# Correlation-Aware and Personalized Privacy-Preserving Data Collection

Dongxiao Yu*, Kaiyi Zhang*, Youming Tao*‡, Wenlu Xu†, Yifei Zou*, Xiuzhen Cheng*
*School of Computer Science and Technology, Shandong University, P.R. China
†Department of Statistics, University of California, Los Angeles (UCLA), U.S.
‡School of Electrical Engineering and Computer Science, TU Berlin, 10587 Berlin, Germany

***Abstract*—Data collection from users is essential for various IoT services. However, privacy concerns may prevent users from sharing their raw data truthfully. The problem becomes more complex when the data and relationships are correlated and the privacy preferences are personalized. In particular, users' data are influenced by social interactions, which implies that others' data can affect users' privacy. Moreover, users care not only about their own privacy leakage, but also about their social contacts' privacy leakage, due to the social ties in reality. Furthermore, different users have different levels of privacy sensitivity for their own data, which poses a challenge for balancing user privacy and data utility. In this paper, we investigate the correlation-aware and personalized private data collection problem. We formulate the private data collection process as a Stackelberg game, where the platform sets its reward policy and users select their noise levels for privacy preservation. To tackle the challenges above, we adopt the Gaussian correlation model to represent the data correlation among users and integrate the relationship correlation and personalization when deriving the optimal strategies for both users and the platform. Notably, we employ mutual information differential privacy for a rigorous quantification of the correlated privacy loss. Through rigorous theoretical analysis, we first establish the connection between users' Nash equilibrium and the payment mechanism, and then optimize the platform's accuracy under a budget constraint by designing the reward policy. We also demonstrate the effectiveness of our proposed framework through extensive numerical experiments.**

***Index Terms*—Data collection, privacy preservation, Stackelberg game**

## I. Introduction

Data collection from users is a key component of many IoT applications, e.g., intelligent transportation [15] and smart city [10], as it enables the platform to understand the users' needs, preferences, and behaviors, and to provide personalized and optimized services. However, data collection poses significant privacy risks for the users who provide their data. Due to pervasive privacy concerns [1], users may be disinclined to share their raw data candidly. For instance, users may not want to divulge their locations, health conditions, or personal preferences to the platform or other third parties. These data can reveal sensitive information about users' identities, behaviors, and lifestyles, which can be exploited for malicious purposes [12]. Specifically, one has to contend with the following inevitable challenges.

One challenge is that users' data are often correlated due to social interactions, which implies that users' privacy can be jeopardized by others' data [19]. For instance, if two users are neighbours and one of them shares his location, the other user's location can be deduced with high probability. Furthermore, with the social relationship forged in reality, users care about the privacy leakage of not only themselves but also those individuals who are socially related to them. This necessitates that the measurement of users' privacy loss incorporates both the data and social correlation with their peers. Thus, it is imperative to safeguard users' privacy under correlation.

Another challenge lies in the diversity of users' privacy sensitivities, intensifying the delicate balance between user privacy and data utility [16]. For instance, certain users may willingly share more data in return for increased rewards or enhanced services, while others may exercise caution and demand higher levels of privacy protection. Consequently, it becomes imperative to tailor the degree of privacy preservation during the data collection process to accommodate varying user preferences and requirements.

Several works have made attempts to tackling these challenges. Due to space limit, here we just discuss the most related works. [13] introduces a two-stage Stackelberg game to analyze the participation level of the mobile users and the optimal incentive mechanism of the crowdsensing service provider. But they neglected the privacy leakage issue. [11] then investigated correlated data collection while protecting users' privacy at the same time. They assume the data collector will perturb the aggregation result by adding noise while having no access to the exact aggregation result, which limits the applicability of the proposed mechanism. Also, in [11], users and the platform are assumed to have the same objective, i.e., maximizing the data aggregation accuracy, which is not practical in general. Based on this insight, [18] focused on a more general non-cooperative game. That is, the platform is interested in maximizing its accuracy, whereas users only care their payoffs, which is also the focus of this paper. This way, the payment mechanism needs to be devised to strike a

Corresponding author: Youming Tao (tao@ccs-labs.org). This work was supported in part by Major Basic Research Program of Shandong Provincial Natural Science Foundation under Grant ZR2022ZD02, National Natural Science Foundation of China (NSFC) under Grant 62122042 and 62102232, and Shandong Science Fund for Excellent Young Scholars (No.2023HWYQ-007).

979-8-3503-7099-7/24/$31.00 ©2024 IEEE

good balance between the different objectives of both sides. However, there are still several defects need to be addressed. Firstly, in both [11] and [18], every user's privacy loss is measured by the mutual informant that indicates its individual information contained by the analysis result. This concept, however, is not rigorous enough compared with the *de facto* standard notion of differential privacy. As demonstrated in [2], [3], differential privacy implies a bound on the mutual information leakage, but not vice-versa. Thus, it is necessary to extend the equilibrium analysis and payment mechanism design to the differential privacy framework. Secondly, none of these work have ever incorporated users' different privacy preferences to provide personalized payments for them.

To address these issues, in this paper, we propose a correlation-aware and personalized privacy-preserving data collection framework. Our framework leverages the two-stage Stackelberg game where the platform determines its reward policy and users choose their noise levels while considering the social influence among users and their personal privacy requirements. We adopt the Gaussian correlation model to characterize the data correlation among users. Specifically, we employ the notion of mutual-information differential privacy (MIDP) [4], a variant of differential privacy for correlated data, to rigorously measure privacy loss of each user for the first time (to the best of our knowledge). Moreover, by taking into account the different levels of personal privacy sensitivity of different users, we conduct a refined analysis for the connection between users' Nash equilibrium and the payment mechanism and provide a personalized payment mechanism for the platform while optimizing its accuracy under the budget constraint. We also conduct extensive numerical experiments to evaluate the performance of our framework over several real-world datasets. Our results demonstrate that our framework can effectively collect useful data for smart city while preserving users' privacy from correlation leakage.

## II. Problem Formulation

### A. Private Data Collection Problem Setting

A platform aims to collect individual data from $n$ users in the set $\iota = \{1, 2, \ldots, n\}$ for analytics. The individual data of user $i$ is $x_i \in \mathbb{R}$. After the collection, the platform aggregates all users' reported data to obtain some statistics. Specifically, we consider the sum aggregation, which serves as the core of many intelligent data analysis tasks, such as training large-scale neural networks and other statistical models, and also federated learning. The aggregation result in principle is

$$y \coloneqq \sum_{i=1}^{n} x_i. \tag{1}$$

Unfortunately, due to privacy concern, users may not report their data truthfully. To protect their privacy, users can perturb their data by adding some random noise. We assume that the noise is a zero-mean Gaussian random variable, just as most of the previous works on private data collection did, such as [11], [14]. This kind of assumption is reasonable since Gaussian noise is effective for privacy preserving as indicated by the Gaussian mechanism of differential privacy [6]. Thus, the actual results obtained by the platform should be

$$\widehat{y} \coloneqq \sum_{i=1}^{n} (x_i + \xi_i) = y + \xi_g, \tag{2}$$

where $\xi_i \sim \mathcal{N}(0, \sigma_i^2)$ is the random noise added by each user $i$, $\sigma_i$ is the standard deviation indicating the noise magnitude, $\xi_g \sim N(0, \lambda^2)$ represents the whole noise with $\lambda^2 = \sum_{i=1}^{n} \sigma_i^2$ for simplicity. Typically, the noise variances of all users are bounded, i.e., $\sigma_i \in [\underline{\sigma}, \overline{\sigma}]$.

The platform provides rewards to users for their data sharing. Specifically, the reward payment $p_i$ for user $i$ is set as a linear function of its noise magnitude, i.e.,

$$p_i \coloneqq r - \theta_i \sigma_i^2, \tag{3}$$

where $r$ is the original reward each user can get if he does not add any noise to his reported data and $\theta_i$ is a positive weight factor to indicate user $i$'s type.

### B. Data and Social Correlation Model

User's data are unavoidably correlated due to social interactions. We use the Gaussian correlation model [17] to capture the data correlation among users, just as [11], [18] did. Gaussian correlation model is a special case of the Markov random field [8] and it presents the correlation among data as a non-negative weighted undirected graph. Specifically, let $\mathcal{G}(\mathcal{V}, \mathcal{W})$ be a weighted undirected graph, where $\mathcal{V}$ is the vertex set and $\mathcal{W}$ is the set of all weighted edges in $G$. Each vertex $v_i \in \mathcal{V}$ represents the user $i$ and each weighted edge $(i, j, w_{ij}) \in \mathcal{W}$ with $w_{ij} \geq 0$ describes the correlation between user $i$ and user $j$. Intuitively, the larger $w_{ij}$ is, the more tightly is $i$ and $j$ correlated. Let $\boldsymbol{W} = (w_{ij}) \in \mathbb{R}^{n \times n}$ be the weighted adjacent matrix and $\boldsymbol{D} = \mathrm{diag}(w_1, w_2 \ldots\ldots w_n)$ be the weighted degree matrix of $\mathcal{G}(\mathcal{V}, \mathcal{W})$ where $w_i \coloneqq \sum_{j \neq i} w_{ij}$. The Laplacian matrix of $\mathcal{G}(\mathcal{V}, \mathcal{W})$ is defined as

$$\boldsymbol{L} \coloneqq \boldsymbol{D} - \boldsymbol{W} = \begin{bmatrix} w_1 & -w_{12} & \cdots & -w_{1n} \\ -w_{21} & w_2 & \cdots & -w_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ -w_{n1} & -w_{n2} & \cdots & w_n \end{bmatrix}. \tag{4}$$

Let $\boldsymbol{x}$ be the vector containing all the user data, i.e., $\boldsymbol{x} \coloneqq (x_1, x_2, \cdots, x_n)^\top$, and $\boldsymbol{x}_{-i}$ be denote all the components in $\boldsymbol{x}$ but $x_i$. For $\forall i$, the conditional joint probability of $\boldsymbol{x}_{-i}$ is

$$p(\mathbf{x}_{-i} | x_i) \propto \exp\left(-\frac{\mathbf{x}^\top \mathbf{L} \mathbf{x}}{2}\right). \tag{5}$$

### C. Privacy Loss and Threat Model

Based on the Gaussian correlation model above, we characterize the privacy loss of each user $i \in \iota$ for the weighted aggregation analysis. Before doing this, we first clarify the threat model. Typically, the strong adversary assumption is widely used for the independent data case. For use $i$, we refer to a strong adversary as one who knows the entire data except for $x_i$. Differential privacy [5] is implicitly designed

as a protection against further information leakage to this adversary. However, in the correlated data case, it has been shown in [17] that a weak adversary who has less background knowledge of the dataset may gain much more information than the strong adversary who knows all data but the target one. Thus, we define each adversary for each user $i$ by exactly specifying the known and unknown data sets. Let $\mathcal{I}^i \subseteq$ be the set of user whose data are known by the adversary. Then, for any user $i \in \iota$, it holds that $\mathcal{I}^i \subseteq \iota \backslash \{i\}$. We denote an adversary as $\mathcal{A}(i, \mathcal{I}^i)$, if he knows the data of users in $\mathcal{I}$ and aims to attack the data of user $i$. For simplicity, we denote all the data in $\mathcal{I}^i$ known by the adversary as $\boldsymbol{x}_{\mathcal{I}^i}$.

Given the threat model above, we adopt the concept of mutual-information differential privacy (MI-DP), proposed in [4], as the measurement for privacy loss. More specifically, we use the generalized version discussed in [4, Section 6.3], which is adapted to protect simultaneously against all adversaries, both strong and weak ones.

**Definition 1** (Mutual Information Differential Privacy (MI-DP) [4])**.** *We say $\epsilon$-mutual-information differential privacy is satisfied for user $i$ if*

$$\sup_{x_i, \mathbf{x}_{\mathcal{I}^i}, i} I(\widehat{y}|\mathbf{x}_{\mathcal{I}^i}; x_i) \leq \epsilon, \tag{6}$$

*where for any random variable $X$ and $Y$, $I(X;Y)$ denotes the mutual information between $X$ and $Y$.*

In contrast to previous works using unconditional mutual information [11], [18], we can see that differential privacy is fundamentally related to conditional mutual information. And as we mentioned before, since differential privacy implies a bound on the mutual information leakage, but not vice-versa [2], [3], our framework can provide more reliable privacy measure for users.

The Proposition 1 below characterizes the variance of the aggregation result in (2) conditional on each user's data.

**Proposition 1.** *The variance of $\widehat{y}$ in* (2) *conditional on user $i$'s data $x_i$ is given by*

$$\mathrm{Var}(\hat{y}|x_i, \mathbf{x}_{\mathcal{I}^i}) = \lambda^2 + \frac{(m-1)^2}{w_i} \tag{7}$$

*Proof (Sketch).* For any given $\mathbf{x}^i_{\mathcal{I}}$, we first calculate the conditional probability density of $\bar{x}_i \coloneqq \left(\sum_{j \in \iota \backslash \mathcal{I}^i \backslash \{i\}} x_i\right) / (n - 1 - |\mathcal{I}^i|)$, i.e., $p(\bar{x}_i|x_i, \mathbf{x}_{\mathcal{I}^i})$ according to (5). With this, the conditional probability density of $y$, i.e., $p(y|x_i, \mathbf{x}_{\mathcal{I}^i})$ can be obtained. Then, considering that the noise $\xi_g$ is independent of $y$, the conditional density of $\widehat{y}$ is just a convolution of $y$ and $\xi_g$. Finally, by an integration, we get the conditional density of $\widehat{y}$, which gives its conditional variance shown in (7). □

Utilizing (7), we now capture the privacy loss for each user $i$, which is based on the concept of MI-DP. Note that $I(\widehat{y}|\mathbf{x}_{\mathcal{I}^i}; x_i) = H(\widehat{y}|\mathbf{x}_{\mathcal{I}^i}) - H(\widehat{y}|\mathbf{x}_{\mathcal{I}}, x_i)$. Here we assume $H(\widehat{y}|\mathbf{x}_{\mathcal{I}^i})$ equals to a constant $\mathcal{C}$ for all $i$, which corresponds that impacts of the adversary for different users are the same to the final aggregation results. And we mainly focus on $H(\widehat{y}|\mathbf{x}_{\mathcal{I}^i}, x_i)$. Since $p(\widehat{y}|x_i)$ has a Gaussian form, we have $H(\widehat{y}|\mathbf{x}_{\mathcal{I}^i}, x_i) = \frac{1}{2}\ln(2\pi e \,\mathrm{Var}(\widehat{y}|\mathbf{x}_{\mathcal{I}^i}, x_i))$. In summary, we define the privacy loss $l_i$ of user $i$ as:

$$l_i = \mathcal{C} - \ln\left(\lambda^2 + \frac{(m-1)^2}{w_i}\right). \tag{8}$$

### D. Utility Functions

*1) Users:* Each user $i$'s utility function $\mathcal{U}_i$ comprises two parts: the reward from the platform and the total loss of privacy. The payment rule has been clearly given in (3). For the total privacy loss for each user $i$, we have to additional take the social relationship among users and their diverge privacy sensitivities into consideration. The total privacy loss for user $i$ is specified as follows:

$$s_i\mathcal{C} - \sum_{j \in \iota} s_{ij} \ln\left(\lambda^2 + \frac{(m-1)^2}{w_j}\right), \tag{9}$$

where $s_{ij}$ is the weight factor that captures the social relationship strength between $i$ and $j$, a larger $s_{ij}$ indicates a closer relationship between user $j$ and $i$, thus $i$ cares more about $j$'s privacy loss. Notably, when $i = j$, $s_{ii}$ reflects user $i$'s privacy sensitivity to his own data, which achieves personalized privacy. Here we use $s_i \coloneqq \sum_{j \in \iota} s_{ij}$ for simplicity. The utility function of user $i$ is formally defined as follows,

$$\mathcal{U}_i(\sigma_i, \sigma_{-i}) = r - \theta_i\sigma_i^2 + \sum_{j=1}^{n} s_{ij} \ln\left(\lambda^2 + \frac{(m-1)^2}{w_j}\right) - s_i\mathcal{C}. \tag{10}$$

*2) Platform:* The gain of platform is determined by two factors. One is the accuracy rate for achieving certain aggregation error $\epsilon$. The other one is whether the price $\mathcal{P} \coloneqq (p_1, p_2, \cdots, p_n)^\top$ he pays for the accuracy rate exceeds the budget $\mathcal{B}$. For any given aggregation error $\epsilon$, the platform wants to maximize the accuracy rate, that is, he wants to find an as large as possible $\alpha$ such that $\mathbb{P}(|y - \widehat{y}| < \epsilon) \geq \alpha$. By using the Bienaymé–Chebyshev inequality [7], we have $Pr(|r - s| < \epsilon) \geq 1 - \frac{\lambda^2}{\epsilon^2}$. Then, for a given error $\epsilon$, the accuracy rate $\alpha$ is upper bounded as follows:

$$\alpha \leq 1 - \frac{\lambda^2}{\epsilon^2}. \tag{11}$$

Next, for the payment, we use $\mathcal{L}_{\mathcal{B}}(\cdot)$ to provide penalty if the budget constraint $\mathcal{B}$ is not respected, i.e.,

$$\mathcal{L}_{\mathcal{B}}(\mathcal{P}) = \begin{cases} 0, & \text{if } \sum_i p_i \leq \mathcal{B}, \\ -\infty, & \text{otherwise.} \end{cases} \tag{12}$$

We define the platform's utility function $\mathcal{U}_p$ as follows,

$$\mathcal{U}_p(\epsilon, \mathcal{B}, \mathcal{P}) = 1 - \frac{\lambda^2}{\epsilon^2} + \mathcal{L}_{\mathcal{B}}(\mathcal{P}). \tag{13}$$

### E. Stackelberg Game Formulation

We model the interaction between the platform and users as a two-stage Stackelberg game. The platform first specifies its reward policy or payment mechanism, and then the users decide their noise-adding strategy.

*Stage I:* The platform gives the payment mechanism $\mathcal{P}^*$:

$$\mathcal{P}^* = \arg\max_{\mathcal{P}} \mathcal{U}_p(\epsilon, \mathcal{B}, \mathcal{P}). \tag{14}$$

*Stage II:* Each use $i$ decides the noise magnitude $\sigma_i$:

$$\sigma_i^* = \arg\max_{\sigma_i} \mathcal{U}_i(\sigma_i, \sigma_{-i}, \mathcal{P}) \tag{15}$$

Through backward induction, We will first analyze users' decisions on given payment mechanism next.Then we maximize the platform's utility base on the users' decisions.

## III. Main Results

### A. Nash-Equilibrium of Users

We first analyze one user's best noise-adding strategy with fixed strategies of others. Then, we derive the relationship between the platform's payment mechanism and users' Nash Equilibrium. For user $i$, given other users' noise magnitude as $\sigma_{-i}$, the optimal strategy $\sigma_i^*$ is given by

$$\sigma_i^* = \arg\max_{\sigma_i} \mathcal{U}_i(\sigma_i, \sigma_{-i}). \tag{16}$$

Take the derivative of $\mathcal{U}_i(\sigma_i, \sigma_{-i})$ with respect to $\sigma_i$, we have,

$$\frac{\partial \mathcal{U}_i}{\partial \sigma_i} = \sum_{j=1}^{n} \frac{2 s_{ij} \sigma_i}{\sum_j \sigma_j^2 + \frac{(m-1)^2}{w_j}} - 2\sigma_i \theta_i. \tag{17}$$

Let $\phi_i$ be the quantity satisfying that

$$\sum_{j=1}^{n} \frac{s_{ij}}{\phi_i + \frac{(m-1)^2}{w_j}} - \theta_i = 0. \tag{18}$$

Note that $\phi_i$ is a constant and can be calculated beforehand. Thus, given $\sigma_{-i}$, the best noise magnitude of user $i$ can be easily calculated as follows

$$\sigma_i^* = \max\left\{\underline{\sigma}, \min\left\{\phi_i - \sum_{j \neq i} \sigma_j, \overline{\sigma}\right\}\right\}. \tag{19}$$

Investigating the optimal noise-adding strategy shown above, we first make the following claim.

**Claim 1.** *For any two users $i$ and $j$ with $\phi_i < \phi_j$, it holds that $\sigma_i \leq \sigma_j$*

With this claim, we provide our main result for the Nash-Equilibrium of users as follows:

**Theorem 1.** *Suppose that $n_1, n_2$ and $n_3$ are three arbitrary non-negative integers such that $n_1 + n_2 + n_3 = n$. The noise-adding strategy $\sigma = (\sigma_1, \sigma_2, \cdots, \sigma_n)$ is a Nash-Equilibrium if and only if there are $n_1$ users with their noise magnitude being $\underline{\sigma}$, $n_2$ users with their noise magnitude falling in the range of $(\underline{\sigma}, \overline{\sigma})$ and their solution of* (18) *being $\sum_{j=1}^{n} \sigma_j$, and $n_3$ users with their noise magnitude being $\overline{\sigma}$.*

*Proof.* As per (19), if there is some user $i$ whose $\sigma_i$ is in $(\underline{\sigma}, \overline{\sigma})$, then $\sigma_i = \phi_i - \sum_{j \neq i} \sigma_j$, which means $\phi_i = \sum_{j=1}^{n} \sigma_j$. This means that, for all users whose noise magnitude is in range of $(\underline{\sigma}, \overline{\sigma})$, their solutions of (18) equal to the same value of $\sum_{j=1}^{n} \sigma_j$. By Claim 1, we know that for any other user $i' \neq i$ that $\sigma_{i'} = \begin{cases} \underline{\sigma}, & \text{if } \phi_{i'} < \phi_i, \\ \overline{\sigma}, & \text{if } \phi_{i'} < \phi_i. \end{cases}$ This concludes the proof. □

From Theorem 1, we can see, the Nash-Equilibrium of users is not unique in general, and $\phi = (\phi_1, \phi_2, \cdots, \phi_n)$ plays a key role in determining the specific Nash-Equilibrium. Define $\overline{\phi} := \max_i \phi_i$ and $\underline{\phi} := \min_i \phi_i$. Next, we specify two special cases, where the Nash-Equilibrium is unique: **Case I:** $\overline{\phi} < n\underline{\sigma}$, **Case II:** $\underline{\phi} > n\overline{\sigma}$.

We make another claim below and then use it to obtain the unique Nash-Equilibrium for each case above.

**Claim 2.** *For user $i$, if $\phi_i \leq n\underline{\sigma}$, then $\sigma_i = \underline{\sigma}$. If $\phi_i \geq n\overline{\sigma}$, then $\sigma_i = \overline{\sigma}$.*

*Proof.* Since for $\forall j$, $\sigma_j \geq \underline{\sigma}$, when $\phi_i \leq n\underline{\sigma}$, we have $\phi_i - \sum_{j \neq i} \sigma_j \leq n\underline{\sigma} - (n-1)\underline{\sigma} = \underline{\sigma}$, which lead to $\sigma_i = \underline{\sigma}$ due to (19). Similarly, because of $\sigma_j \leq \overline{\sigma}$ for $\forall j$, when $\phi_i \geq n\overline{\sigma}$, it holds that $\phi_i - \sum_{j \neq i} \sigma_j \geq n\overline{\sigma} - (n-1)\overline{\sigma} = \overline{\sigma}$, which means $\sigma_i = \overline{\sigma}$ as per (19). □

**Theorem 2.** *For **Case I**, the unique Nash Equilibrium of users' noise magnitude is $\sigma_i = \underline{\sigma}$ for $\forall i$. For **Case II**, the unique Nash Equilibrium of User's noise magnitude is $\sigma_i = \overline{\sigma}$ for $\forall i$.*

### B. Optimal Reward Policy of the Platform

As we mentioned above, the Nash-Equilibrium of users' noise magnitude is determined by $\phi = (\phi_1, \phi_2, \cdots, \phi_n)$, which is affected by $\theta = (\theta_1, \theta_2, \cdots, \theta_n)$, i.e., the payment mechanism of the platform. So the platform needs to elaborately design its payment mechanism $(\theta^*, r^*)$ under a given budget to strike a good balance between users' privacy and the platform's accuracy. Specifically, the accuracy optimization problem can be formulated as

$$\begin{aligned}
\min_{\theta, r} \quad & \lambda^2 = \sum_{i=1}^{n} \sigma_i^2 \\
\text{s.t.} \quad & nr - \sum_{i=1}^{n} \theta_i \sigma_i^2 \leq \mathcal{B} \\
& \mathcal{U}_i(\sigma_i, \sigma_{-i}) \geq 0, && \forall i \in \iota \\
& \sigma_i^* = \max\left\{\underline{\sigma}, \min\left\{\phi_i - \sum_{j \neq i} \sigma_j, \overline{\sigma}\right\}\right\}, && \forall i \in \iota \\
& \sum_{j=1}^{n} \frac{s_{ij}}{\phi_i + \frac{(m-1)^2}{w_j}} - \theta_i = 0, && \forall i \in \iota
\end{aligned}$$

The first two conditions come from the Budget constraint and the rationality of users. And the last two conditions come from the Nash-Equilibrium of users' noise magnitude. By combining the first two conditions, we obtain

$$\mathcal{B} + \sum_{i=1}^{n} \sum_{j=1}^{n} s_{ij} \ln\left(\lambda^2 + \frac{(m-1)^2}{w_j}\right) - \sum_{i=1}^{n} s_i \mathcal{C} \geq 0, \tag{20}$$

which is totally independent of the payment mechanism of the platform, i.e., $\theta$ and $r$. By utilizing this key property, we can solve the optimization problem via a two-step method. In the first step, we first optimize the accuracy under the constraint (20) by finding the optimal noise magnitude of users (i.e.,

TABLE I: Details of Social Network Settings

| Networks | Vertex # | Edges # | $D_{max}$ | $D_{avg}$ | Density |
|---|---|---|---|---|---|
| $S_1$ (Facebook [9]) | 20 | 99 | 14 | 9.9 | 0.521 |
| $S_2$ (Facebook [9]) | 30 | 232 | 22 | 15.45 | 0.533 |
| $S_3$ (Facebook [9]) | 40 | 371 | 25 | 18.55 | 0.476 |
| $S_4$ (Facebook [9]) | 50 | 571 | 27 | 22.48 | 0.466 |

their optimal Nash-Equilibrium). Then, for the optimal Nash-Equilibrium we find, we calculate the desired $\theta$ and $r$ that induce them. In summary, the accuracy optimization problem is converted into the following two sub-problems $P_1$ and $P_2$.

$$\begin{aligned} \min_{\sigma=\{\sigma_1,\sigma_2,\cdots,\sigma_n\}} \quad & \lambda^2 = \sum_{i=1}^{n} \sigma_i^2 \\ \text{s.t.} \quad & \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{s_{ij}\sigma_i^2}{\sum_{j\neq i}\sigma_j^2 + \sigma_i^2 + \frac{(m-1)^2}{w_j}} \geq nr - \mathcal{B} \end{aligned} \tag{$P_1$}$$

$$\begin{cases} \sum_{j=1}^{n} \frac{s_{ij}}{\phi_i + \frac{(m-1)^2}{w_j}} - \theta_i = 0, & \forall i \in \iota \\ \mathcal{U}_i(\sigma_i, \sigma_{-i}) = 0, & \forall i \in \iota \end{cases} \tag{$P_2$}$$

It is easy to see that both $P_1$ and $P_2$ are tractable, as $P_1$ is a convex problem and $P_2$ is a system of equations. To solve $P_2$, we first need to find a set of solutions of (18) based on Theorem 1. Then by solving corresponding equations, we have

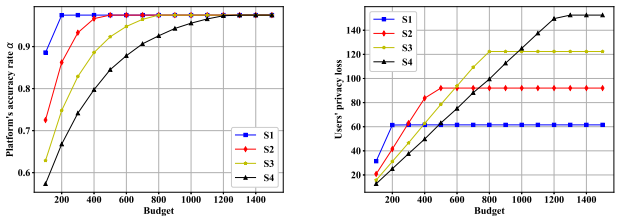
$$\begin{cases} \theta_i = \sum_{j=1}^{n} \frac{s_{ij}}{\phi_i + \frac{(m-1)^2}{w_j}}, \\ r = \theta_i \sigma_i^2 - \sum_{j=1}^{n} s_{ij} \ln\left(\lambda^2 + \frac{(m-1)^2}{w_j}\right) + s_i \mathcal{C}. \end{cases} \tag{21}$$

Note that, in $P_2$, we adopt the same approach as [18] to set each user's final utility to zero so that the platform can fully utilize its budget. This implies that the solution of $\theta^*$ and $r^*$ is not unique. In fact, we can easily generalize to the case where all users have positive utilities. To do this, we can allocate some of the users' utilities and use the adjusted budget to derive the optimal payment mechanism.

## IV. Experiments

In this section, we demonstrate the empirical performance of our framework on a real-world dataset. We evaluate the effectiveness of our framework by analyzing how different factors affect the trade-off between the platform's accuracy and the users' privacy preservation, such as platform's budget, data correlation intensity and social relationship strength (which encompasses the personalized privacy sensitivity).

*1) Experimental Setup:* We construct four social networks, denoted by S1 to S4 respectively, which are all derived from real-world network datasets Facebook [9]. Specifically, for each social network, we extracted denser subgraphs with varying numbers of nodes from the entire dataset. On one hand, different network topology with various vertex degree and graph density reflects the correlation number of users. On the other hand, the correlation weight per edge indicates the correlation intensity between each pair of users. For each undirected edge $(i,j) \in \mathcal{W}, (i<j)$, we generate the correlation weight $w_{ij}$ from a truncated Gaussian distribution with its probability density function $\Psi$ defined as follows,




(a) On platform's accuracy rate (b) On users' privacy loss

Fig. 1: Impact of Platform's Budget

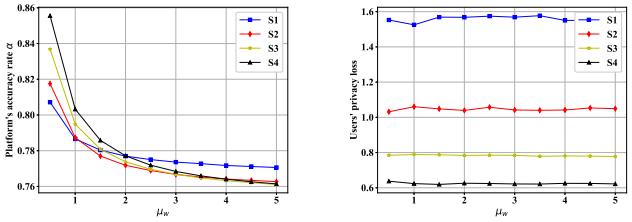
$$\Psi(\mu_w, \sigma_w, 0, +\infty; x) = \begin{cases} 0, & x \leq 0 \\ \frac{\varphi(\mu_w, \sigma_w^2; x)}{1 - \Phi(\mu_w, \sigma_w^2; 0)}, & 0 < x \leq +\infty \end{cases}$$

where $\varphi(\cdot)$ and $\Phi(\cdot)$ are the probability density function and cumulative distribution function of Gaussian distribution, respectively, and the mean of $\Psi$ is proportional to $\mu_w > 0$ given $\sigma_w$, thus $\mu_w$ captures the average intensity of data correlations. For $(i,j) \in \mathcal{W}, (i>j)$, we set $w_{ij} = w_{ji}$ by symmetry. As for users' social relationship strength, for any $i,j \in \mathcal{V}$, $s_{ij}$ is drawn from the truncated normal distribution $\Psi(0, \sigma_s, 0, 1; x)$. Notably, $s_{ij}$ can be different from $s_{ji}$ in general. Throughout our experiments, we set $\sigma_w = 0.02$ $\epsilon^2 = 200$, and $\mathcal{C} = 10$. By default, $\mathcal{B} = 50$, $\mu_w = 2$, $\sigma_s = 0.2$.

*2) Numerical Results:*

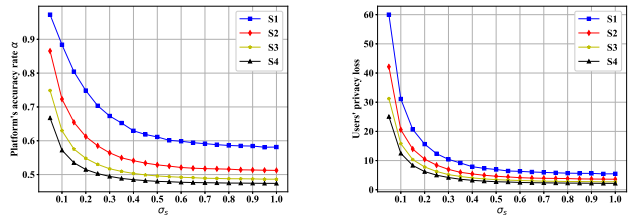
*a) Impact of Platform's Budget:* We explore how the platform's budget $\mathcal{B}$ influences the trade-off between the platform's accuracy and users' privacy loss by varying $\mathcal{B}$ from 100 to 1500 with step 100. We present the platform's accuracy and users' privacy loss for different networks (S1-S4) in Figure 1. It can be observed that, as the platform's budget $\mathcal{B}$ increases, the accuracy rate of the platform also increases for all networks (S1-S4), as shown in Figure 1(a). The accuracy rate reaches a saturation point with different budget thresholds for different network, after which it does not improve any more. On the other hand, users' privacy loss also increases with the increase in budget as depicted in Figure 1(b). Each network (S1-S4) exhibits a different rate of increase in privacy loss, with S4 having the steepest slope. Moreover, each network also has a different saturation point for privacy loss, with S1 reaching it at the lowest budget and S4 at the highest. Therefore, there is a trade-off between the platform's accuracy and users' privacy loss, and the optimal budget depends on the network structure and the desired level of accuracy and privacy.

*b) Impact of Data Correlation Weight:* We examine the effect of the data correlation intensity on both the platform's accuracy and users' privacy loss by varying the average data correlation weight $\mu_w$ from 0.5 to 5 with step 0.5. In this experiment, to make the results more pronounced, we let $\mathcal{B} = 5$, $\mathcal{C} = 6$ and $\epsilon^2 = 40$. As depicted in Figure 2, the platform's accuracy rate decreases as the average data correlation weight ($\mu_w$) increases. All four scenarios (S1 to S4) exhibit a similar trend of declining accuracy, with S4 having the steepest decline. On users' privacy loss, it is observed that all four scenarios remain relatively stable throughout. This




(a) On platform's accuracy rate (b) On users' privacy loss

Fig. 2: Impact of Data Correlation Weight




(a) On platform's accuracy rate (b) On users' privacy loss

Fig. 3: Impact of Data Social Relationship Strength

implies that the data correlation intensity has a negative impact on the platform's accuracy and a limited impact on the users' privacy loss, depending on the network structure and the data distribution.

*c) Impact of Data Social Relationship Strength:* We investigate the influence of users' social relationship strength on the platform's accuracy and users' privacy loss by varying the average data correlation strength $\sigma_s$ from 0.05 to 1 with step 0.05. The results are shown in Figure 3, where two distinct trends are observed. In Figure 3(a), it is evident that as the data correlation strength $\sigma_s$ increases, the platform's accuracy rate decreases for all social relationship strengths (S1 to S4). This indicates a negative correlation between data social relationship strength and platform accuracy. On the other hand, Figure 3(b) illustrates that users' privacy loss decreases with an increase in $\sigma_s$, showing an inverse relationship between these two variables. This implies that the data social relationship strength has a positive impact on the users' privacy and a negative impact on the platform's accuracy, and the optimal trade-off depends on the network structure and the data distribution.

## V. Conclusion

In this paper, we studied the correlation-aware and personalized private data collection problem in IoT services. We proposed a game-theoretic framework that captures the trade-off between user privacy and data utility, taking into account the data correlation, social interactions, and privacy preferences of users. We derived the optimal strategies for both the platform and the users under the Gaussian correlation model and the differential privacy guarantee. We showed that the platform can incentivize users to share their data truthfully by designing a proper reward policy that depends on the correlation structure and the privacy sensitivity of users. We also conducted numerical experiments to validate our theoretical results and to illustrate the performance of our framework in various scenarios. Our work provides a novel and rigorous approach to address the challenges of private data collection in IoT services.

## References


[1] Alessandro Acquisti, Laura Brandimarte, and George Loewenstein. Privacy and human behavior in the age of information. *Science*, 347(6221):509–514, 2015.

[2] Mário S Alvim, Miguel E Andrés, Konstantinos Chatzikokolakis, Pierpaolo Degano, and Catuscia Palamidessi. Differential privacy: on the trade-off between utility and information leakage. In *Formal Aspects of Security and Trust: 8th International Workshop, FAST 2011, Leuven, Belgium, September 12-14, 2011. Revised Selected Papers 8*, pages 39–54. Springer, 2012.

[3] Gilles Barthe and Boris Kopf. Information-theoretic bounds for differentially private mechanisms. In *2011 IEEE 24th Computer Security Foundations Symposium*, pages 191–204. IEEE, 2011.

[4] Paul Cuff and Lanqing Yu. Differential privacy as a mutual information constraint. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 43–54, 2016.

[5] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*, pages 265–284. Springer, 2006.

[6] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

[7] William Feller. An introduction to probability theory and its applications. Technical report, Wiley series in probability and mathematical statistics, 3rd edn.(Wiley, New ..., 1967.

[8] Ross Kindermann and J Laurie Snell. *Markov random fields and their applications*, volume 1. American Mathematical Society, 1980.

[9] Jure Leskovec and Julian Mcauley. Learning to discover social circles in ego networks. *Advances in neural information processing systems*, 25, 2012.

[10] Yibin Li, Wenyun Dai, Zhong Ming, and Meikang Qiu. Privacy protection for preventing data over-collection in smart city. *IEEE Transactions on Computers*, 65(5):1339–1350, 2015.

[11] Guocheng Liao, Xu Chen, and Jianwei Huang. Social-aware privacy-preserving correlated data collection. In *Proceedings of the Eighteenth ACM International Symposium on Mobile Ad Hoc Networking and Computing*, pages 11–20, 2018.

[12] Arvind Narayanan and Vitaly Shmatikov. Myths and fallacies of" personally identifiable information". *Communications of the ACM*, 53(6):24–26, 2010.

[13] Jiangtian Nie, Jun Luo, Zehui Xiong, Dusit Niyato, and Ping Wang. A stackelberg game approach toward socially-aware incentive mechanisms for mobile crowdsensing. *IEEE Transactions on Wireless Communications*, 18(1):724–738, 2018.

[14] Jeffrey Pawlick and Quanyan Zhu. A stackelberg game perspective on the conflict between machine learning and data obfuscation. In *2016 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6. IEEE, 2016.

[15] Jangirala Srinivas, Ashok Kumar Das, Mohammad Wazid, and Athanasios V Vasilakos. Designing secure user authentication protocol for big data collection in iot-based intelligent transportation system. *IEEE Internet of Things Journal*, 8(9):7727–7744, 2020.

[16] Peng Sun, Zhibo Wang, Liantao Wu, Yunhe Feng, Xiaoyi Pang, Hairong Qi, and Zhi Wang. Towards personalized privacy-preserving incentive for truth discovery in mobile crowdsensing systems. *IEEE Transactions on Mobile Computing*, 21(1):352–365, 2020.

[17] Bin Yang, Issei Sato, and Hiroshi Nakagawa. Bayesian differential privacy on correlated data. In *Proceedings of the 2015 ACM SIGMOD international conference on Management of Data*, pages 747–762, 2015.

[18] Guang Yang, Zhiguo Shi, Shibo He, and Junshan Zhang. Socially privacy-preserving data collection for crowdsensing. *IEEE Transactions on Vehicular Technology*, 69(1):851–861, 2019.

[19] Tianqing Zhu, Ping Xiong, Gang Li, and Wanlei Zhou. Correlated differential privacy: Hiding information in non-iid data set. *IEEE Transactions on Information Forensics and Security*, 10(2):229–242, 2014.