Knowledge-Based Systems 54 (2013) 276–287

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

A general framework for privacy preserving data publishing

A.H.M. Sarowar Sattar *, Jiuyong Li *,[1], Xiaofeng Ding, Jixue Liu, Millist Vincent

School of Information Technology and Mathematical Science, University of South Australia, Mawson Lakes, SA 5095, Australia

ARTICLE INFO

Article history:
Received 11 February 2013
Received in revised form 23 September 2013
Accepted 23 September 2013
Available online 8 October 2013

Keywords:
Privacy protection
Data publishing
Anonymization

ABSTRACT

Data publishing is an easy and economic means for data sharing, but the privacy risk is a major concern in data publishing. Privacy preservation is a major task in data sharing for organizations like bureau of statistics, and hospitals. While a large number of data publishing models and methods have been proposed, their utility is of concern when a high privacy requirement is imposed. In this paper, we propose a new framework for privacy preserving data publishing. We cap the belief of an adversary inferring a sensitive value in a published data set to as high as that of an inference based on public knowledge. The semantic meaning is that when an adversary sees a record in a published data set, s/he will have a lower confidence that the record belongs to a victim than not. We design a method integrating sampling and generalization to implement the model. We compare the method with some state-of-the-art methods on privacy-preserving data publishing experimentally, our proposed method provides sound semantic protection of individuals in data and, provides higher data utility.

1. Introduction

Recently, the privacy preservation in data publishing has received considerable attention from researchers. Compared with data publishing through the format of aggregated results or statistical ones, the release of microdata offers an advantage in terms of information availability, which makes it particularly suitable for scientific analysis in a variety of domains such as public health, and demographic studies. However, the release of microdata causes privacy concerns of disclosing sensitive information of individuals. Simply removing explicit identifiers like names or IDs has been shown to be vulnerable to privacy breach, since other personal identifying attributes, such as age, gender and zip code, called quasi-identifier (QID) which usually remain in the published data for data analysis, allow individuals' sensitive information to be revealed when they are linked with publicly available information. For example, by combining a public voter registration list with a released information of health insurance, Sweeney was able to identify the medical record about a former governor of Massachusetts [1]. Many techniques have been proposed to address the problem.

There are generally two types of definitions for privacy.

One type of definitions is microdata based. k-anonymity [1] and l-diversity [2] are two typical examples. k-anonymity requires that a published data set should have at least k rows (called a group) sharing the same QID value. So the probability for identifying an individual in a published data set is $1/k$. The k-anonymity model protects an individual from being identified in a data set with a high confidence. The l-diversity model requires that the number of the sensitive values in a QID group is at least l. So an adversary could not tell which sensitive value belongs to an individual in a group. There are many improved models (definitions) along this line [3–6]. They all associate with one or a few user specified thresholds, like k and l in the above works, and it is difficult for users to set the right thresholds.

Another type of definitions is probabilistic. Differential privacy [7] is a typical example. It assumes that even if an adversary knows all other sensitive values but the victim's, the adversary could not infer victim's sensitive value when knowing the randomized aggregated result with a certain confidence. This requirement is strong and causes a big utility loss.

Most privacy protection principles are to bind the leakage of sensitive information. In general form, the leakage is the difference between the posterior probability and the prior probability. The posterior possibility is easy to be quantified. However, the prior probability is difficult to estimate, and different estimations lead to different privacy protection models. For example, the l-diversity model assumes the uniform distribution of sensitive values. ϵ-differential privacy does not distinguish between sensitive and non-sensitive attributes. One major disadvantage of such models is that the requirement of a small leakage will cause published data set to have little utility due to, for example, too much generalization or too much noise. We will need to search for an alternative approach for sound privacy protection and better data utility.

* Corresponding authors. Tel.: +61 420863356 (A.H.M. Sarowar Sattar).
E-mail addresses: sarowar@gmail.com (A.H.M. Sarowar Sattar), Jiuyong.Li@unisa.edu.au (J. Li).
[1] D3-07, School of Information Technology and Mathematical Science, Mawson Lakes, SA 5095, Australia.

0950-7051/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.knosys.2013.09.022

In this paper, we explore an alternative model that is semantically sound and gives a published data set more utility. When an adversary accesses a published data set, s/he may infer that a record belongs to a victim (adversary knows that the victim's record must be in the published data set). However, if this record is what everyone expects to see in a data set, for example, a 40–50 old male with flu in a medical data set, does this breach the privacy of the victim? We say no, even if the adversary gets the sensitive value of the victim right (note that we do not mean that flu is not sensitive, and we will elaborate this example more later).

We argue that the damage of a privacy breach is not directly associated with whether the adversary obtains the sensitive value right, but is associated with the confidence level of the inference. For example, if an adversary claims that a victim suffers from prostate cancer with a convincing inference in a published data set, but the claim is wrong since the victim actually suffers from bowel cancer. Even though the inference is wrong, the damage has been made to the victim by the claim. Since the claim is convincing, most people believe in it, and this brings damage to the victim. If an adversary alleges a victim suffering from HIV with a weak inference (as strong as a random guess), the victim will not have to do any defence regardless if the allegation is true or not. Few people will believe in the allegation.

Consequently the importance of privacy protection is not to give an adversary strong belief to build an allegation. If the belief of an allegation in a published data set is the same as the confidence of a random guess, this will be a sufficient protection for the privacy of an individual in data since the believability of an allegation is low. The question is how to model a random guess in a published data set. In this paper, we will discuss a model towards such a protection.

Our idea is that the belief of an adversary obtained from a published data set should be at most the same as the belief obtained from the public knowledge. In other words, when an adversary sees a record in a published data set, the adversary should expect to see the same record in a randomly generated data set following the public knowledge. The occurrence of a record in a published data set does not relate to whether the victim's record is in the published data set or not. In the previous example, the 45 year old male patient does not care the claim that he suffers from flu because the adversary sees a record "40–50, male, flu" in the published data set of a hospital where the patient visited because the adversary is expected to see the same record even if the 45 year old male patient's record is not in the published data set (note that in our model, only a sample of records are published). Therefore, the privacy of the patient is protected.

In this paper, we propose a new framework for privacy preserving data publishing based on the above motivations, and propose an effective hybrid method of sampling and generalization for privacy preserving data publishing. Contributions of the work are listed as the following.

- This new model is semantically sound and offers good data utility. Semantically, it provides a strong protection for the privacy of individuals since it does not give an adversary a stronger belief from an inference in a published data set than the belief from an inference on public knowledge. Practically, it allows many records to be published with a light generalization and a large sample rate. The method integrates generalization with sampling. Sampling is essential in our method. We note that good sampling does not reduce the quality of data. The sampling techniques have been used for many rigorous studies for a long time. Furthermore, a major goal for data publishing is to support the shared data analysis in a large community. In data analysis the aggregated results are often derived. When data sets are randomly sampled, the bias in the aggregated results will be low.
- This model controls privacy risk of individuals at the record level. This supports local generalization of each record irrespective with other records. This provides an easy and effective criterion to judge whether a record is publishable. The method only restricts a few records with values of very low frequencies, such as 95 year old male and Huntington's disease, from being published. It provides good data utility for those publishable records. We note that data publishing is not a right means for data sharing with rare values (for example, some rare diseases). If we try to accommodate those rare cases, the overall quality of published data will suffer badly.
- This model links privacy risk to data set size, which is crucial in privacy risk analysis. The data size has not been utilized in previous data publishing models. For example, consider data sets with 100 records and 100,000 records respectively. Intuitively, an individual in the data set of 100 records has higher privacy risk than an individual in the data set of 100,000 records.

The rest of this paper is organized as follows. Section 2 introduces preliminaries and principle of the new privacy framework. Sections 3 and 4 formally define the way of estimating the adversary's expected confidence and observed confidence respectively, followed by a hybrid method to published data sets after satisfying the new privacy criterion in Section 5. Section 6 shows the experimental results, followed by some related works in Section 7. Finally, Section 8 concludes this paper with future direction.

2. Preliminaries and the principle

A data owner has a data set D_1, where each record t contains information about an individual, like 'id', 'age', 'sex', 'zip code', along with the sensitive information, such as a disease or the salary, of that individual. For simplicity, we consider that there is only one sensitive value in each row (multiple sensitive values can be considered as a set of sensitive values). The attributes that uniquely identify an individual are called unique identifiers (IDs), such as social security number and name. The attributes that potentially conjunctively identify an individual are called quasi-identifiers (QID), such as 'age', 'sex' and 'zip code'. Consider that D_1^* is a published data set of D_1, where the attribute ID has been removed, QID and sensitive attributes are kept in D_1^*. Some of the QID attribute's value may be generalized[2] due to legislation [8].

Now we consider an adversary whose goal is to infer whether a victim individual v has a sensitive value s. We assume that an adversary has the following background knowledge.

Definition 1 (*The background knowledge of an adversary*). We assume that a victim is an individual v in D_1. The adversary knows

1. D_1^*, the published version of D_1.
2. the QID values of v.
3. global statistics of the population from which D_1 has been generated.
4. v is in D_1 and v is in D_1^* with a probability because of sampling used in generating D_1^*.

We note that the adversary uses QID values of v to identify a group in D_1^* containing v to narrow down the possible sensitive values of the victim.

Let us assume that the victim v's record is not in the published data set D_1^*. An adversary is still expected to see a record with the

[2] Generalization of an attribute means its current value is replaced by the value of higher level node from its taxonomy. For example, in Fig. 1(a), if the attribute is 'age' and its value is 20, the generalized value can be 13–25.