Information Sciences 348 (2016) 377–393

Contents lists available at ScienceDirect

Information Sciences

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

Generalized bucketization scheme for flexible privacy settings

Ke Wang[a,*], Peng Wang[a], Ada Waichee Fu[b], Raymond Chi-Wing Wong[c]

[a] School of Computing Science, Simon Fraser University, Burnaby, BC, Canada
[b] Department of Computer Science and Engineering, Chinese University of Hong Kong, Hong Kong
[c] Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Hong Kong

ARTICLE INFO

Article history:
Received 16 January 2015
Revised 4 November 2015
Accepted 27 January 2016
Available online 17 February 2016

Keywords:
Anonymity
Bucketization
Data publishing
Disclosure
Privacy

ABSTRACT

Bucketization is an anonymization technique for publishing sensitive data. The idea is to group records into small buckets to obscure the record-level association between sensitive information and identifying information. Compared to the traditional generalization technique, bucketization does not require a taxonomy of attribute values, so is applicable to more data sets. A drawback of previous bucketization schemes is the uniform privacy setting and uniform bucket size, which often results in a non-achievable privacy goal or excessive information loss if sensitive values have variable sensitivity.

In this work, we present a flexible bucketization scheme to address these issues. In the flexible scheme, each sensitive value can have its own privacy setting and buckets of different sizes can be formed. The challenge is to determine proper bucket sizes and group sensitive values into buckets so that the privacy setting of each sensitive value can be satisfied and overall information loss is minimized. We define the *bucket setting problem* to formalize this requirement. We present two efficient solutions to this problem. The first solution is optimal under the assumption that two different bucket sizes are allowed, and the second solution is heuristic without this assumption. We experimentally evaluate the effectiveness of this generalized bucketization scheme.

1. Introduction

1.1. Motivation

Privacy preserving data publishing is concerned with publishing sensitive data for data analysis while ensuring that no sensitive information about individuals is disclosed. The next example illustrates how sensitive information may be disclosed by answering count queries.

Example 1. A hospital provides online services for answering count queries on medical data T containing three attributes, Gender, Zipcode, and Disease. Disease is sensitive and must not be disclosed, and Gender and Zipcode are public. Suppose that the patient Alice has a record in the data and that an adversary tries to learn the value of Disease for Alice. Knowing that Alice's Zipcode is 61434, the adversary issues two queries Q_1 and Q_2:

[*] Corresponding author. Tel.: +1 7787824667.
E-mail addresses: wangk@cs.sfu.ca (K. Wang), pwa22@sfu.ca (P. Wang), adafu@cse.cuhk.edu.hk (A.W. Fu), raywong@cse.ust.hk (R.C.-W. Wong).

http://dx.doi.org/10.1016/j.ins.2016.01.100

Q_1: SELECT COUNT(*) FROM T WHERE Gender=F AND Zipcode=61434
Q_2: SELECT COUNT(*) FROM T WHERE Gender=F AND Zipcode=61434 AND Disease=HIV

Let x and y be the answers for Q_1 and Q_2, i.e., the number of records matching the description in the WHERE clause. Further assume $x \neq 0$. Such answers are building blocks for many data analysis tasks such as constructing Naive Bayes classifiers. However, an adversary could also use these answers to learn that Alice has HIV with the probability $\frac{y}{x}$.

The above example illustrates a sensitive disclosure through *non-independent reasoning*, where Alice's disease information is learnt from *other* people who share the same Gender and Zipcode as Alice, under the assumption that the diseases of those people follow the same underlying distribution. Non-independent reasoning has a great success in many applications, including classification, prediction, direct marketing, and product recommendation. For example, in classification we learn the class information of a new case from a training data set. Preventing sensitive non-independent reasoning has been a focus of *syntactic models* in the literature, where the raw data is modified, usually by generalization and suppression, in order to bound the change of adversary's beliefs after accessing published information. Some representative syntactic models are k-anonymization [17], ρ_1–ρ_2 privacy [9], ℓ-diversity [16], t-closeness [14], β-likeness [3], Δ-growth [18], to name a few. See surveys [1,4,10] for more details.

As an alternative to syntactic models, the *differential privacy* approach [7,8] does not release the data set itself but releases the answers to user queries using the data set, and instead of preventing non-independent reasoning, it seeks to mask the impact of a single individual by releasing a noisy query answer $x + \Delta x$, where x is the true answer and Δx is the added noise such that the distribution of noisy answers changes little with and without the participation of a single individual. A recent study [11,20] showed that differential privacy is in the following dilemma: while a more restricted setting of ϵ-differential privacy (i.e., a small ϵ) helps protect privacy, the noisy answers have a poor utility for data analysis; while a less restricted setting provides a good utility for data analysis, this good utility also permits sensitive disclosures of non-independent reasoning. The root of this dilemma is that both data analysis and sensitive disclosures are making use of the same information, i.e., noisy answers. Wang et al. [20] shows that with the noises Δy and Δx generated by the ϵ-differential privacy mechanism for the query answers y and x in Example 1, $\frac{y+\Delta y}{x+\Delta x}$ arbitrarily approaches the noise-free probability $\frac{y}{x}$ as the data set grows arbitrarily large. The following quotes from two pioneering works on differential privacy also suggest that differential privacy does not address the privacy violation due to non-independent reasoning.

> Page 3 of [2]: "We explicitly consider nonindependent reasoning as a non-violation of privacy; information that can be learned about a row from sources other than the row itself is not information that the row could hope to keep private."
>
> Page 8 of [7]: "Note that a bad disclosure can still occur, but our guarantee assures the individual that it will not be the presence of her data that causes it, nor could the disclosure be avoided through any action or inaction on the part of the user."

According to [2] and [7], the main reason that a disclosure due to non-independent reasoning is not considered as privacy violation is that such disclosures cannot be avoided through any action or inaction on the part of the user. For this reason, the disclosure of Alice's diseases through $\frac{y+\Delta y}{x+\Delta x}$, when it approaches $\frac{y}{x}$, is not considered as privacy violation by differential privacy, though in practice Alice may not agree with this. Therefore, differential privacy is not suitable when non-independent reasoning does violate an individual's privacy.

In this work, we consider non-independent reasoning of sensitive information as a privacy violation. Under this assumption, syntactic methods through data suppression, generalization, and bucketization remain relevant. See surveys [1,4,10] for more discussions on syntactic methods. Data *suppression* has a large information loss for count queries because suppressed records or values cannot be counted by the query. Data *generalization* is applicable only when there is a taxonomy for each public attributes. Also, generalized data cannot be easily analyzed by standard methods. For example, to reduce the information loss in local recoding, the value "Engineer" could be generalized into the high level value "Professional" in some records while it remains unchanged in other records. Consequently, "Professional" and "Engineer" cannot be treated as two distinct values for counting, which makes it impossible to apply any standard counting based data mining methods such as Naive Beyes classifiers. Data *bucketization* does not have the above problems because it does not generalize domain values. In this work, we shall focus on the bucketization approach.

Unfortunately, previous bucketization schemes [22] have some major drawbacks. First, all previous bucketization schemes use a uniform privacy setting for all sensitive values. For example, the Anatomy algorithm in [22] uses ℓ-diversity to specify the privacy criterion that the frequency of HIV and Flu in a bucket is no more than $1/\ell$. Suppose that HIV is more sensitive than Flu, where Flu occurs more frequently than HIV. ℓ-diversity must be set according to the sensitivity of HIV, i.e., a large ℓ so that $1/\ell$ is small enough for HIV. However, often this setting is not satisfiable for Flu that has a higher frequency in the data. Another drawback of previous bucketization schemes is the uniform size for all buckets, i.e., either ℓ or $\ell + 1$ in [22]. In the example of HIV and Flu, ℓ that is set according to the most sensitive HIV can be very large. A larger bucket size means less association of a record with its original sensitive value in the bucket, thus, more information loss. More discussions about these points will be presented in Section 3.