Privacy Preserving Modules for Ontologies

Thomas Studer

Universität Bern, Institut für Informatik und angewandte Mathematik,
Neubrückstrasse 10, CH-3012 Bern, Switzerland,
tstuder@iam.unibe.ch

Abstract. Data privacy is an important application of ontology modularization. The aim is to publish one module while keeping the information of another module private. We show how locality and partitioning - two basic concepts in the theory of modular ontologies - naturally lead to privacy preserving query answering over modular ontologies.

1 Introduction

Recently, big effort has been made to understand modules in the context of ontologies and description logic. The problems studied in that context are to find formalisms for combining OWL ontologies as well as methods for decomposing ontologies. These issues mainly are investigated in order to enable safe ontology reuse and to obtain better reasoning algorithms.

We believe that there is another important application of ontology modularization, namely data privacy for ontologies. If we are given a modular ontology, then it should be possible to publish a module while keeping the information of another module private. We show how concepts of modular ontologies, such as *locality* and *partitioning*, naturally lead to privacy preserving modules.

The privacy notion we study is *provable data privacy* which has been introduced in the context of relational database systems [1]. This notion has later been extended to logic based systems in [2]. Assume we are given a set of axioms T (which can be seen as general public background knowledge, the database schema, or an ontology) and a public view definition V. A view V_I is possible if it may be the answer an agent obtains when issuing the queries of V. We say privacy is preserved for a query C if for no possible view V_I the agent can infer from T and V_I that an individual a belongs to the answer of C. In database systems this is formalized as the set of certain answers to C is empty with respect to T and V_I. For logic based systems this is equivalent to saying that T and V_I do not entail $a : C$ for any a.

This paper is organized as follows. In the next section we introduce the expressive description logic $\mathcal{SHOIQ}$ for which we will state our privacy results. Further we recall the definitions of provable data privacy in the context of description logic. In Section 3, we present a first privacy result which is based in the notion of locality. Intuitively, a concept C is local with respect to a signature S if we can interpret C by the empty set no matter how S is interpreted. This leads immediately to a privacy result since having an interpretation I where

C^I is empty means that $a : C$ cannot be inferred for any a. Then in Section 4 we investigate data privacy based on partitioning of ontologies. This allows us to include in the public view definition other queries than in the locality based approach. Finally we discuss related work and conclude.

2 Technical Preliminaries

In the first part of this section we introduce the description logic $\mathcal{SHOIQ}$, see [3], which underlies modern ontology languages such as OWL. In the second part we recall the notion of provable data privacy from [1].

A $\mathcal{SHOIQ}$ signature S is the disjoint union of a set of role names R, a set of concept names C, and a set of nominals I. A $\mathcal{SHOIQ}$ role is either $R \in \mathsf{R}$ or an inverse role R^- for $R \in \mathsf{R}$. The set of $\mathcal{SHOIQ}$ concepts C is given by the following grammar

$$C ::== A \mid j \mid \neg C \mid C \sqcap C \mid \exists R.C \mid \geq nS.C$$

where $A \in \mathsf{C}$, $j \in \mathsf{I}$, and R, S are roles where S is a simple role[1], and n is a positive integer. We use the abbreviations: $C \sqcup D := \neg(\neg C \sqcap \neg D)$, $\forall R.C := \neg\exists R.\neg C$, and $\leq nS.C := \neg(\geq n+1 S.C)$.

A $\mathcal{SHOIQ}$ TBox is a finite set of role inclusion axioms $R_1 \sqsubseteq R_2$ where R_i are roles, transitivity axioms $\mathsf{trans}(R)$ where $R \in \mathsf{R}$, and general concept inclusion axioms $C_1 \sqsubseteq C_2$ where C_i are concepts. The signature $\mathsf{sig}(T)$ of a TBox T is the set of symbols occurring in T. Similarly, we define the signature of an axiom and of a concept, respectively.

An *interpretation* I for the signature S is a par $(\Delta^I, \cdot^I)$ where Δ^I is a non-empty set (called the domain) and $\cdot^I$ is the interpretation function such that $R^I \subseteq \Delta^I \times \Delta^I$ for each $R \in \mathsf{R}$, $C^I \subseteq \Delta^I$ for each $C \in \mathsf{C}$, and j^I is a singleton subset of Δ^I for each $j \in \mathsf{I}$. The interpretation function extends to complex roles by $(R^-)^I := \{(y,x) : R^I(x,y)\}$ and to concepts by:

$$\begin{aligned}(\neg C)^I &:= \Delta^I \setminus C^I \\ (C \sqcap D)^I &:= C^I \cap D^I \\ (\exists R.C)^I &:= \{x : \exists y(R^I(x,y) \wedge C^I(y))\} \\ (\geq nR.C)^I &:= \{x : \sharp\{y : R^I(x,y) \wedge C^I(y)\} \geq n\}.\end{aligned}$$

We say $I \models R_1 \sqsubset R_2$ iff $R_1^I \subseteq R_2^I$, $I \models \mathsf{trans}(R)$ iff R^I is transitive, and $I \models C \sqsubset D$ iff $C^I \subseteq D^I$. An interpretation I is a *model of a TBox* T ($I \models T$) iff it is a model of all axioms of T. A TBox is *consistent* if it has a model. A TBox T *entails* an axiom α ($T \models \alpha$) iff $I \models T$ implies $I \models \alpha$ for each I.

In this paper we restrict ourselves to the case of data privacy with respect to retrieval queries. Since our ontology language includes nominals, we do not need to introduce individuals. Informally, the statement that an individual a belongs to a concept C can be expressed as $\{a\} \sqsubseteq C$. Therefore we will treat nominals as individuals and write $j : C$ for $j \sqsubseteq C$ when $j \in \mathsf{I}$.

[1] See [3] for a precise definition of simple roles.

Definition 1 (Query, answer, view).

1. *A* retrieval query *is a concept C.*
2. *The* answer *to a query C with respect to a TBox T is the set of all nominals $a \in \mathsf{I}$ for which $T \models a : C$.*
3. *A* view definition *is a finite set of queries.*
4. *A* view *V_I of a view definition V is a finite set of axioms of the form $a : C$ such that if $a : C$ is an element of V_I, then $C \in V$.*
5. *A view V_I is* possible *with respect to a TBox T and a view definition V, if V_I is a view of V and $T \cup V_I$ is consistent.*

In [1] we introduced the notion of provable data privacy. It turned out that for the setting we introduced above, provable data privacy can be reduced to entailment, see [2]. We make use of this fact here to give the following definition of data privacy.

Definition 2 (Data privacy).

1. *Given a TBox T, a view V_I, and a query C, we say that* privacy is preserved *for C with respect to T and V_I if the set of answers to C with respect to $T \cup V_I$ is empty.*
2. *Given a TBox T, a view definition V, and a query C, we say that* privacy is preserved *for C with respect to T and V if for all views V_I that are possible with respect to T and V we have that privacy is preserved for C with respect to T and V_I.*

3 Locality based privacy

We prove a first privacy theorem based on the notion of locality which was first introduced in [4] in order to provide a logical framework for modular ontologies. A similar theorem for subsumption queries and $\mathcal{SHIQ}$ TBoxes is shown in [5].

Definition 3 (Trivial expansion). *An S-interpretation $J = (\Delta^J, \cdot^J)$ is an* expansion *of an S'-interpretation $I = (\Delta^I, \cdot^I)$ if $\mathsf{S}' \subseteq \mathsf{S}$, $\Delta^J = \Delta^I$, and $X^J = X^I$ for every $X \in \mathsf{S}'$. A* trivial expansion *of I to S is an expansion J of I such that $X^J = \emptyset$ for every role name and concept name $X \in \mathsf{S} \setminus \mathsf{S}'$.*

Definition 4 (Locality). *Let S be a signature.*

1. *A concept A is* positively local wrt. S *if for every trivial expansion J of any S-interpretation to any $\mathsf{S}' \supseteq \mathsf{S} \cup \mathsf{sig}(A)$ we have $A^J = \emptyset$.*
2. *An axiom α is* local wrt. S *if every trivial expansion J of any S-interpretation to any $\mathsf{S}' \supseteq \mathsf{S} \cup \mathsf{sig}(\alpha)$ is a model of α.*

Note that the definition of locality implies that an axiom containing a nominal j cannot be local wrt. S if $j \notin \mathsf{S}$.

Grau et al. [6] show how locality can be tested by standard DL reasoners. Although for $\mathcal{SHOIQ}$ this is a NEXPTIME-complete problem, the locality test

will often perform well in practice. However, they also present a tractable approximation to the locality condition which is based on the syntactic structure of concepts.

In order to state our first privacy theorem we make the following assumptions. Let P and S be two signatures with $\mathsf{P} \subseteq \mathsf{S}$. Let T be a TBox over S and let $T_P \subseteq T$ be those axioms of T that are built from the signature P only. Further, we assume that all axioms of $T \setminus T_P$ are local wrt. P.

Theorem 1. *Let C be a positive local query wrt. P. Let V be a view definition which contains only queries over P. Then data privacy is preserved for C with respect to the TBox T and the view definition V.*

Proof. Let

$$V_I \text{ be a possible view with respect to } T \text{ and } V. \tag{1}$$

Since V contains only concepts of P, we find that $\mathsf{sig}(V_I) \setminus \mathsf{P}$ consists of nominals only. Therefore

$$C \text{ is positively local wrt. } \mathsf{P} \cup \mathsf{sig}(V_I), \tag{2}$$

$$\text{all axioms of } T \setminus T_P \text{ are local wrt. } \mathsf{P} \cup \mathsf{sig}(V_I). \tag{3}$$

Because of (1) there exists a $\mathsf{P} \cup \mathsf{sig}(V_I)$-interpretation I such that $I \models T_P$ and $I \models V_I$. Let J be a trivial expansion of I to $\mathsf{S} \cup \mathsf{sig}(C)$. Thus by (3) and the definition of locality we immediately get for each $\alpha \in T \setminus T_P$ that $J \models \alpha$. Therefore we have $J \models T \cup V_I$. Moreover, by (2) we find $C^J = \emptyset$. Since V_I was arbitrary, we conclude that privacy is preserved for C. □

4 Partition based privacy

The assumption in the previous theorem that the view only consists of queries over P may be too restrictive in practice. In this section, we will present a privacy result that is based on partitioning an ontology T in a public part T_P and a private (hidden) part T_H. The public view definition V may now contain queries that access T_H. However, this access will occur only via quantifiers and these quantifiers serve the purpose of information hiding. Therefore privacy will be preserved for positively local concepts of T_H.

Definition 5 (Safe TBox).

1. *A TBox is called* safe *if all its axioms are local with respect to $\emptyset$.*
2. *A concept is* positively local *if it is positively local with respect to $\emptyset$.*

In [7] an algorithm is presented to generate modules from a safe ontology. We use this algorithm to produce a partitioning of a TBox T such that $T = T_P \cup T_H$ where T_H and T_P are disjoint. Moreover this algorithm gives a function $\mathcal{V}$ such that

1. $\mathcal{V}$ assigns to each concept A in $\mathsf{sig}(T)$ either 1 or 2, and

2. $\mathcal{V}$ assigns to each role R in $\mathsf{sig}(T)$ a pair (i, j) with $i, j \in \{1, 2\}$.

The semantic counterpart of the partitioning of a TBox is given by the following construction which is used in the proof of Theorem 3 in [7]. Let $I = (\Delta^I, \cdot^I)$ be a model for the TBox T. We define an interpretation J as follows.

1. For each $x \in \Delta^I$ we generate two new objects x_1 and x_2. We then set $\Delta_1^J := \{x_1 : x \in \Delta^I\}$, $\Delta_2^J := \{x_2 : x \in \Delta^I\}$, and $\Delta^J := \Delta_1^J \cup \Delta_2^J$.
2. For each concept name A with $\mathcal{V}(A) = i$ we set $A^J := \{x_i : x \in A^I\}$.
3. For each role name R with $\mathcal{V}(R) = (i, j)$ we set $R^J := \{(x_i, y_j) : (x, y) \in R^I\}$.

It is easy to see that

1. $\Delta_1^J \cap \Delta_2^J \neq \emptyset$,
2. $A^J \subseteq \Delta_i^J$ for each concept name A with $\mathcal{V}(A) = i$, and
3. $R^J \subseteq \Delta_i^J \times \Delta_j^J$ for each role name R with $\mathcal{V}(R) = (i, j)$.

As in [7] we can show the following lemma.

Lemma 1. *For every concept C with $\mathcal{V}(C) = i$ we have:*

1. *if C is positively local, then $C^J = \{x_i : x \in C^I\}$,*
2. *if C is not positively local, then $C^J = \Delta_{j \neq i}^J \cup \{x_i : x \in C^I\}$.*

From this we immediately get the following theorem, again see [7] for a proof.

Theorem 2. *Let T be a safe TBox and I be a model of T. Let J be the interpretation given above. Then J also is a model of T.*

Next we introduce the notion of an open concept. We will then prove that privacy is preserved for positively local concepts C with $\mathcal{V}(C) = 2$ with respect to the TBox T and any view definition which consists of open concepts only. This privacy result is based on the fact that the view definition (consisting of open concepts) accesses private information only via quantifiers. These quantifiers serve the purpose of information hiding.

Definition 6. *Let T, T_P, T_H, and $\mathcal{V}$ as above. The open concepts are inductively defined by the following clauses.*

1. *A concept C is open if $\mathcal{V}(C) = 1$.*
2. *$C \sqcup D$ and $C \sqcap D$ are open if both C and D are open.*
3. *$\neg C$ is open if C is a positively local concept with $\mathcal{V}(C) = 2$.*
4. *$\exists R.C$ and $\geq nR.C$ are open if $\mathcal{V}(R)) = (1, 2)$ and $\mathcal{V}(C) = 2$.*
5. *$\exists R.C$ and $\geq nR.C$ are open if $\mathcal{V}(R)) = (1, 1)$ and C is an open concept.*
6. *$\forall R.C$ and $\leq nR.C$ are open if $\mathcal{V}(R)) = (1, 2)$ and $\mathcal{V}(C) = 2$.*
7. *$\forall R.C$ and $\leq nR.C$ are open if $\mathcal{V}(R)) = (1, 1)$ and C is an open concept.*

An open view definition is a view definition that consists of open concepts only.

Theorem 3. *Let T be a safe TBox as above. Let V be an open view definition. Let C be a positively local concept with $\mathcal{V}(C) = 2$. Then privacy is preserved for C with respect to T and V.*

Proof. Assume we are given a view V_I based on V and a model I of T and V_I. We define the interpretation J as above where we additionally define

$$a^J := \{x_1 : \{x\} = a^I\} \text{ for each nominal } a \in \mathsf{sig}(V_I). \tag{4}$$

By Theorem 2, we know that J models T. We now show that J also is a model of V_I. Let $a : D$ be an assertion on V_I for an open concept D. We show by induction on the structure of D that $\{x_1 : x \in D^I\} \subseteq D^J$.

1. D is a concept with $\mathcal{V}(D) = 1$. In this case our claim follows from Lemma 1.
2. D is of the form $E \sqcup F$ or $E \sqcap F$ with E and F being open. The claim is an immediate consequence of applying the induction hypothesis to E and F.
3. D is of the form $\neg E$ where E is a positively local concept with $\mathcal{V}(E) = 2$. We find by Lemma 1 that $\Delta_1^J \subseteq D^J$. Therefore we have $\{x_1 : x \in D^I\} \subseteq D^J$.
4. D is of the form $\exists R.E$ or $\geq nR.E$ for (i) a role name R with $\mathcal{V}(R) = (1, 2)$ and a concept E with $\mathcal{V}(E) = 2$ or (ii) R with $\mathcal{V}(R) = (1, 1)$ and an open concept E. Assume there are x, y such that $R^I(x, y)$ and $E^I(y)$. In case (i) we find $R^J(x_1, y_2)$ by the definition of J and by Lemma 1 we find $E^J(y_2)$. In case (ii) we find $R^J(x_1, y_1)$ and applying the induction hypothesis to E yields $E^J(y_1)$. Therefore in both cases we conclude $x_1 \in (\exists R.E)^J$. The cases for $\geq nR.E$ are similar.
5. D is of the form $\forall R.E$ or $\leq nR.E$ for (i) a role name R with $\mathcal{V}(R) = (1, 2)$ and a concept E with $\mathcal{V}(E) = 2$ or (ii) R with $\mathcal{V}(R) = (1, 1)$ and an open concept E. Assume $x \in (\forall R.E)^I$. Let y be such that $R^J(x_1, y)$. In case (i) we have that y is of the form z_2 for some z with $R^I(x, z)$. Thus we have $z \in E^I$ and Lemma 1 yields $E^J(y)$. In case (ii) we have that y is of the form z_1 for some z with $R^I(x, z)$. Thus we have $z \in E^I$ and by the induction hypothesis we obtain $y \in E^J$. Therefore in both cases we conclude $x_1 \in (\forall R.E)^J$. The cases for $\leq nR.E$ are similar.

From $\{x_1 : x \in D^I\} \subseteq D^J$ and (4) we conclude that $J \models a : D$. Thus J is a model of T and V_I such that for each nominal $a \in \mathsf{sig}(V_I)$ we have $a^J \in \Delta_1^J$. Since $C^J \subseteq \Delta_2^J$ by Lemma 1, we conclude that privacy is preserved for C. □

Remark 1. We have to be careful when we try to enlarge the class of open concepts. The following examples show that privacy will be violated if we allow additional open concepts. Let C be a concept with $\mathcal{V}(\neg C) = 1$ and $\mathcal{V}(C) = 1$. Further let D be a positively local concept with $\mathcal{V}(D) = 2$. We consider the following cases:

1. Suppose $E \sqcap F$ is open if E is open. Then $V = \{C \sqcap D\}$ is an open view definition. However, the view $a : C \sqcap D$ entails $a : D$.
2. Suppose $E \sqcup F$ is open if E is open. Then $V = \{C \sqcup D, \neg C\}$ is an open view definition. However, $\{a : C \sqcup D, a : \neg C\}$ is a possible view with respect to V which entails $a : D$.
3. Suppose $\neg E$ is open if E is open. Then $V = \{\neg\neg D\}$ is an open view definition. However, the view $\{a : \neg\neg D\}$ entails $a : D$.

Thus in all three cases, there is a possible view with respect to which the set of answers to D is non-empty. Therefore in all three cases privacy is not preserved for D with respect to V.

5 Related work and conclusion

We have introduced the problem of provable data privacy with respect to *views* in [1, 2]. An investigation of privacy with respect to *view definitions* in the context of $\mathcal{ALC}$ ontologies is provided in [8]. Provable data privacy is a privacy notion which corresponds to entailment. Of course there are also other - more fine grained - notions, most prominently perfect privacy [9]. Unfortunately, lack of space does not permit a discussion of them here.

Locality has been introduced in [4] in order to support safe merging of ontologies. That means an ontology can be integrated with a foreign ontology without changing the meaning of the foreign ontology. Later, locality has also been used to support partial reuse of ontologies [6]. There the problem is to find a fragment of an ontology which captures completely the meaning of some terms. The problem of extracting modules from a given ontology has also been addressed in [7] where the partitioning algorithm is presented which is the core to our results in Section 4. It is worth mentioning that the result of partitioning an ontology can be seen as a knowledge base in the language of E-connections [10]. In fact, all models of an E-connection ontology have the form required for Theorem 3.

A basic notion for the study of modularity is the one of a conservative extension, see for instance [11]. Grau and Horrocks [12] establish a tight connection between conservative extensions and privacy guarantees for logic-based information systems. Privacy aware access to ontologies is also addressed in [13] in the context of view-based query answering over ontologies.

Summing up, we have established two privacy theorems stating that given a modular ontology T, a view definition V, and a query C, privacy is preserved for C wrt. T and any possible view of V. Our first result is based on the notion of locality whereas the second one relies on a partitioning algorithm for ontologies.

References

1. Stoffel, K., Studer, T.: Provable data privacy. In Viborg, K., Debenham, J., Wagner, R., eds.: DEXA 2005. Volume 3588 of LNCS., Springer (2005) 324–332
2. Stouppa, P., Studer, T.: A formal model of data privacy. In Virbitskaite, I., Voronkov, A., eds.: PSI'06. Volume 4378 of LNCS., Springer (2007) 401–411
3. Horrocks, I., Sattler, U.: A tableau decision procedure for $\mathcal{SHOIQ}$. J. Autom. Reason. **39**(3) (2007) 249–276
4. Grau, B.C., Horrocks, I., Kazakov, Y., Sattler, U.: A logical framework for modularity of ontologies. In Veloso, M.M., ed.: IJCAI'07. (2007) 298–303
5. Bao, J., Slutzki, G., Honavar, V.: Privacy-preserving reasoning on the semantic web. In: WI 2007. (2007) 791–797
6. Grau, B.C., Horrocks, I., Kazakov, Y., Sattler, U.: Just the right amount: extracting modules from ontologies. In: WWW '07, ACM (2007) 717–726

7. Grau, B.C., Parsia, B., Sirin, E., Kalyanpur, A.: Modularity and web ontologies. In: KR 2006, AAAI Press (2006) 198–209
8. Stouppa, P., Studer, T.: Data privacy for $\mathcal{ALC}$ knowledge bases. In Artemov, S., Nerode, A., eds.: LFCS 2009. Volume 5407 of LNCS., Springer (2009) 409–421
9. Miklau, G., Suciu, D.: A formal analysis of information disclosure in data exchange. In: SIGMOD. (2004)
10. Kutz, O., Lutz, C., Wolter, F., Zakharyaschev, M.: E-connections of abstract description systems. Artifical Intelligence **156**(1) (2004) 1–73
11. Kontchakov, R., Wolter, F., Zakharyaschev, M.: Modularity in DL-Lite. In: DL 2007. Volume 250 of CEUR Workshop Proceedings. (2007)
12. Cuenca Grau, B., Horrocks, I.: Privacy-preserving query answering in logic-based information systems. In: ECAI 2008. (2008)
13. Calvanese, D., De Giacomo, G., Lenzerini, M., Rosati, R.: View-based query answering over description logic ontologies. In: KR 2008. (2008) 242–251