Journal of Biomedical Informatics xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Small sum privacy and large sum utility in data publishing

Ada Wai-Chee Fu [a,*], Ke Wang [b], Raymond Chi-Wing Wong [c], Jia Wang [d], Minhao Jiang [c]

[a] Department of Computer Science and Engineering, Chinese University of Hong Kong, Hong Kong
[b] Department of Computer Science, Simon Fraser University, Canada
[c] Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Hong Kong
[d] Department of Computer Science, University of Illinois at Urbana-Champaign, USA

ARTICLE INFO

Article history:
Received 4 August 2013
Accepted 1 April 2014
Available online xxxx

Keywords:
Privacy preserving data publishing
Inference attacks
Privacy versus utility

ABSTRACT

While the study of privacy preserving data publishing has drawn a lot of interest, some recent work has shown that existing mechanisms do not limit all inferences about individuals. This paper is a positive note in response to this finding. We point out that not all inference attacks should be countered, in contrast to all existing works known to us, and based on this we propose a model called *SPLU*. This model protects sensitive information, by which we refer to answers for aggregate queries with small sums, while queries with large sums are answered with higher accuracy. Using SPLU, we introduce a sanitization algorithm to protect data while maintaining high data utility for queries with large sums. Empirical results show that our method behaves as desired.

© 2014 Elsevier Inc. All rights reserved.

1. Introduction

In a recent work by Cormode [7], it is shown that despite much progress in two main branches of privacy models for data publishing, namely differential privacy [13], and various *syntactic methods* such as k-anonymity [26] and ℓ-diversity [20], inference-based attacks can still be successful. The study is based on the ability of an attacker to construct accurate classifiers on top of releases protected by state-of-the-art privacy preserving data publishing techniques.

The empirical study result above is in fact consistent with the result from [10]. Following the model in [10], given a dataset $d = (d_1, \ldots, d_n) \in \{0,1\}^n$, a query q is a subset of $\{1, 2, \ldots, n\}$, and its true answer $a_q = \sum_{i \in q} d_i$. Hence, the query q determines a subset of d, and the answer for q is the number of entries in the subset. Given algorithm $\mathcal{A}$ for query response, we say that $\mathcal{A}(q)$ is within ϵ perturbation if it deviates from the true answer by no more than ϵ. $\mathcal{A}$ is within ϵ perturbation if $\mathcal{A}(q)$ is within ϵ perturbation for all q. If an adversary can reconstruct with time complexity $t(n)$ the entire database very accurately, then the database $\mathcal{D} = (d, \mathcal{A})$ is said to be $t(n)$-non-private. The following theorem from [10] says that any privacy preserving algorithm renders the database useless, and conversely utility in the published data implies privacy breach.

Theorem 1 [10]. *Let $\mathcal{D} = (d, \mathcal{A})$ be a database where $\mathcal{A}$ is within $o(\sqrt{n})$ perturbation then $\mathcal{D}$ is poly (n)-non-private.*

The above findings are based on the assumption that all inference attacks are to be defended, and any relatively accurate information derivable from the published data is considered privacy breaching. This is quite inconsistent with the simultaneous requirement of utility whereby minimum distortion is to be introduced so that the published data are as close to the original data as possible. Here we show that the dilemma can be resolved by a segregation of utility and privacy.

The key point as observed by Cormode is that privacy and utility are closely related. As stated in the conclusion in [7], "release of (anonymized) data may reveal hitherto unknown population parameters which compromise individual privacy. ... in some settings, these population statistics may represent exactly the desired utility of the data collection and publication." This remark highlights the issue to be resolved. The key is how to differentiate between utility and privacy. Once we identify the utility of the data and once users agree that this utility has no conflict with their privacy, the proper solution is not to insist on protection for the information related to the utility. We provide a way to differentiate what concepts may be reasonable to be disclosed for utility. If users indeed have concerns about the disclosure of such concepts there is always the option of not releasing any data. This provides for a better alternative for the status quo of releasing the data knowing that certain inference attacks are possible.

To our knowledge, there is no known model for the separation of concepts that need protection and those that need to be

* Corresponding author. Fax: +852 2603 5024.
E-mail addresses: adafu@cse.cuhk.edu.hk (A.W.-C. Fu), wangk@cs.sfu.ca (K. Wang), raywong@cse.ust.hk (R.C.-W. Wong), jwang@cse.cuhk.edu.hk (J. Wang), minhaojiang@gmail.com (M. Jiang).

http://dx.doi.org/10.1016/j.jbi.2014.04.002
1532-0464/© 2014 Elsevier Inc. All rights reserved.

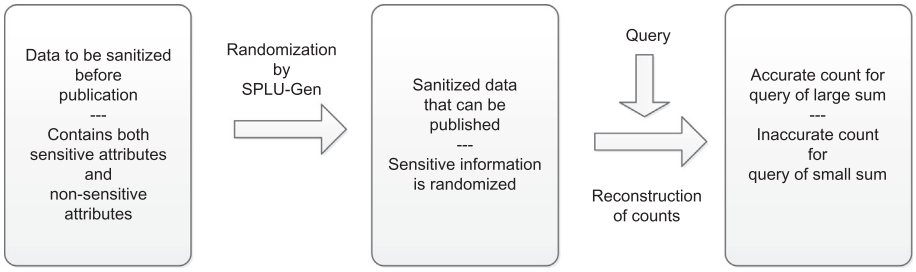

Fig. 1. The *SPLU*-Gen system for data sanitization and query processing.

maintained for utility purposes, in privacy preserving data publishing. Some previous works [11,19] study the adjustment of parameters in the anonymization process for the trade-off between privacy and utility. The problem studied in such works is very different and in their model, all concepts are treated equally in terms of utility and privacy. We assume that aggregate queries of large sums should be answered relatively accurately for utility, while those with very small sums should not. Consider an example from [13] where a dataset D' tells us that almost everyone involved in a dataset is two footed. Knowing with high certainty that an individual is two footed from D' is not considered a privacy issue since it is true for almost everyone in the dataset.[1] Large sum concepts are statistical and of value for utility. In contrast, small count concepts are non-statistical and the protection of small counts has been well-studied in the topic of security in statistical databases [1].

Our main contributions are summarized as follows.

(1) We propose a framework, called SPLU, which allows releasing data for answering large sum queries with high accuracy to provide utility, while offering high inaccuracy for small sum queries in order to ensure privacy. We point out that not all inference attacks should be defended.
(2) To demonstrate the feasibility of the concept of SPLU, we propose a data sanitization mechanism, called *SPLU*-Gen, for achieving this goal. *SPLU*-Gen is based on randomized perturbation on the sensitive values.
(3) We introduce a sophisticated reconstruction algorithm which takes into account the global data distribution. This improves on the known reconstruction approach in syntactic methods and leads to higher data utility.
(4) We have conducted experiments on two real datasets to show that *SPLU*-Gen provides protection for small sums and high utility for large sum queries. We note that existing mechanisms may readily support SPLU, which is an encouraging result.

In Fig. 1, we outline the *SPLU*-Gen mechanism for data sanitization and the query processing based on the sanitized data. The dataset on the left of the figure is passed as input to *SPLU*-Gen. The input is processed and as a result, a sanitized dataset is published. Querying is applied on the sanitized data, and the query result is generated by a reconstruction algorithm. The user will receive relatively accurate results for large sum queries and inaccurate results for queries of small sums.

The rest of the paper is organized as follows. Section 2 presents the SPLU model. Section 3 describes the mechanism *SPLU*-Gen. Section 4 is about count reconstruction and properties of *SPLU*-Gen. Section 5 considers multiple attribute aggregations. Section 6 is on empirical study, and Section 7 is on related works. Section 8 concludes this work.

2. SPLU model

We consider the data model in previous works on k-anonymity [26] and ℓ-diversity [20]. This data model assumes that a set of attributes form a quasi-identifier, the values of which for a target individual can be known to the adversary from other sources, and also one or more sensitive attributes which need to be protected. Hence, there are two kinds of attributes in the dataset, the non-sensitive attributes (*NSA*) and the sensitive attributes (*SA*). In Fig. 2(a) we show a given dataset D. In table D, the attribute *id* is for the tuple id. The attributes Age and Zip-Code are considered non-sensitive attributes and they form a quasi-identifer. The term quasi-identifier indicates that it may be possible to identify an individual based on the respective attribute values. For example, it is possible that Age 90 and Zip-Code [12–17 k] uniquely determine an individual, if there is only one resident aged 90 in the area of Zip-Code [12–17 k]. Such attributes are considered not sensitive. In table D, Disease is a sensitive attribute.

In this model we do not perturb the non-sensitive values but may alter the sensitive values to ensure privacy. This is a commonly used data model and it corresponds to the initial problem settings with real world applications [25,22].

We are given a dataset (table) D which is a set of N tuples that follow the above data model. A concept c in D is a predicator formed by the conjunction of value assignments to a set of attributes in D. Our problem is how to generate and replace the sensitive values for the tuples in D to be published in the output dataset D'. D' should satisfy both utility for large sum querying and privacy protection for small sum queries.

In Fig. 1(b), we show a possible published dataset D', which is a sanitized counterpart of dataset D. We shall discuss in Section 3 about how D' is generated from D.

We define the requirements of our model in the following.

Given a dataset D, an anonymized data set D' generated by sanitization mechanism $\mathcal{A}$, and a concept c involving $s \in SA$, let f_c be the true frequency of c in D and f'_c be the estimated frequency of c from D'.

Definition 1 (*large sum utility*). Concept c has a $(\epsilon, \mathcal{T}_E, \mathcal{T}_f)$ utility guarantee if

$$Pr\left[|f'_c - f_c| \geqslant \varepsilon f_c\right] \leqslant \mathcal{T}_E \text{ for } f_c \geqslant \mathcal{T}_f \quad (1)$$

The above definition says that a concept c has a $(\epsilon, \mathcal{T}_E, \mathcal{T}_f)$ guarantee if whenever the frequency f_c of c is above $\mathcal{T}_f$ in D, then the probability of a relative error of more than ε is at most $\mathcal{T}_E$.

[1] There may be scenarios where our assumption does not hold. That is, even if something is true for most tuples in D', the information is still sensitive. An example would be a dataset containing only information about patients with a certain cancer disease. In such a case knowing that a person is in the dataset is already considered sensitive, and all attributes will be sensitive. Hence, our proposed model becomes irrelevant and does not apply.

Please cite this article in press as: Fu A.W.-C et al. Small sum privacy and large sum utility in data publishing. J Biomed Inform (2014), http://dx.doi.org/10.1016/j.jbi.2014.04.002

Download English Version:

https://daneshyari.com/en/article/6928367

Download Persian Version:

https://daneshyari.com/article/6928367

Daneshyari.com