Privacy-Preserving Correlated Data Publication: Privacy Analysis and Optimal Noise Design

Mingjing Sun, Chengcheng Zhao, Jianping He, Peng Cheng, and Daniel E. Quevedo

***Abstract*—The privacy issue in data publication is critical and has been extensively studied. Correlation is unavoidable in data publication, which universally manifests intrinsic correlations owing to social, physical, behavioral, and genetic relationships. However, most of the existing works assume that private data is independent, i.e., the correlation among data is neglected. 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 and probabilistic correlations on privacy disclosure, respectively. For both kinds of correlations, it is shown that the privacy disclosure with correlations increases compared to the one without correlation knowledge. Meanwhile, a closed-form expression of disclosure probability and a strict bound of privacy disclosure gain are derived, respectively. To minimize the disclosure probability, we provide the optimal noise distribution in the sense of (ε, δ)-MDDP. Extensive simulations on a real dataset verify our analytical results.**

***Index Terms*—Data privacy, correlated data, multi-dimension, optimal distribution, noise adding mechanism.**

I. Introduction

With extensive personal data being generated on a daily basis, data plays a key role in people's lives in various applications ranging from medical treatments to online-social interactions [1]. 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 [2]. 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. Existing works can be categorized as follows: The first type of research focuses on quantitative mechanisms analysis, e.g., differential privacy [3] and data privacy [4]. The second one is protection mechanisms design, e.g., encryption [5], anonymity [6] and noise adding [7]. The third type focuses on optimization, e.g., maximizing the measure of privacy [8]. These representative methods need a basic assumption, namely, that data is one-dimensional or independent [9]. However, 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 [10]. Thus, considering correlated data publication [11], it raises the following problem: Can the attacker use the correlation for analysis? This seems intuitively feasible, but lacks theoretical support. This work contributes to answering this question. For that purpose, we first reveal that the attacker is capable of using the correlation. Then, we provide theoretical tools for privacy analysis and optimal noise design for situations where data to be published is correlated.

There are two main challenges in studying privacy-preserving data publication in correlated settings: 1) The first challenge lies in (ε, δ)-multi-dimensional data privacy analysis under the correlated data model. Some efforts have been made to differential privacy analysis for the data publication with side information [12]. For example, a typical method to characterize the impact of the correlation on differential privacy is to replace the global sensitivity with correlation-based parameters [9]. However, these methods cannot be applied in our setup since the privacy notion is different. Meanwhile, the variation of the disclosure probability, say δ, with correlations compared to that without correlations is difficult to characterize directly. This is due to the fact that δ is usually only known in an integral form and thus too complicated to conduct analysis directly. 2) The second challenge is how to find the optimal distributed noise to maximize (ε, δ)-multi-dimensional data privacy under correlated data constraints. The challenge arises because the optimal noise analysis for (ε, δ)-one-dimension data privacy cannot be applied directly, since the optimization problem is different due to the existence of correlated data constraints. Meanwhile, the existence of correlated data constraints introduces a complex effective integral domain for the disclosure probability and thus the optimal noise analysis is hard to derive.

The proposed privacy notion, i.e., (ε, δ)-MDDP, characterizes to what extent an attacker can infer the true data within a given accuracy. In contrast, differential privacy has been defined and applied for quantifying the degree of individual privacy preservation in a statistical database [3]. It is proposed to maximize query accuracy while maintaining the indistin-

This work was supported by the National Key Research and Development Program of China (Grant No.2016YFB0901900), the Natural Science Foundation of China (NSFC) under Grants 61973218 and 61903328, the China Postdoctoral Science Foundation under Grant 2019M652082, and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) and the National Natural Science Foundation of China (NSFC) through Sino-German Research Project number 392194080.

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.

Peng Cheng is with The State Key Laboratory of Industrial Control Technology, Zhejiang University, Hangzhou, China. E-mail:pcheng@iipc.zju.edu.cn.

Daniel E. Quevedo is with the School of Electrical Engineering and Robotics, Queensland University of Technology, Brisbane, Australia. E-mail: dquevedo@ieee.org.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNSE.2020.3044590, IEEE Transactions on Network Science and Engineering

guishability of each entry. Consequently, when considering practical correlations among data, the existing differential privacy analysis techniques cannot be applied to our problem.

Most importantly, the privacy guarantee by database query problems is different from the privacy demand of data publishers in practice. Data privacy is proposed in [4], where privacy analysis for independent data publication is investigated. More privacy definitions (e.g., identifiability, information-theoretic metrics) have been discussed in [12]–[18]. However, for correlated settings, how to quantify the degree of the data privacy protection in terms of the probability of estimation under a given accuracy remains open.

The privacy analysis of data publication with deterministic correlation has been tackled in our recent conference paper [19], in which the probabilistic correlation was not taken into consideration. In our current paper, we have further studied more general and practical cases where probabilistic correlation and noise dependency are considered. Meanwhile, we have provided more details to enrich the motivation, related works, problem formulation, and the simulation. We have also found the optimal noise distribution by deriving the explicit expression of privacy leakage and optimizing over functional spaces. The main contributions of this paper are summarized as follows.

- We extend the definition of data privacy [4] to multi-dimensional correlated data privacy, i.e., ((ε,δ)-MDDP), where both deterministic and probabilistic correlations are considered. Our new notion quantifies disclosure probability of the published correlated data being jointly estimated with a given accuracy.
- We analyze the effects of both kinds of correlations among data on the privacy disclosure. It is shown that using the correlation, the privacy disclosure increases compared to that without correlation knowledge. Furthermore, a closed-form solution of disclosure probability and a strict bound of privacy gain are derived.
- We propose optimal noise adding strategies, in the sense of (ε,δ)-MDDP, for cases with full couplings and with probabilistic couplings.

The remainder of this paper is organized as follows: Related works are revised in Section II. Section III introduces some preliminaries and formulates the problem of interest. Section IV and Section V present the theoretical results of privacy analysis and optimal noise design for data publication with two kinds of correlations, respectively. Section VI verifies the main results through simulations. Conclusions are given in Section VII.

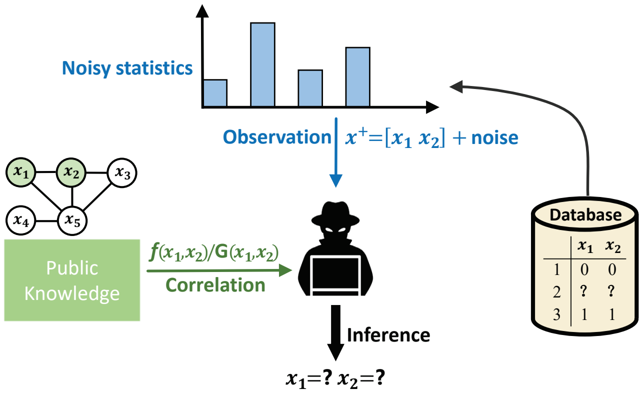


Fig. 1. An attacker attempts to infer the value of x_1, x_2 based on the joint distribution of database x and the published result x^+.

II. Related Works

Many efforts have been devoted to investigating the privacy-preserving data publication problem where the correlation is considered. To solve this problem, a widely used approach is adding random noises to the data to be published [3]. Especially, the differential-privacy-based approach has become a hot research topic, due to its strict indistinguishability guarantee. Three correlation models have been heavily investigated in the privacy analysis for correlated data publication, i.e., attribute correlations, temporal correlations, and spatial correlations.

Specifically, attribute correlations are usually modeled by degree-based methods and the Bayesian network-based methods [12], [20], [21]. For correlation degree-based methods, the correlations among data are usually characterized as a low dimension metric and then used in the important parameters of differential privacy. For example, Zhu *et al.* utilized a correlation coefficient matrix to describe the correlation of a series, and the correlation coefficient was considered as the weight to compute the global sensitivity [20]. In Bayesian network-based methods, the correlations are modeled through Bayesian network and then applied in privacy preservation analysis. In particular, Zhang *et al.* constructed a Bayesian network to model the attribute correlation in high-dimensional data and then synthesized a privacy-preserving dataset in an ad hoc way [21]. Temporal correlations are usually modeled by coupled hidden Markov model-based methods (HMM) [22]–[24]. For example, Xiao *et al.* used HMM to model the temporal correlations between user's locations [23]. Then, they showed that the well known l_1-norm sensitivity in differential privacy fails to capture the geometric sensitivity in multidimensional space and proposed a new notion called, sensitivity hull, based on which the error of differential privacy is bounded.

Spatial correlations are often modeled by the Pearson Correlation Coefficient (PCC). Within this framework, grouping and perturbing the statistics over correlated regions is often applied to avoid noise overdose. As a typical example, Wang *et al.* adopted PCC to measure the similarity of the trend of statistics change [25], and then regions with small statistics and high similarity are grouped together and the same noises are added to reduce errors.

Most of the existing works on noise adding mechanisms are based on differential privacy. The privacy disclosure probability, which characterizes to what extent an attacker can infer the true data within a given accuracy, is neglected. In many application domains, the strict indistinguishability that can be well quantified by differential privacy cannot meet the requirement. For example, in many medical databases, the presence of the record of an individual is not a secret. Instead, the exact value of an individual's disease record

TABLE I
Important Notations

Symbol	Definition
x	the original data vector of N random attributes; variable
x_i	the i-th element of x; variable
χ	the domain of x
χ_i	the domain of x_i
G_X	the full-coupling function of original data
X	the corresponding random vector of N attributes
f_X	the probabilistic coupling among original data
Θ	the domain of θ
Θ_i	the domain of random variable θ_i
f_θ	the joint probability density function of θ
x^+	the observed data vector; variable
ν	the possible original data; variable
$\hat{x}^*$	the optimal estimation of x; variable
ε	the estimation accuracy of x; constant
δ	the joint disclosure probability of x; variable

is. It is worth mentioning that data privacy preserving data publication and its application in network systems have been well studied [4], [26]. However, only one-dimension data privacy preservation is taken into consideration. The key issue on how the correlation affects data privacy and what kind of noise distribution can achieve maximal privacy, remains open. The present work aims to fill this knowledge gap.

III. Preliminaries and Problem Formulation

Consider that one person needs to broadcast his/her real-valued data, which describes N private attributes denoted by random vector X. Hence, there exists a correlation among different terms of the data. Let $x = [x_1, \cdots, x_N]^{\mathrm{T}}$ be a real-valued data vector, i.e., $x \in \mathbb{R}^N$, where N is the dimension of the vector. In this paper, we consider the two common types of correlations, i.e., full coupling and probabilistic coupling. Full coupling is common among numeric data, e.g., the relation between weight in kg (x_1) and weight in lb (x_2) of one person is a classic example of full coupling (i.e., $0.453x_1 - x_2 = 0$), the relation between personal income and personal income tax [9], etc. Probabilistic coupling is usual among attribute data, e.g., the weight and the blood pressure of people are often joint normally distributed [10]. It models the fact that the attributes corresponding to an individual entry are correlated in general and consequently can reveal information about one another. The specific definitions are given below by referring to [9], [10].

*Definition 1 (**Full coupling**):* Full coupling means the original data is correlated by an equation $G(x_1, \cdots, x_N) = 0$ (G_X in abbreviation). Specifically, $G(x_1, \cdots, x_N) = 0$ is a multivariate explicit function, i.e., any variable can be expressed as an explicit function of all other variables.

*Definition 2 (**Probabilistic coupling**):* Probabilistic coupling means the original data is correlated by a joint probability density function $f_X(x_1, \cdots, x_N)$ (f_X in abbreviation).

To preserve the privacy of the sensitive correlated data vector, random noise is added, i.e.,

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

where $\theta = [\theta_1, \cdots, \theta_N]^{\mathrm{T}}$ is a random noise vector, and x^+ is the published data vector. Let $f_\theta(z_1, \cdots, z_N)$ (f_θ in abbreviation) be the joint probability density function (PDF) of N entries of θ. Table I summarizes important notations for easy reference.

A. Attack Model

Suppose that there is an attacker outside the users' data publication group. The attacker is able to eavesdrop the broadcast information x^+ and knows the PDF of noises. The attacker aims to infer the true data vector x, see Fig.1. Let $\hat{x}$ be an estimation of x, where $\hat{x}_i$ represents i-th element of $\hat{x}$. The attacker can infer x 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. We define ε-accurate estimation as follows:

*Definition 3 (**ε-accurate estimation**):* Consider that the attacker knows x^+ and f_θ. Let $\hat{x}$ be an estimate of variable x. If $||\hat{x} - x||_\infty \leq \varepsilon$, where $\varepsilon \geq 0$ is a small constant, then we say $\hat{x}$ is an ε-accurate estimation.

Then, we have

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

Inspired by [4], we provide one important definition below:

*Definition 4 (**Optimal estimation**):* Considering the ε-accurate estimation, given x^+, G_X or f_X, and f_θ, the optimal estimation of x is defined as

$$\hat{x}^* = \arg\max_{\hat{x} \in \chi} \Pr\{\nu + \theta = x^+ | G_X/f_X, ||\nu - \hat{x}||_\infty \leq \varepsilon\}, \tag{3}$$

where the random vector $\nu = [\nu_1, \cdots, \nu_N]^{\mathrm{T}}$ denotes the possible value of x, and χ is the domain of x.

Remark 1: In (3), ν is an arbitrary possible value in χ, and the constraint $||\nu - \hat{x}||_\infty \leq \varepsilon$ limits the estimated value within an ε-accuracy of the arbitrary possible value in χ, i.e., $\hat{x}$ is an ε-accurate estimation. Since the attacker only knows the domain of the published data and the published information, any value in that domain can be the true value from the perspective of the attacker. As a result, the attacker should estimate the true value such that the disclosure probability is the largest, which means that that estimate is mostly like the true one. Hence, we define the optimal estimation of x as (3) to maximize the disclosure probability for the attacker. In (3), the term G_X/f_X (G_X or f_X) is side information of the optimal estimation, and we explore the effects of deterministic and probabilistic correlations on privacy disclosure, respectively. Both the distribution and the estimation range of θ are affected by G_X or f_X, and the details will be discussed later.

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

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

which is consistent with that in [4].

B. Privacy Definition

To quantify the degree of privacy protection of correlated data publication, the relationship between estimation accuracy and privacy is constructed as follows.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNSE.2020.3044590, IEEE Transactions on Network Science and Engineering

*Definition 5 ((ε,δ)-**MDDP**):* A noise adding mechanism (1) satisfies (ε,δ)-multi-dimensional data-privacy, 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 5 quantifies the probability that the attacker can successfully estimate each entry of x in a given interval $[x_i-\varepsilon, x_i+\varepsilon], \forall i \in [1,N]$, using the optimal estimation. A smaller value of ε offers higher accuracy, and a smaller value of δ offers a smaller disclosure probability. Combining (3) and (5), we can obtain δ as a function of G_X and f_X.

C. Problems of Interest

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

1) How do the data correlations affect the disclosure probability compared to that the correlation is unknown? What is the bound of privacy gain (if it exists)?
2) Does there exist a closed-form expression of the optimal estimation and the disclosure probability considering different data correlations?
3) What is the optimal PDF $f_\theta(z_1,\cdots,z_N)$ in the sense of (ε,δ)-MDDP? This optimization problem can be formulated as

$$\begin{aligned} &\min_{f_\theta(z_1,\cdots,z_N)} && \delta \\ &s.t. && \mathrm{E}\{\theta_i\} = 0, \\ & && \mathrm{Var}\{\theta_i\} = \sigma^2, \quad i=1,\cdots,N, \end{aligned} \tag{6}$$

where $\sigma^2 > 0$ is constant, $\mathrm{E}\{\theta_i\}$ and $\mathrm{Var}\{\theta_i\}$ take expectation and variance, respectively. The constraints of zero mean and finite variance aim to protect the utility of the published data.

Note that problem (6) is difficult to solve directly, since the explicit expression of δ is complex considering the data correlation. Furthermore, finding the optimal distribution is difficult, since one needs to optimize over functional spaces.

IV. Multi-dimensional Full-coupled Data Publication

In this section, based on the definition of (ε,δ)-MDDP, we reveal an analytical relationship between the full coupling and the privacy disclosure and design the optimal noise adding strategy. In order to make the problem solution explicit, independent noise adding is considered in this section.

A. Privacy Analysis for the Full-coupled Case

In this subsection, we show that the relationship between δ and G_X is hard to describe directly, and reveal an analytical relationship between full coupling and privacy disclosure by an intuitive conversion.

We note that the attacker is able to infer the domain of θ with the knowledge of G_X and f_θ. It turns out that when G_X is known to the attacker, the effective integral domain Θ will become a subset of the original integral domain $\hat{\Theta}$ without considering G_X, i.e., $\Theta \subseteq \hat{\Theta}$. Then, a case showing that δ will decrease is given below.

Example 1: Given noise distribution $f_{\theta_i}(z_i)$ and estimation accuracy ε approaching zero. Let $\Theta_{i,1}$ and $\Theta_{i,0}$ be two noise domains of x_i, and $f_{\theta_i|\Theta_{i,1}}(z_i)_{\max} = \max_{z_i\in\Theta_{i,1}} f_{\theta_i}(z_i)$, $f_{\theta_i|\Theta_{i,0}}(z_i)_{\max} = \max_{z_i\in\Theta_{i,0}} f_{\theta_i}(z_i)$. Then, if $f_{\theta_i|\Theta_{i,1}}(z_i)_{\max} > f_{\theta_i|\Theta_{i,0}}(z_i)_{\max}$, from Definitions 4 and 5, we have,

$$\begin{aligned} \lim_{\varepsilon\to 0}\frac{\delta|\Theta_{i,1}}{\delta|\Theta_{i,0}} &= \lim_{\varepsilon\to 0}\frac{\max_{\hat{\theta}_i\in\Theta_{i,1}}\int_{\hat{\theta}_i-\varepsilon}^{\hat{\theta}_i+\varepsilon} f_{\theta_i}(z_i)\mathrm{d}z_i}{\max_{\hat{\theta}_i\in\Theta_{i,0}}\int_{\hat{\theta}_i-\varepsilon}^{\hat{\theta}_i+\varepsilon} f_{\theta_i}(z_i)\mathrm{d}z_i} \\ &= \lim_{\varepsilon\to 0}\frac{2\varepsilon * f_{\theta_i|\Theta_{i,1}}(z_i)_{\max}}{2\varepsilon * f_{\theta_i|\Theta_{i,0}}(z_i)_{\max}} > 1, \end{aligned} \tag{7}$$

where $\Theta_{i,0} \subset \Theta_{i,1}$. The reason that the first equality of (7) holds is, based on x_i^+ and $f_{\theta_i}(z_i)$, the attacker can take the maximum probability over the noise domain $\Theta_{i,1}$ to estimate the value of the added noise, i.e.,

$$\begin{aligned} \delta|\Theta_{i,1} &= \max_{\hat{x}_i\in\{x_i^+-\Theta_{i,1}\}} \Pr\{\nu_i+\theta_i = x_i^+ | |\nu_i-\hat{x}_i| \le \varepsilon\} \\ &= \max_{\hat{\theta}_i\in\Theta_{i,1}} \int_{\hat{\theta}_i-\varepsilon}^{\hat{\theta}_i+\varepsilon} f_{\theta_i}(z_i)dz_i. \end{aligned}$$

Example 1 shows that knowing G_X means the attacker knows more domain information (i.e., $\Theta_{i,0} \subset \Theta_{i,1}$). However, the disclosure probability δ, does not consider the effective integral domain of the added noise as the prior knowledge. Having a smaller integral domain explains the fact that δ decreases when G_X is known to the attacker. To further investigate the effect of domain change, we propose the successful disclosure rate, i.e., γ, where the effective integral domain is viewed as the prior knowledge when G_X is available.

*Definition 6 (**Successful disclosure rate**):* With the knowledge of effective integral domain Θ_i and x_i^+, the successful disclosure rate of x_i is defined as

$$\gamma_i = \begin{cases} \dfrac{\int_{\hat{\theta}_i^*-\varepsilon}^{\hat{\theta}_i^*+\varepsilon} f_{\theta_i}(z_i)\mathrm{d}z_i}{\int_{z_i\in\Theta_i} f_{\theta_i}(z_i)\mathrm{d}z_i} \times 100\%, & \text{if } \hat{\theta}_i^* \in \Theta_i, \\ 0, & \text{otherwise,} \end{cases} \tag{8}$$

where $\Theta_i \subseteq \hat{\Theta}_i$ is the effective integral domain of θ_i when G_X is available, and $\hat{\theta}_i^*$ is the optimal estimation of θ_i obtained by using (3). If G_X is unknown to the attacker, use original noise domain $\hat{\Theta}_i$ as integral domain. The successful disclosure rate of x is $\gamma = \prod_{i=1}^N \gamma_i$.

Remark 3: Mathematically, the relation between γ and δ satisfies Bayes theorem. Let A be the event that estimating x with the ε-accuracy under the condition that G_X is not available, and B is the event that the noise is in the effective integral domain. Thus, we have

$$\Pr\{A|B\} = \gamma = \frac{\Pr\{A\} = \delta}{\Pr\{B\}},$$

where $\delta = \Pr\{A\}$ can be viewed as the prior probability, $\gamma = \Pr\{A|B\}$ is the posterior probability, and $\Pr\{B\}$ is the integral of noise distribution function in the effective integral domain. Moreover, from Definition 6, 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 rate 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 what kind of $f_{\theta_i}(z_i)$ is used.

Theorem 1: Consider the mechanism (1). If the original data is full-coupled by G_X, then the successful disclosure rate increases, i.e.,

$$\gamma_{\varnothing} \leq \gamma, \tag{9}$$

where $\gamma_{\varnothing}$ is the successful disclosure rate of x without correlation knowledge.

Proof: By using the full coupling and other elements' domain information, if there exists an element in θ whose domain is narrowed down, we denote such element by θ_i. Then we have $\Theta_i \subset \hat{\Theta}_i$, where $\hat{\Theta}_i$ is the original integral domain of θ_i when G_X is unknown to the attacker, and Θ_i is the effective integral domain of θ_i narrowed down by G_X. Thus,

$$\int_{z_i \in \Theta_i} f_{\theta_i}(z_i)\mathrm{d}z_i < \int_{z_i \in \hat{\Theta}_i} f_{\theta_i}(z_i)\mathrm{d}z_i. \tag{10}$$

Let $\hat{\theta}_i^* \in \hat{\Theta}_i$ be the optimal estimation of θ_i obtained by using (3). Then, we divide it into two cases as follows:

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

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

From Definition 5, we directly obtain

$$\begin{aligned}\gamma_i = \frac{\int_{\hat{\theta}_i^*-\varepsilon}^{\hat{\theta}_i^*+\varepsilon} f_{\theta_i}(z_i)\mathrm{d}z_i}{\int_{z_i \in \Theta_i} f_{\theta_i}(z_i)\mathrm{d}z_i} \times 100\% > \frac{\int_{\hat{\theta}_i^*-\varepsilon}^{\hat{\theta}_i^*+\varepsilon} f_{\theta_i}(z_i)\mathrm{d}z_i}{\int_{z_i \in \hat{\Theta}_i} f_{\theta_i}(z_i)\mathrm{d}z_i} \times 100\% \\ > \gamma_{i,\varnothing},\end{aligned} \tag{11}$$

where γ_i is the successful disclosure rate of x_i using the full coupling, $\gamma_{i,\varnothing}$ is the successful disclosure rate of x_i without correlation knowledge. Since $\gamma_j = \gamma_{j,\varnothing}$, $\forall j \in [1,N]$, $j \neq i$, and $\gamma = \prod_{k=1}^{N} \gamma_k$, (9) holds.

2) $\hat{\theta}_i^* \notin \Theta_i$. This occurs when the optimal estimation of θ_i varies as Θ_i is smaller due to full coupling. Then, we first consider the case $\max_{\hat{\theta}_i \in \Theta_i} \int_{\hat{\theta}_i-\varepsilon}^{\hat{\theta}_i+\varepsilon} f_{\theta_i}(z_i)\mathrm{d}z_i \leq \max_{\hat{\theta}_i \in \hat{\Theta}_i} \int_{\hat{\theta}_i-\varepsilon}^{\hat{\theta}_i+\varepsilon} f_{\theta_i}(z_i)\mathrm{d}z_i$. This case occurs when the optimal estimation $\hat{\theta}_i^*$ without correlation knowledge is not within Θ_i. From Definition 5, $\gamma_{i,\varnothing} = 0$ holds. Thus, we have

$$\gamma_i > 0,\ \gamma_{i,\varnothing} = 0, \tag{12}$$

and (9) holds. The second case is $\max_{\hat{\theta}_i \in \Theta_i} \int_{\hat{\theta}_i-\varepsilon}^{\hat{\theta}_i+\varepsilon} f_{\theta_i}(z_i)\mathrm{d}z_i > \max_{\hat{\theta}_i \in \hat{\Theta}_i} \int_{\hat{\theta}_i-\varepsilon}^{\hat{\theta}_i+\varepsilon} f_{\theta_i}(z_i)\mathrm{d}z_i$. This case is impossible when $\Theta_i \subset \hat{\Theta}_i$.

The equality of (9) 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 with the full coupling, the attacker can obtain more information about the domain, and the uncertainty of x is decreased. Thus, the rate that the attacker can make an accurate estimation is higher. In Theorem 1, γ is a transformation of δ and we have shown that γ increases when G_X is available. As a direct result of Theorem 1, we have

$$\begin{aligned}\delta &= \Pr\{||\hat{x}^* - x||_\infty \leq \varepsilon |\, x^+, \chi\} \\ &\geq \Pr\{||\hat{x}^* - x||_\infty \leq \varepsilon |\, x^+, G_X, \chi\}.\end{aligned} \tag{13}$$

This inequality holds because even though G_X narrows the original integral domain down to a smaller integral domain, δ is not an increasing function of a smaller integral domain. Hence, the correlation G_X will not increase δ.

B. Optimal Noise Design

In this subsection, a closed-form solution of optimal estimation and a strict bound of privacy disclosure gain are derived. The optimal noise adding strategy is proposed to minimize the disclosure probability considering full coupling.

Here we first provide a lemma to show that when data correlation is unknown to the attacker, the optimal noise distribution is the uniform one. It gives the optimal solution to problem (6) when $N = 1$.

Lemma 1: [26] If x_i^+ is the only information available to the attacker, then the optimal solution to problem (6) is

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

Theorem 2: If the original data in (1) is coupled by an explicit function G_X, and independent noises are added, then the attacker can obtain the N-dimensional joint optimal estimation

$$\begin{cases} \hat{x}_j^* = x_j^+ - \arg \max\limits_{\hat{\theta}_j \in \Theta_j} \int_{\hat{\theta}_j-\varepsilon}^{\hat{\theta}_j+\varepsilon} f_{\theta_j}(z_j)\mathrm{d}z_j, \forall j \neq i \\ \hat{x}_i^* = \{\hat{x}_i |\, G(\hat{x}_i, \hat{x}_{-i}^*) = 0, \hat{x}_i \in \chi_i\}, \end{cases} \tag{14}$$

where $i = \arg\min_{\{k=1,\cdots,N\}} \frac{\Pr\{|\hat{x}_k^* - x_k| \leq \varepsilon | \hat{x}_k \in \chi_k\}}{\int_{\inf\{\Theta_k\}}^{\sup\{\Theta_k\}} f_{\theta_k}(z_k)\mathrm{d}z_k}$, and $\hat{x}_{-i}^*$ are all elements of $\hat{x}^*$ except for $\hat{x}_i^*$. In the sense of (ε, δ)-MDDP, $\forall k \in \{1, \cdots, N\}$, the optimal noise distribution satisfies

$$f_{\theta_k}^*(z_k) = \begin{cases} \frac{1}{2\sqrt{3}\sigma}, & \text{if } z_k \in \left[-\sqrt{3}\sigma, \sqrt{3}\sigma\right], \\ 0, & \text{otherwise.} \end{cases} \tag{15}$$

Proof: We prove it from two-dimensional coupled vector $[x_i, x_j]$. Coupling function $G(x_i, x_j) = 0$ can be any explicit function. With the dynamic but observable x^+, the attacker can use the fact that noises added by users satisfy:

$$x_i = x_i^+ - \theta_i,\ x_j = x_j^+ - \theta_j. \tag{16}$$

It directly follows that

$$G(x_i, x_j) = G(x_i^+ - \theta_i, x_j^+ - \theta_j) = 0, \tag{17}$$

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

$$\begin{aligned}\hat{x}^* &= \arg\max_{\hat{x}\in\chi} \Pr\{\nu+\theta=x^+|G_X,||\nu-\hat{x}||_\infty\leq\varepsilon\}\\ &= \arg\max_{\hat{x}\in\chi} \Pr\{\nu_i+\theta_i=x_i^+,\nu_j+\theta_j=x_j^+|G(x_i,x_j)=0,\\ &\qquad\qquad\qquad\qquad ||\nu-\hat{x}||_\infty\leq\varepsilon\}\\ &= x^+ - \arg\max_{\substack{\hat{\theta}\in\{x^+-\chi\}\\ G_X}} \int_{\hat{\theta}_j-\varepsilon}^{\hat{\theta}_j+\varepsilon}\int_{\hat{\theta}_i-\varepsilon}^{\hat{\theta}_i+\varepsilon} f_{\theta_i,\theta_j}(z_i,z_j)\mathrm{d}z_i\mathrm{d}z_j\\ &= x^+ - e_{\theta|G_X}(x^+),\end{aligned} \tag{18}$$

where $e_{\theta|G_X}(x^+)$ is an estimation of θ using x^+ and G_X. Recall 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

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

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

$$\hat{x}^* = x^+ - \arg\max_{\substack{\hat{\theta}_j\in\Theta_j\\ G(x_i^+-\widetilde{\theta}_i,\, x_j^+-\hat{\theta}_j)=0}} \int_{\hat{\theta}_j-\varepsilon}^{\hat{\theta}_j+\varepsilon} f_{\theta_i|\theta_j=z_j}(\widetilde{\theta}_i) f_{\theta_j}(z_j)\mathrm{d}z_j, \tag{21}$$

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

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

where $f_{\theta_i|\theta_j=z_j}(\cdot)$ is the conditional PDF of θ_i under the condition $\theta_j=z_j$ ($z_j\in\Theta_j$), and the value of θ_i is fixed when x^+ is fixed.

Then, the multi-dimensional joint estimation (18) is reduced to two-phase one-dimensional estimation as follows.

i) The attacker uses x_i^+,x_j^+ and full-coupling $G(x_i,x_j)=0$. From the analysis of Theorem 1, if there exists

$$\frac{\Pr\{|\hat{x}_j^*-x_j|\leq\varepsilon|\hat{x}_j\in\chi_j\}}{\int_{\inf\{\Theta_j\}}^{\sup\{\Theta_j\}} f_{\theta_j}(z_j)\mathrm{d}z_j} > \frac{\Pr\{|\hat{x}_i^*-x_i|\leq\varepsilon|\hat{x}_i\in\chi_i\}}{\int_{\inf\{\Theta_i\}}^{\sup\{\Theta_i\}} f_{\theta_i}(z_i)\mathrm{d}z_i}, \tag{23}$$

then, to achieve an accurate estimation of x with a higher possibility, there exists a higher priority for the attacker to target x_j. From (18) and (22), the optimal estimation of x_j is

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

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

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

Then, we investigate the optimal noise adding on data vector x. Since independent noises are added, problem (6) is decoupled as

$$\min_{f_{\theta_k}(z_k)}\max_{\hat{\theta}_k\in\Theta_k}\int_{\hat{\theta}_k-\varepsilon}^{\hat{\theta}_k+\varepsilon} f_{\theta_k}(z_k)\mathrm{d}z_k, \forall k\in\{1,\cdots,N\}, \tag{26}$$

and the solution to problem (26) is obtained in Lemma 1. ■

Remark 5: Note that the optimal noise for the index i in (14) is also uniformly distributed, because the optimal noise in terms of MDDP is only determined by the local noise domain. Furthermore, the index i is not only determined by MDDP ($\Pr\{|\hat{x}_k^*-x_k|\leq\varepsilon|\hat{x}_k\in\chi_k\}$), but also relies on the effective integral domain, which is influenced by G_X. Full coupling is identified to reduce the dimension of joint estimation effectively. In Theorem 2, both the closed-form δ and optimal noise adding strategy are derived. Meanwhile, the term $\sqrt{3}$ comes from the uniform distribution shown in (15), i.e., $\mathrm{Var}\{\theta_j\}=\frac{(\sqrt{3}\sigma-(-\sqrt{3}\sigma))^2}{12}=\sigma^2$. Furthermore, for (25), $\hat{x}_i$ must be selected to satisfy both the coupling function and domain constraints, which are available to the attacker.

*Theorem 3 (**Bound of privacy gain**):* Consider (1), if K pairs of original data are coupled by G_X, given noise distribution satisfying (15), and $x_k\in[-\frac{M}{2},\frac{M}{2}]$, $\forall k\in[1,N]$. Then,

$$\frac{\gamma-\gamma_\varnothing}{\gamma_\varnothing}\leq([\frac{2\varepsilon}{M}]^{-K}-1)(M\gg\varepsilon), \tag{27}$$

holds, where M is a positive number.

Proof: From Theorem 2, it follows that with G_X, 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 (14). Under the optimal noise derived in (15) and given $x_k\in[-\frac{M}{2},\frac{M}{2}],\forall k\in[1,N]$, the γ of x_{-i} is $[\frac{2\varepsilon}{M}]^{N-1}$, i.e., the γ of vector x is $[\frac{2\varepsilon}{M}]^{N-1}$. Without knowing the correlation, the $\gamma_\varnothing$ 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 6: Theorem 1 and Theorem 3 conclude that using the full coupling, the successful disclosure rate γ increases compared to that without correlation knowledge, and a strict promotion bound is derived. To mitigate privacy leakage owing to correlation, there are two protection strategies.

1) We can use 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.
2) We can reselect the noise variance σ^2. A larger 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.

V. Multi-dimensional Probabilistic-coupled Data Publication

In this section, we provide an information-theoretic approach to investigate the analytical relationship between probabilistic coupling and (ε,δ)-MDDP. Then, it is shown that joint uniform noise is optimal in the sense of (ε,δ)-MDDP. It should be pointed out that only two attributes among N attributes are coupled by $f_{\{X_i,X_j\}}$ in this paper. This is a rather simplified assumption. The resulting theoretical conclusion of optimal noise will be of an explicit form. The analysis of effects of probabilistic coupling for two attributes can be applied directly for N attributes using Bayes theorem.

A. Data Inference based on Mutual Information

Considering that the effect of probabilistic coupling on (ε,δ)-MDDP cannot be characterized by the preceding methods directly, we use mutual information to capture the change of noise variance. In the following part, we suppose that x_i and x_j are correlated and the corresponding random variables are denoted by X_i and X_j, respectively. We provide mutual information definition before the details.

*Definition 7 (**Mutual information** [27]):* Consider two random variables $X_i \in \chi_i$ and $X_j \in \chi_j$. The mutual information $I(X_i;X_j)$ is defined as a measure of dependency

$$I(X_i;X_j)=\int_{x_j\in\chi_j}\int_{x_i\in\chi_i} f_{\{X_i,X_j\}}(x_i,x_j)\log\frac{f_{\{X_i,X_j\}}(x_i,x_j)}{f_{X_i}(x_i)f_{X_j}(x_j)}\mathrm{d}x_i\mathrm{d}x_j, \tag{28}$$

where $f_{\{X_i,X_j\}}(x_i,x_j)$ is the joint probability density function, $f_{X_i}(x_i)$ and $f_{X_j}(x_j)$ are the marginal probability density functions.

With $\{x_i^+,x_j^+\}$ available only, from [4], the optimal estimation of data $\{x_i,x_j\}$ satisfies

$$\begin{cases} \hat{x}_i^* = x_i^+ - \arg\max\limits_{\hat{\theta}_i\in\Theta_i} \int_{\hat{\theta}_i-\varepsilon}^{\hat{\theta}_i+\varepsilon} f_{\theta_i}(z_i)\mathrm{d}z_i, \\ \hat{x}_j^* = x_j^+ - \arg\max\limits_{\hat{\theta}_j\in\Theta_j} \int_{\hat{\theta}_j-\varepsilon}^{\hat{\theta}_j+\varepsilon} f_{\theta_j}(z_j)\mathrm{d}z_j, \end{cases} \tag{29}$$

when $\{x_i^+,x_j^+\}$ are released, the values of added noises $\{\theta_i,\theta_j\}$ are fixed. However, for the attacker, $\{\theta_i,\theta_j\}$ are still viewed as random variables with PDF $f_{\theta_i}(z_i)$ and $f_{\theta_j}(z_j)$, respectively.

With $f_{\{X_i,X_j\}}$ known to the attacker, the mutual information $I(X_i;X_j)$ is known. When x_j is information available to the inference of x_i, the uncertainty of x_i decreases due to the correlation among them. Thus, a larger $I(X_i;X_j)$ indicates a higher level privacy leakage of x_i when x_j is disclosed [27]. To make an accurate estimation, $\{\hat{x}_i^*,\hat{x}_j^*\}$ should satisfy the correlation fact. Combine it with (29), i.e., the uncertainty of $\hat{x}_i^*$ stems from the a new noise PDF $f_{\theta_i|I(X_i;X_j)}(z_i)$ with less uncertainty than $f_{\theta_i}(z_i)$, where $f_{\theta_i|I(X_i;X_j)}(z_i)$ is the conditional PDF of θ_i under the condition $\theta_j=\widetilde{\theta}_j$, and $\widetilde{\theta}_j$ is obtained by using (29). Then, we have

$$\Pr\{|\hat{x}_i-x_i|\le\varepsilon|\,x_i^+,\widetilde{\theta}_j\}=\int_{\hat{\theta}_i-\varepsilon}^{\hat{\theta}_i+\varepsilon} f_{\theta_i|I(X_i;X_j)}(z_i)\mathrm{d}z_i. \tag{30}$$

Since $f_{\theta_i|I(X_i;X_j)}(z_i)$ has less uncertainty than $f_{\theta_i}(z_i)$, and if $f_{\theta_i|I(X_i;X_j)}(z_i)$ and $f_{\theta_i}(z_i)$ have the same distribution (but the mean and variance would not be the same), it follows that $f_{\theta_i|I(X_i;X_j)}(z_i)$ will have a smaller variance than $f_{\theta_i}(z_i)$. From the property of PDF, one obtains that disclosure probability increases, i.e.,

$$\begin{aligned}\max_{\hat{x}_i\in\chi_i}\Pr\{|\hat{x}_i-x_i|\le\varepsilon|\,x_i^+,\widetilde{\theta}_j\}&=\max_{\hat{\theta}_i\in\Theta_i}\int_{\hat{\theta}_i-\varepsilon}^{\hat{\theta}_i+\varepsilon} f_{\theta_i|I(X_i;X_j)}(z_i)\mathrm{d}z_i\\ &\ge\max_{\hat{\theta}_i\in\Theta_i}\int_{\hat{\theta}_i-\varepsilon}^{\hat{\theta}_i+\varepsilon} f_{\theta_i}(z_i)\mathrm{d}z_i.\end{aligned} \tag{31}$$

Note that if both θ_i and θ_j are not fixed for joint estimation, the probabilistic coupling cannot reduce the noise uncertainty. Then, from (31), the disclosure probability of data $\{x_i,x_j\}$ generated from $f_{\{X_i,X_j\}}$ satisfies

$$\begin{aligned}&\Pr\{||\hat{x}-x||_\infty\le\varepsilon|\,\{x_i^+,x_j^+\},f_{\{X_i,X_j\}}\}\\ \le&\max_{[\hat{\theta}_i,\hat{\theta}_j]\in\Theta}\int_{\hat{\theta}_j-\varepsilon}^{\hat{\theta}_j+\varepsilon}\int_{\hat{\theta}_i-\varepsilon}^{\hat{\theta}_i+\varepsilon} f_{\theta_i,\theta_j}(z_i,z_j)\mathrm{d}z_i\mathrm{d}z_j\\ \le&\max_{[\hat{\theta}_i,\hat{\theta}_j]\in\Theta}\int_{\hat{\theta}_j-\varepsilon}^{\hat{\theta}_j+\varepsilon}\int_{\hat{\theta}_i-\varepsilon}^{\hat{\theta}_i+\varepsilon} f_{\theta_j}(z_j)f_{\theta_i|\theta_j=z_j}(z_i|z_j)\mathrm{d}z_i\mathrm{d}z_j\\ \le&\max_{\hat{\theta}_i\in\Theta_i}\int_{\hat{\theta}_i-\varepsilon}^{\hat{\theta}_i+\varepsilon} f_{\theta_i}(z_i)\mathrm{d}z_i\cdot\max_{\hat{\theta}_j\in\Theta_j}\int_{\hat{\theta}_j-\varepsilon}^{\hat{\theta}_j+\varepsilon} f_{\theta_j}(z_j)\mathrm{d}z_j\\ \le&\max_{\hat{\theta}_i\in\Theta_i}\int_{\hat{\theta}_i-\varepsilon}^{\hat{\theta}_i+\varepsilon} f_{\theta_i|I(X_i;X_j)}(z_i)\mathrm{d}z_i\cdot\max_{\hat{\theta}_j\in\Theta_j}\int_{\hat{\theta}_j-\varepsilon}^{\hat{\theta}_j+\varepsilon} f_{\theta_j}(z_j)\mathrm{d}z_j.\end{aligned} \tag{32}$$

The last inequality holds when $\hat{\theta}_j\to\hat{\theta}_i$ forms an estimation chain, i.e., the optimally estimated θ_j is used to decrease the uncertainty of θ_i.

Suppose that the probabilistic coupling of N attributes is described by the probability density function f_X. To disclose the information of any dimension, e.g., X_1, we can use Bayes theorem to model all other dimension information X/X_1 as one dimension information, say X_1^-, using the correlation relationship [12], i.e.,

$$\begin{aligned}&\max_{\hat{x}\in\chi}\Pr\{||\hat{x}-x||_\infty\le\varepsilon|\,x^+,f_X\}\\ =&\max_{\hat{\theta}_1\in\Theta_1}\int_{\hat{\theta}_1-\varepsilon}^{\hat{\theta}_1+\varepsilon} f_{\theta_1|I(X_1;X_1^-)}(z_1)\mathrm{d}z_1\cdot\max_{\hat{\theta}\in\Theta}\int_{\hat{\theta}_2-\varepsilon}^{\hat{\theta}_2+\varepsilon}\cdots\int_{\hat{\theta}_N-\varepsilon}^{\hat{\theta}_N+\varepsilon}\\ &f_{\theta_2,\cdots,\theta_N}(z_N,\cdots,z_2)\mathrm{d}z_N\cdots\mathrm{d}z_2.\end{aligned} \tag{33}$$

As a result, the analysis of the effect of probabilistic coupling for two attributes can be applied directly for N attributes.

Remark 7: In (32), it is hard to obtain the analytical expression directly. Luckily, we can use mutual information to characterize the change of δ under $f_{\theta_i}(z_i)\to f_{\theta_i|I(X_i;X_j)}(z_i)$. As $I(X_i;X_j)$ is known to the attacker, the uncertainty reduction of PDF $f_{\theta_i}(z_i)$ is fixed and equivalent to $I(X_i;X_j)$. Then, if $f_{\theta_i|I(X_i;X_j)}(z_i), f_{\theta_i}(z_i)$ have the same distribution, the uncertainty reduction leads to variance reduction (computable). Thus, we have the last inequality hold, which means that the disclosure probability will not decrease with probabilistic coupling.

B. Privacy Analysis for the Probabilistic-coupled Case

In this subsection, we reveal an analytical relationship involving probabilistic coupling, disclosure probability and a bound of disclosure probability.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNSE.2020.3044590, IEEE Transactions on Network Science and Engineering

Theorem 4: Suppose that in (1) x_i and x_j are coupled by $f_{\{X_i,X_j\}}$, and that $f_{\theta_i|I(X_i;X_j)}(z_i), f_{\theta_i}(z_i)$ have the same distribution. Then the disclosure probability increases, i.e.,

$$\delta_{\varnothing} \leq \delta, \tag{34}$$

where $\delta_{\varnothing}$ is the disclosure probability of x without correlation knowledge. Further, δ is upper bounded by

$$\max_{\hat{\theta}_i\in\Theta_i}\int_{\hat{\theta}_i-\varepsilon}^{\hat{\theta}_i+\varepsilon} f_{\theta_i|I(X_i;X_j)}(z_i)\mathrm{d}z_i \cdot \max_{\hat{\theta}_j\in\Theta_j}\int_{\hat{\theta}_j-\varepsilon}^{\hat{\theta}_j+\varepsilon} f_{\theta_j}(z_j)\mathrm{d}z_j, \tag{35}$$

where $\hat{\theta}_j \to \hat{\theta}_i$ forms an estimation chain.

Proof: Given the estimation chain, where x_j is to be inferred first, the probability that the attacker can accurately infer the value of x_j from $f_{\theta_j}(z_j), \forall j \in [1,N]$ is characterized by δ_j. Then, from the analysis (29), (30) and (31), we have $\delta_i \geq \delta_{i,\varnothing}$, where δ_i is the disclosure probability of x_i using $f_{\{X_i,X_j\}}$ and $\delta_{i,\varnothing}$ is the disclosure probability of x_i without correlation knowledge. Then we conclude that (34) holds. The expression (35) is obtained from (32), and the upper bound reaches when $\hat{\theta}_j \to \hat{\theta}_i$ forms an estimation chain. ■

Remark 8: In Theorem 4, since the disclosure probability changes with the specific form of noise distribution, $f_{\{X_i,X_j\}}$, and ε, we cannot provide a more explicit form bound of disclosure probability here. However, once the above information is given, then the expression of $f_{\theta_i|I(X_i;X_j)}(z_i)$ is known, and the bound value is easily obtained using (35).

C. Optimal Noise Distribution

In this subsection, we find the optimal joint distribution of noises in the sense of achieving the highest (ε,δ)-MDDP.

The optimization problem is given below.

$$\begin{aligned} &\min_{f_\theta(z_1,\cdots,z_N)} && \delta \\ &s.t. && \mathrm{E}\{\theta_i\} = 0, \\ & && \mathrm{Var}\{\theta_i\} = \sigma^2, \quad i = 1,\cdots,N, \end{aligned} \tag{36}$$

and the covariance is unconstrained.

Theorem 5: If $\{x_i^+, x_j^+\}$ is the only information available to the attacker, then the optimal solution of (36) is

$$f^*_{\theta_i,\theta_j}(z_i,z_j) = \begin{cases} \frac{1}{4\pi\sigma^2}, & \text{if } z_i^2+z_j^2 \leq 4\sigma^2, \\ 0, & \text{otherwise}, \end{cases} \tag{37}$$

where the covariance matrix is $\begin{bmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{bmatrix}$, i.e., given finite noise variance, the joint uniform distribution is optimal in the sense of (ε,δ)-MDDP.

Proof: We prove the optimality by contradiction. The mean vector $\mathrm{E}\{\theta\} = [0,0]^{\mathrm{T}}$, and the covariance matrix

$$\Sigma_\theta = \begin{bmatrix} \sigma^2 & \mathrm{cov}(\theta_i,\theta_j) \\ \mathrm{cov}(\theta_j,\theta_i) & \sigma^2 \end{bmatrix}.$$

Without loss of generality, we assume that $\sigma^2 = \frac{1}{4}$. Let $f_1(z_i,z_j)$ and $f_2(z_i,z_j)$ be the joint PDF of two random variables both with mean 0 and variance $\sigma^2 = \frac{1}{4}$. Suppose they follow a joint uniform and non-uniform distribution, respectively. From (37), we have

$$f_1(z_i,z_j) = \begin{cases} \frac{1}{\pi}, & \text{if } (z_i,z_j) \in G, \\ 0, & \text{otherwise}, \end{cases} \tag{38}$$

where $G = \{(z_i,z_j) | z_i^2+z_j^2 \leq 1\}$.

Suppose that the non-uniform distribution $f_2(z_i,z_j)$ is the optimal distribution, i.e., there exists an ε_1, such that

$$\begin{aligned} &\max_{a,b\in R}\int_{a-\varepsilon}^{a+\varepsilon}\int_{b-\varepsilon}^{b+\varepsilon} f_1(z_i,z_j)\mathrm{d}z_i\mathrm{d}z_j \\ &\qquad > \max_{a,b\in R}\int_{a-\varepsilon}^{a+\varepsilon}\int_{b-\varepsilon}^{b+\varepsilon} f_2(z_i,z_j)\mathrm{d}z_i\mathrm{d}z_j \end{aligned} \tag{39}$$

holds for $\forall\varepsilon \in (0,\varepsilon_1]$. As a result, it follows that

$$\max_{z_i,z_j\in R} f_1(z_i,z_j) > \max_{z_i,z_j\in R} f_2(z_i,z_j).$$

Since $f_1(z_i,z_j)$ is a joint uniform distribution satisfying (38), we have

$$f_1(z_i,z_j) - f_2(z_i,z_j) > 0, \quad z_i^2+z_j^2 \leq 1.$$

It follows that

$$\int\limits_G f_1(z_i,z_j)\mathrm{d}z_i\mathrm{d}z_j - \int\limits_G f_2(z_i,z_j)\mathrm{d}z_i\mathrm{d}z_j > 0. \tag{40}$$

Using the PDF property that $\int\limits_G f_1(z_i,z_j)\mathrm{d}z_i\mathrm{d}z_j = 1$, then, it follows from (40) that

$$\int\limits_G f_2(z_i,z_j)\mathrm{d}z_i\mathrm{d}z_j < 1. \tag{41}$$

Since both $f_1(z_i,z_j)$ and $f_2(z_i,z_j)$ have mean 0 and variance $\sigma^2 = \frac{1}{4}$, we have

$$\begin{cases} \int\limits_{\mathbb{R}^2} z_i^2 f_1(z_i,z_j)\mathrm{d}z_i\mathrm{d}z_j = \int\limits_{\mathbb{R}^2} z_i^2 f_2(z_i,z_j)\mathrm{d}z_i\mathrm{d}z_j, \\ \int\limits_{\mathbb{R}^2} z_j^2 f_1(z_i,z_j)\mathrm{d}z_i\mathrm{d}z_j = \int\limits_{\mathbb{R}^2} z_j^2 f_2(z_i,z_j)\mathrm{d}z_i\mathrm{d}z_j. \end{cases} \tag{42}$$

This implies:

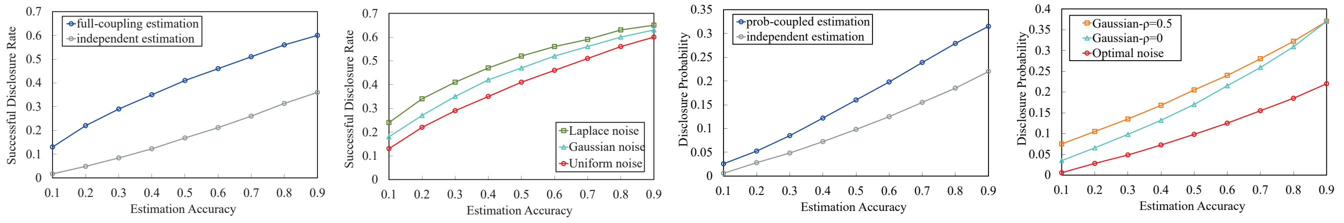
$$\int\limits_{\mathbb{R}^2} (z_i^2+z_j^2) f_1(z_i,z_j)\mathrm{d}z_i\mathrm{d}z_j = \int\limits_{\mathbb{R}^2} (z_i^2+z_j^2) f_2(z_i,z_j)\mathrm{d}z_i\mathrm{d}z_j, \tag{43}$$

which means

$$\begin{aligned} &\int\limits_G (z_i^2+z_j^2)(f_1(z_i,z_j) - f_2(z_i,z_j))\mathrm{d}z_i\mathrm{d}z_j \\ &\qquad = \int\limits_\Omega (z_i^2+z_j^2) f_2(z_i,z_j)\mathrm{d}z_i\mathrm{d}z_j, \end{aligned} \tag{44}$$

where $\Omega = \mathbb{R}^2 - G$. For the left hand side of (44), we have

$$\begin{aligned} &\int\limits_G (z_i^2+z_j^2)(f_1(z_i,z_j) - f_2(z_i,z_j))\mathrm{d}z_i\mathrm{d}z_j \\ &< \int\limits_G (f_1(z_i,z_j) - f_2(z_i,z_j))\mathrm{d}z_i\mathrm{d}z_j \\ &= 1 - \int\limits_G f_2(z_i,z_j)\mathrm{d}z_i\mathrm{d}z_j. \end{aligned} \tag{45}$$



(a) γ and $\gamma_{\varnothing}$

(b) γ under noise distributions with mean 0 and $\sigma^2 = \frac{1}{4}$

(c) δ and $\delta_{\varnothing}$

(d) δ under noise distributions with mean 0 and $\sigma^2 = \frac{1}{4}$

Fig. 2. Privacy comparison on different data correlations and noise distributions

For the right hand side of (44), since $\int\limits_{\mathbb{R}^2} f_2(z_i,z_j)\mathrm{d}z_i\mathrm{d}z_j = 1$ and (41), we have

$$\begin{aligned}\int\limits_{\Omega}(z_i^2+z_j^2)f_2(z_i,z_j)\mathrm{d}z_i\mathrm{d}z_j >& \int\limits_{\Omega} f_2(z_i,z_j)\mathrm{d}z_i\mathrm{d}z_j \\ =&1-\int\limits_{G} f_2(z_i,z_j)\mathrm{d}z_i\mathrm{d}z_j.\end{aligned} \tag{46}$$

Combine (44), (45) and (46), we achieve a contradiction:

$$\begin{aligned}&1-\int\limits_{G} f_2(z_i,z_j)\mathrm{d}z_i\mathrm{d}z_j \\ <&\int\limits_{G}(z_i^2+z_j^2)(f_1(z_i,z_j)-f_2(z_i,z_j))\mathrm{d}z_i\mathrm{d}z_j \\ <&1-\int\limits_{G} f_2(z_i,z_j)\mathrm{d}z_i\mathrm{d}z_j.\end{aligned} \tag{47}$$

Hence, we cannot find a joint noise distribution $f_2(z_i,z_j)$, such that the value of δ is smaller than that under $f_1(z_i,z_j)$. We conclude that, given the finite variance, the noise distribution (37) is optimal in the sense of (ε,δ)-MDDP. ■

Remark 9: It is pointed out that the conclusion of Lemma 1 can be viewed as a special case of Theorem 5. The reason is that when the dimension of optimization problem solved in Theorem 5 is equal to 1, the correlated data constraints will no longer exist, and the disclosure probability δ will not be in a coupled form. Thus, when $N=1$, the problem becomes the same one solved by Lemma 1. Moreover, the optimal noise distribution derived in (37) provides (ε,δ)-MDDP with $\delta = \frac{\varepsilon^2}{3\sigma^2}$, which is the minimal disclosure probability theoretically. As shown in (32), by adding the joint distributed noises, the estimation chain cannot be applied in the joint estimation, and the privacy disclosure gain under f_X is prevented. Thus, using (37), the disclosure probability for the case with probabilistic coupling is minimized in the sense of (ε,δ)-MDDP. Interestingly, Theorem 5 implies that all non-diagonal elements of covariance matrix of optimal noises are zero. This means given any upper-bounded constraints on non-diagonal elements of the covariance matrix, this will not change the optimal noise analysis result.

VI. Simulation Results

A. Simulation Scenario

In this section, we conduct simulations to verify our theoretical results through a publicly available data set taken from the Cardiovascular Disease Dataset of Kaggle platform [28]. For the full coupled case, the dataset for simulation is from the height and weight information of 10,000 men aged 30-39. G_X is obtained by using the polynomial regression, that is, $\widehat{\text{Weight}} = 261.88 - 7.35 * \text{Height} + 0.083 * \text{Height}^2$, i.e., weight is a explicit function of height (and height is optimally estimated with a higher priority since its smaller domain). For the probabilistic coupled case, the dataset contains two continuous attributes, the weight and blood pressure, which approximately follow a joint normal distribution with correlation coefficient $\rho = 0.612$. For the full coupled case, 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 1,2 randomly generate a noise vector θ with Laplacian noise distribution, Gaussian noise distribution and the optimal noise distribution derived in (15), respectively. In the simulation, 10000 runs are conducted. Laplacian noise is popular because its mathematical property fits well with differential privacy [3], [29].

The attacker uses the proposed optimal estimation approach derived in (14) 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 $||\hat{\theta}-\theta||_\infty \le \varepsilon$ is obtained in each run, the value of δ is computed by using the maximum probability among these probabilities in all runs. Since both the ranges of height and weight are known, Θ is known. Thus, one can obtain γ by dividing δ by the probability of the estimated $\hat{\theta}$ within Θ. For the probabilistic coupled case, the simulation steps remain the same, but the optimal estimation policy and optimal noise distribution changes to (35) and (37), respectively.

B. Verification

1) Privacy Disclosure vs Data Correlation: We first compare γ with $\gamma_{\varnothing}$ under the optimal noise adding strategy derived in Theorem 2. From Fig. 2(a), one observes that using G_X, the successful disclosure rate increases effectively compared to that the correlation is unknown. It is because using G_X, the attacker is able to decrease the uncertainty of noise adding effectively. Furthermore, in Fig. 2(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, in Fig. 2(c), we compare δ with $\delta_{\varnothing}$ under the independent uniform noise

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNSE.2020.3044590, IEEE Transactions on Network Science and Engineering

added. Since the correlation among weight and blood pressure is strongly positive, once the estimation of weight fixed, the estimation on blood pressure will be consistent with the trend of strongly positive correlation (e.g., if the $\widehat{\text{Weight}} > 100\,\text{kg}$, it will infer Blood pressure$> 130\,\text{mmHg}$ with high possibility). Thus, for the uniform noise added on blood pressure, the noise variance is reduced by f_X, and the disclosure probability increases, which verifies Theorem 4.

2) Privacy Disclosure vs Noise Distribution: First, in Fig. 2(b), we compare the privacy of full-coupled data publication for Laplacian noise distribution, Gaussian noise distribution and the optimal noise distribution derived in (15). It is observed from Fig. 2(b) that under the G_X, the optimal noise distribution derived in (15) performs better than the extensively-used Laplacian noise distribution or Gaussian noise distribution in the sense of (ε,δ)-MDDP. Then, we compare the privacy of probabilistic data publication with joint uniform noise distribution derived in (37) and joint Gaussian noise distribution. Comparing Fig. 2(c) with Fig. 2(d), we note that the privacy disclosure gain under f_X is prevented in the joint estimation, which is consistent with (32). Further, the optimal joint distribution of noises derived in Theorem 5 achieves a higher (ε,δ)-MDDP than others.

VII. 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. We then quantified the privacy disclosure variation of correlated data publication. We obtained a closed-form expression of the optimal estimation for the correlated data publication. It is shown that for deterministic and probabilistic correlations, the original data can be restored with a higher privacy disclosure. Moreover, a strict bound of privacy gain is derived. We designed the optimal noise adding strategy to minimize the disclosure probability in the sense of (ε,δ)-MDDP. We showed that when original data is correlated, uniform noise distribution achieves higher (ε,δ)-MDDP than Laplacian or Gaussian noise distribution.

There are still many open issues worth further investigation. For example, the tradeoff between privacy degree and data utility. Meanwhile, the comparison with optimal privacy-preserving policies when using mutual information as a measure of privacy needs further investigation. In addition, it has been pointed by [10] that adding noise is not suitable for non-real-valued data publication, e.g., categorical data (social security numbers, postal codes, etc.). How to characterize the privacy analysis of multi-dimensional discrete data remains open. Lastly, how to extend our result to data generated by dynamical systems is also interesting.

References

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

[2] S. J. Rizvi and J. R. Haritsa, "Maintaining data privacy in association rule mining," in *Proc. VLDB*, 2002, pp. 682–693.

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

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

[5] E. Fujisaki and T. Okamoto, "Secure integration of asymmetric and symmetric encryption schemes," in *Proc. CRYPTO*, 1999, pp. 537–554.

[6] L. Sweeney, "k-anonymity: A model for protecting privacy," *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, vol. 10, no. 05, pp. 557–570, 2002.

[7] X. Wang, J. He, P. Cheng, and J. Chen, "Differentially private maximum consensus: Design, analysis and impossibility result," *IEEE Trans. Netw. Sci. Eng.*, vol. 6, no. 4, pp. 928–939, 2018.

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

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

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

[11] 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.

[12] B. Yang, I. Sato, and H. Nakagawa, "Bayesian differential privacy on correlated data," in *Proc. ACM SIGMOD*, 2015, pp. 747–762.

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

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

[15] 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 Trans. Inf. Forensics Secur.*, vol. 13, no. 9, pp. 2151–2166, 2018.

[16] F. du Pin Calmon and N. Fawaz, "Privacy against statistical inference," in *Proc. IEEE Annu. Allerton Conf. Commun., Control, Comput.*, 2012, pp. 1401–1408.

[17] P. Cuff and L. Yu, "Differential privacy as a mutual information constraint," in *Proc. ACM CCS*, 2016, pp. 43–54.

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

[19] M. Sun, C. Zhao, and J. He, "Privacy-preserving correlated data publication with noise adding mechanism," in *Proc. IEEE ICCA*. Accepted, 2020.

[20] T. Zhu, P. Xiong, G. Li, and W. Zhou, "Correlated differential privacy: Hiding information in non-iid data set," *IEEE Trans. Inf. Forensics Secur.*, vol. 10, no. 2, pp. 229–242, 2015.

[21] J. Zhang, G. Cormode, C. M. Procopiuc, D. Srivastava, and X. Xiao, "Privbayes: Private data release via bayesian networks," in *Proc. ACM SIGMOD*, 2014, p. 1423–1434.

[22] Y. Cao, M. Yoshikawa, Y. Xiao, and L. Xiong, "Quantifying differential privacy in continuous data release under temporal correlations," *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 7, pp. 1281–1295, 2019.

[23] Y. Xiao and L. Xiong, "Protecting locations with differential privacy under temporal correlations," in *Proc. ACM CCS*, 2015, p. 1298–1309.

[24] N. M. Oliver, B. Rosario, and A. P. Pentland, "A bayesian computer vision system for modeling human interactions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 831–843, 2000.

[25] Q. Wang, Y. Zhang, X. Lu, Z. Wang, Z. Qin, and K. Ren, "Real-time and spatio-temporal crowd-sourced social network data publishing with differential privacy," *IEEE Trans. Dependable Secure Comput.*, vol. 15, no. 4, pp. 591–606, 2018.

[26] J. He, L. Cai, C. Zhao, P. Cheng, and X. Guan, "Privacy-preserving average consensus: privacy analysis and algorithm design," *IEEE Trans. Signal Inf. Proces. Netw.*, vol. 5, no. 1, pp. 127–138, 2018.

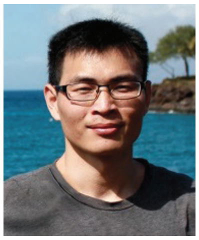
[27] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. NY, USA: Wiley-Interscience, 2006.

[28] S. Ulianova, "Cardiovascular disease dataset," https://www.kaggle.com/sulianova/cardiovascular-disease-dataset, 2019, online. Accessed 6 March 2020.

[29] C. Dwork, M. Naor, T. Pitassi, and G. N. Rothblum, "Differential privacy under continual observation," in *Proc. ACM STOC*, 2010, p. 715–724.

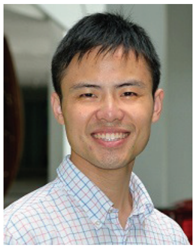
Mingjing Sun received the B.E. degree from the College of Automation, Northwestern Polytechnical University, Xi'an, China, in 2018. He is currently pursuing the master's degree with the Department of Automation, Shanghai Jiao Tong University, Shanghai, China. His current research interests include analysis and optimization for data privacy preservation.

Chengcheng Zhao received the PhD degree in control science and engineering from Zhejiang University, Hangzhou, China, in 2018. She is currently a research fellow in the Department of Electrical and Computer Engineering, University of Victoria. Her research interests include consensus and distributed optimization, distributed energy management in smart grids, vehicle platoon, and security and privacy in network systems. She received IEEE PESGM 2017 best conference papers award, and one of her paper was shortlisted in IEEE ICCA 2017 best student paper award finalist. She is a peer reviewer for Automatica, IEEE Transactions on Information Forensics and Security, IEEE Transactions on Industrial Electronics and etc. She was the TPC member for IEEE GLOBECOM 2017, 2018, and IEEE ICC 2018.



Jianping He (M'15-SM'19) is currently an associate professor in the Department of Automation at Shanghai Jiao Tong University, Shanghai, China. He received the Ph.D. degree in control science and engineering from Zhejiang University, Hangzhou, China, in 2013, and had been a research fellow in the Department of Electrical and Computer Engineering at University of Victoria, Canada, from Dec. 2013 to Mar. 2017. His research interests mainly include the distributed learning, control and optimization, security and privacy in network systems.

Dr. He serves as an Associate Editor for IEEE Open Journal of Vehicular Technology and KSII Trans. Internet and Information Systems. He was also a Guest Editor of IEEE TAC, International Journal of Robust and Nonlinear Control, etc. He was the winner of Outstanding Thesis Award, Chinese Association of Automation, 2015. He received the best paper award from IEEE WCSP'17, the best conference paper award from IEEE PESGM'17, and was a finalist for the best student paper award from IEEE ICCA'17.



Peng Cheng (M'10) received the B.E. degree in Automation, and the Ph.D. degree in Control Science and Engineering in 2004 and 2009 respectively, from Zhejiang University, Hangzhou, China. Currently he is Professor with College of Control Science and Engineering, Zhejiang University. He serves as Associate Editor of IEEE Transactions on Control of Network Systems, Wireless Networks, and International Journal of Communication Systems. He also serves/served as the Guest Editor for IEEE Transactions on Automatic Control, IEEE Transactions on Signal and Information Processing over Networks, and IEEE Transactions on Control of Network Systems. He served as the TPC co-chair of IEEE IOV 2016, local arrangement co-chair for ACM MobiHoc 2015, and the publicity co-chair for IEEE MASS 2013. His research interests include networked sensing and control, cyber-physical systems, control system security.

Daniel E. Quevedo (S'97–M'05–SM'14–F'21) received Ingeniero Civil Electrónico and M.Sc. degrees from Universidad Técnica Federico Santa María, Valparaíso, Chile, in 2000, and in 2005 the Ph.D. degree from the University of Newcastle, Australia. He is Professor of Cyberphysical Systems at the School of Electrical Engineering and Robotics, Queensland University of Technology (QUT), in Australia. Before joining QUT, he established and led the Chair in Automatic Control at Paderborn University, Germany. In 2003 he received the IEEE Conference on Decision and Control Best Student Paper Award and was also a finalist in 2002. He is co-recipient of the 2018 IEEE Transactions on Automatic Control George S. Axelby Outstanding Paper Award.

Prof. Quevedo currently serves as Associate Editor for *IEEE Control Systems* and in the Editorial Board of the *International Journal of Robust and Nonlinear Control*. From 2015–2018 he was Chair of the IEEE Control Systems Society *Technical Committee on Networks & Communication Systems*. His research interests are in networked control systems, control of power converters and cyberphysical systems security.