Data Anonymity in Multi-Party Service Model

Shinsaku Kiyomoto, Kazuhide Fukushima, and Yutaka Miyake
KDDI R & D Laboratories Inc.
2-1-15 Ohara, Fujimino-shi, Saitama, 356-8502, Japan
kiyomoto@kddilabs.jp

Abstract. Existing approaches for protecting privacy in public database consider a service model where a service provider publishes public datasets that consist of data gathered from clients. We extend the service model to the multi-service providers setting. In the new model, a service provider obtains anonymized datasets from other service providers who gather data from clients and then publishes or uses the anonymized datasets generated from the obtained anonymized datasets. We considered a new service model that involves more than two data holders and a data user, and proposed a new privacy requirement. Furthermore, we discussed feasible approaches searching a table that satisfies the privacy requirement and showed a concrete algorithm to find the table.

Keywords: k-Anonymity, Privacy, Public DB, Multi-Party.

1 Introduction

Privacy is an increasingly important aspect of data publishing. Sensitive data, such as medical records in public databases, are recognized as a valuable source of information for the allocation of public funds, medical research and statistical trend analysis [1]. Furthermore, secondary-use of personal data has been considered a new market for personalized services. A service provider makes an anonymized dataset from original data, such as records of service use, and distributes the anonymized datasets to other service providers. The service providers can improve their services of using anonymized datasets.

However, if personal private information is leaked from the database, the service will be regarded as unacceptable by the original owners of the data [1]. Thus, anonymization methods have been considered a possible solution for protecting personal information [8]. One class of models, called global-recoding, maps the values of attributes to other values [38] in order to generate an anonymized dataset. Generalization methods modify the original data to avoid identification of the records. These methods generate a common value for some records and replace identifying information in the records with the common value.

Existing approaches consider a service model where a service provider publishes datasets that consist of data gathered from clients. We extend the service model to the multi-service providers setting. That is, a service provider obtains anonymized datasets from other service providers who gather data from clients.
and then publishes or uses the anonymized datasets generated from the obtained anonymized datasets.

In this paper, we considered a new service model that involves more than two data holders and a data user, and proposed a new privacy requirement. Furthermore, we discussed feasible approaches searching a table that satisfies the privacy requirement and showed a concrete algorithm to find the table.

The rest of the paper is organized as follows; section 2 provides related articles. Privacy definitions are summarized in section 3. We presented a new service model in section 4, and then an new adversary model and privacy protection schemes are proposed in section 5. We conclude this paper in section 6.

2 Related Work

Samarati and Sweeney [32,31,35] proposed a primary definition of privacy that is applicable to generalization methods. A data set is said to have \(k\)-anonymity if each record is indistinguishable from at least \(k-1\) other records with respect to certain identifying attributes called quasi-identifiers [10]. Minimizing this information loss thus presents a challenging problem in the design of generalization algorithms. The optimization problem is referred to as the \(k\)-anonymity problem. Meyerson reported that optimal generalization in this regard is an NP-hard problem [29]. Aggarwal et al. proved that finding an optimal table including more than three attributes is NP-hard [2]. Nonetheless, \(k\)-anonymity has been widely studied because of its conceptual simplicity [4,26,27,39,37,33]. Machanavajjhala et al. proposed another important definition of privacy in a public database [26]. The definition, called \(l\)-diversity assumes a strong adversary having certain background knowledge that allows the adversary to identify the object persons in the public database.

There are several methods of generating \(k\)-anonymization tables. Samarati proposed a simple binary search algorithm for finding a \(k\)-anonymous table [31]. A drawback of Samarati’s algorithm is that for arbitrary definitions of minimality, it is not always guaranteed that this binary search algorithm can find the minimal \(k\)-anonymity table. Sun et al. presented a hash-based algorithm that improves the search algorithm [33]. Aggarwal et al. proposed an \(O(k)\)-approximation algorithm [3] for the \(k\)-anonymity problem. A greedy approximation algorithm [23] proposed by LeFevre et al. searches optimal multi-dimensional anonymization. A genetic algorithm framework [19] was proposed because of its flexible formulation and its ability to allow more efficient anonymization. Utility-based anonymization [40] makes \(k\)-anonymous tables using a heuristic local recoding anonymization. Moreover, the \(k\)-anonymization problem is viewed as a clustering problem. Clustering-based approaches [5,36,25,41] search a cluster with \(k\)-records.

Differential Privacy [12,13] is a notion of privacy for perturbative methods based on the statistical distance between two database tables differing by at most one element. The basic idea is that, regardless of background knowledge, an adversary with access to the data set draws the same conclusions, whether