Journal of Network and Computer Applications 109 (2018) 78–88

Contents lists available at ScienceDirect

Journal of Network and Computer Applications

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

# Privacy-aware data publishing against sparse estimation attack

Xuangou Wu [a], Panlong Yang [b], Shaojie Tang [c], Xiao Zheng [a,*], Xiaolin Wang [a]

[a] School of Computer Science and Technology, Anhui University of Technology, Maanshan, China
[b] School of Computer Science and Technology, University of Science and Technology of China, Hefei, China
[c] Naveen Jindal School of Management, University of Texas at Dallas, Texas, USA

ARTICLE INFO

*Keywords:*
Privacy protection
Network data security
Compressive sensing
Sparse estimation

ABSTRACT

Recently a number of privacy-preserving data publishing techniques have been proposed to protect the privacy of released data. In this paper, we study how to protect unreleased data privacy during data publishing. Our experiment results show that the unreleased data could be well estimated when an attacker leverages sparse estimation techniques, even when a large amount of noise is randomly added to the released data. To address unreleased data privacy while guaranteeing the utility of released data, we propose a privacy-aware structural data publishing framework against sparse estimation attack. Specifically, we present a nonzero element Gaussian random noise addition strategy, which is realized by maximizing global information loss between original data and noisy data. Furthermore, we deduce the upper bound of the number of released data that could be published, which acts as the criterion to guarantee the unreleased data privacy. Our experiment results show that the proposed framework is able to protect unreleased data privacy with desirable performance.

## 1. Introduction

With the proliferation of various sensors, smart devices, cloud computing and so on, many organizations and individuals generate or own huge amounts of datasets from various applications, such as environment monitoring, cloud platform, mobile crowdsensing, just to name a few (Khan, 2016). Although publishing these data have enormous social and individual benefits, several security and privacy concerns arise. Recently, the problem of privacy-preserving data publishing has received wide research interests, which aims to eliminate privacy threats and preserve useful information of released data at the same time (Fung et al., 2010). Existing researches include well-known techniques such as *k*-anonymity (Sweeney, 2002), *l*-diversity (Machanavajjhala et al., 2007), differential privacy (Dwork, 2011), graph-based privacy-preserving data publication (Li et al., 2016) and so on (e.g., Wu et al., 2015; Wang et al., 2015). To the best of our knowledge, these techniques mainly focus on protecting the released data privacy.

However, the privacy of unreleased data is the same or more important compared with the released data. For example, the data owner of wireless sensor networks may publish some public area monitoring data and is unwilling to release some sensitive area monitoring data with privacy consideration (Mao et al., 2012). In this scenario, the public area monitoring data are considered as released data, and the sensitive area monitoring data are considered as unreleased data. In smart grid, the consumers may be allowed to query a part of their monitoring traces of smart meter rather than all the monitoring traces, because it might reveal the personal habits and behavior of consumers, which would bring serious privacy threats (Mármol et al., 2012; Huang et al., 2014). In general, data owner would think that the unreleased data could not be well estimated because there are more unreleased data than released data. However, the real-world data is usually structural, which means it could be transformed into a sparse form (Tosic and Frossard, 2011). In fact, a large number of unknowns could be well estimated by exploiting sparsity and a small number of observations, especially with the development of compressive sensing based sparse estimation techniques (Candèset al, 2006).

Although noise addition techniques have been studied for many years in secure data publishing (e.g., Dwork, 2011; Bianchi et al., 2011), existing work mainly focuses on preventing the adversary from getting the accurate released individual data values, which lacks of global structural information consideration. Unfortunately, our experiment results show that the unreleased data could be well estimated, when an attacker leverages sparse estimation techniques with a small number of released data. Moreover, the unreleased data still could be well estimated, even when a large amount of noise is added to released data with existing noise addition strategies. Therefore,

* Corresponding author.
*E-mail addresses:* wxgou@mail.ustc.edu.cn (X. Wu), xzheng@seu.edu.cn (X. Zheng).

https://doi.org/10.1016/j.jnca.2018.02.016
Received 14 April 2017; Received in revised form 8 July 2017; Accepted 27 February 2018
Available online 13 March 2018

unreleased data would confront with serious privacy threats during data publishing.

To protect unreleased data privacy while guaranteeing the utility of released data, we propose a privacy-aware structural data publishing framework against sparse estimation attack. Our framework contains data owner and data user. For data user, the requested data could be allowed to contain a certain amount of noise. Data owner needs to prevent the unreleased data to be well estimated against sparse estimation attack. Data owner is responsible for adding noise to released data while guaranteeing the utility of released data. Meanwhile, it decides whether the user's requested data is allowed. If the unreleased data could be estimated within the range of unacceptable errors, data owner would reject the data user's request. Our work needs to maximize global information loss by noise addition and find the upper bound of number of released data. Therefore, we face two challenging problems to protect unreleased data privacy. (1) How to maximize unreleased data information loss with noise addition, as well as guaranteeing the utility for released data? Because released data protection focus on maximizing released data distortion without unreleased data privacy protection, our goal is to protect unreleased data form global information loss consideration. (2) How to determine the upper bound of the number of released data while meeting unreleased data privacy requirement? Due to the fact that existing sparse estimation techniques focus on finding the lower bound of observations to estimate unknowns successfully. We should find the upper bound of noisy observations to estimate unknowns, and the estimated errors are out of the unacceptable range.

To address the aforementioned challenges, we present a $k$-nonzero Gaussian noise addition strategy, which is realized by maximizing global information loss between original data and noisy data. Meanwhile, we deduce the upper bound of released data with our noise addition strategy by compressive sensing based sparse estimation. The contributions of this paper are in three folds.

- We propose a privacy-aware structural data publishing framework, which could protect unreleased data privacy effectively as well as guaranteeing the utility of released data. To the best of our knowledge, this is the first work to study unreleased data privacy protection during data publishing.
- We propose a nonzero element Gaussian random noise addition strategy to protect unreleased data privacy. Meanwhile, we discuss the upper bound of the number of released data with compressive sensing based sparse estimation analysis.
- We conduct an extensive set of experiments on real-world datasets which shows that the proposed framework is able to protect unreleased data privacy with desirable performance.

The rest of this paper is organized as follows. Section 2 presents the preliminaries and security threats. The problem statement is given in Section 3. The privacy-preserving framework and noise addition issue are presented in Section 4. In Section 5 and 6, we propose Gaussian random noise addition and discuss the upper bound of released data in detail, respectively. Section 7 reports our experiment results. We present a literature review of existing work in Section 8 and make a conclusion in Section 9.

## 2. Preliminaries and security threats

### 2.1. Sparse estimation

The estimation of a sparse vector is a fundamental problem in signal processing, which focuses on how to recover unknowns from a few sampling values. It lies at the growing field of compressive sensing (CS) (e.g., Candèset al, 2006; Candes and Davenport, 2013; Candès et al., 2006). The CS asserts that a relatively small number linear combination of a sparse vector could contain most of its salient information (Donoho, 2006). Assuming that $\mathbf{s}$ is a sparse vector ($\mathbf{s} \in \mathbb{R}^n$). If the number of nonzero elements of $\mathbf{s}$ is not greater than $k$, $\mathbf{s}$ is called $k$-sparse vector. For a $k$-sparse vector, the process of obtaining information could be expressed as

$$\mathbf{y} = A\mathbf{s} \tag{1}$$

where $\mathbf{y}$ is the sampling vector, $A$ is an $m \times n$ measurement matrix ($m \ll n$). Then the original data $\mathbf{s}$ can be reconstructed with an overwhelming probability when $A$ satisfies the restricted isometric property (RIP) and $m \geq O(k \log(n/k))$ (Candes and Tao, 2005). The recovery process is

$$\mathbf{s}^* = \arg\min_{\mathbf{s}\in\mathbb{R}^n} \|\mathbf{s}\|_1 \quad s.t. \quad \|\mathbf{y} - A\mathbf{s}\|_2^2 \leq \epsilon \tag{2}$$

where $\mathbf{s}^*$ is the reconstructed sparse vector of $\mathbf{s}$, $\epsilon$ is recovery error.

The real-world data usually do not meet sparsity in original spatial-temporal domain, but it can be transformed into sparse vector with an orthogonal basis. The transformed basis could be obtained by dictionary learning or general orthogonal basis such as discrete concise transform (DCT) basis, discrete fourier transform (DFT) basis (Tosic and Frossard, 2011). In following parts, we assume that a structural data could be transformed into a sparse vector by a suitable orthogonal basis.

### 2.2. Security threats

In this subsection, we display our experiment results to show that the unreleased data could be well-estimated when an attacker exploits sparse estimation technique. In our experiments, we used OMP (Tropp and Gilbert, 2007) algorithm to estimate unreleased data, because it is a typical sparse estimation algorithm (Candes and Tao, 2005). To display the experimental results efficiently, we consider a synthesis data to implement sparse estimation. Our experiment data contains 256 values, it could be transformed into a 5-sparse data under DCT basis, and the mean-squared of data elements is 0.8297 as shown in Fig. 1.

First, we display the experiment results of unreleased data estimation without noise. Fig. 1(a) displays the original data, 25 random released values and the estimated data. It shows the unreleased values are estimated exactly. In fact, we can consider the 5 sparse data as the structural information of the 256 original values. According to CS theory (Candèset al, 2006), the unreleased data could be recovered exactly

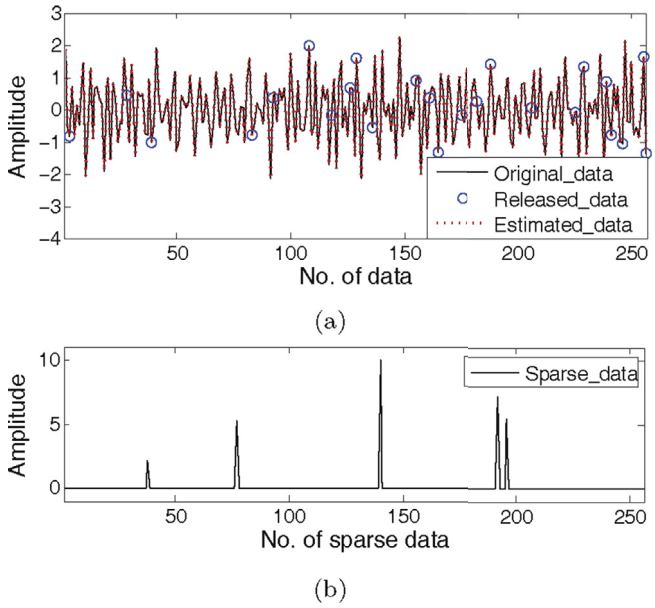


**Fig. 1.** Experiment results without noise. (a) Original data, Released data and Estimated data. (b) Sparse data with discrete concise transform.