deduced by the boss. Thus, you let the employees use a pseudonym when answering the survey. However, the survey does ask some personal data from the participants, and you are concerned that the boss can use the provided answers, in combination with the employee database and general knowledge about how things work in the company, to deduce that a certain pseudonym corresponds to a specific employee. For example, assume that in the survey the anonymous individual x states that she is female and has expertise in logic and privacy. The boss knows that all

* Funded by DFG within the Research Training Group 1907 "RoSI".

employees with expertise logic belong to the formal verification task force and all employees with expertise privacy belong to the security task force. In addition, the employee database contains the information that the members of the first task force are John, Linda, Paul, Pattie and of the second Jim, John, Linda, Pamela. Since Linda is the only female employee belonging to both task forces, the boss can deduce that Linda hides behind the pseudonym x. The question is now whether you can use an automated system to check whether such a breach of privacy can occur in your survey.

The purpose of this paper is to show that ontology reasoners can in principle be used for this purpose. We assume that both the information provided in the survey and the employee database are represented in a DL ABox $\mathcal{A}$, where the employees from the database are represented as known individuals in $\mathcal{A}$ and the pseudonyms used in the survey are represented as anonymous individuals in $\mathcal{A}$. Background information (such as disjointness of the concepts Male and Female, or the connection between expertise and task forces) are represented in a DL TBox $\mathcal{T}$. In order to detect a breach of privacy, we then need to check whether the ontology $\mathfrak{O}$ consisting of $\mathcal{T}$ and $\mathcal{A}$ implies an identity between some anonymous individual x and a known individual a. We call the underlying reasoning task the *identity problem* for $\mathfrak{O}$, x, and a.

In Section 2 we formally introduce the identity problem and show that, for a large class of DLs, this problem is trivial in the sense that no identities between distinct individuals can be deduced from consistent ontologies formulated in these DLs. Not surprisingly, this class consists of the DLs that are fragments of first-order logic without equality. In Section 3, we introduce three DLs for which the identity problem is non-trivial, i.e., the DL $\mathcal{ALCO}$ [10], where nominals allow us to derive identities; $\mathcal{ALCQ}$ [8], where number restrictions allow us to derive identities; and $\mathcal{CFD}_{nc}$ [17], where functional dependencies allow us to derive identities. In Section 4 we show that the identity problem can be reduced in polynomial time to the instance problem, and that for the three DLs mentioned above this actually yields an optimal procedure w.r.t. worst-case complexity. Section 5 considers the identity problem in the context of rôle-based access control [9] to ontologies. Basically, we assume that a user rôle $\hat{r}$ is associated with access to a subset $\mathfrak{O}_{\hat{r}}$ of the ontology.[1] While having rôle $\hat{r}$, the user can access $\mathfrak{O}_{\hat{r}}$ through queries, and can then store the result in a view $V_{\hat{r}}$. In a setting where rôles can dynamically change, the user may have collected (and stored) a sequence of views for different rôles. The question is then whether it is possible to derive the identity of an anonymous individual with a known one using these views. We will show that answering this question can be reduced to the identity problem investigated in the previous sections.

Similar privacy scenarios have been considered for databases [2], but also in the context of ontology-based data access [5,13,4]. In particular, [13] introduces a setting with sub-ontologies and views that is similar to what we consider in Section 5. However, the main difference between these works and ours is that we

[1] To distinguish user rôles from DL roles, we write them with "ô" and also denote specific such rôles with letters with a hat.

concentrate on hiding the *identity* of an anonymous individual with a known one. In contrast, the other works are trying to hide *properties* of known individuals, i.e., the membership of an individual (or a tuple of individuals) in the answers to certain queries.

2 The Identity Problem

We assume that the reader is familiar with the basic notions of Description Logics, as e.g. introduced in [1]. We denote the set of concept names by N_C, the set of role names by N_R, and the set of individual names by N_I. TBoxes and ABoxes as well as their models are assumed to be defined in the standard way. Note, however, that we do *not* make the unique name assumption (UNA) for individual names. An ontology is of the form $\mathfrak{O} = (\mathcal{T}, \mathcal{A})$ where $\mathcal{T}$ is a TBox and $\mathcal{A}$ is an ABox. The *identity problem* asks whether two individuals are equal w.r.t. a given ontology. Since anything (also identities) follows from an inconsistent ontology, we consider this problem only for the case where the ontology is consistent.

Definition 1. *Let $a, b \in N_I$ be distinct individual names and $\mathfrak{O}$ a consistent ontology. Then a is* equal *to b w.r.t. $\mathfrak{O}$ (denoted by $\mathfrak{O} \models a \doteq b$) iff $a^{\mathcal{I}} = b^{\mathcal{I}}$ for all models $\mathcal{I}$ of $\mathfrak{O}$. The* identity problem *for $\mathfrak{O}, a, b$ asks whether $\mathfrak{O} \models a \doteq b$.*

Not all DLs are able to derive equality of individuals. We call those that can *DLs with equality power.*

Definition 2. *$\mathcal{L}$ is a* description logic without equality power *if there is no consistent ontology $\mathfrak{O}$ formulated in $\mathcal{L}$ and two distinct individual names $a, b \in N_I$ such that $\mathfrak{O} \models a \doteq b$. Otherwise we say that $\mathcal{L}$* has equality power.

It is well-known (see, e.g., Chapter 6 in [1]) that many DLs can be translated into first-order predicate logic (FOL). Basically, concept names and role names are translated into unary and binary predicates, respectively, and complex concept descriptions are translated into FOL formulas with one free variable. Individual names are translated into constant symbols and TBoxes and ABoxes into closed formulas. For the translation of some DLs, FOL without equality is sufficient whereas for others equality is needed.

Theorem 1. *If the DL $\mathcal{L}$ can be translated into* FOL without equality*, then it is a DL* without equality power*.*

Proof. Let $\mathfrak{O} = (\mathcal{T}, \mathcal{A})$ be a consistent ontology of $\mathcal{L}$ and $a, b \in N_I$ be two distinct individual names. We must show that $\mathfrak{O} \not\models a \doteq b$. According to our assumption on $\mathcal{L}$, there is an FOL formula ϕ not containing the equality symbol that is equivalent to $\mathfrak{O}$. Consequently, it is sufficient to show that $\phi \not\models a = b$ according to the semantics of FOL, where the equality symbol $=$ is interpreted as equality. Since $\mathfrak{O}$ is consistent, the formula ϕ is satisfiable.

Using well-known approaches and results regarding FOL, we can transform ϕ into a formula ϕ' in Skolem form containing additional function symbols such that (i) ϕ is satisfiable iff ϕ' is satisfiable, and (ii) any model of ϕ' is a model of ϕ. Thus, ϕ' is satisfiable and since it is in Skolem form it has a Herbrand model $\mathcal{I}_H$. Since ϕ' does not contain equality, distinct terms (and thus in particular distinct constants) are interpreted by distinct elements in $\mathcal{I}_H$. Finally, we know that $\mathcal{I}_H$ is also a model of ϕ, which shows that there is a model of ϕ in which a and b are not interpreted by the same domain element. This proves $\phi \not\models a = b$. □

As a consequence of this theorem, we conclude that the basic DL $\mathcal{ALC}$ and its fragments, but also more expressive DLs such as $\mathcal{SRI}$, do not have equality power, and thus the identity problem is trivial for these DLs.

3 Three DLs with Equality Power

In this section, we introduce three DLs that are able to derive equalities between individuals, and for which thus the identity problem is non-trivial. The first two DLs are $\mathcal{ALCO}$, which extends $\mathcal{ALC}$ by *nominals*, and $\mathcal{ALCQ}$, which extends $\mathcal{ALC}$ by qualified number restrictions. Since these DLs are standard, we refer the reader to [1] for the definition of their syntax and semantics. The third DL, called $\mathcal{CFD}_{nc}$ [17], derives its equality power from so-called functional dependencies. We leave it to the reader to verify that expressing these logics in FOL really requires the equality symbol.

Since the DL $\mathcal{CFD}_{nc}$ is less standard and not introduced in [1], we describe it in more detail. Instead of roles, this logic uses attributes, which are interpreted as total functions. We use the symbol N_A to denote the set of all attributes, replacing the set N_R. Concept descriptions C, D of $\mathcal{CFD}_{nc}$ are defined using the following syntax rules:

$$C, D ::= A \mid \neg A \mid C \sqcap D \mid \forall \mathtt{Pf}.C \mid A : \mathtt{Pf}_1, \ldots, \mathtt{Pf}_k \to \mathtt{Pf},$$

where $A \in N_C$, $k \geq 1$, and the *path functions* $\mathtt{Pf}, \mathtt{Pf}_i$ are words in N_A^* with the convention that the empty word is denoted by *id*. A concept descriptions of the form $A : \mathtt{Pf}_1, \ldots, \mathtt{Pf}_k \to \mathtt{Pf}$ is called a *path functional dependency* (PFD). In $\mathcal{CFD}_{nc}$ there is an additional restriction on PFDs to ensure that reasoning in this logic is polynomial: for any PFD of the form above there is an $i, 1 \leq i \leq k$ such that

1. $\mathtt{Pf}$ is a prefix of $\mathtt{Pf}_i$, or
2. $\mathtt{Pf} = \mathtt{Pf}'f$ for $f \in N_A$ and $\mathtt{Pf}'$ is a prefix of $\mathtt{Pf}_i$.

Note that PFDs whose right-hand side $\mathtt{Pf}$ has length ≤ 1 trivially satisfy this restriction.

The interpretation of attributes as total functions is extended to path functions by using composition of functions and interpreting *id* as the identity function. The semantics of atomic negation ($\neg A$) and conjunction ($C \sqcap D$) is defined in the usual way. For the constructors involving path functions, it is defined as follows:

$$(\forall \texttt{Pf}.C)^{\mathcal{I}} := \{d \in \Delta^{\mathcal{I}} \mid \texttt{Pf}^{\mathcal{I}}(d) \in C^{\mathcal{I}}\},$$

$$(A : \texttt{Pf}_1, \ldots, \texttt{Pf}_k \to \texttt{Pf})^{\mathcal{I}} :=$$

$$\{d \in \Delta^{\mathcal{I}} \mid \forall e \in A^{\mathcal{I}}. \left(\bigwedge_{1 \leq i \leq k} \texttt{Pf}_i^{\mathcal{I}}(d) = \texttt{Pf}_i^{\mathcal{I}}(e) \right) \Rightarrow \texttt{Pf}^{\mathcal{I}}(d) = \texttt{Pf}^{\mathcal{I}}(e)\}$$

A TBox $\mathcal{T}$ in $\mathcal{CFD}_{nc}$ consists of a finite set of inclusion dependencies $A \sqsubseteq C$, and an ABox $\mathcal{A}$ consists of a finite set of concept assertions $A(a)$ and path function assertions $\texttt{Pf}_1(a) = \texttt{Pf}_2(b)$, where $A \in N_C$, C is a $\mathcal{CFD}_{nc}$ concept description, $a, b \in N_I$, and $\texttt{Pf}_i \in N_A^*$.

Theorem 2. *The DLs $\mathcal{ALCO}$, $\mathcal{ALCQ}$, and $\mathcal{CFD}_{nc}$ have equality power.*

This theorem is an immediate consequence of the following three examples, which each shows for the respective DL that it can derive equality between different individuals.

Example 1. Here we formulate the example from the introduction in the DL $\mathcal{ALCO}$. Let $\mathfrak{O} = (\mathcal{T}, \mathcal{A})$ where

$$\begin{aligned}\mathcal{T} := \{&\exists \mathsf{expertise}.\{\mathsf{LOGIC}\} \sqsubseteq \mathsf{VerTF}, \quad \exists \mathsf{expertise}.\{\mathsf{PRIVACY}\} \sqsubseteq \mathsf{SecTF},\\ &\mathsf{VerTF} \equiv \{\mathsf{JOHN}, \mathsf{LINDA}, \mathsf{PAUL}, \mathsf{PATTIE}\},\\ &\mathsf{SecTF} \equiv \{\mathsf{JIM}, \mathsf{JOHN}, \mathsf{LINDA}, \mathsf{PAMELA}\}, \quad \mathsf{Female} \sqsubseteq \neg\mathsf{Male}\},\\ \mathcal{A} := \{&\mathsf{Female}(x), \mathsf{expertise}(x, \mathsf{LOGIC}), \mathsf{expertise}(x, \mathsf{PRIVACY}),\\ &\mathsf{Female}(\mathsf{LINDA}), \mathsf{Female}(\mathsf{PATTIE}), \mathsf{Female}(\mathsf{PAMELA}),\\ &\mathsf{Male}(\mathsf{JOHN}), \mathsf{Male}(\mathsf{JIM}), \mathsf{Male}(\mathsf{PAUL})\}.\end{aligned}$$

It is easy to see that $\mathfrak{O} \models x \doteq \mathsf{LINDA}$ since x's expertise implies that she belongs to both the verification and the security task force, but the only female employee belonging to both is Linda.

For the sake of brevity, we use abstract examples to show that $\mathcal{ALCQ}$ and $\mathcal{CFD}_{nc}$ have equality power. It would, however, be easy to provide intuitive examples also for these two DLs.

Example 2. Consider the $\mathcal{ALCQ}$ ontology $\mathfrak{O} = (\mathcal{T}, \mathcal{A})$ where

$$\mathcal{T} := \{A \sqsubseteq \leqslant 1\, r.B\} \quad \text{and} \quad \mathcal{A} := \{A(a), r(a,b), r(a,x), B(a), B(x)\}.$$

Obviously, we have $\mathfrak{O} \models x \doteq b$.

Example 3. Consider the $\mathcal{CFD}_{nc}$ ontology $\mathfrak{O} = (\mathcal{T}, \mathcal{A})$ where

$$\mathcal{T} := \{A \sqsubseteq A : f \to id\} \quad \text{and} \quad \mathcal{A} := \{A(a), f(a) = b, A(x), f(x) = b\}.$$

Since both x and a belong to A and have the same value b for the path function f, the path functional dependency in $\mathcal{T}$ implies that they must be equal, i.e., we have $\mathfrak{O} \models x \doteq a$.

4 The Complexity of the Identity Problem

In this section, we first show that the identity problem can be polynomially reduced to the *instance problem* for all DLs with equality power. Note that the instance problem is one of the basic inference problems for DLs, and thus instance checking facilities are available in most DL reasoners.

Lemma 1. *Let $\mathcal{L}$ be a DL with equality power, $\mathfrak{O} = (\mathcal{T}, \mathcal{A})$ an $\mathcal{L}$ ontology and a, b two distinct individual names. If B is a concept name not occurring in $\mathfrak{O}$, then we have*

$$\mathfrak{O} \models a \doteq b \;\; \textit{iff} \;\; (\mathcal{T}, \mathcal{A} \cup \{B(a)\}) \models B(b).$$

Proof. The direction from left to right is trivial. We show the other direction by contraposition. Thus, assume that $\mathfrak{O} \not\models a \doteq b$. Let $\mathcal{I}$ be a model of $\mathfrak{O}$ such that $a^{\mathcal{I}} \neq b^{\mathcal{I}}$. Let $\mathcal{I}'$ be the interpretation that coincides with $\mathcal{I}$ on all role names, individual names, and concept names different from B. For B we define $B^{\mathcal{I}'} := \{a^{\mathcal{I}}\}$. Since B does not occur in $\mathfrak{O}$, the interpretation $\mathcal{I}'$ is still a model of $\mathcal{T}$ and $\mathcal{A}$, and it satisfies $B(a)$ by our definition of $B^{\mathcal{I}'}$. However, it does not satisfy $B(b)$ since $b^{\mathcal{I}'} = b^{\mathcal{I}} \neq a^{\mathcal{I}}$ does not belong to $B^{\mathcal{I}'}$. □

This lemma shows that the identity problem is at most as complex as the instance problem for all DLs with equality power that allow instance assertions for concept names in the ABox. Since the instance problem is polynomial for $\mathcal{CFD}_{nc}$ [17], this implies that also the identity problem is polynomial for this DL. In [17] it is mentioned that P-hardness of the consistency problem for $\mathcal{CFD}_{nc}$ ontologies is an easy consequence of P-hardness of satisfiability of propositional Horn formulas [3]. It is easy to see that the same is true also for the identity problem.

Theorem 3. *The identity problem is P-complete for $\mathcal{CFD}_{nc}$ ontologies.*

For $\mathcal{ALCO}$ and $\mathcal{ALCQ}$, the instance problem is ExpTime-complete [10,16]. Thus, we obtain exponential-time upper bounds for the identity problem in these DLs. To show that these upper bounds are optimal, we prove that there are polynomial-time reductions of the instance problem in $\mathcal{ALC}$ to the identity problem in these logics. In fact, the instance problem is already ExpTime-hard for the common sub-logic $\mathcal{ALC}$ of $\mathcal{ALCO}$ and $\mathcal{ALCQ}$ [11].

Lemma 2. *Let $\mathcal{L} \in \{\mathcal{ALCO}, \mathcal{ALCQ}\}$, $\mathfrak{O}$ be an $\mathcal{ALC}$ ontology, C an $\mathcal{ALC}$ concept description, and a an individual name. Then we can construct in polynomial time an $\mathcal{L}$ ontology $\mathfrak{O}'$ and individuals a', b' such that*

$$\mathfrak{O} \models C(a) \;\; \textit{iff} \;\; \mathfrak{O}' \models a' \doteq b'.$$

Proof. Let $\mathfrak{O} = (\mathcal{T}, \mathcal{A})$. We consider the two DLs separately.

1.) $\mathcal{L} = \mathcal{ALCO}$:
We define $\mathfrak{O}' := (\mathcal{T} \cup \{C \sqsubseteq \forall r.\{b'\}\}, \mathcal{A} \cup \{r(a, a'), r(a, b')\})$, where a', b' are distinct individual names and r is a role name such that a', b', r do not occur

in $\mathfrak{O}$. The direction from left to right is again trivial. The other direction is shown by contraposition. Let $\mathcal{I}$ be a model of $\mathfrak{O}$ such that $a^{\mathcal{I}} \notin C^{\mathcal{I}}$. We can assume without loss of generality that the domain of $\mathcal{I}$ contains at least two distinct elements $d_1 \neq d_2$.[2] We construct an interpretation $\mathcal{I}'$ that coincides with $\mathcal{I}$ on all concept, role, and individual names occurring in $\mathfrak{O}$, and thus is also a model of $\mathfrak{O}$. In addition, $\mathcal{I}'$ interprets r as $r^{\mathcal{I}'} := \{(a^{\mathcal{I}}, d_1), (a^{\mathcal{I}}, d_2)\}$ and the new individual names as $a'^{\mathcal{I}'} := d_1$ and $b'^{\mathcal{I}'} := d_2$. By construction, $\mathcal{I}'$ satisfies the assertional part of $\mathfrak{O}'$. To see that it also satisfies the GCI $C \sqsubseteq \forall r.\{b'\}$, note that $a^{\mathcal{I}} = a^{\mathcal{I}'}$ is the only element of $\mathcal{I}'$ that has successors w.r.t. the role r. Since it does not belong to $C^{\mathcal{I}} = C^{\mathcal{I}'}$, the elements of $C^{\mathcal{I}'}$ trivially satisfy the value restriction $\forall r.\{b'\}$. Thus, $\mathcal{I}'$ is a model of $\mathfrak{O}'$ in which the individuals a', b' are interpreted by different elements, which shows $\mathfrak{O}' \not\models a' \doteq b'$.

2.) $\mathcal{L} = \mathcal{ALCQ}$:
We define $\mathfrak{O}' := (\mathcal{T} \cup \{C \sqsubseteq \leqslant 1\, r.\top\}, \mathcal{A} \cup \{r(a, a'), r(a, b')\})$, where a', b' are distinct new individuals and r is a new role name not occurring in $\mathfrak{O}$. The direction from left to right is again trivial. To show the other direction, assume that $\mathcal{I}$ is a model of $\mathfrak{O}$ such that $a^{\mathcal{I}} \notin C^{\mathcal{I}}$. Again, we assume without loss of generality that the domain of $\mathcal{I}$ contains at least two distinct elements $d_1 \neq d_2$. We construct an interpretation $\mathcal{I}'$ in the same way as in case 1. above. Also, the argument why $\mathcal{I}'$ is a model of $\mathfrak{O}'$ in which a', b' are interpreted by different elements is identical to the one above. □

As an easy consequence of Lemma 1 and Lemma 2 we obtain the exact complexity of the identity problem in $\mathcal{ALCO}$ and $\mathcal{ALCQ}$. In fact, Lemma 1 yields ExpTime upper bounds. To show that Lemma 2 indeed yields ExpTime lower bounds, we need to take into account the fact that we have defined the identity problem with only consistent ontologies as possible input. However, it is easy to see that ExpTime-hardness of the instance problem in $\mathcal{ALC}$ also holds if we consider the instance problem only for consistent $\mathcal{ALC}$ ontologies $\mathfrak{O}$. In addition, if $\mathfrak{O}$ is a consistent $\mathcal{ALC}$ ontology, then so are the ontologies $\mathfrak{O}'$ constructed from it in the proof of Lemma 1.

Theorem 4. *The identity problem is ExpTime-complete for $\mathcal{ALCO}$ and $\mathcal{ALCQ}$ ontologies.*

For the three DLs with equality power considered in this paper, the identity problem has the same complexity as the instance problem. A natural question to ask is whether this is always the case. A simple example shows that the answer to this question is negative. In fact, let $\mathcal{ALC}^{=}$ be the DL $\mathcal{ALC}$, with the only difference that $\mathcal{ALC}^{=}$ ABoxes may contain equality assertions $a \doteq b$ between individual names. It is easy to see that the identity problem in this DL is non-trivial, but it can be solved in polynomial time. In fact, to check whether a

[2] This follows from the fact that models of $\mathcal{ALC}$ ontologies are closed under disjoint union. Note that this assumption could not be made without loss of generality for $\mathcal{ALCO}$ ontologies. For example, $\mathfrak{O} = (\{\top \sqsubseteq \{a\}\}, \emptyset)$ has only models of size 1.

consistent $\mathcal{ALC}^=$ ontology implies an equality $a \doteq b$, we only need to construct the reflexive, transitive, and symmetric closure of the explicitly stated equalities. However, since $\mathcal{ALC}$ is a sub-logic of $\mathcal{ALC}^=$, the instance problem in this DL is ExpTime-hard (and it is easy to show that it is also in ExpTime).

One may also wonder whether the complexity of the instance problem can be transferred to the identity problem also for DLs where the instance problem has a higher complexity than ExpTime. For example, the DL $\mathcal{ALCOIQ}$, which extends both $\mathcal{ALCO}$ and $\mathcal{ALCQ}$ and additionally allows the use of inverse roles, has a coNExpTime-complete instance problem [15]. Since it contains $\mathcal{ALCO}$, it has equality power and can force models to have cardinality 1. Lemma 1 implies that the identity problem in $\mathcal{ALCOIQ}$ is in coNExpTime. Regarding hardness, the reductions employed in the proof of Lemma 2 can in principle both be used since the constructors employed in them are available in $\mathcal{ALCOIQ}$. However, Lemma 2 uses an $\mathcal{ALC}$ ontology $\mathfrak{O}$ in the reduction, which yields only an ExpTime lower bound. Simply using an $\mathcal{ALCOIQ}$ ontology instead does not work since the proof depends on the fact that $\mathfrak{O}$ has models refuting the instance relation of cardinality at least 2. However, by looking at the coNExpTime-hardness proof for the instance problem in $\mathcal{ALCOIQ}$, it is easy to see that the following modified instance problem is also coNExpTime-hard in $\mathcal{ALCOIQ}$: is a an instance of C in all models of $\mathfrak{O}$ of cardinality ≥ 2. Thus, one can without loss of generality restrict the attention to models of cardinality ≥ 2 when reducing the instance problem for $\mathcal{ALCOIQ}$ to the identity problem for this logic.

Theorem 5. *The identity problem is coNExpTime-complete for $\mathcal{ALCOIQ}$ ontologies.*

5 The View-based Identity Problem

In this section, we will adapt the approach of [13,12] for view-based information hiding such that it can formalize the rôle-based access control scenario sketched in the introduction. We assume that ontologies are written using some DL $\mathcal{L}$ with equality power.

To define what kind of information is to be hidden, we divide the set of individual names into the disjoint sets N_{AI} and N_{KI} consisting of anonymous and known individuals, respectively. As before, we do not make the unique name assumption for these individuals. Given an anonymous individual $x \in N_{AI}$ and an ontology $\mathfrak{O}$, we define the *identity* of x w.r.t. $\mathfrak{O}$ as

$$idn(x, \mathfrak{O}) := \{b \in N_{KI} \mid \mathfrak{O} \models x \doteq b\}.$$

Note that $b, b' \in idn(x, \mathfrak{O})$ implies that $\mathfrak{O} \models b' \doteq b$. Thus, if the cardinality of $idn(x, \mathfrak{O})$ is greater 1, this does not mean that x is equal to one of these individuals, but rather that it is equal to all of them (and thus that all of them are equal). We say that x is a *hidden* if $idn(x, \mathfrak{O}) = \emptyset$.

In the rôle-based access control scenario we assume that there is a "large" *input ontology* $\mathfrak{O}_I$ that is always consistent, but users can only see a part of it depending on which rôle they currently have. More formally, we assume that there is a finite set of *user rôles* $\mathfrak{R}$, and that playing the rôle $\hat{r} \in \mathfrak{R}$ gives access to a subset $\mathfrak{O}_{\hat{r}} \subseteq \mathfrak{O}_I$ of the input ontology. Here "access" does not mean that a user with rôle $\hat{r}$ can download the ontology $\mathfrak{O}_{\hat{r}}$. Instead, the users can ask queries to $\mathfrak{O}_{\hat{r}}$, where a *subsumption query* is of the form $C \sqsubseteq D$ for concept descriptions C, D and a *retrieval query* is of the form C for concept descriptions C or r for role names r.

Definition 3. *Let $\mathfrak{O}_I$ be the input ontology, $\mathfrak{O}_{\hat{r}} \subseteq \mathfrak{O}_I$ the ontology accessible by users with rôle $\hat{r} \in \mathfrak{R}$, and q be a query. The* answer to q w.r.t. $\hat{r}$, *denoted by $\mathit{ans}(q, \hat{r})$, is defined as follows:*

- $\mathit{ans}(q, \hat{r}) := \{\mathtt{true}\}$, *if* $q = C \sqsubseteq D$ *and* $\mathfrak{O}_{\hat{r}} \models C \sqsubseteq D$,
- $\mathit{ans}(q, \hat{r}) := \emptyset$, *if* $q = C \sqsubseteq D$ *and* $\mathfrak{O}_{\hat{r}} \not\models C \sqsubseteq D$,
- $\mathit{ans}(q, \hat{r}) := \{a \in N_I \mid \mathfrak{O}_{\hat{r}} \models C(a)\}$, *if* $q = C$,
- $\mathit{ans}(q, \hat{r}) := \{(a, b) \in N_I \times N_I \mid \mathfrak{O}_{\hat{r}} \models r(a, b)\}$, *if* $q = r$.

Since $\mathfrak{O}_{\hat{r}} \subseteq \mathfrak{O}_I$, *positive answers* to queries, i.e., $\mathit{ans}(C \sqsubseteq D, \hat{r}) = \{\mathtt{true}\}$, $a \in \mathit{ans}(C, \hat{r})$, or $(a, b) \in \mathit{ans}(r, \hat{r})$, imply that this subsumption, instance, or role relationship also holds in $\mathfrak{O}_I$. In contrast, negative answers do not tell us anything about what holds in $\mathfrak{O}_I$ since the inclusion may be strict. Answers to queries w.r.t. rôle $\hat{r}$ can be stored in a view.

Definition 4. *A* view *is a total function $V : \mathit{dom}(V) \to 2^{N_I} \cup 2^{N_I \times N_I} \cup \{\{\mathtt{true}\}\}$ where the* view definition *$\mathit{dom}(V)$ is a finite set of queries and $V(q)$ is a finite set for all $q \in \mathit{dom}(V)$. This view is a* view for $\hat{r} \in \mathfrak{R}$ *(written $\hat{r} \models V$) if $V(q) = \mathit{ans}(q, \hat{r})$ holds for all $q \in \mathit{dom}(V)$.*

In a setting where user rôles can dynamically change, a user may successively play rôles $\hat{r}_1, \hat{r}_2, \ldots, \hat{r}_k$, in each rôle $\hat{r}_i$ generating (and storing) a view $V_{\hat{r}_i}$ for $\hat{r}_i$ by asking queries. The question is now whether these views can be used to find out the identity of a given anonymous individual $x \in N_{AI}$. Assume that the user wants to know whether there is a $b \in N_{KI}$ such that $b \in \mathit{idn}(x, \mathfrak{O}_I)$. However, the user cannot access $\mathfrak{O}_I$ as a whole, all she knows is that the positive answers to the queries in the views $V_{\hat{r}_i}$ are justified by subsets of $\mathfrak{O}_I$. Consequently, instead of one (unknown) ontology $\mathfrak{O}_I$, the user needs to consider all possible ontologies, i.e., all ontologies that are compatible with the positive answers in the views.

Definition 5. *The ontology $\mathfrak{P}$ is a* possible ontology *for the sequence of views $V_{\hat{r}_1}, \ldots, V_{\hat{r}_k}$ if $\mathfrak{P}$ is consistent and compatible with all positive answers in these views, where $\mathfrak{P}$ is* compatible *with*

- $V_{\hat{r}_i}(C \sqsubseteq D) = \{\mathtt{true}\}$ *if* $\mathfrak{P} \models C \sqsubseteq D$,
- $a \in V_{\hat{r}_i}(C)$ *if* $\mathfrak{P} \models C(a)$, *and* $(a, b) \in V_{\hat{r}_i}(r)$ *if* $\mathfrak{P} \models r(a, b)$.

We denote the set of all possible ontologies *for $V_{\hat{r}_1}, \ldots, V_{\hat{r}_k}$ with $\mathtt{Poss}(V_{\hat{r}_1}, \ldots, V_{\hat{r}_k})$. The* certain identity *of x w.r.t. $V_{\hat{r}_1}, \ldots, V_{\hat{r}_k}$ is defined as*

$$\mathit{cert\_idn}(x, V_{\hat{r}_1}, \ldots, V_{\hat{r}_k}) := \bigcap_{\mathfrak{P} \in \mathtt{Poss}(V_{\hat{r}_1}, \ldots, V_{\hat{r}_k})} \mathit{idn}(x, \mathfrak{P}).$$

We say that x is hidden *w.r.t.* $V_{\hat{r}_1}, \ldots, V_{\hat{r}_k}$ *if* $\mathit{cert\_idn}(x, V_{\hat{r}_1}, \ldots, V_{\hat{r}_k}) = \emptyset$.

Since $\mathfrak{O}_I \in \mathtt{Poss}(V_{\hat{r}_1}, \ldots, V_{\hat{r}_k})$, we know that $b \in \mathit{cert\_idn}(x, V_{\hat{r}_1}, \ldots, V_{\hat{r}_k})$ implies that $b \in \mathit{idn}(x, \mathfrak{O}_I)$. Thus, if $\mathit{cert\_idn}(x, V_{\hat{r}_1}, \ldots, V_{\hat{r}_k}) \neq \emptyset$, the identity of x in $\mathfrak{O}_I$ is no longer hidden. Conversely, if $\mathit{cert\_idn}(x, V_{\hat{r}_1}, \ldots, V_{\hat{r}_k}) = \emptyset$, then for all $b \in N_{KI}$ there is a $\mathfrak{P} \in \mathtt{Poss}(V_{\hat{r}_1}, \ldots, V_{\hat{r}_k})$ such that $\mathfrak{P} \not\models x \doteq b$. Since, according to the information available to the user, $\mathfrak{O}_I$ could be this $\mathfrak{P}$, she cannot conclude for any $b \in N_{KI}$ that $\mathfrak{O}_I \models x \doteq b$. This shows that $\mathit{cert\_idn}(x, V_{\hat{r}_1}, \ldots, V_{\hat{r}_k}) = \emptyset$ indeed corresponds to the fact that the views $V_{\hat{r}_1}, \ldots, V_{\hat{r}_k}$ do not disclose the identity of x.

Since the set $\mathtt{Poss}(V_{\hat{r}_1}, \ldots, V_{\hat{r}_k})$ consists of infinitely many ontologies, the definition of $\mathit{cert\_idn}(x, V_{\hat{r}_1}, \ldots, V_{\hat{r}_k})$ does not directly yield an approach for computing this set. We will now show that we can reduce this computation to the identity problem for the *canonical ontology* of $V_{\hat{r}_1}, \ldots, V_{\hat{r}_k}$. Basically, this ontology consists of the GCIs, concept assertions, and role assertions obtained from the positive answers in the views.

Definition 6. *The* canonical ontology $\mathcal{C}(V_{\hat{r}_1}, \ldots, V_{\hat{r}_k})$ *of* $V_{\hat{r}_1}, \ldots, V_{\hat{r}_k}$*is defined as* $\mathcal{C}(V_{\hat{r}_1}, \ldots, V_{\hat{r}_k}) := (\mathcal{T}, \mathcal{A})$ *where*

$$\begin{aligned}
\mathcal{T} &:= \{C \sqsubseteq D \mid V_{\hat{r}_i}(C \sqsubseteq D) = \{\textit{\texttt{true}}\} \textit{ for some } i, 1 \leq i \leq k\} \\
\mathcal{A} &:= \{C(a) \mid a \in V_{\hat{r}_i}(C) \textit{ for some } i, 1 \leq i \leq k\} \cup \\
& \{r(a,b) \mid (a,b) \in V_{\hat{r}_i}(r) \textit{ for some } i, 1 \leq i \leq k\}.
\end{aligned}$$

Since $\mathcal{C}(V_{\hat{r}_1}, \ldots, V_{\hat{r}_k})$ consists of all positive answers in the views $V_{\hat{r}_1}, \ldots, V_{\hat{r}_k}$, it clearly implies them, and thus $\mathcal{C}(V_{\hat{r}_1}, \ldots, V_{\hat{r}_k}) \in \mathtt{Poss}(V_{\hat{r}_1}, \ldots, V_{\hat{r}_k})$. Conversely, every ontology $\mathfrak{P} \in \mathtt{Poss}(V_{\hat{r}_1}, \ldots, V_{\hat{r}_k})$ implies all these positive answers, and thus all the GCIs, concept assertions, and role assertions in $\mathcal{C}(V_{\hat{r}_1}, \ldots, V_{\hat{r}_k})$. This implies that every consequence of $\mathcal{C}(V_{\hat{r}_1}, \ldots, V_{\hat{r}_k})$ is also a consequence of $\mathfrak{P}$.

Theorem 6. *Given views* $V_{\hat{r}_1}, \ldots, V_{\hat{r}_k}$ *and an anonymous individual* $x \in N_{AI}$*, we have* $\mathit{cert\_idn}(x, V_{\hat{r}_1}, \ldots, V_{\hat{r}_k}) = \mathit{idn}(x, \mathcal{C}(V_{\hat{r}_1}, \ldots, V_{\hat{r}_k}))$.

Proof. First assume that $b \in \mathit{cert\_idn}(x, V_{\hat{r}_1}, \ldots, V_{\hat{r}_k})$. Then we have $\mathfrak{P} \models x \doteq b$ for all $\mathfrak{P} \in \mathtt{Poss}(V_{\hat{r}_1}, \ldots, V_{\hat{r}_k})$. Since $\mathcal{C}(V_{\hat{r}_1}, \ldots, V_{\hat{r}_k}) \in \mathtt{Poss}(V_{\hat{r}_1}, \ldots, V_{\hat{r}_k})$, this yields $\mathcal{C}(V_{\hat{r}_1}, \ldots, V_{\hat{r}_k}) \models x \doteq b$, and thus $b \in \mathit{idn}(x, \mathcal{C}(V_{\hat{r}_1}, \ldots, V_{\hat{r}_k}))$.

Conversely, assume $b \in \mathit{idn}(x, \mathcal{C}(V_{\hat{r}_1}, \ldots, V_{\hat{r}_k}))$, and thus $\mathcal{C}(V_{\hat{r}_1}, \ldots, V_{\hat{r}_k}) \models x \doteq b$. We must show that, for all $\mathfrak{P} \in \mathtt{Poss}(V_{\hat{r}_1}, \ldots, V_{\hat{r}_k})$, we have $\mathfrak{P} \models x \doteq b$. This is an immediate consequence of the fact that all the consequences of $\mathcal{C}(V_{\hat{r}_1}, \ldots, V_{\hat{r}_k})$ are also consequences of $\mathfrak{P}$. □

This theorem shows that, to check whether x is *hidden* w.r.t. $V_{\hat{r}_1}, \ldots, V_{\hat{r}_k}$, it is sufficient to compute $\mathit{idn}(x, \mathcal{C}(V_{\hat{r}_1}, \ldots, V_{\hat{r}_k}))$. In case the employed ontology language $\mathcal{L}$ allows for unrestricted GCIs, concept assertions, and role assertions, the set $\mathit{idn}(x, \mathcal{C}(V_{\hat{r}_1}, \ldots, V_{\hat{r}_k}))$ can clearly be computed using an algorithm that solves the identity problem for $\mathcal{L}$ ontologies a polynomial number of times. Note that this applies to the DLs $\mathcal{ALCO}$, $\mathcal{ALCQ}$, and $\mathcal{ALCOIQ}$ considered in the previous sections, but not to $\mathcal{CFD}_{nc}$ since there GCIs and concept assertions need to satisfy certain restrictions.

Corollary 1. *For $\mathcal{L} \in \{\mathcal{ALCO}, \mathcal{ALCQ}\}$ we can check in exponential time whether an anonymous individual x is hidden w.r.t. views $V_{\hat{r}_1}, \ldots, V_{\hat{r}_k}$. For $\mathcal{L} = \mathcal{ALCOIQ}$, this problem can be solved in NExpTime.*

The ExpTime upper bound for $\mathcal{ALCO}$ and $\mathcal{ALCQ}$ is obvious. For $\mathcal{ALCOIQ}$, one considers all the (polynomially many) known individuals $a_1, \ldots, a_p$. Using a NExpTime procedure for the complement of the identity problem, one then checks whether x is not identical to a_1. The non-successful paths of this nondeterministic computation stop with failure whereas the successful ones continue with the same test for a_2, etc. It is easy to see that this yields the desired NExpTime procedure. In fact, any path of this procedure has only exponential length, and a successful path indicates that inequality with x holds for all known individuals.

6 Conclusions and Future Work

In this paper, we have provided some initial definitions and results regarding the identity problem in DL ontologies, i.e., the question whether the ontology implies that a given anonymous individual is equal to a known individual. We have also considered a more involved rôle-based access control scenario where users can access parts of the ontology depending on their rôle. In a setting where users can change rôles dynamically, the question is then whether, by changing rôles and asking queries in these rôles, the user can find out the identity of an anonymous individual although this may not be possible for a single rôle. We have shown how to use the identity problem to address this question.

Until now, we have only investigated how to find out whether the identity of an anonymous individual is disclosed in a certain situation. We have not considered what to do when this is the case. One possibility would be to additionally anonymize the available information, e.g., by replacing some of the known individuals in assertions by new anonymous ones, similar to what is done in [6].

Another direction for future research could be to look at k-anonymity [14] rather than identity. In principle, our identity problem is concerned with 1-anonymity, i.e., we want to avoid that one can deduce from the given information that an anonymous individual belongs to a singleton set consisting of only one known individual. In many applications, one also wants to ensure that the set of known individuals to which the anonymous one is known to belong has a large enough cardinality, i.e., one $> k$. Of course, in this setting additional anonymization (as mentioned above) is also relevant in cases where k-anonymity is not given.

Finally, we intend to consider cases where the information about known and anonymous individuals holds only with a certain probability, e.g., using ontologies with subjective probability as introduced in [7]. In this setting, equality can also only be derived with a certain probability.

References

1. F. Baader, D. Calvanese, D. L. McGuinness, D. Nardi, and P. F. Patel-Schneider, editors. *The Description Logic Handbook: Theory, Implementation, and Applications.* Cambridge University Press, New York, NY, USA, 2003.
2. J. Biskup and P. A. Bonatti. Controlled query evaluation for enforcing confidentiality in complete information systems. *Int. J. Inf. Sec.*, 3(1):14–27, 2004.
3. S. Cook and P. Nguyen. *Logical Foundations of Proof Complexity.* Cambridge University Press, 1st edition, 2010.
4. B. C. Grau. Privacy in ontology-based information systems: A pending matter. *Semantic Web*, 1:137–141, 2010.
5. B. C. Grau and I. Horrocks. Privacy-preserving query answering in logic-based information systems. In *In Proc. of the 18th European Conference on Artificial Intelligence*, pages 40–44, 2008.
6. B. C. Grau and E. V. Kostylev. Logical foundations of privacy-preserving publishing of linked data. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pages 943–949. AAAI Press, 2016.
7. V. Gutiérrez-Basulto, J. C. Jung, C. Lutz, and L. Schröder. Probabilistic description logics for subjective uncertainty. *J. Artif. Intell. Res. (JAIR)*, 58:1–66, 2017.
8. B. Hollunder and F. Baader. Qualifying number restrictions in concept languages. In *Proc. of the 2nd Int. Conf. on the Principles of Knowledge Representation and Reasoning (KR'91)*, pages 335–346, 1991.
9. R. S. Sandhu, E. J. Coyne, H. L. Feinstein, and C. E. Youman. Role-based access control models. *Computer*, 29(2):38–47, Feb. 1996.
10. A. Schaerf. Reasoning with individuals in concept languages. *Data Knowl. Eng.*, 13(2):141–176, 1994.
11. K. Schild. A correspondence theory for terminological logics: Preliminary report. In *Proc. of the 12th Int. Joint Conf. on Artificial Intelligence (IJCAI'91)*, pages 466–471, 1991.
12. P. Stouppa and T. Studer. A formal model of data privacy. In I. Virbitskaite and A. Voronkov, editors, *Proceedings of Perspectives of System Informatics*, volume 4378 of *Lecture Notes in Computer Science*, pages 401–411. Springer, 2007.
13. P. Stouppa and T. Studer. Data privacy for $\mathcal{ALC}$ knowledge bases. In *Proc. of the International Symposium on Logical Foundations of Computer Science*, volume 5407 of *Lecture Notes in Computer Science*, pages 409–421. Springer, 2009.
14. L. Sweeney. K-anonymity: A model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, 10(5):557–570, 2002.
15. S. Tobies. The complexity of reasoning with cardinality restrictions and nominals in expressive description logics. *J. of Artificial Intelligence Research*, 12:199–217, 2000.
16. S. Tobies. Complexity results and practical algorithms for logics in knowledge representation. *CoRR*, cs.LO/0106031, 2001. PhD thesis, RWTH Aachen.
17. D. Toman and G. E. Weddell. Conjunctive query answering in $\mathcal{CFD}_{nc}$: A PTIME description logic with functional constraints and disjointness. In *Proc. of the 26th Australasian Joint Conference on Advances in Artificial Intelligence*, volume 8272 of *Lecture Notes in Computer Science*, pages 350–361. Springer, 2013.