Social-Aware Privacy-Preserving Correlated Data Collection

Guocheng Liao
The Chinese University of Hong Kong
lg016@ie.cuhk.edu.hk

Xu Chen
Sun Yat-sen University
chenxu35@mail.sysu.edu.cn

Jianwei Huang
The Chinese University of Hong Kong
jwhuang@ie.cuhk.edu.hk

ABSTRACT

We study a privacy-preserving data collection problem, by jointly considering data reporters' data correlation and social relationship. A data collector gathers data from individuals to perform a certain analysis with a privacy-preserving mechanism. Due to data correlation, the data analysis based on the reported data can cause privacy leakage to other individuals (even if they do not report data). The data reporters will take such a privacy threat into account, owing to the social relationship among individuals. This motivates us to formulate a two-stage Stackelberg game: In Stage I, the data collector selects some individuals as data reporters and designs a privacy-preserving mechanism for a sum query analysis. In Stage II, the selected data reporters contribute their data with possible perturbations (through adding noise). By analyzing the data reporters' equilibrium decisions in Stage II, we show that given any fixed reporter set, only one data reporter with the most significant joint consideration of the social relationship and data correlation may add noise to his reported data. The rest of the data reporters will truthfully report their data. In Stage I, we derive the data collector's optimal privacy-preserving mechanism and propose an efficient algorithm to select the data reporters. We conclude that the data collector should jointly capture the impact of data correlation and social relation to ensure all data reporters truthfully reporting their data. We conduct extensive simulations based on random network and real-world social data to investigate the impact of data correlation and social network on the system. We find that the availability of social network information is more critical to the data collector compared with data correlation information.

CCS CONCEPTS

• **Theory of computation → Social networks**; **Network games**;
• **Security and privacy → Data anonymization and sanitization**;

KEYWORDS

Privacy-Preserving mechanism, social network, data correlation

ACM Reference Format:
Guocheng Liao, Xu Chen, and Jianwei Huang. 2018. Social-Aware Privacy-Preserving Correlated Data Collection. In *Mobihoc '18: The Eighteenth ACM International Symposium on Mobile Ad Hoc Networking and Computing, June 26–29, 2018, Los Angeles, CA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3209582.3209584

1 INTRODUCTION

With extensive personal data generated in various applications ranging from medical treatments to online-social interactions, effective and accurate data collection from individuals is becoming important for institutions to develop practical data-driven analysis. The analysis based the collected data, however, would compromise personal privacy, since the analysis result may expose personal sensitive information. From a data reporter's perspective, despite the appealing data-based service benefit, he may still want to exert an effective control over the reported data (for example, by adding noise to the true data) to protect his privacy [2]. From the data collector's perspective, with the legal obligation and ethical responsibility to protect the individual privacy as well as the desire to obtain accurate data and perform high-quality analysis, she[1] is also interested in implementing a privacy-preserving mechanism.

When tackling this privacy-preserving data collection problem, some of the existing work (e.g., [19, 22]) assumed independence among the data from different individuals. Under this assumption, one individual's data contribution will not cause privacy leakage to others. However, this is not always true in practice. For example, in a social network, individuals who have close relationships may share some common interests. An adversary may infer individual private information by exploiting such data correlations.

Nevertheless, a person will not set aside the privacy threat to people he cares about in reality. With the social relationship created by both offline and online interactions, a data reporter will consider the privacy protection of those individuals who are socially connected with him with correlated data.

The existing studies that consider the privacy protection with correlated data (e.g. [9, 13, 17, 20]) often did not capture the impact of social relationship. The consideration of social relationship will not only make data reporters more conservative but also encourage the data collector to design a more effective privacy-preserving mechanism.

Motivated by the above discussions, we will focus on the privacy-preserving data collection problem considering the impact of both data correlation and social relationship. The study of this problem involves several challenges, including properly characterizing the individuals' privacy loss under data correlation and deriving the

[1] In this paper, we use "he" to refer to a data reporter and "she" to refer to the data collector.

This work is supported by the National Key Research and Development Program of China under grant No.2017YFB1001703, the National Science Foundation of China under Grant No. U1711265, the Fundamental Research Funds for the Central Universities under grant No.17lgjc40, the Program for Guangdong Introducing Innovative and Enterpreneurial Teams (No. 2017ZT07X355), and the General Research Funds (Project Number CUHK 14219016) established under the University Grant Committee of the Hong Kong Special Administrative Region, China.

Mobihoc '18, June 26–29, 2018, Los Angeles, CA, USA

ACM ISBN 978-1-4503-5770-8/18/06…$15.00
https://doi.org/10.1145/3209582.3209584

optimal strategies of the data collector and reporters. We want to answer the following key questions: *(1) How should a data reporter report his data considering data correlation and social relationship? (2) How should the data collector design an optimal privacy-preserving mechanism accordingly?*

The main results and contributions of this paper are as follows:

- *Joint modeling of social relationship and data correlation.* To the best of our knowledge, this is the first paper that proposes a theoretical framework for the joint consideration of the impact of data correlation and social relationship on privacy-preserving data collection. As the first step, we focus on the commonly considered sum query data analysis in this paper.
- *Equilibrium of the data reporters' reporting decisions.* For the data reporters' reporting decision equilibrium, we show that under the sum query analysis, only one data reporter with the most significant joint consideration may add a sufficiently large noise to his reported data (while others would report their data truthfully). We derive the condition under which all the data reporters will truthfully report their data.
- *The data collector's optimal strategy.* We derive the data collector's optimal privacy-preserving mechanism design under the sum query analysis, and propose an efficient algorithm to construct the reporter set. We conclude that the data collector should jointly consider the impact of data correlation and social relationship when maximizing her utility maximization, and ensure all data reporters truthfully reporting their data.
- *Impact of data correlation network and social network on the system.* We conduct extensive simulations based on random network and real-world social relationship. We find that both the data collector and data reporters experience lower utilities under higher data correlation and closer social relationship. In addition, comparing with the data correlation information, accurately knowing the social network information is more critical to the data collector.

The rest of this paper is organized as follows. In Section 2, we discuss the related work. In Section 3, we introduce the utility functions of the data reporters and the data collector, and present the two-stage game formulation. In Section 4, we derive the data reporters' reporting equilibrium in Stage II. In Section 5, we solve the data collector's utility optimization problem in Stage I. In Section 6, we provide some simulation results and discuss the corresponding insights. We conclude the paper in Section 7.

2 RELATED WORK

Privacy-preserving data collection (e.g., [5–7, 19, 22]) has been extensively investigated under the notion of ϵ-differential privacy [4], which is a powerful privacy protection framework. However, the framework of ϵ-differential privacy fails to guarantee the desirable privacy under correlated data [12]. In order to solve this problem, some studies have focused on developing more general frameworks and mechanisms for privacy protection with correlated data. For example, Kifer et al. in [13] proposed a novel framework called Pufferfish, which ensures that the potential secrets are not distinguishable under correlated data evolution scenarios. Song et al. in [20] proposed the Markov Quilt mechanism for interdependent data expressed by a Bayesian network.

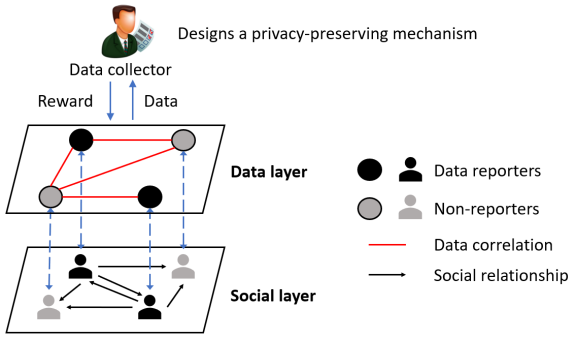


Figure 1: System model

The above work concentrated on the data collector's perspective. However, we focus on a different approach by incorporating the data reporters' decisions and studying the interaction between the data collector and reporters. More specifically, we formulate the interaction as a Stackelberg game, and study the data collector's optimal privacy-preserving mechanism design considering the data reporters' equilibrium decisions.

One feature of our model is that data reporters' decisions are coupled with each other, due to the data correlation and social relationship. This naturally leads to a game theoretical formulation among data reporters to study their strategic decisions.

There exist some studies (e.g., [3, 21, 23]) that exploited game theory to study the individuals' choices considering privacy protection. For example, Chessa et al. in [3] considered the case where the individuals are interested in the analysis which depends on all the reporters' reporting precision, and studied the reporters' reporting precision strategies with a game-theoretic model. But they didn't consider data correlation and the individuals' social privacy concerns. Wu et al. in [23] studied the case where a publisher's privacy leakage due to data publication depends on the privacy protection choices of the publishers whose data is correlated with him. They exploited a game theoretic model to study the publishers' privacy choices. But they didn't consider the impact of publishers' social relationship. Our work differs from previous work in that we jointly consider the data correlation and social relationship.

3 SYSTEM MODEL

In this section, we present our system model of privacy-preserving data collection problem with data correlation and social relationship consideration, as illustrated in Fig 1. In this model, a data collector wants to gather data to perform analysis with a privacy protection promise. She selects a subset (including selecting the whole set as a special case) of the target individual pool as the reporter set. She provides analysis-related reward to the data reporters. For example, Apple Inc. has been collecting user data from Apple devices to improve the quality of services or products such as Health Type Usage and Lookup Hints [10]. The users can enjoy better services if they agree to provide personal information. We use two layers, the data layer and the social layer, to capture the data correlation and social relationship. The data reporters can perturb the data (by adding noise) to alleviate their privacy concerns.

Section 3.1 introduces the basic setting of the problem. We then characterize the privacy loss of individuals based on the data correlation model in Section 3.2. We discuss the data collector's and data reporters' benefit from an accurate data analysis in Section 3.3. Based on these, we derive the utility functions of the data reporters and the data collector in Section 3.4. Finally, we formulate a Stackelberg game in Section 3.5.

3.1 Problem Setting

In this subsection, we introduce the decision-making problems of the data reporters and the data collector.

We first focus on the data reporters' decision-making problem. We consider a target individual pool denoted by $\mathcal{T}$ with a size of T. We denote the set of data reporters selected by the data collector as $\mathcal{M}(\subset \mathcal{T})$ with a size of $M(\leq T)$. Due to data correlation, the data reporting by an individual in set $\mathcal{M}$ may cause privacy leakage of other individuals in set $\mathcal{T}$. To reduce the impact of such privacy leakage, the data reporters can perturb their data by adding some random noise. Similar to [19], we assume that the noise is a Gaussian random variable with a zero mean. More specifically, we can represent a data reporter j's reported data as

$$y_j = x_j + v_j, \tag{1}$$

where x_j is the true data and $v_j \sim N(0, \sigma_j^2)$ is the noise. His reporting strategy is his choice of σ_j^2. A higher variance means that the reported data is more likely to deviate from the true data, hence the potential privacy leakage is smaller. Let $\boldsymbol{\sigma}^2 \triangleq (\sigma_j^2, \forall j \in \mathcal{M})$ denote the strategy profile of all data reporters.

Each data reporter will decide his reporting strategy to maximize his utility, which consists of his privacy concern and his analysis benefit from the computation result. We will elaborate the utility function design in Sections 3.2 an 3.3.

We then focus on the data collector's decision problem. The data collector's decision has two parts. First, she selects data reporters from the whole set $\mathcal{T}$ to form the reporter set $\mathcal{M}$. Later we will discuss how to efficiently select the data reporters. Then she performs analysis based on the reported data from set $\mathcal{M}$, and returns the analysis result to the data reporters as a reward. We take sum query (e.g., [7, 13, 24]) as an example throughout this paper, which is a typical analysis in data analytics (e.g., the sum of the salaries and aggregation of business data). To satisfy the privacy requirements of the data reporters, she designs a privacy-preserving mechanism, which adds a Gaussian noise [19] v_g with a zero mean and a variance σ_g^2 to the sum. So the analysis result will be

$$z = \sum_{j \in \mathcal{M}} y_j + v_g, \tag{2}$$

where y_j is the data reporter j's reported data in (1) and $v_g \sim N(0, \sigma_g^2)$ is the noise added by the data collector. So in the second part, the data collector's strategy is her choice of variance σ_g^2. Intuitively, a higher value of σ_g^2 provides a better privacy protection to all the individuals. Similar to previous work (e.g., [5, 7, 19]), we emphasize that the data collector obtains the analysis result $z = \sum_{j \in \mathcal{M}} y_j + v_g$ without the access to $\sum_{j \in \mathcal{M}} y_j$ (i.e., a database stores the data and returns the result with random noise when she requests a query). The privacy-preserving mechanism takes effect prior to the data collector's viewing the result $\sum_{j \in \mathcal{M}} y_j$.

The data collector will construct the reporter set and design a privacy-preserving mechanism accordingly to maximize her utility, which is her analysis benefit from the analysis result. We will characterize the benefit in Section 3.3.

3.2 Privacy Loss Based on the Data Correlation Model

In this subsection, we discuss the privacy loss caused by the data reporting due to data correlation.

3.2.1 Data Correlation Model. We treat the data from each individual in set $\mathcal{T}$ as a random variable, and multiple variables depend on each other. We capture the data correlation with the Gaussian correlation model [24], which is a special case of the Markov random field [14]. This model presents the correlation among data as a weighted undirected graph. A vertex represents an individual's data, and the edge weight represents the correlation between two vertices. All the conditional distributions have the Gaussian form, which is a widely used distribution [1, 11]. Notice that this model is not applicable to discrete random variables and does not capture the causality. Since we are considering sum query analysis for continuous random variables and don't require edge direction (causality), so Gaussian correlation model is a proper choice to model the data correlation. Next, we introduce the definition of Gaussian correlation model rigorously.

Definition 1 (Gaussian Correlation Model). *Let $G(\mathcal{V}, \mathcal{E})$ be a weighted undirected graph. Each vertex $i \in \mathcal{V}$ denotes the data x_i of individual $i \in \mathcal{T}$. Each edge $\{i, j\} \in \mathcal{E}$ with weight $w_{ij} \geq 0$ denotes the data correlation between individual i and individual j. Let $\mathbf{W} = (w_{ij}) \in \mathbb{R}^{T \times T}$ be the weighted adjacent matrix and $\mathbf{D} = diag(w_1, ..., w_T)$ be the weighted degree matrix where $w_i = \sum_{j:j \neq i} w_{ij}$. Recall that T is the size of the individual pool, i.e., the total amount of data. Let $\mathbf{x}$ be the vector containing all the data, i.e., $\mathbf{x} = [x_1, \ldots, x_T]^T$, and $\mathbf{x}_{-i}$ denote all the elements in $\mathbf{x}$ but x_i. Then $\forall i \in \mathcal{T}$, the conditional joint probability of $\mathbf{x}_{-i}$ given x_i is*

$$p(\mathbf{x}_{-i}|x_i) \propto \exp\left(-\frac{\mathbf{x}^\mathbf{T}\mathbf{L}\mathbf{x}}{2}\right). \tag{3}$$

Here $\mathbf{L}$ is Laplacian matrix , i.e.,

$$\mathbf{L} = \mathbf{D} - \mathbf{W} = \begin{bmatrix} w_1 & -w_{12} & \cdots & -w_{1T} \\ -w_{21} & w_2 & \cdots & -w_{2T} \\ \cdots & \cdots & \ddots & \cdots \\ -w_{T1} & -w_{T2} & \cdots & w_T \end{bmatrix}. \tag{4}$$

The exponential formula together with the quadratic form in (3) captures the Gaussian distribution.[2] In [24], the authors emphasized that the formula in (3) could be more general if it contains a linear term, such as $\exp\left(-(\mathbf{x}-\mathbf{u})^\mathbf{T}\mathbf{L}(\mathbf{x}-\mathbf{u})/2\right)$. They eliminated the linear term by replacing $\mathbf{x} - \mathbf{u}$ with an auxiliary variable $\mathbf{x}'(= \mathbf{x} - \mathbf{u})$. Following a similar argument, we assume that the mean term $\mathbf{u}$ is

[2]The coefficient of the exp term in a Gaussian distribution can be obtained by an integration. We omit the coefficient and use the notation $\propto$ for simplicity.

known from a public trend and can be removed in advance. For example, the mean $\mathbf{u}$ can be inferred from the aggregation of extensive historical data under a stationary assumption [10].

3.2.2 Characterizing the Privacy Loss. Based on the data correlation model, we will characterize the privacy loss of the individuals in set $\mathcal{T}$ due to a sum query analysis based on the data reporting. Note that the privacy loss of an individual is from the announced analysis result in (2) rather than the reported data.

Based on the Gaussian correlation model, we calculate the variance of the analysis result conditional on an individual's data to investigate his privacy loss. We assume that the individuals' data have the Gaussian property. Given the strategies of the data collector and the data reporters, i.e., $\mathcal{M}$, σ_g^2, and $\boldsymbol{\sigma}^2$, Proposition 1 below characterizes the variance of the sum analysis in (2) conditional on each individual's data.

Proposition 1. *The variance of z in (2) conditional on individual i's data $x_i, i \in \mathcal{T}$, is given by*

$$Var(z|x_i) = \sum_{j\in\mathcal{M}, j\neq i} \frac{1}{a_{ji}} + \sum_{j\in\mathcal{M}} \sigma_j^2 + \sigma_g^2. \tag{5}$$

Here

$$a_{ji} = 1/Var(x_j|x_i) = w_{ji} + \mathbf{w}_{jc}\left(\mathbf{W}_c^{-1}\right)^T \mathbf{w}_{ic}^T \neq 0. \tag{6}$$

The vector $\mathbf{w}_{jc} \in \mathbb{R}^{1\times(T-2)}$ contains all the entries of the jth row in matrix $\mathbf{L}$ except w_j and $-w_{ji}$, and the vector $\mathbf{w}_{ic} \in \mathbb{R}^{1\times(T-2)}$ is defined similarly. The matrix $\mathbf{W}_c \in \mathbb{R}^{(T-2)\times(T-2)}$ contains all the entries in matrix $\mathbf{L}$ except the jth row, the jth column, the ith row, and the ith column.

Proof (sketch): The proof of Proposition 1 consists of five steps. Conditional on individual i's data, we first calculate $p(x_j|x_i), \forall j \in \mathcal{M}$, and $j \neq i$ based on $p([x_j, \mathbf{x}_{-ji}]|x_i)$ in (3) by marginalizing $\mathbf{x}_{-ji}$. The vector $\mathbf{x}_{-ji}$ contains all data except x_i and x_j. Second, we calculate $p(y_j = x_j + v_j|x_i)$ where $v_j \propto \exp(-v_j^2/2\sigma_j^2)$ through convolution. Third, we calculate $p(\sum_{j\in\mathcal{M}} y_j|x_i)$ through the convolution of M random variables. Fourth, we calculate $p(z = \sum_{j\in\mathcal{M}} y_j + v_g|x_i)$ where $v_g \propto \exp(-v_g^2/2\sigma_g^2)$ through the convolution. Finally, based on the conditional distribution $p(z|x_i)$, we are able to obtain the conditional variance $Var(z|x_i)$. For the detailed proof, see [16].

In Proposition 1, $1/a_{ji}$ is the variance of x_j conditional on x_i. It not only depends on the direct correlation w_{ji} between individual j and individual i, but also depends on the influence propagated from others, i.e., $\mathbf{w}_{jl}(\mathbf{W}_l^{-1})^T\mathbf{w}_{il}^T$. Furthermore, the conditional variance of analysis result in (5) is jointly affected by the data correlation term a_{ji} as well as the strategies of the data collector (σ_g^2) and the data reporters, i.e., $\boldsymbol{\sigma}^2$.

We utilize the conditional variance $Var(z|x_i)$ to capture the privacy loss of individual i. The justification is based on the mutual information. The mutual information $I(Z;X_i)\,(= H(Z) - H(Z|X_i))$ between the data random variable X_i and the result random variable Z indicates how much personal information that the analysis result contains. If $I(Z;X_i) = 0$, then the result z does not contain any information regarding x_i. As $H(Z)$ is the same for all x_i, we are able to focus on $H(Z|X_i)$. As $p(z|x_i)$ has a Gaussian form, we have $H(Z|X_i) = 1/2\ln(2\pi e Var(z|x_i))$, i.e., a higher conditional variance $Var(z|x_i)$ corresponds to a higher uncertainty, thus a higher conditional entropy $H(Z|X_i)$. In other words, if the conditional variance $Var(z|x_i)$ is higher, the mutual information $I(Z;X_i)$ is lower, thus the privacy loss caused to individual i is less severe.

To summarize, the privacy loss of an individual $i \in \mathcal{T}$ is a decreasing function of the conditional variance $Var(z|x_i)$. We consider a positive and convex function $h(\cdot)$,[3] i.e., absolute marginal change of privacy loss diminishes when the conditional variance $Var(z|x_i)$ increases such that the privacy loss will not decrease to $-\infty$. We obtain the individual i's privacy loss $l_i(\cdot)$ given the data collector's strategy $\mathcal{M}$ and σ_g^2, and the data reporters' strategy profile $\boldsymbol{\sigma}^2$:

$$\begin{aligned} l_i\left(\mathcal{M}, \sigma_g^2, \boldsymbol{\sigma}^2\right) &= h\left(Var(z|x_i)\right) \\ &= h\left(\sum_{j\in\mathcal{M}, j\neq i} \frac{1}{a_{ji}} + \sum_{j\in\mathcal{M}} \sigma_j^2 + \sigma_g^2\right), i \in \mathcal{T}. \end{aligned} \tag{7}$$

3.3 Data Analysis Benefit of Reporters and Data Collector

The data collector and data reporters will obtain the data analysis result, and we will discuss how the analysis accuracy affects them.

We first focus on the data reporters' data analysis benefit. For example, the evaluation of a certain disease based on reported medical records enables better medical treatments. So the data reporters desire an accurate analysis.

We characterize the data reporters' expected accuracy loss due to the noise added by both the data collector and the data reporters. We define $p(v)$ as the accuracy loss [8] when the noise is v. We consider the quadratic form as in [8], i.e., $p(v) = r_d v^2$ where $r_d > 0$ is a parameter. A higher value of r_d means that the data reporter cares more about the accuracy. The random noise is the aggregation of independent Gaussian noise from the data collector and the data reporters. So we have $v \sim N(0, \sum_{j\in\mathcal{M}} \sigma_j^2 + \sigma_g^2)$. Let $f(v)$ be the probability density function of this Gaussian noise. Hence the expected accuracy loss of a data reporter is as follows:

$$E\left(p(n)\right) = \int_{-\infty}^{\infty} p(v)f(v)dv = r_d\left(\sum_{j\in\mathcal{M}} \sigma_j^2 + \sigma_g^2\right). \tag{8}$$

Furthermore, we consider the data reporter j's analysis benefit $B_j(\mathcal{M})$ when no noise is added, which is a function of the reporter set $\mathcal{M}$. As $\mathcal{M}$ is the choice of the data collector instead of a data reporter, $B_j(\mathcal{M})$ will be a constant as far as data reporter j is concerned. The specific function form is not important for reporter j. Then we have the analysis benefit for the data reporter $j \in \mathcal{M}$ as follows, which is his absolute benefit minus expected accuracy loss:

$$R_j\left(\mathcal{M}, \sigma_g^2, \boldsymbol{\sigma}^2\right) = B_j(\mathcal{M}) - r_d\left(\sum_{j\in\mathcal{M}} \sigma_j^2 + \sigma_g^2\right). \tag{9}$$

The analysis benefit is a linear function of the variances of noise, and the parameter r_d is the accuracy decreasing rate.

We then derive the data collector's analysis benefit similarly. Recall that her accuracy loss is affected by the noise added by herself. Since the data collector also cares about the accuracy of the collected data (not just the analysis result), the decreasing rate

[3] In simulations (Section 6), we will use a concrete function that have these features.

with respect to σ_j^2 ($j \in \mathcal{M}$) is larger than that with respect to σ_g^2. So the data collector's analysis benefit is given as follows:

$$R_g\left(\mathcal{M}, \sigma_g^2, \boldsymbol{\sigma}^2\right) = B_g(\mathcal{M}) - r_g \sum_{j\in\mathcal{M}} \sigma_j^2 - r_a \sigma_g^2, \tag{10}$$

where $r_g > r_a$ and $B_g(\mathcal{M})$, a function of the reporter set $\mathcal{M}$, is the data collector's benefit when no noise is added. We will discuss $B_g(\mathcal{M})$ later in Section 5.1.

3.4 Utility Functions

After analyzing the privacy loss of individuals (in set $\mathcal{T}$) and the analysis benefit of the data reporters (in set $\mathcal{M}$) and the data collector, we are able to define the utility functions of the data reporters and the data collector.

We first focus on a data reporter' utility function. The utility function consists of two components, analysis benefit and privacy concern, which is given as follows:

$$U_j\left(\mathcal{M}, \sigma_g^2, \boldsymbol{\sigma}^2\right) = R_j\left(\mathcal{M}, \sigma_g^2, \boldsymbol{\sigma}^2\right) - \sum_{i\in\mathcal{T}} s_{ji} l_i\left(\mathcal{M}, \sigma_g^2, \boldsymbol{\sigma}^2\right). \tag{11}$$

The social relationship parameter s_{ji} is the social strength maintained by j to i, capturing how much j cares about i's privacy loss. If this value is higher, then i's privacy loss will play a critical role in the data reporter j's decision-making. We do not assume symmetry of the social relationship, i.e., s_{ji} could be different from s_{ij}.

We then focus on the data collector's utility function. Since the data collector does not have the social relationship with the individuals, her utility function equals to the analysis benefit:

$$U_g\left(\mathcal{M}, \sigma_g^2, \boldsymbol{\sigma}^2\right) = R_g\left(\mathcal{M}, \sigma_g^2, \boldsymbol{\sigma}^2\right). \tag{12}$$

3.5 Stackelberg Game Formulation

We model the interaction between the data collector and the data reporters as a two-stage Stackelberg game as follows:

- Stage I: the data collector selects the set of data reporters $\mathcal{M}$ and the noise variance σ_g^2 to maximize her utility:

$$\left(\mathcal{M}^*, \sigma_g^{2*}\right) = \arg\max_{\mathcal{M}, \sigma_g^2} U_g\left(\mathcal{M}, \sigma_g^2, \boldsymbol{\sigma}^2\right).$$

- Stage II: each data reporter $j \in \mathcal{M}$ chooses his reporting strategy σ_j^2 to maximize his utility, given the data collector's and other data reporters' strategies:

$$\sigma_j^{2*} = \arg\max_{\sigma_j^2 \in \Sigma_j} U_j\left(\mathcal{M}, \sigma_g^2, \sigma_j^2, \boldsymbol{\sigma}_{-j}^2\right).$$

In Sections 4 and 5, we study this two-stage game through backward induction [18]. We first analyze the data reporters' equilibrium decisions in Stage II, and then analyze the data collector's optimal strategy in Stage I.

4 STAGE II: PRIVACY PROTECTION EQUILIBRIUM OF DATA REPORTERS

In this section, we will characterize the privacy protection equilibrium of the data reporters in Stage II. To achieve that, we first formulate the data reporting game in Section 4.1. Then we analyze the best response in Section 4.2. Finally, we obtain the privacy protection equilibrium of the data reporters in Section 4.3.

4.1 Game Formulation

In this subsection, we formulate the interaction among all the data reporters as a game.

The data reporter desires privacy protection for the sum query analysis (rather than for his own data which is unknown to others). We can see from (11) that one data reporter's utility not only depends on his own strategy, but also depends on other reporters' strategies. This naturally leads to a data reporting game as follows:

Definition 2. *The data reporting game $\left\langle \mathcal{M}, (\Sigma_j), (U_j)\right\rangle$ in Stage II is defined as follows:*

- *Players: the individuals in the data reporter set $\mathcal{M}$.*
- *Strategies: a data reporter $j \in \mathcal{M}$ chooses his strategy, i.e., the variance of the noise σ_j^2, from the strategy set $\Sigma_j = [0, \infty]$. Let $\boldsymbol{\sigma}_{-j}^2$ denote the strategy profile of all other data reporters in $\mathcal{M}$ except the data reporter j. Recall that $\boldsymbol{\sigma}^2 = (\sigma_j^2, \boldsymbol{\sigma}_{-j}^2)$.*
- *Utilities: the utility function of a data reporter $j \in \mathcal{M}$ is given in (11).*

4.2 Best Response Analysis

In this subsection, we will analyze a data reporter's best response, which is the strategy that maximizes his utility given all other data reporters' strategies.

We can show that the data reporter's utility function in (11) is a concave function. Based on the first-order condition, we obtain the data reporter j' best response function given as follows:

$$\begin{aligned} BR_j\left(\boldsymbol{\sigma}_{-j}^2\right) &= \arg\max_{\sigma_j^2} U_j\left(\mathcal{M}, \sigma_g^2, \sigma_j^2, \boldsymbol{\sigma}_{-j}^2\right) \\ &= \max\left\{0, -\sum_{m\in\mathcal{M}, m\neq j} \sigma_m^2 - \sigma_g^2 + \beta_j\right\}, \end{aligned} \tag{13}$$

where β_j satisfies that

$$-\sum_{i\in\mathcal{T}} s_{ji} h'\left(\sum_{m\in\mathcal{M}, m\neq i} \frac{1}{a_{mi}} + \beta_j\right) = r_d. \tag{14}$$

Here $h'(\cdot) < 0$ denotes the first-order derivative of the decreasing function $h(\cdot)$ in (7). We can see that the best response function in (13) is a decreasing linear function of other data reporters' strategies σ_m^2 ($m \in \mathcal{M}\backslash\{j\}$).

The parameter β_j in (14) captures the data reporter j's joint consideration of social relationship and data correlation. If his social relationship with others is stronger (i.e., s_{ji} is larger) while other parameters are fixed, the parameter β_j will be larger. If the data correlation is stronger (i.e., a_{mi} is larger), the parameter β_j will also be larger.

4.3 Nash Equilibrium Analysis

Next, we characterize the Nash Equilibrium by finding the fixed point of all the data reporters' best response functions.

Definition 3. *The Nash Equilibrium (NE) of the data reporting game is a strategy profile $\boldsymbol{\sigma}^{2*}$ such that $\forall j \in \mathcal{M}$,*

$$\sigma_j^{2*} = \arg\max_{\sigma_j^2 \in \Sigma_j} U_j\left(\mathcal{M}, \sigma_g^2, \sigma_j^2, \boldsymbol{\sigma}_{-j}^{2*}\right). \tag{15}$$

In order to ensure the uniqueness of the NE, we make the following Assumption 1 with respect to β_j defined previously in (14)

Assumption 1. *The sequence $\{\beta_j\}, \forall j \in \mathcal{M}$, has a unique maximum value.*

We denote this unique maximum value as $\beta_{\hat{k}}$, where

$$\hat{k} = \arg\max_{j \in \mathcal{M}} \beta_j. \tag{16}$$

Assumption 1 is not restrictive in practice, since the value of $\beta_j, j \in \mathcal{M}$, is continuous. The probability of having at least two maximum values in this continuous sequence is zero. Assumption 1 corresponds to the situation where there is only one data reporter who has the most significant joint consideration of social relationship and data correlation.

Theorem 1. *Under Assumption 1, there is a unique NE for the data reporting game given by*

$$\boldsymbol{\sigma}^{2*} = \left(\sigma_{\hat{k}}^{2*}, \boldsymbol{\sigma}_{-\hat{k}}^{2*}\right) = \left(\max\left\{0, -\sigma_g^2 + \beta_{\hat{k}}\right\}, \mathbf{0}\right). \tag{17}$$

The proof is given in Appendix (Section 8.1).

When we have $\sigma_j^{2*} = 0, j \in \mathcal{M}$, we have a Gaussian distribution with zero variance, which indicates that the data reporter j will truthfully report his data without adding noise.

We can see that at the unique NE, the data reporter with the highest β_j may add noise, while all other data reporters will truthfully report data. This implies that the data reporter who has the most significant joint consideration of social relationship and data correlation will have the motivation to add a sufficiently large noise, which leaves no need for others to add noise.

We further discuss $\beta_{\hat{k}}$ in (17). According to how we obtain the parameter in (14), we can see that both social relationship parameter s_{ji} and data correlation term a_{mi} will affect the variance of the noise. This implies that a stronger social relationship or data correlation will lead to a larger noise added by the data reporter at the equilibrium.

Next, we focus on σ_g^2 in (17), i.e., the data collector's strategy. The noise variance $\sigma_{\hat{k}}^{2*}$ from the data reporter decreases with the data collector's strategy σ_g^2. This suggests that a stricter privacy-preserving mechanism imposed by the data collector (with a larger σ_g^2) can lead to a more accurate reporting. Moreover, considering the max operation in (17), we can see that if σ_g^2 is high enough, i.e., $\sigma_g^2 \geq \beta_{\hat{k}}$, the NE will be a strategy profile of all zeros. Later in Section 5.1, we can see such an all-zero equilibrium corresponds to the global equilibrium of our two-stage Stackelberg game.

Corollary 1. *Under Assumption 1, when $\sigma_g^2 \geq \beta_{\hat{k}}$, there is a unique NE of the data reporting game where each data reporter truthfully reports the data, i.e.,*

$$\sigma_j^{2*} = 0, \forall j \in \mathcal{M}. \tag{18}$$

Recall that the data reporter desires to restrict privacy leakage due to the announced sum query analysis (rather than due to his own data reporting alone). Corollary 1 implies that a good enough privacy-preserving mechanism with a high noise variance added by the data collector to the result can mitigate the data reporters' privacy concern and result in truthful reporting.

5 STAGE I: OPTIMAL STRATEGY OF THE DATA COLLECTOR

In this section, we study the data collector's optimal data reporter selection and privacy-preserving mechanism in Stage I, based on the privacy protection equilibrium of the data reporters derived in Section 4. We consider the scenario where the data collector has *complete* information of social relationship and data correlation. Later in the simulation (Section 6), we will look at the impact of her incomplete information on the overall system performance.

According to the utility function in (12) (and (10)) and the NE in (17), we can formulate the data collector's utility maximization problem as follows. She decides the optimal strategy to maximize his utility as follows:

$$\begin{aligned} \max_{\mathcal{M}, \sigma_g^2} \; & U_g\left(\mathcal{M}, \sigma_g^2, \boldsymbol{\sigma}^{2*}\right) = B_g(\mathcal{M}) - r_g \sum_{j \in \mathcal{M}} \sigma_j^{2*} - r_a \sigma_g^2 \\ \text{subject to} \; & \sigma_g^2 \in [0, +\infty], \\ & \mathcal{M} \subset \mathcal{T}, \\ & |\mathcal{M}| \geq M_{\min}. \end{aligned} \tag{19}$$

There are two decision variables for the data collector, the reporter set $\mathcal{M}$ and the noise variance σ_g^2 of the privacy-preserving mechanism. The reporter set $\mathcal{M}$ is a subset of the target individual pool $\mathcal{T}$, and has a minimum size requirement $M_{\min}$.

We first analyze the optimal noise variance σ_g^2 for a given data reporter set $\mathcal{M}$ in Section 5.1. Then we develop an algorithm to identify a reporter set to achieve a global optimum of Problem (19) in Section 5.2.

5.1 Optimal Privacy-Preserving Mechanism under a Fixed Reporter Set $\mathcal{M}$

In this subsection, we analyze the data collector's optimal privacy-preserving mechanism for a given reporter set $\mathcal{M}$.

We can see from Corollary 1 that $\beta_{\hat{k}}$ is a threshold of the data collector's strategy. All the data reporters will truthfully report their data if $\sigma_g^2 \geq \beta_{\hat{k}}$. It turns out that it is optimal for the data collector to add just enough noise to reach such a threshold, as shown in Theorem 2.

Theorem 2. *Given the selected data reporter set $\mathcal{M}$, the data collector's optimal strategy is given by*

$$\sigma_g^{2*}(\mathcal{M}) = \beta_{\hat{k}}. \tag{20}$$

Proof By applying (17) to the objective function in the utility maximization problem (19), we have

$$\begin{aligned} & U_g\left(\mathcal{M}, \sigma_g^2, \boldsymbol{\sigma}^{2*}\right) \\ & = \begin{cases} B_g(\mathcal{M}) - r_g \beta_{\hat{k}} + (r_g - r_a)\sigma_g^2, & \text{if } \sigma_g^2 \in [0, \beta_{\hat{k}}], \\ B_g(\mathcal{M}) - r_a \sigma_g^2, & \text{otherwise.} \end{cases} \end{aligned} \tag{21}$$

Note that $r_g - r_a > 0$. We can see that the objective function is increasing in $[0, \beta_{\hat{k}}]$ and is decreasing in $(\beta_{\hat{k}}, +\infty]$. So the optimal solution is $\sigma_g^{2*}(\mathcal{M}) = \beta_{\hat{k}}$. □

Theorem 2 implies that if the social relation is stronger or the data correlation is stronger (in both cases $\beta_{\hat{k}}$ will be higher), the data collector will enforce a stricter privacy-preserving mechanism accordingly to guarantee truthful reporting.

5.2 Optimal Data Reporter Selection

After obtaining the optimal privacy-preserving mechanism for a given reporter set, we then discuss how the data collector should construct the reporter set to maximize her utility in this subsection.

By plugging the data collector's optimal strategy in (20) for a fixed reporter set $\mathcal{M}$ into her utility function in (12), we obtain her optimal utility for a fixed reporter set $\mathcal{M}$ in the following (22). Recall that $\hat{k} = \arg\max_{j\in\mathcal{M}} \beta_j$.

$$U_g^*(\mathcal{M}) = B_g(\mathcal{M}) - r_a\beta_{\hat{k}}. \tag{22}$$

For the data collector's pure benefit $B_g(\mathcal{M})$ in (22), we make the following general assumption.

Assumption 2.

$$B_g(\mathcal{M}) \le B_g(\mathcal{M}') \quad if \quad \mathcal{M} \subset \mathcal{M}'. \tag{23}$$

Assumption 2 states that if the data collector is able to include more data reporters from the current set, her benefit will not decrease. This is satisfied for most practical scenarios.

Algorithm 1: Data reporter selection algorithm

Input: Data correlation $a_{j,i}$ in (6), $\forall i \neq j \in \mathcal{T}$; social relationship $s_{j,i}$, $\forall i \neq j \in \mathcal{T}$; data collector's utility function $U_g(\cdot)$.

Output: Optimal data reporter set $\mathcal{M}^*$.

1 $\mathcal{M}^* \leftarrow \mathcal{T}, U_g^* \leftarrow U_g\left(\mathcal{M}^*, \sigma_g^{2*}(\mathcal{M}^*)\right)$;
2 $\mathcal{M} = \mathcal{T}$;
3 **for** $i = 1$ *to* $T - M_{\min}$ **do**
4 $k_i = \arg\max_{j\in\mathcal{M}} \beta_j$;
5 $\mathcal{M} \leftarrow \mathcal{M}\backslash\{k_i\}$;
6 **if** $U_g\left(\mathcal{M}, \sigma_g^{2*}(\mathcal{M})\right) \ge U_g^*$ **then**
7 $\mathcal{M}^* = \mathcal{M}$;
8 $U_g^* = U_g\left(\mathcal{M}, \sigma_g^{2*}(\mathcal{M})\right)$;

We then propose Algorithm 1 to construct the optimal reporter set $\mathcal{M}^*$ to maximize the utility in (22). Algorithm 1 initializes the optimal set $\mathcal{M}^*$ as the whole set $\mathcal{T}$ and the optimal utility U_g^* as the whole set's optimal utility $U_g\left(\mathcal{M}^*, \sigma_g^{2*}(\mathcal{M}^*)\right)$ (line 1). Then in each iteration, we remove the data reporter with the highest β_j (line 5), and update the optimal set $\mathcal{M}^*$ and the optimal utility U_g^* after comparing the resulting utility after such a removal with the current optimal utility (line 6-8). The algorithm stops after only $M_{\min}$ individuals remain in the current set $\mathcal{M}$.

Theorem 3. *Under Assumption 2, the pair* $\left(\mathcal{M}^*, \sigma_g^{2*}(\mathcal{M}^*)\right)$ *(with* $\mathcal{M}^*$ *obtained from Algorithm 1) is the optimal solution of the data collector's utility maximization problem (19).*

The proof is in the Online-Appendix [16]. A key step of the proof is to show that removing a data reporter that is not with the highest β_j in any iteration will never be an optimal choice. Recall that the parameter β_j satisfies $-\sum_{i\in\mathcal{T}} s_{ji}h'\left(\sum_{m\in\mathcal{M}, m\neq i} 1/a_{mi} + \beta_j\right) = r_d$ and it is different in different reporter sets. Consider a current reporter set $\mathcal{M}$. If the data collector removes a data reporter $k' \neq \hat{k}$ from current set $\mathcal{M}$, we will have $\sum_{m\in\mathcal{M}\backslash\{k'\}, m\neq i} 1/a_{mi} < \sum_{m\in\mathcal{M}, m\neq i} 1/a_{mi}$ and thus the reporter $\hat{k}$'s parameter $\beta'_{\hat{k}}$ in the remained set $\mathcal{M}\backslash\{k'\}$ will be higher than $\beta_{\hat{k}}$ in the current set $\mathcal{M}$, i.e., $\beta'_{\hat{k}} > \beta_{\hat{k}}$. Combining Assumption 2 ($B_g(\mathcal{M}\backslash\{k'\}) \le B_g(\mathcal{M})$), we can see the overall utility in (22) after removing k' is lower than that without removal. So the only *possible* choice to have a higher utility than that without removal is to remove the data reporter with the highest β_j, i.e., $\hat{k}$. We compare the utility after removing $\hat{k}$ with the utility without removal to verify whether this removal of $\hat{k}$ is better off, which is iteratively implemented in Algorithm 1.

Compared with the naive approach of exhaustive traverse over all possible sets that has an exponential complexity of $O(2^T)$, Algorithm 1 has a polynomial complexity of $O(T^4)$. More specifically, computing each $\beta_j, j \in \mathcal{T}$ in line 4 has complexity of $O(T^2)$ and finding the maximum one has complexity of $O(T)$. With the for loop operation, the overall complexity becomes $O(T^4)$.

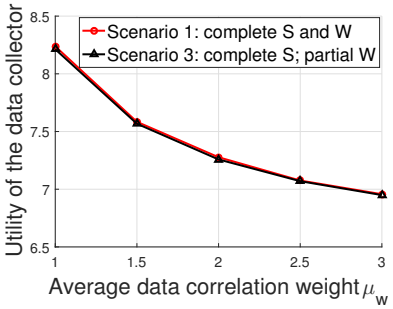
6 SIMULATION RESULTS

In this section, we provide some simulation results to investigate the impact of data correlation and social relationship on the overall system performance. We also examine the performance of the system in some practical scenarios where the data collector fails to have complete network information.

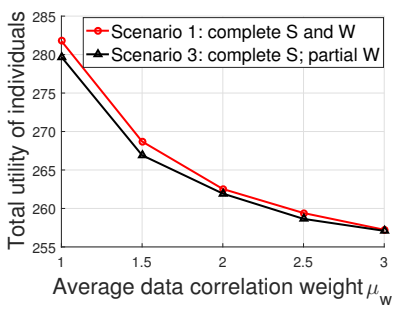
6.1 Simulation Setup

We first construct the randomized social network **S**. We exploit Facebook social data [15]. Specifically, the data set we used contains 61 users and shows the social connections between users (i.e., whether there exists an edge between them). However, it does not indicates the social strength s_{ij} between individual i and j (if there is an edge between them). So for each individual i, we generate the social relationship s_{ij} through a truncated normal distribution with a mean μ_i, a standard deviation σ, and a support set of $(0, \infty)$. We consider an asymmetric social relationship, where it is possible that $s_{ij} \neq s_{ji}$. To capture the diversity of social relationship among different individuals, we do not set the means μ_i, $\forall i \in \mathcal{T}$ to be equal. Instead, we generate mean μ_i through a truncated normal distribution with a mean μ_s, a standard deviation σ_s, and a support set $(0, \infty)$. A higher σ_s means a greater diversity of individuals' consideration of others' privacy. We set $p_s = 0.8$, $\sigma = 0.1$, $\mu_s = 0.5$, and $\sigma_s = 0.5$, unless specified otherwise. Furthermore, we set $s_{ii} = 1, \forall i \in \mathcal{T}$.

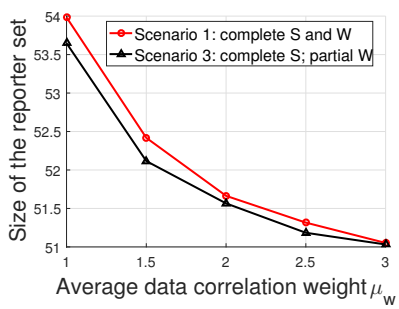
We then construct the randomized data correlation network **W**. The direct correlation exists between i and j ($i, j \in \mathcal{T}$) with a probability p_w. If there exits a direct correlation, we generate the correlation weight w_{ij} through a truncated normal distribution with a mean μ_w, a standard deviation σ_w, and a support set $(0, \infty)$.



(a) Utility of the data collector vs average data correlation weight μ_w.

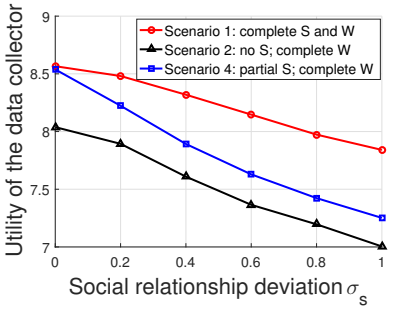


(b) Total utility of individuals vs average data correlation weight μ_w.

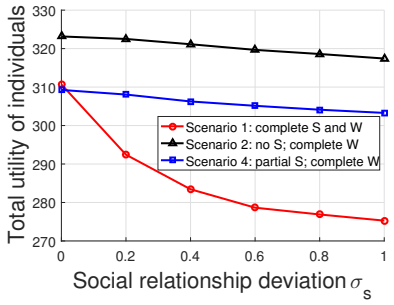


(c) Size of the reporter set vs average data correlation weight μ_w.

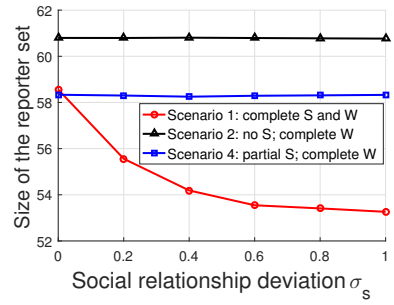
Figure 2: Impact of the average weight μ_w of data correlation network under different network information knowledge.



(a) Utility of the data collector vs social relationship deviation σ_s



(b) Total utility of individuals vs social relationship deviation σ_s.



(c) Size of the reporter set vs social relationship deviation σ_s.

Figure 3: Impact of the social relationship deviation σ_s of social network under different network information knowledge.

We have $w_{ij} = w_{ji}$ due to the symmetry of data correlation. We set $p_w = 0.8$, $\mu_w = 1$, and $\sigma_w = 0.4$, unless specified otherwise.

For the data collector's utility function, we set $B_g(\mathcal{M}) = 10 + |\mathcal{M}| \times 0.01$, which satisfies Assumption 2. We further set $M_{\min} = 50$, $r_a = 0.9$, and $r_g = 1 > r_a$. For the function in (7) to characterize privacy loss, we consider a concrete form $h(Var(z|x_i)) = \exp(-Var(z|x_i))$, which is decreasing, convex, and positive. Under this function, we have the data reporter j's parameter as follows:

$$\beta_j = \sum_{i \in \mathcal{T}} s_{ji} \exp\left(- \sum_{m \in \mathcal{M}, m \neq i} \frac{1}{a_{mi}} \right), \forall j \in \mathcal{M}. \tag{24}$$

For the data reporters' utility function, we set $B_j(\mathcal{M}) = 5 + |\mathcal{M}| \times 0.01$ and $r_d = 0.1$. We assume that the utility function parameters are the same for all individuals for computational convenience. In the simulation, we will calculate the total utility of all the individuals including non-reporters. We define the utility function of a non-reporter as $U_j(\mathcal{M}, \sigma_g^2, \boldsymbol{\sigma}^2) = -\sum_{i \in \mathcal{T}} s_{ji} l_i(\mathcal{M}, \sigma_g^2, \boldsymbol{\sigma}^2), \forall j \in \mathcal{T} \backslash \mathcal{M}$, which reflects only the privacy concern without reward.

For each set of system parameters, we implement the simulations 500 times with different random realizations of network parameters, and calculate the average results.

We examine the performance of the system in some practical scenarios where the data collector fails to have complete network information. More specifically, we consider the following four scenarios with different network information availabilities. She adopts the optimal strategy according to her available knowledge, respectively. We assume that the data reporters always have complete network information in order to realize the derived NE in (17).

- **Scenario 1: complete information of S and W.** The data collector has complete information of networks **S** and **W**.
- **Scenario 2: not aware of social relationship; complete information of W.** The data collector has complete information of data correlation network **W**. However, she has no information of individuals' social relationship. Hence she assumes that the social network is a unit matrix **E**, i.e., each individual only cares about his own privacy loss.
- **Scenario 3: complete information of S; partial information of W.** The data collector has complete information of social network **S**. However, she only has partial information of data correlation in terms of the average value $\mu_w p_w$. In other words, she considers a data correlation network $\mathbf{W}' = [w'_{ij}]_{T \times T}$ where $w'_{ii} = 0, \forall i \in \mathcal{T}$ and $w'_{ij} = \mu_w p_w, \forall i, j \in \mathcal{T}, i \neq j$.
- **Scenario 4: partial information of S; complete information of W.** The data collector has partial information of social network **S** but complete information of data correlation network **W**. She can obtain the social connections from the Facebook data [15]. But she only has partial information of social strength s_{ij} in terms of the average value μ_s. In other words, she considers a social network $\mathbf{S}'$ as follows.

For each individual i: (i) if there does not exist an edge between i and j based on Facebook data, she sets $s'_{ij} = 0$; (ii) if there exists an edge between i and j, she set $s'_{ij} = \mu_s$; (iii) as for individual i himself, she set $s'_{ii} = 1$. Note that she fails to capture the social diversity of the individuals, since s'_{ij} are the same for all i and j (if there is an edge between them).

6.2 Impact of Data Correlation Network and Social Relationship Network

We study the impact of the data correlation network and social relationship network. More specifically, we investigate how the utility of the data collector, the total utility of the individuals in set $\mathcal{T}$, and the size of the reporter set will change as the network information availabilities to the data collector change. To study the impact of data correlation network $\mathbf{W}$, we show the results under different information availabilities of $\mathbf{W}$ with same complete information of $\mathbf{S}$, i.e., Scenarios 1 and 3. To study the impact of social network $\mathbf{S}$, we show the results under different information availabilities of $\mathbf{S}$ with same complete information of $\mathbf{W}$, i.e., Scenarios 1, 2 and 4.

Impact of the average weight μ_w of data correlation network (Figure 2). Figure 2a and Figure 2b show that both the data collector's utility and the individuals' total utility decrease as the average weight μ_w increases. This is because *stronger data correlation network indicates more serious privacy loss*, which enforces a stricter privacy-preserving mechanism with a higher noise variance. Furthermore, Figure 2c shows that the data collector will also tend to select fewer data reporters under a stronger data correlation. From Figure 2a, we can see that *the data collector with partial information of $\mathbf{W}$ (Scenario 3) is able to achieve a near optimal utility comparing with the full information scenario (Scenario 1). This implies that symmetric partial information of $\mathbf{W}$ is enough for the data collector's utility maximization.*

Impact of social relationship deviation σ_s of social network (Figure 3). Recall that the deviation σ_s captures that diversity of social consideration among individuals. Figure 3a and Figure 3b show that both the data collector's utility and the individuals' total utility decrease as the variance σ_s increases. Recall that data collector's optimal noise variance σ_g^{2*} is closely related to the individual with the largest $\beta_j = \sum_{i\in\mathcal{T}} s_{ji} \exp(-\sum_{m\in\mathcal{M}, m\neq i} 1/a_{mi})$. A higher social diversity will lead to more diverse values of s_{ji}, hence may lead to a higher value of the largest β_j. *This forces the data collector to add a higher noise variance considering individuals' higher privacy concerns.* Figure 3c (Scenario 1) shows that as the social deviation σ_s increases, the data collector tends to select fewer data reporters to release the pressure to add larger noise.

Importance of the awareness of social relationship (Figure 3). Figure 3a shows that, unlike the comparison for the data correlation network (i.e., Figure 2a), the data collector experiences great utility loss due to incomplete information of $\mathbf{S}$ (by comparing Scenarios 1 and 4 here), and the gap increases with the deviation σ_s. *This implies that having access to the accurate asymmetric social network information is more critical than accessing the symmetric data correlation network information for the data collector.* The utility loss is even more significant if she completely ignores the social relationship as in scenario 2. This is because data reporters' concern of others' privacy plays an important role in their privacy protection strategies. From Figure 3c, we observe that *the data collector would select more data reporters under incomplete information of $\mathbf{S}$ (in Scenarios 2 and 4) compared with complete information (in Scenario 3).* She thinks that data reporters have the similar privacy concerns under incomplete information of $\mathbf{S}$ and she is less likely to remove data reporters from the whole set $\mathcal{T}$. So individuals benefit more from the data collector's incomplete information (as shown in Figure 3b) since more individuals get rewards.

6.3 Summary of the Observations

We summarize the main insights as follows: (i) Both the data collector and data reporters experience lower utilities under a stronger data correlation network or a strong social network under all information scenarios. (ii) With partial data correlation network information (and complete social network information), both the data collector and individuals can achieve utilities that are closed to the optimum under complete data correlation network information. (iii) With partial or no social network information, the data collector suffers some utility loss but the data reporters enjoy higher utilities, compared with the complete social network information.

7 CONCLUSION

In this paper, we study the privacy-preserving data collection problem with the consideration of data correlation and social relationship. We utilize the Gaussian correlation model, a special Markov random field model, to characterize the data correlation and the corresponding individuals' privacy loss due to data analysis. We further exploit a game theoretic model to study data reporters' social interactions. Our analysis shows that at the equilibrium only the data reporter with the most significant consideration may have the motivation to add a sufficiently large noise, while other data reporters will truthfully report their data. Anticipating such an equilibrium, the data collector should jointly capture the impact of the data correlation and social relationship by adding enough noise to the analysis result, to ensure that all the data reporters will truthfully report the data. We further develop an algorithm for the data collector to optimally choose the set of data reporters.

For the future work, we can consider other data analysis methods (e.g., logistic regression, linear regression, etc). Specifically, we need to develop a relevant criterion to characterize individuals' privacy loss for the corresponding analysis application. We can also consider the a challenging scenario where each data reporter has incomplete information of the social network and the data correlation, which leads to a Bayesian game model regarding their interactions.

8 APPENDIX

8.1 Proof of Theorem 1

First, we can prove this strategy profile (17) is a NE by verifying that every individual is playing his best response as in (13). For the data reporter $\hat{k}$, we have his best response $BR_{\hat{k}}(\mathbf{0}) = \max\left\{0, -\sigma_g^2 + \beta_j\right\}$. For other data reporter k, $\forall k \neq \hat{k}$, we also have his best response $BR_k\left((\sigma_{\hat{k}}^2, \mathbf{0})\right) = \max\left\{0, -\max\left\{\sigma_g^2 - \beta_k, \beta_{\hat{k}} - \beta_k\right\}\right\} = 0$. The second equality holds since $\beta_{\hat{k}} > \beta_k, \forall k \neq \hat{k}$. So this strategy profile is a NE.

Second, we can prove the uniqueness of the NE (17) by verifying that none of other strategy profiles is an NE. First, we can show that for the strategy profiles $(\sigma_{\hat{k}}^{2'}, \mathbf{0})$ where $\sigma_{\hat{k}}^{2'} \neq \sigma_{\hat{k}}^{2*}$, and $(\sigma_{\hat{k}}^{2*}, \boldsymbol{\sigma}_{-\hat{k}}^{2'})$ where $\boldsymbol{\sigma}_{-\hat{k}}^{2'} \neq \mathbf{0}$, they can't be NE because $(\sigma_{\hat{k}}^{2*}, \mathbf{0})$ is NE.

It remains to consider the strategy profile $\boldsymbol{\sigma}^{2'} = (\sigma_{\hat{k}}^{2'}, \boldsymbol{\sigma}_{-\hat{k}}^{2'})$ where $\sigma_{\hat{k}}^{2'} \neq \sigma_{\hat{k}}^{2*}$ and $\boldsymbol{\sigma}_{-\hat{k}}^{2'} \neq \mathbf{0}$. We prove this through contradiction. Let us assume that such a strategy profile is a NE. Let $k \neq \hat{k}$ be one of those data reporters whose strategy is non-zero, i.e., $\sigma_k^{2'} > 0$. Since the data reporter k is playing his best response, we have

$$BR_k(\boldsymbol{\sigma}_{-k}^{2'}) = \max\{0, -\sigma_{\hat{k}}^{2'} - \sum_{m\in\mathcal{M}, m\neq\hat{k},k} \sigma_m^{2'} + \beta_k\} = \sigma_k^{2'} > 0,$$

i.e.,

$$\sigma_k^{2'} = -\sigma_{\hat{k}}^{2'} - \sum_{m\in\mathcal{M}, m\neq\hat{k},k} \sigma_m^{2'} + \beta_k. \tag{25}$$

As for the data reporter $\hat{k}$, we have his best response

$$\begin{aligned} BR_{\hat{k}}(\boldsymbol{\sigma}_{-\hat{k}}^{2'}) &= \max\{0, -\sigma_k^{2'} - \sum_{m\in\mathcal{M}, m\neq\hat{k},k} \sigma_m^{2'} + \beta_{\hat{k}}\} \\ &= \max\{0, \sigma_{\hat{k}}^{2'} - \beta_k + \beta_{\hat{k}}\}. \end{aligned}$$

The second equality is obtained by applying (25). Since $-\beta_k + \beta_{\hat{k}} > 0$, then $\max\{0, \sigma_{\hat{k}}^{2'} - \beta_k + \beta_{\hat{k}}\} > \sigma_{\hat{k}}^{2'}$ This means that the data reporter $\hat{k}$ is not playing his best response, which contracts with the assumption. So the strategy profile $\boldsymbol{\sigma}^{2'}$ is not NE. This concludes the proof of the uniqueness. □

8.2 Proof of Theorem 3

We need to prove that Algorithm 1 returns the optimal reporter set $\mathcal{M}^*$ that maximizes the utility in (22). We can verify that removing the data reporter that is not with the highest β_j can't achieve a higher utility than not removing does.

Consider an arbitrary current reporter set as $\mathcal{M}$ associated with $\beta_j, j \in \mathcal{M}$. The optimal utility in (22) under this reporter set is

$$U_g^*(\mathcal{M}) = \bar{B}_g(\mathcal{M}) - r_a\beta_{\hat{k}}. \tag{26}$$

Let $\mathcal{M}' = \mathcal{M}\backslash\mathcal{K}$ ($\hat{k} \notin \mathcal{K}$) be the remained set after removing a set of reporters that doesn't contain $\hat{k}$, and the associated parameter be $\beta_j', j \in \mathcal{M}'$. The optimal utility under this reporter set is

$$U_g^*(\mathcal{M}') = \bar{B}_g(\mathcal{M}') - r_a \max_{j\in\mathcal{M}'} \beta_j'. \tag{27}$$

We have

$$\beta_{\hat{k}} < \beta_{\hat{k}}' \leq \max_{j\in\mathcal{M}'} \beta_j'. \tag{28}$$

The first inequality holds since

$$\sum_{m\in\mathcal{M}, m\neq i} \frac{1}{a_{mi}} > \sum_{m\in\mathcal{M}', m\neq i} \frac{1}{a_{mi}}.$$

Recall that $\beta_{\hat{k}}$ satisfies $-\sum_{i\in\mathcal{T}} s_{ji} h'\left(\sum_{m\in\mathcal{M}, m\neq i} \frac{1}{a_{mi}} + \beta_j\right) = r_d$ and $\beta_{\hat{k}}'$ satisfies $-\sum_{i\in\mathcal{T}} s_{ji} h'\left(\sum_{m\in\mathcal{M}', m\neq i} \frac{1}{a_{mi}} + \beta_j\right) = r_d$. The second inequality holds due to max operation. Finally, we have $\beta_{\hat{k}} < \max_{j\in\mathcal{M}'} \beta_j'$. According to Assumption 2, we have $\bar{B}_g(\mathcal{M}) \geq \bar{B}_g(\mathcal{M}')$. So after comparing (26) and (27), we have $U_g^*(\mathcal{M}) \geq U_g^*(\mathcal{M}')$, i.e., removing a reporter set that does not contain $\hat{k}$ can't achieve a higher utility than not removing does.

The only *possible* choice to have a higher utility than that without removal is to remove the data reporter with the highest β_j, i.e., $\hat{k}$. So we initialize the current set as the whole set $\mathcal{T}$. We iteratively remove the one with the highest parameter in the current set and calculate the corresponding utility. After traversing all the possible removals, we are able to find the optimal reporter set that maximizes the data collector's utility. □

REFERENCES

[1] P. Boyle and M. Frean. 2005. Dependent gaussian processes. In *Advances in neural information processing systems*. 217–224.

[2] F. Brunton and H. Nissenbaum. 2015. *Obfuscation: A user's guide for privacy and protest*. Mit Press.

[3] M. Chessa, J. Grossklags, and P. Loiseau. 2015. A game-theoretic study on non-monetary incentives in data analytics projects with privacy implications. In *IEEE Computer Security Foundations Symposium (CSF)*. IEEE, 90–104.

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

[5] L. K. Fleischer and Y. Lyu. 2012. Approximately optimal auctions for selling privacy when costs are correlated with data. In *Proceedings of ACM Conference on Electronic Commerce*.

[6] A. Ghosh, K. Ligett, A. Roth, and G. Schoenebeck. 2014. Buying private data without verification. In *Proceedings of the fifteenth ACM conference on Economics and computation*. ACM, 931–948.

[7] A. Ghosh and A. Roth. 2015. Selling privacy at auction. *Games and Economic Behavior* 91 (2015), 334–346.

[8] A. Ghosh, T. Roughgarden, and M. Sundararajan. 2012. Universally utility-maximizing privacy mechanisms. *SIAM J. Comput.* 41, 6 (2012), 1673–1693.

[9] X. He, A. Machanavajjhala, and B. Ding. 2014. Blowfish privacy: Tuning privacy-utility trade-offs using policies. In *Proceedings of the ACM international conference on Management of data*.

[10] Apple Inc. 2016. Differential Privacy Overview. Online: https://images.apple.com/privacy/docs/Differential_Privacy_Overview.pdf.

[11] H. Jin, L. Su, and K. Nahrstedt. 2017. Theseus: Incentivizing Truth Discovery in Mobile Crowd Sensing Systems. In *Proceedings of the 18th ACM International Symposium on Mobile Ad Hoc Networking and Computing*.

[12] D. Kifer and A. Machanavajjhala. 2011. No free lunch in data privacy. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*. 193–204.

[13] D. Kifer and A. Machanavajjhala. 2014. Pufferfish: A framework for mathematical privacy definitions. *ACM Transactions on Database Systems* 39, 1 (2014), 3.

[14] R. Kindermann and J. L. Snell. 1980. *Markov random fields and their applications*. Vol. 1. American Mathematical Society.

[15] J. Leskovec and A. Krevl. 2014. SNAP Datasets: Stanford Large Network Dataset Collection. Online: http://snap.stanford.edu/data.

[16] G. Liao, X. Chen, and J. Huang. 2018. Online Appendix, available at http://jianwei.ie.cuhk.edu.hk/publication/AppendixMobihoc18Privacy.pdf.

[17] S. Liu, C.and Chakraborty and P. Mittal. 2016. Dependence Makes You Vulnberable: Differential Privacy Under Dependent Tuples.. In *Proceeding of Network Distributed System Security Symposium*.

[18] A. Mas-Colell, M. D. Whinston, J. R Green, et al. 1995. *Microeconomic theory*. Vol. 1. Oxford university press New York.

[19] J. Pawlick and Q. Zhu. 2016. A Stackelberg game perspective on the conflict between machine learning and data obfuscation. In *IEEE International Workshop on Information Forensics and Security*.

[20] S. Song, Y. Wang, and K. Chaudhuri. 2017. Pufferfish Privacy Mechanisms for Correlated Data. In *Proceedings of ACM International Conference on Management of Data*.

[21] W. Wang, L. Ying, and J. Zhang. 2015. A game-theoretic approach to quality control for collecting privacy-preserving data. In *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. 474–479.

[22] W. Wang, L. Ying, and J. Zhang. 2016. The value of privacy: Strategic data subjects, incentive mechanisms and fundamental limits. *ACM International Conference on Measurement and Modeling of Computer Science* (2016).

[23] X. Wu, T. Wu, M. Khan, Q. Ni, and W. Dou. 2017. Game Theory Based Correlated Privacy Preserving Analysis in Big Data. *IEEE Transactions on Big Data* PP (2017). Issue 99.

[24] B. Yang, I. Sato, and H. Nakagawa. 2015. Bayesian differential privacy on correlated data. In *Proceedings of the ACM International Conference on Management of Data*. 747–762.