Abstract. Information systems support data privacy by granting access only to certain (public) views. The data privacy problem is to decide whether hidden (private) information may be inferred from the public views and some additional general background knowledge. We study the problem of provable privacy in the context of \textit{ALC} knowledge bases. First we show that the \textit{ALC} privacy problem wrt. concept retrieval and subsumption queries is ExpTime-complete. Then we provide a sufficient condition for data privacy that can be checked in PTime.

1 Introduction

In the context of information systems, the problem of data privacy is to verify whether the confidential information that is stored in a system is not provided to unauthorized users and therefore, personal and other sensitive data remain private. Data privacy issues are particularly critical in environments where sharing and reuse of information are constantly applied.

Such an area is, for example, the semantic web. There, knowledge is represented by ontologies which provide formalizations of concept definitions for an application domain. These ontologies are expressed in an ontology language. OWL (Web Ontology Language) is the W3C endorsed standard language for this purpose. The underlying formal framework of OWL are the so-called description logics [1]. In the present paper we will study the privacy problem with respect to the basic description logic \textit{ALC} which is the simplest description logic that is boolean closed.

It was always clear that privacy issues have to be considered in the context of ontology languages. Let us cite the OWL Language Guide [2]: ‘...the capability to merge data from multiple sources, combined with the inferential power of OWL, does have potential for abuse. Users of OWL should be alert to the potential privacy implications.’

The present paper is the continuation of our work started in [3,4]. There, we introduced the problem of provable data privacy on views as follows. Assume that some agent has access to a view provided by an information system. Additionally, there is some background knowledge that is publicly available. The privacy problem under this setting is to decide whether the user is not able to infer - from the view and the background knowledge - any answer to a given query $q$. That one cannot infer any answer to $q$ is formalized as the set of certain
answers to q is empty. If the problem is answered positively, we say that privacy is preserved for q.

We will now use the notion of provable privacy to study a more general problem: namely, the problem of deciding data privacy on view definitions. The new problem is now the following: given only a view definition instead of a complete view, decide whether privacy is preserved on all possible views of that view definition. We investigate the new problem for the case of ALC knowledge bases with general concept inclusion axioms (GCIs). In such a knowledge base the domain is only partially known (incomplete), background knowledge is formalized as a part of the knowledge base, and for the view and the privacy condition we allow for concept retrieval and subsumption queries.

Let us now illustrate the difference between privacy on views and privacy on view definitions. Our running example will be a business information system storing information about account managers and their salaries.

Example 1. The background knowledge states that an account manager gets a high or a low salary:

\[ \text{account\_manager} = \text{high} \cup \text{low}. \]

Assume that an agent has access to the views defined by

\{ account\_manager, \neg \text{high} \}

and that for some reason the extension of low should be hidden.

For the privacy problem on views, we assume that we are given the answers to the views. For instance, assume \{a\} is the answer of the query account\_manager and \{b\} is the answer to the query \neg \text{high}. In this case, privacy for low is preserved with respect to the given view, since for no individual we can infer that it belongs to low.

For the privacy problem on view definitions, we do not assume that the answers to the views are given. Rather the question is whether privacy is preserved for all possible sets of answers. In our example, privacy is not preserved on the view definition. Consider the following possibility: the answer to the query account\_manager might be \{a, b\} and the answer to the query \neg \text{high} might be \{b\}. In this case b must belong to low. Thus privacy is not preserved for low with respect to the view definition.

In the next section, we present the syntax and the semantics of ALC, explain how a query is answered on an ALC knowledge base, and recall from [3,4] the problem of provable data privacy on a given view. Then, in Section 3 we define data privacy on a view definition. We show that in order to decide this problem it is enough to consider a finite number of possible views. As a corollary we obtain that the problem is ExpTime-complete. Moreover, we present a syntactic condition on the knowledge base and the view definition which is sufficient for data privacy. This condition can be checked in PTime. We discuss related work in Section 4. Then we conclude and give some directions for further work.