Privacy-Preserving Schema Reuse

Nguyen Quoc Viet Hung, Do Son Thanh, Nguyen Thanh Tam, and Karl Aberer

École Polytechnique Fédérale de Lausanne
{quocviethung.nguyen, sonthanh.do, tam.nguyenthanh, karl.aberer}@epfl.ch

Abstract. As the number of schema repositories grows rapidly and several web-based platforms exist to support publishing schemas, schema reuse becomes a new trend. Schema reuse is a methodology that allows users to create new schemas by copying and adapting existing ones. This methodology supports to reduce not only the effort of designing new schemas but also the heterogeneity between them. One of the biggest barriers of schema reuse is about privacy concerns that discourage schema owners from contributing their schemas. Addressing this problem, we develop a framework that enables privacy-preserving schema reuse. Our framework supports the contributors to define their own protection policies in the form of privacy constraints. Instead of showing original schemas, the framework returns an anonymized schema with maximal utility while satisfying these privacy constraints. To validate our approach, we empirically show the efficiency of different heuristics, the correctness of the proposed utility function, the computation time, as well as the trade-off between utility and privacy.

1 Introduction

Schema reuse is a new trend in creating schemas by allowing users to copy and adapt existing ones. The key driving forces behind schema reuse are the slight differences between schemas in the same domain; thus making reuse more realistic. Reusing existing schemas supports to reduce not only the effort of creating a new schema but also the heterogeneity between schemas. Moreover, as the number of publicly available schema repositories (e.g. schema.org[2], Factual[3]) grows rapidly and several web-based platforms (e.g. Freebase[14], Google Fusion Tables[27]) exists to support publishing schemas, reusing them becomes a great interest in both academic and industrial worlds.

One of the biggest barriers of reuse is about privacy concerns that discourage contributors from contributing their schemas[11]. In traditional approaches, all original schemas and their own attributes are presented to users[17]. However, in practical scenarios, the linking of attributes to their containing schemas, namely attribute provenance, is dangerous because of two reasons. First, providing the whole schemas (and all of their attributes) leads to privacy risks and potential attacks on the owner database. Second, since some attributes are the source of revenue and business strategy, the schema contributors want to protect their sensitive information to maintain the competitiveness. As a result, there is a need of developing new techniques to cope with these requirements.

In this paper, we develop a privacy-preserving schema reuse framework that protects the attribute provenance from being disclosed. To this end, our framework enables schema owners to define their own protection policies in terms of privacy constraints. Unlike previous works[17,20,36], we do not focus on finding and ranking schemas
relevant to a user search query. Instead, our framework takes as input these relevant schemas and visualizes them in a unified view, namely *anonymized schema*, which satisfies pre-defined privacy constraints. Constructing such an anonymized schema is challenging because of three reasons. First, defining the representation for an anonymized schema is non-trivial. The anonymized schema should be concise enough to avoid overwhelming but also generic enough to provide comprehensive understanding. Second, for the purpose of comparing different anonymized schemas, we need to define a utility function to measure the amount of information they carry. The utility value must reflect the conciseness and the completeness of an anonymized schema. Third, finding an anonymized schema that maximizes the utility function and satisfies privacy constraints is NP-complete.

The main goal of this paper is to construct an anonymized schema with maximal utility while preserving the privacy constraints, which are defined to prevent an adversary from linking the shown attributes back to the original schemas (and to the schema owners). Our key contributions are summarized as follows.

- We model the setting of schema reuse with privacy constraints by introducing the concept of *affinity matrix* (represents a group of relevant schemas) and *presence constraint* (is a privacy constraint that translates human-understandable policies into mathematical standards).
- We develop a quantitative metric for assessing the utility of an anonymized schema by capturing two important aspects: (i) *attribute importance*—which reflects the popularity of an attribute—and (ii) *completeness*—which reflects the diversity of attributes in the anonymized schema.
- We show the intractability result for the problem of finding an anonymized schema with maximal utility, given a set of privacy constraints. We propose a heuristic-based algorithm for this problem. Through experiments on real and synthetic data, we show the effectiveness of our algorithm.

The paper is organized as follows. Section 2 gives an overview of our approach. Section 3 formally introduces the notion of schema group, anonymized schema and privacy constraint. Section 4 demonstrates the intractability result and the heuristic-based algorithm to the *maximizing anonymized schema* problem. Section 5 empirically shows the efficiency of our framework. Section 6 and 7 present related work and conclusions.

## 2 Overview

**System Overview.** Fig. 1 illustrates a schema reuse system, in which there is a repository of schemas that are willingly contributed by many participants in the same domain. We focus on the scenario in which end-users want to design a new database and reuse the existing schemas as hints, by exploring the repository through search queries [17,20,36] to find schemas of relevance to their applications. In traditional approaches, all the relevant schemas and their whole attributes are shown to users. Nevertheless, it is important to support the contributors to preserve the privacy of their schemas [18] because of several privacy issues. One possible issue is the threats of being attacked and unprivileged accesses to the owner database systems. For example, knowing the schema information (e.g. schema name, attribute names), an adversary can use SQL injection [28] to extract the data without sufficient privileges (details are described in the report [41]). Another possible issue is the policies of schema owners that require hiding