2020 IEEE 16th International Conference on Control & Automation (ICCA)
Oct 9-11, 2020 (Virtual) Sapporo, Hokkaido, Japan

Privacy-Preserving Correlated Data Publication with a Noise Adding Mechanism

Mingjing Sun, Chengcheng Zhao, and Jianping He

Abstract— **The privacy issue in data publication is critical and has been extensively studied. However, most of the existing works assume the data to be published is independent, i.e., the correlation among data is neglected. The correlation is unavoidable in data publication, which universally manifests intrinsic correlations owing to social, behavioral, and genetic relationships. In this paper, we investigate the privacy concern of data publication where deterministic and probabilistic correlations are considered, respectively. Specifically, (ε, δ)-multi-dimensional data-privacy (MDDP) is proposed to quantify the correlated data privacy. It characterizes the disclosure probability of the published data being jointly estimated with the correlation under a given accuracy. Then, we explore the effects of deterministic correlations on privacy disclosure. For deterministic correlations, it is shown that the successful disclosure rate with correlations increases compared to the one without knowing the correlation. Meanwhile, a closed-form solution of the optimal disclosure probability and the strict bound of privacy disclosure gain are derived. Extensive simulations on a real dataset verify our analytical results.**

I. Introduction

With extensive personal data generated, data plays a key role in people's lives in various applications ranging from medical treatments to online-social interactions [1], [2]. Before using data for statistical analysis, users need to publish data. For data publication, how to protect individual privacy while obtaining accurate data analysis is an increasingly crucial issue [3]. For example, when users broadcast electrical usages to the data fusion center, the exact individual's data is fuzzy in the broadcasting process for privacy concern, while the aggregated result should be accurate.

Many efforts have been devoted to investigating privacy-preserving data publication. The existing works can be summarized as three aspects. The first type of research focuses on quantitative mechanisms analysis, e.g., differential privacy [4] and data privacy [5]. The second one is protection mechanisms design, e.g., encryption [6], anonymity [7] and noise adding [4]. The third type focuses on optimization, e.g., maximizing the measure of privacy [8]–[10]. These representative methods basically assumed that data is one-dimensional or independent [11]. However, the real-world data often exhibits strong coupling relations, e.g., medical data such as weight and blood pressure are often assumed to be normally distributed [12]. Thus, considering the correlated data publication [13], it rises a natural problem that, can the attacker use the correlation for analysis. It is intuitively feasible but lacks theoretical support. Noting this, we first reveal that the attacker is capable of using the correlation. Then, we theoretically provide privacy analysis and optimal noise design for the correlated data publication.

There are two main challenges in studying privacy-preserving data publication in correlated settings: 1) The first challenge is, for the data publication scenes, a proper and rigorous privacy metric should be proposed to guarantee the users' privacy. Note that different from database query problems, the key privacy concern of users is to ensure that their original data cannot be accurately estimated, rather than the indistinguishability that can be well quantified by differential privacy [4]. For example, in an individual deposit dataset, for the rich, it is an indisputable fact that his property is far more than others. It is more important for the rich that the attacker cannot infer the specific value of his deposit in a bank, rather than whether he deposits in a bank. 2) The second challenge lies in correlation modeling and privacy analysis. The analytical relationship between general correlations and privacy leakage remains unclear. To achieve efficient privacy protection for correlated data publication, it is essential to conduct rigorous theoretical studies to understand the analytical relationship between privacy leakage and general correlations.

Differential privacy has been defined and applied for quantifying the degree of individual privacy preservation in a statistical database [4]. It is proposed to maximize query accuracy while maintaining indistinguishability of each entry. For preserving privacy in a correlated database, differential privacy may result in redundant noise derived from both records and queries [11], [14]. Most importantly, the privacy guarantee by database query problems is different from the privacy demand of data publishers in reality. Data privacy is proposed in [5], where privacy analysis for independent data publication is investigated. More privacy definitions (e.g., identifiability, information-theoretic metrics) have been discussed in [15]–[21]. Most of the existing works (e.g., [22], [23]) need the assumption that data is independent. How to quantify the privacy for correlated data publication and what kind of noise distribution can achieve maximal privacy are remained open.

The main contributions are summarized as follows.

- We extend the definition of data privacy [5] to multi-

This work was supported in part by the Natural Science Foundation of China (NSFC) under Grants 61973218 and 61903328, and in part by the China Postdoctoral Science Foundation under Grant 2019M652082.

Mingjing Sun and Jianping He are with the Department of Automation, Shanghai Jiao Tong University, and Key Laboratory of System Control and Signal Processing, Ministry of Education of China, Shanghai 200240, China. E-mail: mj.sun@sjtu.edu.cn, jphe@sjtu.edu.cn.

Chengcheng Zhao is with The State Key Laboratory of Industrial Control Technology and Institute of Cyberspace Research, Zhejiang University, China, and the Department of Electrical and Computer Engineering, University of Victoria, BC, Canada. E-mail: zccsq90@gmail.com.

978-1-7281-9093-8/20/$31.00 ©2020 IEEE

dimensional correlated data privacy defined by (ε, δ)-multi-dimensional data-privacy ((ε, δ)-MDDP), quantifying disclosure probability of the published data being jointly estimated with correlations under a given accuracy. To the best of our knowledge, this is the first study to investigate the noise adding on multi-dimensional correlated data in the sense of (ε, δ)-MDDP.
- We analyze the effects of the correlation among data on the privacy disclosure. It is shown that using the correlation, the successful disclosure rate increases compared to that the correlation is unknown. Furthermore, the closed-form solution of disclosure probability and the strict bound of privacy gain are derived.
- We propose the optimal noise adding strategy in the sense of (ε, δ)-MDDP for the case with full couplings.

The remainder of this paper is organized as follows. Section II formulates the problem. The main results are presented in Section III. Moreover, Section IV evaluates the theoretical results through simulations. Finally, Section V concludes the paper.

II. Preliminaries and Problem Formulation

Consider that there are N users with unique ID, indexed by $1, 2, \dots, N$, and each user broadcasts his/her real-valued data to the data fusion center. Let $x = [x_1, \dots, x_N]^{\mathrm{T}}$ denote a private real-valued data vector, where x_i represents private data held by i-th user. Users preserve the privacy of their sensitive correlated data via adding random noises, i.e.,

$$x^{+} = x + \theta, \tag{1}$$

where $\theta = [\theta_1, \dots, \theta_N]^{\mathrm{T}}$ is a random noise vector, and x^{+} is a data vector to be published. Assume that the entries of the noise vector θ are uncorrelated. By referring to [11], [12], we consider two common types of correlations among original data as follows.

Definition 1: Full coupling means the original data is correlated in the form of multivariate explicit function $G(x_1, \dots, x_N)$ (G_x).

Definition 2: Probabilistic coupling means the original data is correlated in the form of multivariate joint probability density function $f(x_1, \dots, x_N)$ (f_x).

Full coupling is common among numeric data, e.g., the relation between height and weight of male can be obtained by polynomial regression [24]. Probabilistic coupling is usual among attribute data, e.g., the attributes such as weight and blood pressure are often normally distributed [12].

A. Attacker Model

Suppose that there is an attacker outside the users' data publication group. It is able to eavesdrop the broadcast information x^{+} and knows the noise adding mechanism (1). The attacker aims to infer the true data vector x. Let $\hat{x}$ be an estimation of x, where $\hat{x}_i$ represents i-th element of $\hat{x}$. The attacker can infer x by using the difference between the observed value and the estimated value of the added noise, i.e., $\hat{x} = x^{+} - \hat{\theta}$, where $\hat{\theta}$ is the estimation of the added noises θ. The infinite norm is denoted as $||x||_\infty = \max\{|x_i|\}$, which is the maximum absolute value of elements in vector x. Then, we have

$$\Pr\{||\hat{x} - x||_\infty \le \varepsilon\} = \Pr\{||\hat{\theta} - \theta||_\infty \le \varepsilon\}, \tag{2}$$

where $\hat{x}$ is an ε-accurate estimation if $||\hat{x} - x||_\infty \le \varepsilon$, $\varepsilon > 0$ and it is a given error threshold. Inspired by [5], we provide one important definition as follows.

TABLE I

Important Notations

Symbol	Definition
x_i	user i's original data
x	the original data vector of all users
χ	the vector of the possible value sets of x
$G(\cdot)$	the full-coupling among original data
$f(\cdot)$	the probabilistic coupling among original data
Θ_i	the set of random variable θ_i
$f_\theta(\cdot)$	the joint probability density function of θ
$f_{\theta_i\|\cdot}(\cdot)$	the marginal PDF of random variable θ_i
x^{+}	the observed data vector
ν	the variable denoting the possible original data
ε	the estimation accuracy of x
δ	the joint disclosure probability of x
$e_{\theta\|G_x(\cdot)}(x^{+})$	the estimation of θ using x^{+} and the correlation

Definition 3: Given ε-accurate estimation, and $x^{+} = [x_1^{+}, \dots, x_N^{+}]^{\mathrm{T}}$, G_x or f_x are the information available to the attacker, the optimal estimation of x is defined as

$$\hat{x}^{*} = \arg\max_{\hat{x} \in \chi} \Pr\{\nu + \theta = x^{+} | \, \forall ||\nu - \hat{x}||_\infty \le \varepsilon, G_x(f_x)\}, \tag{3}$$

where ν is variable, denoting the possible value of original data. $\chi = [\chi_1, \dots, \chi_N]^{\mathrm{T}}$, and χ_i is the set of the possible values of x_i.

Remark 1: In (3), it should be pointed out that both the distribution and the estimation range of θ are affected by G_x, and the details will be discussed later.

When the correlation is unknown to the attacker, then (3) is simplified to

$$\hat{x}^{*} = \arg\max_{\hat{x} \in \chi} \Pr\{\nu + \theta = x^{+} | \, \forall ||\nu - \hat{x}||_\infty \le \varepsilon\}, \tag{4}$$

which is consistent with [5].

B. Privacy Definition

Under the above attacker model, to quantify the degree of privacy protection of correlated data publication, the relationship between estimation accuracy and privacy is defined as follows.

Definition 4: A noise adding mechanism (1) satisfies (ε, δ)-multi-dimensional data-private, iff,

$$\delta = \Pr\{||\hat{x}^{*} - x||_\infty \le \varepsilon\}, \tag{5}$$

where ε is the estimation accuracy and δ is the disclosure probability.

Remark 2: Definition 4 quantifies the probability that the attacker can successfully estimate each entry of x in a given interval $[x_i - \varepsilon, x_i + \varepsilon]$, $\forall i \in [1, N]$ is equal to δ, using the optimal estimation. A smaller value of ε offers higher accuracy, and a smaller value of δ offers a lower disclosure probability.

C. Problem of Interests

In this paper, we are mainly concerned about the following issues:

i) How the data correlations affect the attacker's optimal estimation compared to that the correlation is unknown, and what the bound of privacy gain is if it exists.

ii) Whether there is a closed-form expression of the optimal estimation and the disclosure probability considering data correlations.

iii) For the case with full couplings, which kind of noise probability density distribution $f_{\theta_i}(z)$, $i = 1, 2, \ldots, N$ is optimal in the sense of (ε, δ)-MDDP, that is,

$$\begin{aligned} &\min_{f_{\theta_i}(z), i=1,2,\ldots,N} \quad \delta \\ s.t. \quad &\mathrm{E}\{\theta_i\} = 0, \\ &\mathrm{Var}\{\theta_i\} = \sigma^2, \quad i = 1, \ldots, N. \end{aligned} \tag{6}$$

where σ^2 is noise variance. Problem (6) is difficult to solve directly, since the explicit expression of δ is difficult to obtain considering the data correlation.

III. Main Results

A. Multi-dimensional Full-coupled Data Publication

Consider a powerful attacker not only has the observation x^+ but also knows the full-coupling G_x. First, we investigate how full-coupling affects δ. From Definition 3, 4, one sees that δ depends on $f_\theta(z_1, \ldots, z_N)$, ε and the estimation policy (3). Here we have,

$$\begin{aligned} \delta &= \Pr\{||\hat{x}^* - x||_\infty \le \varepsilon |\, x^+, \chi\} \\ &\ge \Pr\{||\hat{x}^* - x||_\infty \le \varepsilon |\, x^+, G_x, \chi\}. \end{aligned} \tag{7}$$

The inequality (7) holds because the full-coupling may narrow the real value set down to a small set within the initial set. Then, according to (3) and (5), δ varies and it is no larger than that before domain changing. Therefore, although there exists extra information G_x available to the attacker rather than only x^+, extra knowledge of full-coupling does not necessarily increase δ. A case of δ decrease is provided as follows.

Example 1: Given the noise distribution $f_{\theta_i}(z)$ and the estimation accuracy ε approaching zero, for $\Theta_i = \{x_i^+ - \chi_i\}$, let $\Theta_{i,1}$, $\Theta_{i,0}$ be the noise sets of x_i, $f^*_{\theta_i|\Theta_{i,1}}(z) = \max_{z \in \Theta_{i,1}} f_{\theta_i}(z)$. If $f^*_{\theta_i|\Theta_{i,1}}(z) > f^*_{\theta_i|\Theta_{i,0}}(z)$, from Definition 3, 4, we have,

$$\begin{aligned} \lim_{\varepsilon \to 0} \frac{\delta|\Theta_{i,1}}{\delta|\Theta_{i,0}} &= \lim_{\varepsilon \to 0} \frac{\max_{\hat{\theta}_i \in \Theta_{i,1}} \int_{\hat{\theta}_i - \varepsilon}^{\hat{\theta}_i + \varepsilon} f_{\theta_i}(z) \mathrm{d}z}{\max_{\hat{\theta}_i \in \Theta_{i,0}} \int_{\hat{\theta}_i - \varepsilon}^{\hat{\theta}_i + \varepsilon} f_{\theta_i}(z) \mathrm{d}z} \\ &= \lim_{\varepsilon \to 0} \frac{2\varepsilon * f^*_{\theta_i|\Theta_{i,1}}(z)}{2\varepsilon * f^*_{\theta_i|\Theta_{i,0}}(z)} > 1, \end{aligned} \tag{8}$$

where $\Theta_{i,1} \subset \{\Theta_{i,1} \cup \Theta_{i,0}\}$.

In order to quantify the disclosure probability under correlation G_x and construct a relationship between δ and Θ, we define successful disclosure rate as follows.

Definition 5: With the knowledge of noise set $\hat{\Theta}_i$ and x_i^+, the successful disclosure rate of x_i is defined as

$$\gamma_i = \begin{cases} \frac{\int_{\hat{\theta}_i^* - \varepsilon}^{\hat{\theta}_i^* + \varepsilon} f_{\theta_i}(z) \mathrm{d}z}{\int_{z \in \hat{\Theta}_i} f_{\theta_i}(z) \mathrm{d}z} \times 100\%, & \text{if } \hat{\theta}_i^* \in \Theta_i, \\ 0, & \text{otherwise}, \end{cases} \tag{9}$$

where $\Theta_i = \{x_i^+ - \chi_i\}$, $\Theta_i \subset \hat{\Theta}_i$ is the true set narrowed down by full coupling, and $\hat{\theta}_i^*$ is the optimal estimation of θ_i obtained by using (3). And the successful disclosure rate of x is $\gamma = \prod_{i=1}^N \gamma_i$.

Remark 3: From Definition 5, it is observed that if $\hat{\theta}_i^* \in \Theta_i$, then, γ_i increases with $\hat{\Theta}_i$ being narrowed down, i.e., the attacker can successfully estimate the real value with a higher probability when $\hat{\Theta}_i$ is smaller. Note that, ε should be constrained as $\varepsilon < \frac{\sup\{\Theta_i\} - \inf\{\Theta_i\}}{2}$, otherwise, $\gamma_i = 1$ may holds no matter which kind of $f_{\theta_i}(z)$ is used. We conclude that when $\delta_i = \int_{\hat{\theta}_i^* - \varepsilon}^{\hat{\theta}_i^* + \varepsilon} f_{\theta_i}(z) \mathrm{d}z$ is fixed, γ_i increases as the coverage of $\hat{\Theta}_i$ be smaller.

Theorem 1: Consider the mechanism (1), if the original data is full-coupled by $G(x_1, \ldots, x_N)$, then, the successful disclosure rate increases, i.e.,

$$\hat{\gamma} \le \gamma, \tag{10}$$

where $\hat{\gamma}$ is the successful disclosure rate of x without knowing the correlation.

Proof: By using the full coupling and other elements' domain information, there exists an element in x whose real value set is narrowed down, we denote such element by x_i, and other elements by x_j, $\forall j \in [1, N]$, $j \neq i$. Thus, we have $\chi_i \subset \hat{\chi}_i$, i.e., $\Theta_i \subset \hat{\Theta}_i$, and

$$\int_{z \in \Theta_i} f_{\theta_i}(z) \mathrm{d}z < \int_{z \in \hat{\Theta}_i} f_{\theta_i}(z) \mathrm{d}z, \tag{11}$$

where $\hat{\Theta}_i$ is the noise set of i-th element in x when the correlation is unknown to the attacker, and Θ_i is the noise set of i-th element narrowed down in x. Let $\hat{\theta}_i^* \in \hat{\Theta}_i$ be the optimal estimation of θ_i obtained by using (3). Then, if

i) $\hat{\theta}_i^* \in \Theta_i$. It occurs when the optimal estimation of θ_i without using the full coupling is also within the real domain Θ_i. It follows that,

$$\max_{\hat{\theta}_i \in \Theta_i} \int_{\hat{\theta}_i - \varepsilon}^{\hat{\theta}_i + \varepsilon} f_{\theta_i}(z) \mathrm{d}z = \max_{\hat{\theta}_i \in \hat{\Theta}_i} \int_{\hat{\theta}_i - \varepsilon}^{\hat{\theta}_i + \varepsilon} f_{\theta_i}(z) \mathrm{d}z.$$

From Definition 5, we directly obtain

$$\begin{aligned} \gamma_i &= \frac{\int_{\hat{\theta}_i^* - \varepsilon}^{\hat{\theta}_i^* + \varepsilon} f_{\theta_i}(z) \mathrm{d}z}{\int_{z \in \Theta_i} f_{\theta_i}(z) \mathrm{d}z} \times 100\% \\ &> \frac{\int_{\hat{\theta}_i^* - \varepsilon}^{\hat{\theta}_i^* + \varepsilon} f_{\theta_i}(z) \mathrm{d}z}{\int_{z \in \hat{\Theta}_i} f_{\theta_i}(z) \mathrm{d}z} \times 100\% \\ &> \hat{\gamma}_i, \end{aligned} \tag{12}$$

where γ_i is the successful disclosure rate of x_i using the full coupling, $\hat{\gamma}_i$ is the successful disclosure rate of x_i without knowing the correlation. For $\gamma_j = \hat{\gamma}_j$, $\forall j \in [1,N]$, $j \neq i$, and $\gamma = \prod_{k=1}^{N} \gamma_k$, (10) holds.

ii) $\hat{\theta}_i^* \notin \Theta_i$. It occurs when the optimal estimation of θ_i varies as Θ_i be smaller due to full coupling. Then, if

a) $\max_{\hat{\theta}_i \in \Theta_i} \int_{\hat{\theta}_i-\varepsilon}^{\hat{\theta}_i+\varepsilon} f_{\theta_i}(z)\mathrm{d}z \leq \max_{\hat{\theta}_i \in \hat{\Theta}_i} \int_{\hat{\theta}_i-\varepsilon}^{\hat{\theta}_i+\varepsilon} f_{\theta_i}(z)\mathrm{d}z$

This case occurs when the optimal estimation $\hat{\theta}_i^*$ without using the full coupling is not within the real domain Θ_i. From Definition 5, $\hat{\gamma}_i = 0$ holds. Thus, we have

$$\gamma_i > 0,\ \hat{\gamma}_i = 0, \tag{13}$$

and (10) holds.

b) $\max_{\hat{\theta}_i \in \Theta_i} \int_{\hat{\theta}_i-\varepsilon}^{\hat{\theta}_i+\varepsilon} f_{\theta_i}(z)\mathrm{d}z > \max_{\hat{\theta}_i \in \hat{\Theta}_i} \int_{\hat{\theta}_i-\varepsilon}^{\hat{\theta}_i+\varepsilon} f_{\theta_i}(z)\mathrm{d}z$

This case is impossible when $\Theta_i \subset \hat{\Theta}_i$.

The equality of (10) holds when Θ_i can not be narrowed down by G_x, e.g, when $\hat{\Theta} = \mathbb{R}^N$. The proof is completed. ■

Remark 4: Theorem 1 implies that because the full coupling exists, we can obtain more information about the real domain, and the uncertainty of x is decreased. Thus, there exists a higher probability that the attacker can make an accurate estimation on data vector x.

Lemma 1: Consider (1), if x_i and x_j are independent for $\forall i,j \in [1,N]$, $i \neq j$, $x \in \mathbb{R}^N$. Then, with zero mean and finite variance σ^2, uniform distribution of $f_{\theta_i}^*(z)$ reaches the maximum privacy in the sense of (ε,δ)-MDDP, i.e.,

$$f_{\theta_i}^*(z) = \begin{cases} \frac{1}{2\sqrt{3}\sigma}, & \text{if } z \in \left[-\sqrt{3}\sigma, \sqrt{3}\sigma\right], \\ 0, & \text{otherwise}, \end{cases} \tag{14}$$

where $i = 1,\ldots,N$.

Proof: The lemma follows from Theorem 4.4 in [25]. ■

Remark 5: Lemma 1 means that, to minimize the disclosure probability of multi-dimensional independent data publication under the optimal estimation, independently adding uniform noise on each dimension of x is optimal in the sense of (ε,δ)-MDDP.

Theorem 2: Consider (1), if the original data is full-coupled by $G(x_1,\ldots,x_N)$, and add independent noises, then, the attacker is able to reduce the N-dimensional joint optimal estimation to

$$\begin{cases} \hat{x}_j^* = x_j^+ - \arg \max\limits_{\hat{\theta}_j \in \{x_j^+ - \chi_j\}} \displaystyle\int_{\hat{\theta}_j-\varepsilon}^{\hat{\theta}_j+\varepsilon} f_{\theta_j}(z)\mathrm{d}z, \forall j \neq i \\ \hat{x}_i^* = \{\hat{x}_i \,|\, G(\hat{x}_i, \hat{x}_{-i}^*) = 0, \hat{x}_i \in \chi_i\}, \end{cases} \tag{15}$$

where $i = \arg\min_{\{k=1,\ldots,N\}} \frac{\Pr\{|\hat{x}_k^* - x_k| \leq \varepsilon | \hat{x}_k \in \chi_k\}}{\int_{\inf(\Theta_k)}^{\sup(\Theta_k)} f_{\theta_k}(z)\mathrm{d}z}$, and $\hat{x}_{-i}^*$ are entries in $\hat{x}^*$ but $\hat{x}_i^*$. And users cooperate to add the uniform noise on $(N-1)$ dimensions of x_j is optimal in the sense of (ε,δ)-MDDP.

Proof: We prove it from two-dimensional coupled vector $y = [y_i, y_j]$. Coupling function $G(y_i,y_j)$ can be any explicit function, since it is able to represent y_j in an algebraic form of y_i. Without loss of generality, we assume $G(y_i,y_j) = y_i + y_j - c = 0$. With the dynamic but observable y^+, the attacker can use the fact that noises added by users satisfies:

$$y_i + y_j = y_i^+ - \theta_i + y_j^+ - \theta_j. \tag{16}$$

It directly follows that

$$\theta_i + \theta_j = (y_i^+ + y_j^+) - c, \tag{17}$$

which means the sum value of all possible noises added on y is known to the attacker, based on observation and correlation knowledge. Let $\hat{y}^*$ be the optimal estimation under y^+ and $G(y_i,y_j) = 0$. Using (3), we have

$$\begin{aligned} \hat{y}^* &= \arg\max_{\hat{y} \in \chi} \Pr\{\nu + \theta = y^+ | G(y_i, y_j = 0), \\ &\quad \forall ||\nu - \hat{y}||_\infty \leq \varepsilon\} \\ &= \arg\max_{\hat{y} \in \chi} \Pr\{\nu_i + \theta_i = y_i^+, \nu_j + \theta_j = y_j^+ | G(y_i,y_j) = 0, \\ &\quad \forall ||\nu - \hat{y}||_\infty \leq \varepsilon\} \\ &= y^+ - \arg \max_{\substack{\hat{\theta} \in \{y^+ - \chi\} \\ \hat{y}_i + \hat{y}_j = c}} \int_{\hat{\theta}_i-\varepsilon}^{\hat{\theta}_i+\varepsilon} \int_{\hat{\theta}_j-\varepsilon}^{\hat{\theta}_j+\varepsilon} f_{\theta_i,\theta_j}(z,h)\mathrm{d}z\mathrm{d}h \\ &= y^+ - e_{\theta|G_y}(y^+), \end{aligned} \tag{18}$$

where $\hat{\theta} = y^+ - \hat{y}$. Note that the joint distribution of any two random variables M and N satisfies

$$f_{M,N}(m,n) = f_{M|N}(m|n) f_N(n). \tag{19}$$

Using (17), we have

$$\begin{aligned} &\int_{\hat{\theta}_i-\varepsilon}^{\hat{\theta}_i+\varepsilon} \int_{\hat{\theta}_j-\varepsilon}^{\hat{\theta}_j+\varepsilon} f_{\theta_i,\theta_j}(z,h)\mathrm{d}z\mathrm{d}h \\ &= \int_{\hat{\theta}_j-\varepsilon}^{\hat{\theta}_j+\varepsilon} f_{\theta_i,\theta_j}(c-h,h)\mathrm{d}h \\ &= \int_{\hat{\theta}_j-\varepsilon}^{\hat{\theta}_j+\varepsilon} f_{\theta_i|\theta_j=h}(\widetilde{\theta}_i|\theta_j = h)\, f_{\theta_j}(h)\mathrm{d}h. \end{aligned} \tag{20}$$

Substituting (20) to the right side of (18), the 2-dimensional joint optimal estimation is as follows,

$$\hat{y}^* = y^+ - \arg \max_{\substack{\hat{\theta}_j \in \{y_j^+ - \chi_j\} \\ \widetilde{\theta}_i = c - \hat{\theta}_j}} \int_{\hat{\theta}_j-\varepsilon}^{\hat{\theta}_j+\varepsilon} f_{\theta_i|\theta_j=h}(\widetilde{\theta}_i) f_{\theta_j}(h)\mathrm{d}h, \tag{21}$$

which depends on the joint distribution of θ_i and θ_j, the values of y^+, $\widetilde{\theta}_i$ and χ_j. When θ_i and θ_j are independent, it follows from (20) that

$$\begin{aligned} &\int_{\hat{\theta}_j-\varepsilon}^{\hat{\theta}_j+\varepsilon} f_{\theta_i|\theta_j=h}(\widetilde{\theta}_i|\theta_j = h) f_{\theta_j}(h)\mathrm{d}h \\ &= f_{\theta_i}(\widetilde{\theta}_i) \int_{\hat{\theta}_j-\varepsilon}^{\hat{\theta}_j+\varepsilon} f_{\theta_j}(h)\mathrm{d}h, \end{aligned} \tag{22}$$

where $f_{\theta_i|\theta_j=h}$ is the conditional PDF of θ_i under the condition $\theta_j = h$ ($h \in \Theta_j$). θ_j is a variable, and θ_i is a fixed value based on variable θ_j when y^+ is fixed.

Then, the multi-dimensional joint estimation (18) is reduced to two-phase one-dimensional estimation as follows, i) the attacker use y_i^+, y_j^+ and full-coupling $G(y_i,y_j) = 0$, from Theorem 1, if there exists

$$\frac{\Pr\{|\hat{y}_j^* - y_j| \leq \varepsilon | \hat{y}_j \in \chi_j\}}{\int_{\inf(\Theta_j)}^{\sup(\Theta_j)} f_{\theta_j}(h)\mathrm{d}h} > \frac{\Pr\{|\hat{y}_i^* - y_i| \leq \varepsilon | \hat{y}_i \in \chi_i\}}{\int_{\inf(\Theta_i)}^{\sup(\Theta_i)} f_{\theta_i}(z)\mathrm{d}z}, \tag{23}$$

then, to achieve an accurate estimation of y with a higher probability, there exists a higher priority for the attacker to target y_j. Using (3), the optimal estimation of y_j is

$$\begin{aligned}\hat{y}_j^* &= \arg\max_{\hat{y}_j\in\chi_j} \int_{y_j^+-\hat{y}_j-\varepsilon}^{y_j^+-\hat{y}_j+\varepsilon} f_{\theta_i|\theta_j=h}(\widetilde{\theta}_i|\theta_j=h) \\ &\quad\times f_{\theta_j}(h)\mathrm{d}h \qquad (24)\\ &= \arg\max_{\hat{y}_j\in\chi_j} \int_{y_j^+-\hat{y}_j-\varepsilon}^{y_j^+-\hat{y}_j+\varepsilon} f_{\theta_j}(h)\mathrm{d}h.\end{aligned}$$

ii) To minimize the error of $|\hat{y}_i^*-y_i|$, the attacker infers y_i by using the full-coupling fact $G(y_i,y_j)=0$ and the $\hat{y}_j^*$ estimated in (24), i.e.,

$$\hat{y}_i^* = \{\hat{y}_i|\ G(\hat{y}_i,\hat{y}_j^*)=0, \hat{y}_i\in\chi_i\}. \qquad (25)$$

Then, we investigate the optimal noise adding on data vector y. From the two-phase estimation, the dimension of the optimization problem is reduced,

$$\min_{f_{\theta_j}(h)}\ \max_{\hat{\theta}_j\in\Theta_j} \int_{\hat{\theta}_j-\varepsilon}^{\hat{\theta}_j+\varepsilon} f_{\theta_j}(h)\mathrm{d}h, \forall j\neq i, \qquad (26)$$

the solution of (26) is given in Lemma 1.

Based on the closed-form expression of disclosure probability, in the sense of (ε,δ)-MDDP, the optimal noise adding strategy is

$$f_{\theta_j}^*(h) = \begin{cases} \frac{1}{2\sqrt{3}\sigma}, & \text{if } z\in\left[-\sqrt{3}\sigma,\sqrt{3}\sigma\right], \\ 0, & \text{otherwise,}\end{cases} \qquad (27)$$

where $\forall j\in[1,N]$, $j\neq i$, i.e., users cooperate to add uniform distribution noise on y_j selected is optimal in the sense of (ε,δ)-MDDP. ■

Remark 6: Full coupling is identified to reduce the dimension of joint estimation effectively, the closed-form δ and optimal noise adding strategy are derived. Users are suggested to cooperatively protect those data with less uncertainty by adding optimal noise. Thus, the disclosure probability is minimized while less noise is added contrasted with (14).

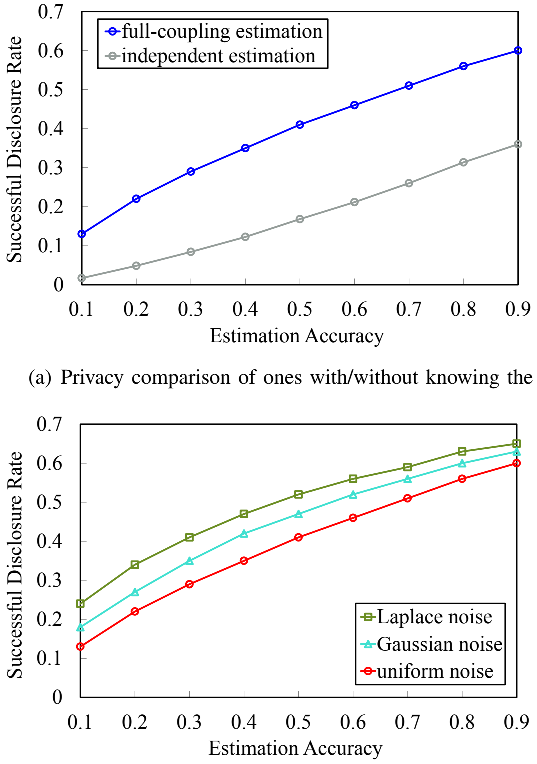
Theorem 3: (**Bound of privacy gain**) Consider (1), if K pairs of original data are coupled by G_x, given noise distribution satisfying (27), and $x_k\in[-\frac{M}{2},\frac{M}{2}]$, $\forall k\in[1,N]$. Then,

$$\frac{\gamma-\hat{\gamma}}{\hat{\gamma}} \leq ([\frac{2\varepsilon}{M}]^{-K}-1)(M\gg\varepsilon), \qquad (28)$$

holds.

Proof: From Theorem 2, it infers that with full coupling, if the attacker makes an accurate estimation of x_{-i}, where x_{-i} are entries in x but x_i, then, it is able to make an accurate estimation of x_i by using (15). Under the optimal noise derived in (27) and given $x_k\in[-\frac{M}{2},\frac{M}{2}]$, $\forall k\in[1,N]$, γ of x_{-i} is $[\frac{2\varepsilon}{M}]^{N-1}$, i.e., γ of vector x is $[\frac{2\varepsilon}{M}]^{N-1}$. Without knowing the correlation, $\hat{\gamma}$ of x based on x^+ is $[\frac{2\varepsilon}{M}]^{N}$ under the optimal uniform noise. The upper bound of γ is $[\frac{2\varepsilon}{M}]^{N-K}$, which depends on the maximum number of dimensionality reductions. ■

Remark 7: Theorem 1 and Theorem 3 conclude that using the full coupling, the successful disclosure rate γ increases compared to that the correlation is unknown, and the strict promotion bound is derived.



(a) Privacy comparison of ones with/without knowing the G_x

(b) Privacy under x^+ and G_x with different noise distribution

Fig. 1. The privacy comparison with full-coupling

To mitigate privacy leakage owing to correlation, there are two protection strategies.

i) Using the optimal noise adding strategy derived in Theorem 2. Despite considering the privacy leakage caused by correlation, both the disclosure probability and the successful disclosure rate are minimized.

ii) Reselect the noise variance σ^2. A larger value of σ^2 means that the published data is more likely to deviate from the true data, hence the privacy leakage δ is smaller. Such a method will make the utility of the published data decrease.

IV. Simulation Results

A. Simulation Scenario

In this section, we conduct simulations to verify the obtained theoretical results. The dataset for simulation is from the height and weight information of 10,000 women aged 30-39 in the R platform. The full coupling G_x is obtained by using the polynomial regression, that is, $\widehat{Weight} = 261.88-7.35*Height+0.083*Height^2$. Then, 10000 two-dimensional data vectors x are generated from the dataset, and each vector contains one paired weight and height data. For each run, users indexed by 1 and 2 randomly generate a noise vector θ with Laplacian noise distribution, Gaussian noise distribution and the optimal noise distribution derived in (27), respectively. In the simulation, 10000 runs are conducted.

The attacker uses the proposed optimal estimation approach derived in (15) to estimate the value of x. Specifically, it generates two sets of estimated value in each run, each set contains 10000 random numbers with the same distribution of users and uses them as the estimation of θ. Then, the probability of $||\hat{\theta}-\theta||_\infty \leq \varepsilon$ is obtained in each run, the value of δ is computed by using the maximum probability among these probabilities in all runs. When both the ranges of height and weight are known, Θ is known. Thus, one can obtain γ through dividing δ by the probability of the estimated $\hat{\theta}$ within Θ.

B. Verification

We first compare γ with $\hat{\gamma}$ under the optimal noise adding strategy derived in Theorem 2. From Fig. 1(a), one observes that using the G_x, the successful disclosure rate increases effectively compared to that when the correlation is unknown. It is because using the G_x, the attacker is able to decrease the uncertainty of noise adding effectively. Furthermore, in Fig. 1(a), the privacy gain is consistent with the bound in Theorem 3. The value of γ in simulation matches its theoretic result, which illustrates the correctness of theoretical results.

Then, we compare the privacy of full-coupled data publication for Laplacian noise distribution, Gaussian noise distribution and the optimal noise distribution derived in (27). It is observed from Fig. 1(b) that under G_x, the optimal noise distribution derived in (27) performs better than the extensively-used Laplacian noise distribution or Gaussian noise distribution in the sense of (ε,δ)-MDDP.

V. Conclusion

In this paper, we investigated the privacy-preserving correlated data publication problem. We proposed (ε,δ)-multi-dimensional data privacy to characterize the probability of the published data being restored with the correlation under a given accuracy, then, the privacy disclosure variation of correlated data publication can be quantified. We obtained the closed-form expression of the optimal estimation for the correlated data publication. It is shown that using the correlation, the original data vector can be restored with a higher successful disclosure rate. Moreover, the strict bound of privacy gain is derived. Lastly, we designed a cooperative noise adding strategy to minimize the disclosure probability for full-coupled data publication in the sense of (ε,δ)-MDDP. The privacy analysis and optimal noise design for multi-dimensional probabilistic data publication is to be our future work.

References

[1] C. Lutz, "The role of privacy concerns in the sharing economy," *Information, Communication & Society*, vol. 21, no. 10, pp. 1472–1492, 2018.

[2] C. Zhao, J. Chen, J. He, and P. Cheng, "Privacy-preserving consensus-based energy management in smart grids," *IEEE Transactions on Signal Processing*, vol. 66, no. 23, pp. 6162–6176, 2018.

[3] S. J. Rizvi and J. R. Haritsa, "Maintaining data privacy in association rule mining," in *Proceedings of the 28th International Conference on Very Large Data Bases*. VLDB Endowment, 2002, pp. 682–693.

[4] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory of Cryptography Conference*. Springer, 2006, pp. 265–284.

[5] J. He, L. Cai, and X. Guan, "Preserving data-privacy with added noises: Optimal estimation and privacy analysis," *IEEE Transactions on Information Theory*, vol. 64, no. 8, pp. 5677–5690, 2018.

[6] E. Fujisaki and T. Okamoto, "Secure integration of asymmetric and symmetric encryption schemes," in *Annual International Cryptology Conference*. Springer, 1999, pp. 537–554.

[7] L. Sweeney, "k-anonymity: A model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 05, pp. 557–570, 2002.

[8] F. Farokhi and H. Sandberg, "Ensuring privacy with constrained additive noise by minimizing fisher information," *Automatica*, vol. 99, pp. 275–288, 2019.

[9] Q. Geng and P. Viswanath, "The optimal noise-adding mechanism in differential privacy," *IEEE Transactions on Information Theory*, vol. 62, no. 2, pp. 925–951, Feb. 2016.

[10] W. Liao, J. He, S. Zhu, C. Chen, and X. Guan, "On the tradeoff between data-privacy and utility for data publishing," in *2018 IEEE 24th International Conference on Parallel and Distributed Systems (ICPADS)*, 2018, pp. 779–786.

[11] T. Zhu, G. Li, W. Zhou, and S. Y. Philip, *Differential privacy and applications*. Springer, 2017.

[12] L. Sankar, S. R. Rajagopalan, and H. V. Poor, "Utility-privacy tradeoffs in databases: An information-theoretic approach," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 6, pp. 838–852, 2013.

[13] R. Chen, B. C. Fung, S. Y. Philip, and B. C. Desai, "Correlated network data publication via differential privacy," *The VLDB Journal*, vol. 23, no. 4, pp. 653–676, 2014.

[14] D. Kifer and A. Machanavajjhala, "No free lunch in data privacy," in *Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data*, ser. SIGMOD '11. New York, NY, USA: Association for Computing Machinery, 2011, p. 193–204. [Online]. Available: https://doi.org/10.1145/1989323.1989345

[15] W. Wang, L. Ying, and J. Zhang, "On the relation between identifiability, differential privacy, and mutual-information privacy," *IEEE Transactions on Information Theory*, vol. 62, no. 9, pp. 5018–5029, 2016.

[16] B. Yang, I. Sato, and H. Nakagawa, "Bayesian differential privacy on correlated data," in *Proceedings of the SIGMOD international conference on Management of Data*. ACM, 2015, pp. 747–762.

[17] E. Nekouei, T. Tanaka, M. Skoglund, and K. H. Johansson, "Information-theoretic approaches to privacy in estimation and control," *Annual Reviews in Control*, 2019.

[18] X. Ren, C. Yu, W. Yu, S. Yang, X. Yang, J. A. McCann, and P. S. Yu, "LoPub: High-dimensional crowdsourced data publication with local differential privacy," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 9, pp. 2151–2166, 2018.

[19] F. du Pin Calmon and N. Fawaz, "Privacy against statistical inference," in *IEEE Annual Allerton Conference on Communication, Control, and Computing*. IEEE, 2012, pp. 1401–1408.

[20] P. Cuff and L. Yu, "Differential privacy as a mutual information constraint," in *Proceedings of the SIGSAC Conference on Computer and Communications Security*. ACM, 2016, pp. 43–54.

[21] F. Farokhi and H. Sandberg, "Optimal privacy-preserving policy using constrained additive noise to minimize the fisher information," in *IEEE Annual Conference on Decision and Control*. IEEE, 2017, pp. 2692–2697.

[22] W. Wang, L. Ying, and J. Zhang, "The value of privacy: Strategic data subjects, incentive mechanisms and fundamental limits," *ACM SIGMETRICS Performance Evaluation Review*, vol. 44, no. 1, pp. 249–260, 2016.

[23] C. Murguia, I. Shames, F. Farokhi, and D. Nešić, "On privacy of quantized sensor measurements through additive noise," in *IEEE Conference on Decision and Control*. IEEE, 2018, pp. 2531–2536.

[24] T. Furukawa, M. Inoue, and H. Abe, "Assessment of biological age by multiple regression analysis," *Journal of Gerontology*, vol. 30, no. 4, pp. 422–434, 1975.

[25] J. He, L. Cai, C. Zhao, P. Cheng, and X. Guan, "Privacy-preserving average consensus: privacy analysis and algorithm design," *IEEE Transactions on Signal and Information Processing over Networks*, vol. 5, no. 1, pp. 127–138, 2018.