Sparse Mobile Crowdsensing With Differential and Distortion Location Privacy

Leye Wang, Daqing Zhang, *Fellow, IEEE*, Dingqi Yang, Brian Y. Lim, Xiao Han, and Xiaojuan Ma

***Abstract*— Sparse Mobile Crowdsensing (MCS) has become a compelling approach to acquire and infer urban-scale sensing data. However, participants risk their location privacy when reporting data with their actual sensing positions. To address this issue, we propose a novel location obfuscation mechanism combining ϵ-*differential-privacy* and δ-*distortion-privacy* in Sparse MCS. More specifically, differential privacy bounds adversaries' relative information gain regardless of their prior knowledge, while distortion privacy ensures that the expected inference error is larger than a threshold under an assumption of adversaries' prior knowledge. To reduce the data quality loss incurred by location obfuscation, we design a differential-and-distortion privacy-preserving framework with three components. First, we learn a *data adjustment* function to fit the original sensing data to the obfuscated location. Second, we apply a linear program to select an *optimal location obfuscation* function. The linear program aims to minimize the uncertainty in data adjustment under the constraints of ϵ-differential-privacy, δ-distortion-privacy, and evenly-distributed obfuscation. We also design an *approximated* method to reduce the required computation resources. Third, we propose an *uncertainty-aware inference* algorithm to improve the inference accuracy for the obfuscated data. Evaluations with real environment and traffic datasets show that our optimal method reduces the data quality loss by up to 42% compared to the state-of-the-art methods with the same level of privacy protection; the approximated method incurs <3% additional quality loss than the optimal method, but only needs <1% of the computation time.**

***Index Terms*— Mobile crowd-sensing, location privacy, differential privacy, distortion privacy, compressive sensing.**

Manuscript received November 13, 2019; accepted February 12, 2020. Date of publication February 24, 2020; date of current version March 9, 2020. This work was supported in part by the NSFC under Grant 61972008, Grant 61572048, and Grant 71601106, in part by the State Language Commission of China Key Program under Grant ZDI135-18, in part by the Hong Kong ITF under Grant ITS/391/15FX, and in part by the ERC Consolidator under Grant 683253 (GraphInt). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Lejla Batina. *(Corresponding author: Daqing Zhang.)*

Leye Wang is with the Key Laboratory of High Confidence Software Technologies (Peking University), Ministry of Education, Beijing 100871, China, and also with the Department of Computer Science and Technology, Peking University, Beijing 100871, China (e-mail: leyewang@pku.edu.cn).

Daqing Zhang is with the Key Laboratory of High Confidence Software Technologies (Peking University), Ministry of Education, Beijing 100871, China, also with the Department of Computer Science and Technology, Peking University, Beijing 100871, China, and also with the Telecom SudParis, IP Paris, 91120 Paris, France (e-mail: dqzhang@sei.pku.edu.cn).

Dingqi Yang is with the eXascale Infolab, University of Fribourg, 1700 Fribourg, Switzerland (e-mail: dingqi@exascale.info).

Brian Y. Lim is with the Department of Computer Science, National University of Singapore, Singapore 119077 (e-mail: brianlim@comp.nus.edu.sg).

Xiao Han is with the School of Information Management and Engineering, Shanghai University of Finance and Economics, Shanghai 200433, China (e-mail: xiaohan@mail.shufe.edu.cn).

Xiaojuan Ma is with the Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Hong Kong (e-mail: mxj@cse.ust.hk).

Digital Object Identifier 10.1109/TIFS.2020.2975925

I. Introduction

With the recent proliferation of sensor-equipped smartphones, Mobile CrowdSensing (MCS) [1]–[3] has become an emerging paradigm to engage mobile users to sense and collect urban-scale environment information, such as noise [4], air quality [5] and traffic conditions [6]. In a typical MCS environment monitoring application, the organizer often needs to specify a target sensing area consisting of a set of locations or regions, and recruit a group of mobile users according to an incentive budget and the mobility pattern of users to perform the sensing and data collection task [7], [8]. However, the target sensing area can sometimes be so large that it might be challenging to get sufficient spatial coverage of mobile users due to budget or time constraints. One solution is to use *Sparse Mobile Crowdsensing* (Sparse MCS) to impute information of the uncovered regions by combining historical records with available sensing data from nearby regions [9], [10].

In Sparse MCS tasks, the participants report not only the sensing data, but also their corresponding location and time. This may introduce serious privacy risks [11]. Therefore, ensuring location privacy is an essential aspect of MCS to attract and retain participants.

Location privacy has been widely studied in location-based systems (LBS). There are two general mechanisms to protect users' location privacy [12]: (i) protecting users' identities through *anonymity*, so that their location traces cannot be linked to specific individuals, and (ii) using location *obfuscation* to alter users' actual locations exposed to the service provider. In *anonymity* mechanisms, to differentiate each individual's contributed data for compensation, it is usually necessary to keep a copy of the mapping between the participants' actual and anonymous identities on the MCS server. Unfortunately, if the server is attacked, the participants' true identities and location information would be at risk. In contrast, most location obfuscation mechanisms can be pre-configured to operate on the mobile clients, and thus no one, except users themselves, can know their actual locations. To avoid the dependence on trustful and secure servers, we thus focus on location *obfuscation*.

Various obfuscation-based protection mechanisms have been studied. One of the most popular mechanisms is *cloaking*. It represents a user's location as a cloaked region containing multiple fine-grained cells instead of a specific place or cell.

Usually, the number of cells in a cloaked region is a measure of its privacy protection level [13]. However, the efficacy of the cloaking mechanisms can be greatly impaired if the adversary has some prior knowledge about the target user's location distribution [14]. For example, if the cloaked region where a user appears consists of a school and a government office and it is known in advance that this user is a student, the adversary may conclude with high confidence that the user would be at the school rather than the government office. This violates the intended protection effect of cloaking.

To address this problem, *differential privacy* [15], [16] has been introduced to ensure that the chance of users being mapped to one specific obfuscated location from any of the actual locations is similar [14]. In conventional LBS, the data loss introduced by applying differential privacy is measured by the distance between the actual and the obfuscated locations. However, in Sparse MCS, the data quality loss is determined by the difference of sensing data between the actual and the obfuscated locations, instead of the geographic distance. In other words, a participant's location may be mapped to a place far away, as long as the sensing values of the two locations are close enough. Therefore, instead of directly using the existing algorithms for LBS [14], [17], we need to redesign the obfuscation mechanisms for Sparse MCS.

However, with differential privacy protection, users are still not clear about how close that an adversary's estimated location will be to their real locations, i.e., *inference error*, which is the *de-facto* location privacy metric nowadays [18]. To mitigate this concern, we introduce *distortion privacy* [19] along with differential privacy into our privacy-preserving mechanism design. Distortion privacy defines the *minimum expected inference error* under the *optimal* location attack that tries to minimize this error (i.e., the expected distance between inferred location and actual one). Note that to apply distortion privacy, we need to specify prior knowledge possessed by adversaries [19], a practical example of which is the location distribution learned from a user's public actions such as check-ins [20]. With this prior knowledge assumption, distortion privacy actually bounds the expected location inference error against any inference attack, as no attack can achieve a smaller inference error than the *optimal* one by definition (under a pre-specified prior knowledge). As distortion privacy itself is not robust to adversaries with arbitrary prior knowledge, it also needs to be used with differential privacy together to provide more comprehensive protection.

The main contributions of this work are:

1) This is the first work to provide both *differential* and *distortion* location privacy protection to Sparse MCS without involving a trustful third-party server.

2) To reduce the data quality loss incurred by the location privacy protection, we propose a privacy-preserving framework including three components. (a) A *data adjustment* function is learned to fit the original sensing data to the obfuscated location. (b) An *optimal* obfuscation function is selected by a linear program, *DU-Min-*$\epsilon\delta$, which aims to minimize the uncertainty in data adjustment under the constraints of ϵ-differential-privacy, δ-distortion-privacy and evenly-distributed obfuscation. A *fast approximated* method, *FDU-Min-*$\epsilon\delta$, is further proposed to reduce the number of constraints from $O(n^3)$ to $O(n^2)$. (c) An *uncertainty-aware* inference algorithm is designed to improve the inference accuracy for the obfuscated data.

3) We evaluate our framework using real-world environment and traffic monitoring datasets. Our results show that compared to the existing differential privacy mechanisms [14], [16], *DU-Min-*$\epsilon\delta$ can reduce data quality loss by up to 42% and also ensure δ-distortion-privacy protection. Compared to *DU-Min-*$\epsilon\delta$, *FDU-Min-*$\epsilon\delta$ increases the quality loss by $<3\%$, but needs $<1\%$ of the computation time.

This paper extends our preliminary conference version [21] by adding the distortion privacy protection in addition to the differential privacy protection. Moreover, in this article, we mathematically analyze the approximation ratio of our proposed fast algorithm. We prove that the approximation ratio is theoretically bounded by a constant value (see Sec. VI-C).

II. Related Work

We review the related work about Sparse MCS, location privacy in LBS, and location privacy in MCS.

A. Sparse Mobile Crowdsensing Applications

In recent years, MCS has become an effective way to engage mobile users in sensing and collecting urban-scale environmental information, such as noise [4], air quality [5] and traffic conditions [6]. Ideally, MCS applications require *all* regions in a spatial area to be measured at *regular* time intervals. However, due to reasons such as a limited incentive budget or a large sensing area, the MCS task organizer might be unable to recruit a sufficient number of mobile participants to cover all the regions within the target sensing area for all the time intervals. Consequently, the missing values in those uncovered regions should be inferred using the historical data records and the sensing values of covered regions [4], [6], [22], [23].

In these Sparse MCS tasks [9], techniques such as compressive sensing [24], multichannel singular spectrum analysis [25] and expectation maximization [26], have been developed to infer the missing data. We note that compressive sensing has been proven to be more effective than other methods [6], [27]. Therefore, in this work, we adopt compressive sensing as the inference algorithm.

B. Location Privacy in Location-Based Services

Location privacy has been widely studied in recent years because of the growing popularity of LBS applications [12]. As discussed earlier, instead of privacy protection through *anonymity*, we limit our scope to location *obfuscation* mechanisms which seek to confuse the adversary with either *inaccurate* or *imprecise* locations [12], [13]. *Inaccuracy* means giving a different location from the actual one, and *imprecision* means giving a plurality of possible locations [12]. Among the obfuscation mechanisms, a popular mechanism is *cloaking* [13], [28], [29]. It employs *imprecision* to process location-based queries relative to a larger cloaked region, compared to the smaller regions or cells, where a user can be uniquely located. However, cloaking is sensitive to the adversary's prior knowledge about the user's location distribution.

First proposed in [30] and later extended in [19], *distortion privacy* adopts the concept of *inaccuracy* to alter or transform the user's actual location to an obfuscated location. Its goal is to ensure the inference error under adversaries' optimal location attack (with an assumption of adversaries' prior knowledge) should be larger than a threshold while minimizing the LBS quality loss (e.g., the distance of the obfuscated and actual locations) [19].

Differential privacy [15], [16] is recently introduced into LBS [14], [17], [19], [31], [32] to provide location privacy protection independent of an adversary's prior knowledge. Like distortion privacy, differential location privacy also employs *inaccuracy* to provide protection. Inspired by these studies, our work leverages the basic idea of differential privacy for location obfuscation, but also considers the sensing data quality. To suit Sparse MCS, we propose a new way to achieve optimal location obfuscation via linear programming, which can reduce the loss in data quality significantly. Moreover, the differential location privacy mechanisms in LBS do not need to deal with the sensing data discrepancy issue, and thus do not perform *data adjustment* or *uncertainty-aware inference*.

C. Location Privacy in Mobile Crowdsensing

According to a recent survey on the MCS privacy issues [33], cloaking is still a widely used startegy in location privacy protection for MCS, e.g., [34]–[37]. Such works all have the same drawback of being sensitive to the adversary's prior knowledge. More recently, [38]–[40] proposed differential location privacy protection frameworks to assign tasks to participants in MCS. Our work has at least three differences from these works: (i) [38]–[40] aim to reduce the distances that assigned participants need to travel, while we try to reduce the data quality loss caused by location obfuscation; (ii) in [38]–[40], the actual sensing locations of the uploaded data are visible to the organizer, whereas our mechanism hides the participants' actual sensing positions from the organizers; (iii) [38]–[40] do not consider distortion privacy.

III. Sparse MCS Concepts

In this section, we introduce Sparse MCS.

A. Use Case: Temperature Monitoring

Suppose an MCS temperature monitoring task in a target urban area divided into fine-grained regions, with the sensing map updating once every hour (sensing cycle). Many candidate participants can be recruited to sense the target regions in Figure 1 (1). In each cycle, the organizer will select some candidates as the actual participants. Then each selected participant uploads the sensed temperature from her smartphone to the server, along with her region information (see Figure 1 (2-Top)). Typically, with a limited incentive budget, the selected participants in each cycle cannot fully cover all the regions, leading to some unsensed regions. Thus, the MCS server infers the temperature values of the unsensed regions by considering both temporal and spatial correlations

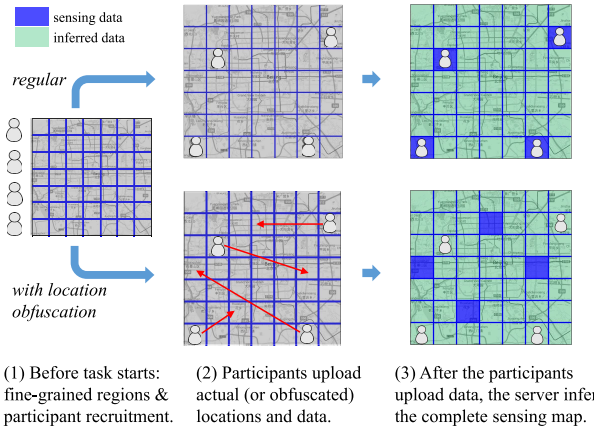


Fig. 1. Regular data reporting for Sparse MCS (Top) and with location privacy protection (Bottom).

(see Figure 1 (3-Top)). This inference is the core method of Sparse MCS which we detail next.

B. Collected Sensing Matrix

The data inference issue is modeled as a matrix completion problem: let *Collected Sensing Matrix* (C) be a matrix to record all the sensing data collected from the participants, such that $C[r, t]$ represents the data of region r in sensing cycle t. If no participant uploads data from region r in cycle t, then $C[r, t]$ is unknown. The key to a successful Sparse MCS task is to determine a high quality, low uncertainty inference algorithm to infer such missing data.

C. Inferring Missing Data in Sparse MCS

There are several methods to infer missing data in Sparse MCS, such as multichannel singular spectrum analysis [25] and expectation maximization [26]. Recently, *compressive sensing* [41] has been proven effective in inferring urban sensing data such as temperature and traffic [6], [27]. Given its improved accuracy over other methods, we use compressive sensing as the inference method in this paper.

As a corollary from compressive sensing theory, Candes and Plan [42] have proven that *one can recover an unknown matrix of low rank, given a small number (compared to the size of the matrix) of noisy entries* ***uniformly sampled****, with* ***an error which is proportional to the noise level****.* This means that, applying compressive sensing to the problem of matrix completion makes two inherent assumptions:

1) Even Data Distribution: To ensure the data inference algorithm performs effectively, uniform distribution of the observed data is required. In Sparse MCS, this means that the sensed regions in the target sensing area should be evenly distributed. If the distribution is biased, e.g., one row of a matrix contains no observation (i.e., one region is not covered by participants in all sensing cycles), then it is impossible to infer the missing data for this row [24], [42].

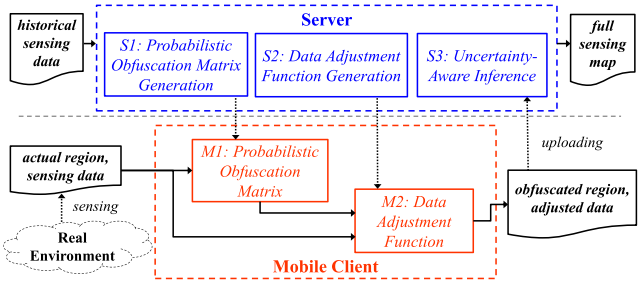


Fig. 2. Location privacy-preserving framework.

2) Small Data Uncertainty: When there is no noise or uncertainty in the sampled entries, the missing values in the matrix can be accurately inferred as long as the previous assumption also holds. When the sampled entries have noise or uncertainty involved, the total inference error is proportional to the uncertainty level of the sampled entries [42]. The smaller the uncertainty of the sampled entries, the better the overall inference performance.

The two assumptions are critical and we will revisit them when designing *quality-optimized* location privacy-preserving mechanisms later.

IV. Location Privacy-Preserving Framework for Sparse MCS

Regular data collection in Sparse MCS needs participants to report their actual regions. Using obfuscation to add location privacy protection can allay participants' privacy concerns, but can lead to data quality loss as the sensing data of the original region may not be representative of the obfuscated region. Therefore, we design a location privacy-preserving framework, which incorporates two unique components: *location obfuscation* and *data adjustment*.

Figure 1 (Bottom) illustrates the privacy-preserving process of Sparse MCS for the temperature monitoring use case. In each sensing cycle, a participant reports her obfuscated region to the server, so that the server never knows her actual region. However, as the sensing data from the actual region is usually not the same as that of the obfuscated region, the data to report also needs to be adjusted before uploading. After receiving all the (adjusted) data from the participants, the server infers the complete sensing map.

Figure 2 shows the overview of our proposed location privacy-preserving framework for Sparse MCS. It consists of two tiers — server side and mobile client side. On the server side, before a Sparse MCS task starts, based on the historical sensing data, it generates the *probabilistic obfuscation matrix* (Step S1) and *data adjustment function* (Step S2) in an offline manner. The probabilistic obfuscation matrix encodes the probabilities of obfuscating any one region to another one. More specifically, the matrix entry $[i, j]$ refers to the probability of mapping the region i to the region j. By carefully selecting the probabilities in the matrix, we can guarantee the participant's location privacy so that the actual region cannot be accurately inferred from the obfuscated region, *even if the adversary knows the obfuscation matrix*. The data adjustment function is used to reduce the uncertainty in the sensing data due to the region obfuscation. It is learned by studying the correlation between any two regions' sensing data in the historical log.

Before task execution, participants pre-download both obfuscation matrix and data adjustment function to their mobile phones. When executing a task, the workflow on the mobile phone client is as follows. First, each mobile phone senses its actual location region. Then, based on the probabilistic obfuscation matrix, it obfuscates the actual region to another region locally with the corresponding probability (Step M1). Afterward, knowing both the actual region and obfuscated region, the data adjustment function adjusts the original sensing data to fit the obfuscated region (Step M2). The mobile phone then uploads the obfuscated region and adjusted data to the server. Later, the server infers the full sensing map from the obfuscated regions collectively, which contains a certain degree of uncertainty compared to the actual data (Step S3).

With this framework in mind, in the following sections, we introduce the notion of differential location privacy for Sparse MCS. In particular, we describe how to design the probabilistic obfuscation matrix, the data adjustment function, and the uncertainty-aware inference algorithm.

V. Differential & Distortion Location Privacy

This section defines differential and distortion location privacy for Sparse MCS.

A. Differential Location Privacy

For Sparse MCS where the sensing area is divided into regions, we show how to apply differential location privacy to provide protection regardless of the adversary's prior knowledge about the participant's location distribution.

1) Bayesian Adversary Model: To illustrate differential privacy, we focus on Bayesian attack [14], [17]. In Sparse MCS, as the target sensing area is usually divided into several regions, thus a Bayesian attack seeks to predict the user's actual region based on the obfuscated region. Specifically, suppose an adversary has some *prior* knowledge about the probabilistic distribution of a user's actual region r, $\pi(r)$; also, the adversary is assumed to know the location obfuscation probability $P(r^*|r)$ for any source region r and target region r^*.[1] Then, if the adversary observes the user's obfuscated region r^*, he can predict a *posterior* distribution of the user's location, noted as $\sigma(r|r^*)$, based on Bayes' rule:

$$\sigma(r|r^*) = \frac{P(r^*|r) \cdot \pi(r)}{\sum_{r' \in \mathcal{R}} P(r^*|r') \cdot \pi(r')} \tag{1}$$

2) Definition of Differential Location Privacy: With the above adversary model, our intention of defining differential privacy in Sparse MCS is to bound the improvement of the adversary's posterior knowledge over the prior knowledge, i.e., $\sigma(r|r^*)/\pi(r)$ for any r^*. Intuitively, if two regions r

[1]The adversary can obtain P in several ways, e.g., hacking the communication between the server and the target user, or spoofing to be a participant to receive P from the server directly.

TABLE I
TWO OBFUSCATION MATRICES SATISFYING ln(2)-DIFFERENTIAL-PRIVACY. THE ENTRY $[i, j]$ REFERS TO THE PROBABILITY OF OBFUSCATING REGION i TO REGION j

region	*1*	*2*	*3*
1	0.50	0.25	0.25
2	0.25	0.50	0.25
3	0.25	0.25	0.50

region	*1*	*2*	*3*
1	0.25	0.25	0.50
2	0.25	0.50	0.25
3	0.50	0.25	0.25

and r' have *similar* probabilities of being obfuscated to r^*, then an adversary, if observing a user in r^*, will be unable to distinguish whether the user is actually in r or r'. With this insight, the differential location privacy attempts to formally define such *similarity* for *any* two source regions r & r', and *any* target region r^*.

Definition 1 (ϵ-Differential-Privacy): Suppose the sensing area consists of a set of regions $\mathcal{R}$, then a probabilistic obfuscation matrix P satisfies ϵ-differential-privacy iff:

$$P(r^*|r) \leq e^{\epsilon} \cdot P(r^*|r'), \quad \forall r, r', r^* \in \mathcal{R} \tag{2}$$

where ϵ indicates the level of privacy, and $P(r^*|r)$ denotes the probability of obfuscating r to r^*.

For an obfuscated region r^*, Eq. 2 bounds the maximum ratio difference between any two entries of $P(r^*|r)$, $\forall r$. The smaller ϵ, the higher privacy. Table I shows two examples of obfuscation matrices for three regions, both satisfying ln(2)-differential-privacy. From this example, we point out that there exist *multiple* probability matrices satisfying ϵ-differential-privacy, given a predefined ϵ.

ϵ-differential-privacy can theoretically limit the relative knowledge gain, $\sigma(r|r^*)/\pi(r) \in [1/e^{\epsilon}, e^{\epsilon}]$ for any r^*, whatever the adversary's prior knowledge $\pi(r)$ is (Theorem 3.2 in [14]).

B. Distortion Location Privacy

While differential privacy bounds adversaries' relative information gain regardless of their prior knowledge, users are still not clear about how close the adversary's estimated location is to their actual locations, i.e., adversaries' *inference error* [18]. To mitigate this concern, we incorporate *δ-distortion-privacy*, i.e., *the inference error of the optimal attack must be larger than δ*, given the public prior leakage of a user's location distribution [19].

1) Optimal Attack Model for Minimizing Error: We use $\sigma^*(\hat{r}|r^*)$ to denote an inference attack, i.e., the posterior probability of a victim's location being inferred as $\hat{r}$ by adversaries when observing the obfuscated location r^*. Suppose that the user's actual location is r, the expected inference error is,

$$\sum_{r^* \in \mathcal{R}} P(r^*|r) \sum_{\hat{r} \in \mathcal{R}} \sigma^*(\hat{r}|r^*) \cdot d(\hat{r}, r) \tag{3}$$

where $d(\hat{r}, r)$ is the distance between $\hat{r}$ and r.

Now, suppose there is a public knowledge on prior leakage of the target user u's location distribution $\eta_u(r)$. This prior leakage may be due to the user's previous actions such as social media check-ins [20].[2] With this public knowledge, an adversary can conduct an *optimal attack* to the target user by minimizing the expected inference error averaging over η_u,

$$\arg\min_{\sigma^*} \sum_{r \in \mathcal{R}} \eta_u(r) \sum_{r^* \in \mathcal{R}} P(r^*|r) \sum_{\hat{r} \in \mathcal{R}} \sigma^*(\hat{r}|r^*) \cdot d(\hat{r}, r) \tag{4}$$

2) Definition of Distortion Location Privacy:

Definition 2 (δ-Distortion-Privacy): A probabilistic obfuscation matrix P satisfies δ-distortion-privacy iff:

$$\sum_{r \in \mathcal{R}} \eta_u(r) \sum_{r^* \in \mathcal{R}} P(r^*|r) \sum_{\hat{r} \in \mathcal{R}} \sigma^*(\hat{r}|r^*) \cdot d(\hat{r}, r) \geq \delta \tag{5}$$

where σ^* represents the adversary's optimal inference attack, δ indicates a user's privacy requirement of the lower bound of the inference error, and η_u is the public prior leakage of the user u's location distribution.

Remark: Public prior leakage of u's location distribution is *not* always equivalent to adversary's prior knowledge. In fact, no mechanism can guarantee distortion privacy for adversaries with an arbitrary prior knowledge [19]. In an extreme case where the adversary foreknows a victim's exact location from some auxiliary data sources, the inference error will always be *zero*. To this end, the distortion privacy definition makes a mild assumption that the adversary only holds the user's public leakage η_u [19].

C. Combing Differential and Distortion Privacy

Here, we clarify the advantage of combining differential and distortion privacy. For differential privacy, the adversary's chance of *exactly picking the user's right region is limited* (not larger than a factor times the prior). For distortion privacy, its definition indicates that *the adversary's predicted regions, even wrong, should be far from the right region* (otherwise the inference error distance could be small). In other words, differential privacy ensures the adversary's low probability of guessing the right region, and distortion privacy makes the regions near the right region also hold low prediction probability. Hence, differential and distortion privacy schemes are from two complementary perspectives, which can be combined to provide more comprehensive protection.

VI. DATA QUALITY OPTIMIZATION UNDER DIFFERENTIAL & DISTORTION PRIVACY

A. Data Quality Requirements for Obfuscation

Recall that in regular Sparse MCS tasks, to infer the complete sensing matrix, compressive sensing theory assumes that (1) the participants report from *evenly* distributed regions, and (2) their reported sensing data are *accurate* [24], [42]. However, introducing differential location privacy may compromise these two requirements:

[2]If a user does not leak any public information, then $\eta(r)$ can be modeled as a uniform distribution or overall population distribution (e.g., estimated by mobile phone call records [8]).

1) Even Obfuscated Region Distribution: While participants' actual location distribution is even, that of the obfuscated regions may be uneven. Consider an extreme case that no region can be obfuscated to region i. Then in the collected sensing matrix, all the values of the ith row will be unknown; recovering the values in this row is hard.

2) Small Data Uncertainty in Obfuscated Regions: The participant's actual sensing data corresponds to the original region, not the obfuscated region. Although the data adjustment step can reduce the discrepancy in the reported data, there is still a baseline uncertainty due to the choice of the obfuscation matrix. We thus want to select an obfuscation matrix can minimize this baseline uncertainty.

B. Optimized Obfuscation Matrix Generation

We seek to reduce the data uncertainty and control the distribution evenness of the obfuscated regions which arise due to location obfuscation. To reduce data uncertainty, we optimally select an obfuscation matrix that can minimize the *expectation of data uncertainty between the reported and true data* in the obfuscated regions. To keep the obfuscated region evenly distributed, we introduce an *evenness constraint* to the obfuscation matrix. Finally, we propose a linear program, *DU-Min-*$\epsilon\delta$, to combine these two aspects, ϵ-differential-privacy, and δ-distortion-privacy, to obtain a quality-optimized obfuscation matrix for Sparse MCS.

1) Objective: Data Uncertainty Minimization: The first step to reduce the data uncertainty is applying a *data adjustment* function to fit the original sensing data to the obfuscated region. As environmental data (e.g., temperature, humidity) usually has high spatial correlations [27], we learn a *linear regression* model for data adjustment, based on the historical sensing data of the original and obfuscated regions. More specifically, suppose the original region is r_1, sensing data is D_1 and the obfuscated region is r_2, the final uploaded data is as follows,

$$D_2^* = A \cdot D_1 + B \tag{6}$$

where A and B are learned based on the historical sensing data of r_1 and r_2 using linear regression. In our framework (Figure 2), the linear regression function is learned on the server (Step S2), while the linear fit estimation is performed on the mobile client (Step M2).[3]

All data adjustment models have intrinsic error or uncertainty, we then define an *uncertainty matrix*, U, to represent this uncertainty, where $U[r, r^*]$ is the data uncertainty incurred by obfuscating region r to r^*. Note that the uncertainty matrix U is intrinsic to the mapping $\langle r \rightarrow r^* \rangle$ and is independent of the obfuscation matrix P. With linear regression as the data adjustment model, the uncertainty $U[r, r^*]$ can be computed by the *residual standard error* [44]. $U[r, r] = 0$ is because there is no uncertainty if the obfuscated and actual regions are same.

[3]The data adjustment function may be enhanced by additional region contexts, e.g., point-of-interests may improve traffic sensing [43]. As not the focus of this paper, we leave this study to the future work.

Intuitively, the smaller uncertainty incurred by the obfuscation, the better quality can be achieved. Then, given a data adjustment function, we aim to find an obfuscation matrix P that can minimize the overall *expectation* of data uncertainty in U, denoted as $\overline{U}$.

$$\overline{U} = \sum\nolimits_{r \in \mathcal{R}} p(r) \cdot \sum\nolimits_{r^* \in \mathcal{R}} U[r, r^*] \cdot P(r^*|r) \tag{7}$$

where $p(r)$ is the probability that a user will appear at the region r ($\sum_{r \in \mathcal{R}} p(r) = 1$). As the MCS server does not know exactly a user's actual location probability distribution, in practice we can set $p(r)$ to a user's public prior leakage of the location distribution $\eta_u(r)$ (Sec. V-B).

Then, the objective of the optimization problem is minimizing Eq. 7, with the following constraints.

Constraint 1: ϵ-Differential-Privacy

The first constraint is *ϵ-differential-privacy*:

$$P(r^*|r) \leq e^{\epsilon} \cdot P(r^*|r'), \quad \forall r, r', r^* \in \mathcal{R} \tag{8}$$

Constraint 2: δ-Distortion-Privacy

The second constraint is *δ-distortion-privacy*. Note that we cannot directly apply Eq. 5 as it self involves a linear program to obtain the adversary's optimal inference attack σ^*. We hence convert Eq. 5 to the following linear constraints (please refer to [19] for more details),

$$\sum_{r \in \mathcal{R}} \eta_u(r) P(r^*|r) d(\hat{r}, r) \geq x(r^*), \quad \forall \hat{r}, r^* \in \mathcal{R} \tag{9}$$

$$\sum_{r^* \in \mathcal{R}} x(r^*) \geq \delta \tag{10}$$

Constraint 3: Even Obfuscated Region Distribution

This constraint concerns data quality. In addition to minimizing the uncertainty, the obfuscated regions need to be evenly distributed to ensure the inferred data quality, i.e.,

$$\psi(r^*) = \sum\nolimits_{r \in \mathcal{R}} \eta_u(r) \cdot P(r^*|r) = 1/|\mathcal{R}| \tag{11}$$

In the evaluation, we will also verify that the evenly-distributed obfuscation outperforms unevenly-distributed obfuscation by getting better data quality.

2) Linear Optimization: DU-Min-$\epsilon\delta$*:* With the objective of reducing data quality loss, by considering data uncertainty, location privacy and obfuscated region distribution, we formulate the linear program, ***D**ata **U**ncertainty-**Min**imization under constraints of ϵ-differential-privacy, δ-distortion-privacy, and evenly-distributed obfuscation* (*DU-Min-*$\epsilon\delta$), to obtain P:

$$\arg\min_{P} \ \bar{\mathbf{U}}(P) = \sum_{r \in \mathcal{R}} \eta_u(r) \sum_{r^* \in \mathcal{R}} U[r, r^*] P(r^*|r) \tag{12}$$

$$\text{s.t.} \ P(r^*|r) \leq e^{\epsilon} \cdot P(r^*|r'), \quad \forall r, r', r^* \in \mathcal{R} \tag{13}$$

$$\sum_{r \in \mathcal{R}} \eta_u(r) P(r^*|r) d(\hat{r}, r) \geq x(r^*), \quad \forall \hat{r}, r^* \in \mathcal{R} \tag{14}$$

$$\sum_{r^* \in \mathcal{R}} x(r^*) \geq \delta \tag{15}$$

$$\sum_{r \in \mathcal{R}} \eta_u(r) \cdot P(r^*|r) = 1/|\mathcal{R}|, \quad \forall r^* \in \mathcal{R} \tag{16}$$

$$P(r^*|r) \geq 0, \quad \forall r, r^* \in \mathcal{R} \tag{17}$$

$$\sum_{r^* \in \mathcal{R}} P(r^*|r) = 1, \quad \forall r \in \mathcal{R} \tag{18}$$

Setting of ϵ: Theoretically, ϵ can be set to any value larger than zero according to users' privacy requirement. But if too large, ϵ will not get a reasonable protection effect. The original paper proposing the location differential privacy [14] recommends setting ϵ for a city-scale location protection to a value around ln(4).

Setting of δ: A larger δ can provide stronger protection, but δ cannot be set to a too large value since it represents the expected location inference error. For example, it cannot exceed the largest distance between any two regions within the target sensing area. We then use the following linear program to get the maximum possible value of δ:

$$\max \delta \quad \text{s.t.} \quad Eq.14 \text{ to } 18 \tag{19}$$

In practice, a valid setting of δ should close to (but smaller than) the one of Eq. 19 to ensure strong protection.

It is also worth noting that ϵ, δ, and η_u are user-dependent. Hence, for each participant, we can learn a personalized obfuscation function P according to her/his own privacy requirements and public prior leakage.

3) Complexity Analysis:

a) Data adjustment functions learning: Before learning the obfuscation matrix, we need to learn the data adjustment functions and the uncertainty matrix from historical data using linear regression. As one linear regression is needed for each pair of regions, the running time is up to $O(|\mathcal{R}|^2)$.

b) Obfuscation matrix learning: The computation complexity of state-of-the-art linear program methods is usually proportional to the number of constraints. For example, the most popular algorithm *simplex* is proved to take expected time polynomial in the number of constraints [45]. Thus, we use the number of constraints in *DU-Min-$\epsilon\delta$* to estimate its computation complexity, which is up to $O(|\mathcal{R}|^3)$.

C. Approximation of Optimal Obfuscation Matrix

As *DU-Min-$\epsilon\delta$* has $O(|\mathcal{R}|^3)$ constraints (Eq. 13), it is hard to be scaled up to a large number of regions in practice. We thus approximate *DU-Min-$\epsilon\delta$* to reduce the number of the constraints to $O(|\mathcal{R}|^2)$. The basic idea is: instead of comparing the probabilities of obfuscating all pairs of regions r, r' to a given region r^* (Eq. 13), we restrict the comparison between only some specific region pairs.

To mark which two regions need to be compared, we define a *region-comparison* graph $\mathcal{G}(\mathcal{R}, \mathcal{E})$ where $\mathcal{E}$ represents all edges in the graph, and each vertex $r \in \mathcal{R}$ represents a region. Two regions r_1, r_2 are required to be compared if the edge $\langle r_1, r_2 \rangle \in \mathcal{E}$. For *DU-Min-$\epsilon\delta$*, $\mathcal{G}$ is a complete graph as every two regions should be compared.

Now, to describe the approximation mechanism, we introduce the *diameter-2-critical* graph [46], whose diameter (the maximum distance between any pair of vertexes) is 2 and the deletion of any edge increases its diameter (examples in Figure 3). Then, the following theorem holds (proof in the supplemental file):

Theorem 1: If $\mathcal{G}(\mathcal{R}, \mathcal{E})$ is a diameter-2-critical graph, and an obfuscation matrix P satisfies:

$$P(r^*|r) \leq e^{\frac{\epsilon}{2}} \cdot P(r^*|r'), \quad \forall \langle r, r' \rangle \in \mathcal{E}, r^* \in \mathcal{R} \tag{20}$$

Then, P satisfies *ϵ-differential-privacy*.

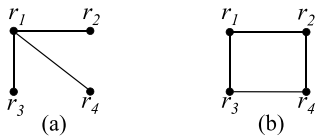


Fig. 3. Two examples of diameter-2-critical graphs.

Proof: As $\mathcal{G}(\mathcal{R}, \mathcal{E})$ is a diameter-2-critical graph, for any two regions $r, r' \in \mathcal{R}$, we can find a region r'' and $\langle r, r'' \rangle, \langle r'', r' \rangle \in \mathcal{E}$, then for any region $r^* \in \mathcal{R}$:

$$\begin{aligned} & P(r^*|r) \leq e^{\frac{\epsilon}{2}} \cdot P(r^*|r'') \\ \Longrightarrow \quad & P(r^*|r) \leq e^{\frac{\epsilon}{2}} \cdot (e^{\frac{\epsilon}{2}} \cdot P(r^*|r')) \\ \Longrightarrow \quad & P(r^*|r) \leq e^{\epsilon} \cdot P(r^*|r') \end{aligned}$$

■

The number of comparisons in Eq. 20 is $O(|\mathcal{E}||\mathcal{R}|)$. To minimize $O(|\mathcal{E}||\mathcal{R}|)$, we aim to find the diameter-2-critical graph with the minimal number of edges. Referring to [47], the minimal diameter-2-critical graph can be constructed: one vertex is joined by an edge with all others. For the minimal diameter-2-critical graph, $O(|\mathcal{E}|) = O(|\mathcal{R}|)$, so the number of comparisons is $O(|\mathcal{R}|^2)$. We then propose the linear program to approximate *DU-Min-$\epsilon\delta$* by replacing Eq. 13 with Eq. 20, called *Fast DU-Min-$\epsilon\delta$* (*FDU-Min-$\epsilon\delta$*). To create the minimal diameter-2-critical graph, any region can be chosen as the "central" vertex that connects to all others. For example, Figure 3a is an example of diameter-2-critical graph constructed for four regions where r_1 is selected as the "central" vertex connected to all the other regions. Our empirical study shows that the selection of "central" vertex has negligible effect on the data quality.

1) Theoretical Performance Analysis: Here, we prove that the approximation ratio of *FDU-Min-$\epsilon\delta$* to *DU-Min-$\epsilon\delta$* is within a constant factor.

Lemma 1: If P satisfies ϵ-differential privacy, then

$$\frac{u_{min} \cdot (|\mathcal{R}| - 1)}{e^{\epsilon} + |\mathcal{R}| - 1} \leq \bar{\mathbf{U}}(P) \leq \frac{u_{max} \cdot (|\mathcal{R}| - 1)}{e^{-\epsilon} + |\mathcal{R}| - 1} \tag{21}$$

where u_{min} is the minimum positive value in the uncertainty matrix U and u_{max} is the maximum value in U.

Proof: Note that in practice, uncertainty matrix U includes non-negative values, and the values in the diagonal, $U[r, r]$, are zero (i.e., obfuscating a region to itself incurs zero uncertainty). Then

$$\bar{\mathbf{U}}(P) \geq \sum_{r \in \mathcal{R}} \eta_u(r) \cdot \sum_{r^* \in \mathcal{R}, r^* \neq r} u_{min} \cdot P(r^*|r) \tag{22}$$

where u_{min} is the minimum positive value in U. In other words, we loosen all the positive values in U (except zero) to the minimum positive value in U. As P satisfies ϵ-differential-privacy, we get,

$$\sum_{r^* \in \mathcal{R}, r^* \neq r} P(r^*|r) \geq \frac{|\mathcal{R}| - 1}{e^{\epsilon} + |\mathcal{R}| - 1} \tag{23}$$

where the right-side value is obtained when we put the highest possible probability to $P(r|r)$ and the lowest to

$P(r^*|r), r^* \neq r$. Then,

$$\bar{\mathbf{U}}(P) \geq \frac{u_{min} \cdot (|\mathcal{R}|-1)}{e^{\epsilon}+|\mathcal{R}|-1} \cdot \sum_{r \in \mathcal{R}} \eta_u(r) = \frac{u_{min} \cdot (|\mathcal{R}|-1)}{e^{\epsilon}+|\mathcal{R}|-1} \quad (24)$$

Then, the lower bound is proven. For the upper bound, the proof is similar by loosening non-zero values in U to its maximum value u_{max}, and then assign the largest possible probability to $P(r^*|r), r^* \neq r$ and the smallest probability to $P(r|r)$. ∎

Theorem 2: Denote the objective values ($\bar{\mathbf{U}}$) obtained by DU-Min-$\epsilon\delta$ and FDU-Min-$\epsilon\delta$ as v^* and $\widetilde{v}$, respectively, then

$$\frac{\widetilde{v}}{v^*} \leq C, \quad C = \frac{u_{max}}{u_{min}} \cdot \frac{e^{\epsilon}+|\mathcal{R}|-1}{e^{-\epsilon}+|\mathcal{R}|-1}, \quad \forall \epsilon > 0 \quad (25)$$

Proof: As DU-Min-$\epsilon\delta$ and FDU-Min-$\epsilon\delta$ both satisfy ϵ-differential-privacy, their obtained objective value ratio will not exceed the ratio of the upper bound over the lower bound in *Lemma 1*. ∎

Theorem 2 tells that the ratio between the objective values obtained by *FDU-Min-*$\epsilon\delta$ and *DU-Min-*$\epsilon\delta$ is bounded within a constant. We also have some observations: *(i) if the maximum and minimum positive values in U are closer to each other, or (ii) if* ϵ *becomes smaller, then the bound is smaller*. In other words, the approximation may work well when (i) the data adjustment uncertainty values have low variance among all the region pairs, and (ii) the required differential privacy protection level is high. Later in the experiment, we will also evaluate the performance between *FDU-Min-*$\epsilon\delta$ and *DU-Min-*$\epsilon\delta$ in practice.

2) FDU-Min vs. Spanner-Based Approximation: In fact our diameter-2-critical graph-based approximation method *FDU-Min-*$\epsilon\delta$ is a special case of the set of spanner-based approximation methods. A spanner (or called spanning graph) $\mathcal{S}$ of a graph $\mathcal{G}$ is a subgraph of $\mathcal{G}$ covering all the vertices in $\mathcal{G}$ and the edge weight is same as $\mathcal{G}$ (number of edges can be reduced). A key property of $\mathcal{S}$ is *dilation*, defined as the maximum ratio of the distance between two vertices in $\mathcal{S}$ compared to their distance in $\mathcal{G}$ [17],

$$\alpha = \max_{r \neq r'} \frac{d_{\mathcal{S}}(r, r')}{d_{\mathcal{G}}(r, r')} \quad (26)$$

We call a spanner with dilation of α as *α-spanner*. We then explain why *FDU-Min-*$\epsilon\delta$ is in fact one *2-spanner-based approximation* of *DU-Min-*$\epsilon\delta$.

As aforementioned, we can see the differential privacy constraints as a region-comparison graph and the original *DU-Min-*$\epsilon\delta$ has a complete graph $\mathcal{G}$ (every two regions to compare); besides, as the comparison formula for any two regions is same in our problem (with a factor of e^{ϵ}), so the edge weights of $\mathcal{G}$ can all be set to one (unweighted graph). Generally, for an α-spanner $\mathcal{S}$, we can have:

Theorem 3: If $\mathcal{S}(\mathcal{R}, \mathcal{E})$ is an α-spanner of the complete graph $\mathcal{G}(\mathcal{R})$, and an obfuscation matrix P satisfies:

$$P(r^*|r) \leq e^{\frac{\epsilon}{\alpha}} \cdot P(r^*|r'), \quad \forall \langle r, r' \rangle \in \mathcal{E}, r^* \in \mathcal{R} \quad (27)$$

Then, P satisfies *ϵ-differential-privacy*.

The proof is similar to Theorem 1. In fact, Theorem 1 is a special case of Theorem 3 when $\alpha = 2$. A good spanner-based approximation method should lead to better data quality and shorter running time. Hence, it may hold the below properties:

- **Small** α: The smaller α is, the more flexible the constraints of Eq. 27 are, which may lead to a better optimization result of data quality.
- **Small** $|\mathcal{E}|$: To accelerate the problem solving speed, the number of the constraints of Eq. 27 should be as small as possible.

Then, we analyze our proposed diameter-2-critical graph-based approximation:

- $\alpha = 2$, *the* ***smallest*** *dilation among all the spanners of* $\mathcal{G}$. Apparently, if we delete any edge from the complete graph $\mathcal{G}$, then the distance of the two nodes connected by the deleted edge will be increased from 1 to 2.
- $|\mathcal{E}| = |\mathcal{R}| - 1$, *the* ***smallest*** *value among all the spanners of* $\mathcal{G}$. As a spanner connects all the vertices in $\mathcal{G}$, so the edge number ($|\mathcal{E}|$) is at least equal to the number of vertices minus one ($|\mathcal{R}| - 1$).

With both α and $|\mathcal{E}|$ equaling the smallest possible values, our diameter-2-critical graph-based approximation *FDU-Min-*$\epsilon\delta$ can thus lead to good approximation performance in both data quality and running speed.

D. Uncertainty-Aware Inference Algorithm

Ordinary compressive sensing inference for matrix completion treats all the collected data instances equally in the learning process [6], [27]. However, as the privacy-preserving data instances inherently have uncertainties, we propose an *uncertainty-aware* inference algorithm by assigning higher weights to the uploaded (adjusted) data with lower uncertainty as an indicator of trust. More specifically, we extend the *stochastic gradient descent* [48] learning process and give different sampling weights to different entries in the collected sensing matrix. The weight assignment is based on the overall uncertainty $\bar{u}(r^*)$ of the obfuscated region r^*, which considers all the possible source regions and their associated obfuscation probabilities.

$$\bar{u}(r^*) = \sum\nolimits_{r \in \mathcal{R}} \eta(r) \cdot P(r^*|r) \cdot U[r, r^*] \quad (28)$$

As higher weights should be assigned to lower-uncertainty regions, we compute the sampling weight $w(r^*)$ as follows:

$$w(r^*) = w_0 + (1 - w_0) \cdot \frac{\bar{u}_{\max} - \bar{u}(r^*)}{\bar{u}_{\max} - \bar{u}_{\min}} \quad (29)$$

where $\bar{u}_{\max}$ and $\bar{u}_{\min}$ are the maximum and minimum overall uncertainties among all the regions, respectively; $w_0 \in [0, 1]$ is the basic sampling weight for the region with the highest uncertainty, which is set to 0.75 according to our empirical study (see later in the evaluation).

Complexity Analysis:

In each iteration of stochastic gradient descent, we select one observation in the collected sensing matrix (according to the weight specified above), and update the matrix factorization results according to [48]. Theoretically, the running time for stochastic gradient descent to converge depends on many parameters, such as the number of observations in the matrix and the pre-set learning rate. In practice, this process can be

rather fast. Through our empirical experiments, we find that if there are N observations in the collected sensing matrix, by setting the learning rate of the stochastic gradient descent to 0.01, this algorithm needs only around $10N$ iterations and takes less than 1 second to converge (detailed running time is shown in the evaluation).

VII. Evaluation

A. Experiment Configurations

1) Baselines: To the best of our knowledge, no existing privacy mechanism in Sparse MCS can guarantee ϵ-differential-privacy and δ-distortion-privacy simultaneously, and thus the baselines are the state-of-the-art ϵ-differential-privacy mechanisms. Under the same level of ϵ-differential-privacy, we will show that our methods can outperform baselines by getting higher data quality, and also offer extra distortion privacy guarantee.

Self [49]: The *Self* obfuscation matrix assigns higher probability to self-obfuscation pairs, i.e., $P(r|r)$. Formally, the *Self* obfuscation matrix satisfying ϵ-differential-privacy is: $P_{self}(r^*|r) \propto e^{\epsilon}$, if $r^* = r$; 1, otherwise.

Laplace (LAP) [14]: Laplace mechanisms are widely used to implement differential privacy [14], [31]. We apply the method in [14] to Sparse MCS. *Laplace* tends to obfuscate a region to its nearby regions with a high probability.

Exponential (EXP) [16]: Exponential mechanism is also widely-used to achieve differential privacy [16], [50]. In the design of Exponential mechanism, a scoring function needs to be modeled so as to obtain good data quality. Given an original region r, a 'better' obfuscation should be assigned a higher score. In Sparse MCS, a higher score is preferred for the obfuscation that leads to lower uncertainty. With this idea, we design the following Exponential baseline,

$$P_{exp}(r^*|r) \propto e^{\frac{\epsilon}{2}\cdot(1-\frac{U[r,r^*]}{\max_{r'\in\mathcal{R}} U[r,r']})} \tag{30}$$

2) Evaluation Scenarios: We evaluate two scenarios.

Environment monitoring: We use the temperature and humidity sensing datasets from *SensorScope* [51], which deployed sensors across the EPFL campus (300m×500m). We divided the area into 100 equal-sized regions (30m×50m), and got 57 regions with sensors (i.e. having ground truth). The sensing data spanned one week, with a sensing cycle of 30 minutes. As this is a static sensor dataset, we use SWIM with the 'Dartmouth' setting [52] to generate the simulated moving traces for 1000 candidate participants.

Traffic monitoring: We use a four-day trajectory dataset generated by ∼30,000 taxis in Beijing [43]. The sensing cycle is set to 1 hour. Road segments are seen as 'regions'. According to [6], we keep a subset of road segments to make the ground truth sensing matrix have as fewer vacancies as possible. We thus choose the top 100-500 road segments which are covered by the taxis most frequently to construct the target sensing area.

For both scenarios, the data of the first day is used for training the data adjustment function and region obfuscation matrix, while the rest days are for the test. Table VII-A.3 summarizes the the statistics of the two evaluation scenarios.

TABLE II
Statistics of Two Evaluation Datasets

	Environment	**Traffic**
City	Lausanne (Switzerland)	Beijing (China)
Data	temperature, relative humidity	traffic speed
Region size	50m*30m	one road segment
Number of regions	57	100-500
Cycle length	0.5 hour	1 hour
Duration	7 days	4 days
Mean±Std.	6.04±1.87 °C (temperature) 84.52±6.32 % (relative humidity)	13.01±6.97 m/s

TABLE III
Evaluation Parameters

	Default	**Description**
k	15	number of participants selected in one cycle
η_u	uniform	public prior leakage of user u's location distribution
ϵ	ln(4)	differential privacy level
δ	0.12 km	distortion privacy level (environment)
	6 km	distortion privacy level (traffic)
$\lvert\mathcal{R}\rvert$	57	number of regions (environment)
	100-500	number of road segments (traffic)

3) Experiment Parameters: We set up the experiment with the differential privacy level, ϵ, distortion privacy level, δ, and the number of participants selected in each cycle, k, as independent variables. ϵ is usually chosen by participants. For simplicity, we assume all the participants require the same ϵ, ranging from ln(2) to ln(8), as in the original paper on location differential privacy [14]. The organizer typically decides on k, based on the incentive budget and the expected data quality. We set δ to a value close to the maximum δ obtained from Eq. 19 so as to provide strong distortion privacy. For the environment and traffic monitoring, we set δ to 0.12 and 6 *km*, respectively (maximum possible values of δ are around 0.128 and 6.1 *km*, respectively). Users' prior leakage of locations (i.e. η_u) is set to be uniform distribution (no explicit leakage). Table III summarizes these parameters.

4) Data Quality Metric: We calculate the *Mean Absolute Error (MAE)* between the server-side data and ground truth data. For each experiment setting of k and ϵ, *MAE* is calculated over 5 repeated trials. Since we focus on the *data quality* ***loss*** due to privacy protection, we define $Loss_{MAE}$ for the ease of presentation:

$$Loss_{MAE}(DU\text{-}Min\text{-}\epsilon\delta) = MAE(DU\text{-}Min\text{-}\epsilon\delta) - MAE(No\text{-}Privacy)$$

We also compute *Root Mean Squared Error* and the results are similar to MAE; for brevity, we only show MAE.

B. Performance Evaluation

1) Computation Resources: We use *CPLEX*[4] to solve *DU-Min-*$\epsilon\delta$ or *FDU-Min-*$\epsilon\delta$ to obtain the obfuscation matrix.

[4]http://www-01.ibm.com/software/commerce/optimization/cplex-optimizer

TABLE IV
COMPUTATION TIME FOR SERVER AND MOBILE SIDES

Stage 1: Server-Side Function Generation	
S1: obfuscation matrix	$17.2s$
S2: linear regression function	$8.3s$
Stage 2: Mobile Client-Side Real-time Running	
M1: region obfuscation	$3.6 \times 10^{-4}s$
M2: data adjustment	$5.7 \times 10^{-6}s$
Stage 3: Sever-Side Data Inference	
S3: uncertainty-aware inference	$0.45s$

TABLE V
COMPUTATION TIME AND MEMORY USAGE OF THE OBFUSCATION MATRIX GENERATION FOR 100-500 ROAD SEGMENTS

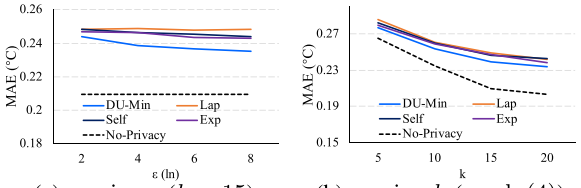
#segments	*DU-Min-$\epsilon\delta$*	*FDU-Min-$\epsilon\delta$*
100	$337s$ / 1000MB	$2s$ / 28MB
150	$1119s$ / 1433MB*	$4s$ / 60MB
200	N/A	$8s$ / 108MB
300	N/A	$22s$ / 236MB
400	N/A	$48s$ / 410MB
500	N/A	$121s$ / 633MB

*turn on the 'memory reduction switch' in CPLEX; otherwise N/A

Recall that our proposed privacy-preserving framework includes three stages: (1) the server-side process to learn both obfuscation matrix and data adjustment function, (2) the mobile client-side process to real-time obfuscate a participant's actual location and adjust her raw sensing data, and (3) the server-side process to infer missing data. More specifically, the first stage is the offline process, only needing to run once before the crowdsensing campaign starts. In comparison, the second and third stages are online computation processes. Table IV shows the computation time of each stage on the temperature dataset with *DU-Min-$\epsilon\delta$*. The first stage is the most computation-intensive, especially for solving the linear program *DU-Min-$\epsilon\delta$* to get the optimal obfuscation matrix. This lasts for ~17 seconds on our test computer (CPU: Intel Core i7-3612QM@2.10GHz, RAM: 8 GB, OS: Windows 7). The first stage is an offline process, so this computation duration is acceptable. For the second and third stages, the running time is quite short that meets the online computation requirements.

To further study the impact of problem scale on the running performance of *DU-Min-$\epsilon\delta$* and *FDU-Min-$\epsilon\delta$*, we show the computation time and memory usage for traffic monitoring in Table V. *DU-Min-$\epsilon\delta$* can only run on 100 or 150 road segments in our test computer. In comparison, *FDU-Min-$\epsilon\delta$* can deal with the scenario up to 500 road segments. Moreover, the computation time of *FDU-Min-$\epsilon\delta$* is less than 1% of the computation time of *DU-Min-$\epsilon\delta$*, if *DU-Min-$\epsilon\delta$* is available. With this running efficiency, *FDU-Min-$\epsilon\delta$* can meet many realistic crowdsensing scenarios. For example, air quality monitoring tasks usually require a region to be 1 $km \times 1$ km [53]. Then, 400 regions can cover the area of most cities (20 $km \times 20km$), while it only needs less than 1 minute computation according to Table V.

2) Data Quality: In general, our results show that *DU-Min-$\epsilon\delta$* can reduce data quality loss by up to 42% compared to baseline mechanisms, and *FDU-Min-$\epsilon\delta$* achieves similar data quality as *DU-Min-$\epsilon\delta$* (<3% additional quality loss). Next we elaborate the two experiment scenarios in detail.



(a) varying ϵ ($k = 15$) (b) varying k ($\epsilon = \ln(4)$)

Fig. 4. MAE in temperature monitoring.

a) Environment monitoring: Figure 4a shows MAE of temperature monitoring under varying differential privacy level ϵ, with a fixed number of participants $k = 15$ and distortion privacy level $\delta = 0.12km$. As expected, *No-Privacy* achieves the best data quality. Among the privacy-preserving mechanisms, *DU-Min-$\epsilon\delta$* incurs the smallest $Loss_{MAE}$ at each privacy level. For example, when ϵ is ln(4), $Loss_{MAE}$(*DU-Min-$\epsilon\delta$*) = 0.030, smaller than $Loss_{MAE}$(*LAP*) = 0.040, $Loss_{MAE}$(*Self*) = 0.037, and $Loss_{MAE}$(*EXP*) = 0.037. Generally, when varying ϵ in [ln(2), ln(8)], *DU-Min-$\epsilon\delta$* can reduce $Loss_{MAE}$ by 11-32%, 11-24% and 7-21% compared to *LAP*, *Self* and *EXP*, respectively. We also see that the MAE of *DU-Min-$\epsilon\delta$* decreases more sharply with increasing ϵ than three baselines. Hence, loosening the differential privacy level leads to more improvements in data quality for *DU-Min-$\epsilon\delta$*.

Figure 4b shows that MAE decreases with more participants. When $\epsilon = \ln(4)$, $\delta = 0.12$, by varying k from 5 to 20, $Loss_{MAE}$ of *DU-Min-$\epsilon\delta$* is always the smallest among all the privacy-preserving mechanisms. Specifically, $Loss_{MAE}$ of *DU-Min-$\epsilon\delta$* is smaller than three baselines by 13-42%. Furthermore, we can see that if the organizer requires a certain data quality level, he will have to trade-off between smaller, more manageable recruitment populations (k) and the participants' privacy level (ϵ). For example, suppose the organizer requires MAE $\leq$ 0.235, this can be achieved with *No-Privacy* by recruiting 10 participants; with *DU-Min-$\epsilon\delta$*, the organizer needs to recruit 20 participants for ln(4)-differential-privacy; loosening to ln(8)-differential-privacy allows a smaller recruitment size of 15.

Figure 5 plots MAE by varying δ of *DU-Min-$\epsilon\delta$*. Note that in the environment monitoring case, the maximum δ that we can achieve is about 0.128 *km* (by Eq. 19). Figure 5 illustrates a general upward trend of MAE with the increase of δ, because a large δ means a strong protection w.r.t. distortion privacy. In fact, even if we set the δ to the maximum possible value 0.128, the obtained MAE will be 0.241, still smaller than baselines LAP, Self, and EXP. This verifies that *DU-Min-$\epsilon\delta$* significantly outperforms baselines in not only getting a better data quality with the same level of differential privacy protection, but also offering the additional protection of distortion privacy.

To verify the effectiveness of our proposed *Uncertainty-Aware inference* (UA) in the Step S3, we compare it with *Ordinary Compressive sensing* (OC), as shown in Figure 6. We experiment with $w_0 = 0.25$, 0.5 and 0.75 for UA.

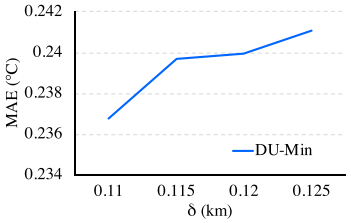


Fig. 5. MAE when varying δ (temperature).

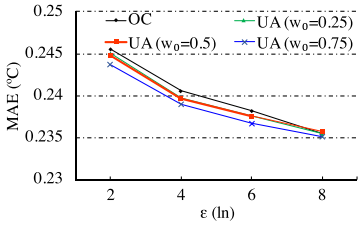


Fig. 6. UA vs. OC inference (temperature).

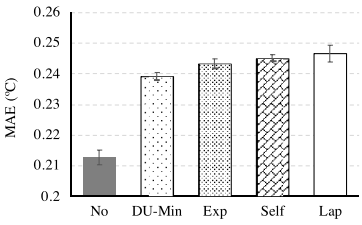


Fig. 7. MAE when users have different leakage η_u (temperature).

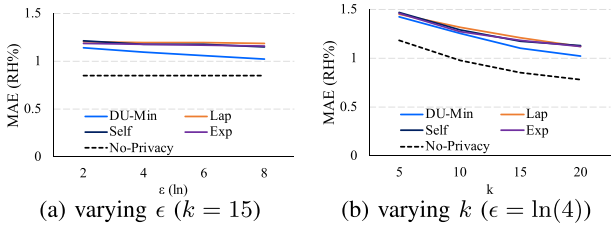


Fig. 8. MAE in relative humidity (RH) monitoring.

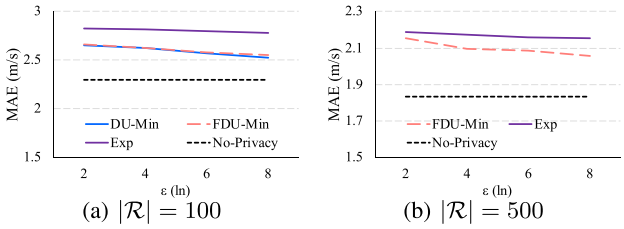


Fig. 9. MAE in traffic monitoring ($k = 0.3|\mathcal{R}|$, varying ϵ).

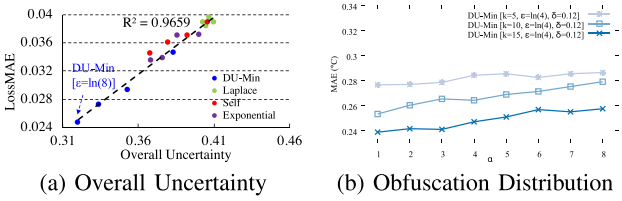


Fig. 10. Relations of MAE and two quality requirements.

Recall that w_0 is the basic sampling weight of UA for the region with the highest uncertainty. We find that 0.75 performs the best. Overall, UA achieves a smaller MAE than OC when ϵ is low. However, for large $\epsilon = \ln(8)$, UA does not improve accuracy. This could be because at higher ϵ (lower privacy), the obfuscation leads to less uncertainty, and thus there is not much data quality loss for UA to recover.

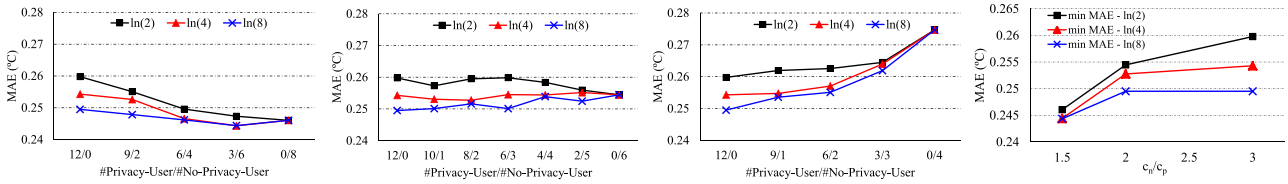
We also conduct an experiment when users leak certain public information on their location distributions η_u (e.g., leaked by public check-ins [54]). More specifically, we assume that a user's 'home' location is leaked and the probability of visiting the 'home' location to be ten times higher than the other locations (i.e., in η_u, the probability of u's 'home' location is 10x lager than the others).[5] Note that different users' η_u are different as their 'home' locations are not same. For the evaluation purpose, we randomly select a region as a participant's 'home' location. We repeat the experiment for five trials and Figure 7 shows the results. *DU-Min-*$\epsilon\delta$ is still consistently better than baselines by achieving a lower MAE with the same privacy protection.

The results on humidity are similar to temperature, as shown in Figure 8. Generally, *DU-Min-*$\epsilon\delta$ outperforms baselines by reducing the data quality loss by 9-32%, with the same level of differential privacy.

b) Traffic monitoring: Figure 9 shows the data quality with 100 and 500 road segments when k is fixed to $0.3|\mathcal{R}|$ with varying ϵ in traffic monitoring, we set δ to 6 *km* (the maximum possible δ is 6.1). For clarity, among all the baselines, we only show the best one, *EXP*, and focus on comparing *DU-Min-*$\epsilon\delta$ to *FDU-Min-*$\epsilon\delta$. Generally, *DU-Min-*$\epsilon\delta$ and *FDU-Min-*$\epsilon\delta$ achieve much better data quality than *EXP*, and *FDU-Min-*$\epsilon\delta$ achieves quite similar data quality as *DU-Min-*$\epsilon\delta$. For example, *FDU-Min-*$\epsilon\delta$ degrades data quality only by <3% compared to *DU-Min-*$\epsilon\delta$ when 30 taxis are randomly selected on 100 road segments with different privacy levels. More specifically, when ϵ is small, the MAEs of *DU-Min-*$\epsilon\delta$ and *FDU-Min-*$\epsilon\delta$ are almost same; with the increase of ϵ, the difference becomes a bit obvious. This is also consistent with Theorem 2: when ϵ is smaller, the approximation is better. When the number of road segments are 500 (Figure 9b), *DU-Min-*$\epsilon\delta$ cannot work due to the large problem scale, but *FDU-Min-*$\epsilon\delta$ still can incur significantly lower data quality loss compared to the baseline.

c) Validation of data quality requirements: Our quality optimization requires minimizing the overall uncertainty while keeping an even obfuscated region distribution. Here we elaborate on the relations between these requirements and data

[5]We also try other settings of the visiting probability of the 'home' location, and the results are similar.



(a) $c_n = 1.5c_p$ (b) $c_n = 2c_p$ (c) $c_n = 3c_p$ (d) min MAE, varying c_n/c_p

Fig. 11. MAE of different budget partition between no-privacy and privacy-concerned participants (temperature). c_n is the cost per no-privacy participant; c_p is the cost per privacy-concerned participant; the budget is $12c_p$ per cycle.

quality in temperature monitoring, to verify their importance in ensuring the data quality.

Figure 10a shows the overall uncertainty ($\overline{U}$, Eq. 7) of all the baselines in comparison with *DU-Min-*$\epsilon\delta$ under different ϵ settings ranging from ln(2) to ln(8) (i.e., experiment settings in Figure 4a), with respect to their $Loss_{MAE}$. We observe a strong correlation between $\overline{U}$ and $Loss_{MAE}$, indicating that $\overline{U}$ of a privacy mechanism can be used to estimate the actual $Loss_{MAE}$ of a Sparse MCS task to a large extent. It suggests that our objective of minimizing the overall uncertainty reflects the core of quality optimization.

Figure 10b shows how the obfuscated region distribution affects the data quality. We replace the original *evenness* constraint (Eq. 16) in *DU-Min-*$\epsilon\delta$ with: $\forall r_1^*, r_2^* \in \mathcal{R}$, $\psi(r_1^*) \leq \alpha \cdot \psi(r_2^*)$, where α is a constant ($\alpha \geq 1$) to tune the evenness. The smaller α is, the more even the distribution is. Note $\alpha = 1$ yields the original evenness constraint (Eq. 16). As shown in the figure, even distribution achieves the best data quality.

C. Budget Trade-Off: Privacy vs. No-Privacy

In reality, users may be willing to sacrifice their location privacy when given enough incentives [55]. Furthermore, less stringent privacy can allow better data quality due to less obfuscation. With this in mind, we examine a practical issue: if the organizer can access both *privacy-concerned* and *no-privacy* participants (the latter with a higher cost), how can the organizer recruit participants to achieve the best data quality within a fixed budget?

Intuitively, this depends on the cost for recruiting a privacy-concerned user (c_p) versus a no-privacy user (c_n), and the privacy level. For clarity, we only change ϵ (differential privacy) to represent users' different privacy requirements and fix δ (distortion privacy) to 0.12 *km*. We calculate the data quality (MAE) under different configurations of budget spending as shown in Figure 11. Evaluating with the 57 regions of the environment monitoring scenario, we consider a total budget of $12c_p$ per cycle; c_n equals to 1.5, 2 and 3 times of c_p, respectively. The x axis m/n means that m privacy-concerned users (uploading obfuscated locations) and n no-privacy users (uploading actual locations) contribute data in each cycle. We can see that when c_n is relatively low ($1.5c_p$), the organizer can recruit more no-privacy users within the budget. This increases the percentage of actual data without greatly reducing the total number of data instances available, and thus achieves higher data quality (Figure 11a). In contrast, when c_n is relatively high ($3c_p$), recruiting more privacy-concerned users is better, as it ensures a sufficient number of data getting submitted (Figure 11c). More importantly, the results reveal an interesting insight — sometimes recruiting a mixed group of privacy-concerned and no-privacy users may achieve the best data quality. For example, when $\epsilon = \ln(4)$ and $c_n = 1.5c_p$, the best data quality is achieved when the budget is spent on three privacy-concerned users and six no-privacy users. Such an observation indicates that organizers need to carefully evaluate the cost and benefit of each type of participants under actual conditions. We also summarize the minimal MAE achieved by each setting of c_n/c_p in Figure 11d. Generally, the minimal MAE decreases with a decreasing c_n, as the organizer can recruit more no-privacy participants with the budget, reducing the uncertainty incurred by obfuscation. Such an improvement is more significant for lower ϵ that provides stronger privacy protection at the cost of higher data uncertainty.

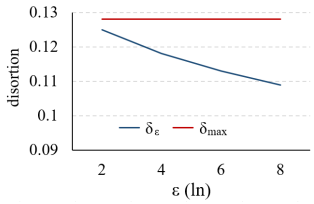


Fig. 12. Distortion range (temperature).

To sum up, recruiting privacy-concerned versus no-privacy participants under a budget constraint reflects the trade-off between the *data quantity* and the *data accuracy*. It requires careful analysis to achieve the optimal data quality (inference accuracy) in actual Sparse MCS applications.

D. Differential Privacy vs. Distortion Privacy

Here, we conduct empirical experiments to investigate the relationship between differential and distortion privacy.

First, we note that the distortion privacy requirement δ cannot be set to an arbitrarily large value, as the target sensing area has a spatial range, which we have described in Eq. 19 (denoted as δ_{max}). Also, for any obfuscation matrix P, it implicitly indicates a certain δ that we can calculate with P by Eq. 9–10. If we ignore the distortion privacy constraint

when solving (F)DU-Min-$\epsilon\delta$ (i.e., set $\delta = 0$), we can obtain an optimal ϵ-differential-privacy obfuscation matrix P_ϵ. With P_ϵ, we can calculate its satisfied distortion privacy δ_ϵ. Then, in (F)DU-Min-$\epsilon\delta$, if we set $\delta < \delta_\epsilon$, we will still get P_ϵ. That is, only when we set $\delta \in [\delta_\epsilon, \delta_{max}]$, the distortion privacy constraint can really take effect. Figure 12 shows the range of $[\delta_\epsilon, \delta_{max}]$ for different ϵ in the temperature case. In particular, for $\epsilon = \ln(8)$, the tuning range of δ is $[0.109, 0.128]$. This means that for a certain level of differential privacy, we can still enhance it with stronger distortion privacy (set $\delta > \delta_\epsilon$) by *DU-Min-*$\epsilon\delta$.

VIII. Discussion and Future Work

A. Repeated-Observations Trajectory Attack

In the future, we aim to extend this work from the *snapshot* attack to the *trajectory* attack, where an adversary observes a user's obfuscated regions over multiple or repeated sensing cycles, $\langle r_1^*, r_2^*, \ldots, r_n^* \rangle$. Specifically, if P satisfies ϵ-differential-privacy under the snapshot case, then it just satisfies $n\epsilon$-differential-privacy under the trajectory case. Future research can be done to address this limitation.

B. Auxiliary Knowledge Beyond Location Distribution

If an adversary has auxiliary knowledge beyond the location distribution (which cannot be written in the form of $\pi(r)$), then the differential privacy protection mechanism may fail. For example, if the adversary foreknows that one participant is "in the hottest region" and he gets the temperature sensing map from the MCS task, then he may infer the participant's actual location. Some sensing data perturbation may also be needed for this attack.

C. Need for Historical Sensing Data

Our mechanism requires some initial ground truth historical sensing data to generate the data adjustment functions and the optimal location obfuscation matrix. To get this initial data, before the privacy-preserving Sparse MCS task begins, the MCS organizer needs to employ a group of workers to cover the target sensing area and collect accurate sensing data. After the data adjustment functions and obfuscation matrix are learned, we can stop the initialization stage and start privacy-preserving Sparse MCS. We will study whether we can preserve workers' privacy in the initialization stage.

IX. Conclusion

In this paper we present a differential-and-distortion location privacy framework for Sparse MCS. It takes into account the desired level of privacy protection, the prior knowledge about participants' location distribution, and the data quality loss due to location obfuscation. Particularly, our framework can provide a guaranteed level of differential and distortion privacy with reduced data quality loss in Sparse MCS applications. Experiments on real data validate the effectiveness of our framework.

References

[1] R. Ganti, F. Ye, and H. Lei, "Mobile crowdsensing: Current state and future challenges," *IEEE Commun. Mag.*, vol. 49, no. 11, pp. 32–39, Nov. 2011.

[2] D. Zhang, L. Wang, H. Xiong, and B. Guo, "4W1H in mobile crowd sensing," *IEEE Commun. Mag.*, vol. 52, no. 8, pp. 42–48, Aug. 2014.

[3] B. Guo *et al.*, "Mobile crowd sensing and computing: The review of an emerging human-powered sensing paradigm," *ACM Comput. Surv.*, vol. 48, no. 1, pp. 1–31, Aug. 2015.

[4] R. K. Rana, C. T. Chou, S. S. Kanhere, N. Bulusu, and W. Hu, "Ear-phone: An end-to-end participatory urban noise mapping system," in *Proc. 9th ACM/IEEE Int. Conf. Inf. Process. Sensor Netw. (IPSN)*, Apr. 2010, pp. 105–116.

[5] D. Hasenfratz, O. Saukh, S. Sturzenegger, and L. Thiele, "Participatory air pollution monitoring using smartphones," in *Proc. Int. Workshop Mobile Sens. (IPSN)*, Apr. 2012, pp. 1–5.

[6] Y. Zhu, Z. Li, H. Zhu, M. Li, and Q. Zhang, "A compressive sensing approach to urban traffic estimation with probe vehicles," *IEEE Trans. Mobile Comput.*, vol. 12, no. 11, pp. 2289–2302, Nov. 2013.

[7] H. Xiong, D. Zhang, L. Wang, and H. Chaouchi, "EMC3: Energy-efficient data transfer in mobile crowdsensing under full coverage constraint," *IEEE Trans. Mobile Comput.*, vol. 14, no. 7, pp. 1355–1368, Jul. 2015.

[8] H. Xiong, D. Zhang, G. Chen, L. Wang, V. Gauthier, and L. E. Barnes, "ICrowd: Near-optimal task allocation for piggyback crowdsensing," *IEEE Trans. Mobile Comput.*, vol. 15, no. 8, pp. 2010–2022, Aug. 2016.

[9] L. Wang, D. Zhang, Y. Wang, C. Chen, X. Han, and A. M'hamed, "Sparse mobile crowdsensing: Challenges and opportunities," *IEEE Commun. Mag.*, vol. 54, no. 7, pp. 161–167, Jul. 2016.

[10] L. Wang *et al.*, "SPACE-TA: Cost-effective task allocation exploiting intradata and interdata correlations in sparse crowdsensing," *ACM Trans. Intell. Syst. Technol.*, vol. 9, no. 2, pp. 1–28, Oct. 2017.

[11] J. E. Dobson and P. F. Fisher, "Geoslavery," *IEEE Technol. Soc. Mag.*, vol. 22, no. 1, pp. 47–52, 2003.

[12] J. Krumm, "A survey of computational location privacy," *Pers. Ubiquitous Comput.*, vol. 13, no. 6, pp. 391–399, Oct. 2008.

[13] M. Duckham and L. Kulik, "A formal model of obfuscation and negotiation for location privacy," in *Pervasive Computing*. Berlin, Germany: Springer, 2005, pp. 152–170.

[14] M. E. Andrés, N. E. Bordenabe, K. Chatzikokolakis, and C. Palamidessi, "Geo-indistinguishability: Differential privacy for location-based systems," in *Proc. CCS*, 2013, pp. 901–914.

[15] C. Dwork, "Differential privacy," in *Proc. ICALP*, 2006, pp. 1–12.

[16] F. McSherry and K. Talwar, "Mechanism design via differential privacy," in *Proc. 48th Annu. IEEE Symp. Found. Comput. Sci. (FOCS)*, Oct. 2007, pp. 94–103.

[17] N. E. Bordenabe, K. Chatzikokolakis, and C. Palamidessi, "Optimal geo-indistinguishable mechanisms for location privacy," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur. (CCS)*, 2014, pp. 251–262.

[18] R. Shokri, G. Theodorakopoulos, J.-Y. Le Boudec, and J.-P. Hubaux, "Quantifying location privacy," in *Proc. IEEE Symp. Secur. Privacy*, May 2011, pp. 247–262.

[19] R. Shokri, "Privacy games: Optimal user-centric data obfuscation," *Proc. Privacy Enhancing Technol.*, vol. 2015, no. 2, pp. 299–315, Jun. 2015.

[20] D. Yang, D. Zhang, Z. Yu, and Z. Yu, "Fine-grained preference-aware location search leveraging crowdsourced digital footprints from LBSNs," in *Proc. ACM Int. Joint Conf. Pervas. Ubiquitous Comput. (UbiComp)*, 2013, pp. 479–488.

[21] L. Wang, D. Zhang, D. Yang, B. Y. Lim, and X. Ma, "Differential location privacy for sparse mobile crowdsensing," in *Proc. IEEE 16th Int. Conf. Data Mining (ICDM)*, Dec. 2016, pp. 1257–1262.

[22] L. Xu, X. Hao, N. D. Lane, X. Liu, and T. Moscibroda, "More with less: Lowering user burden in mobile crowdsourcing through compressive sensing," in *Proc. UbiComp*, Sep. 2015, pp. 659–670.

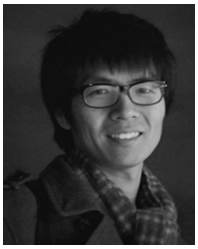
[23] L. Wang *et al.*, "CCS-TA: Quality-guaranteed online task allocation in compressive crowdsensing," in *Proc. UbiComp*, 2015, pp. 683–694.

[24] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Found. Comput. Math.*, vol. 9, no. 6, pp. 717–772, 2009.

[25] D. Kondrashov and M. Ghil, "Spatio-temporal filling of missing points in geophysical data sets," *Nonlinear Processes Geophys.*, vol. 13, no. 2, pp. 151–159, May 2006.

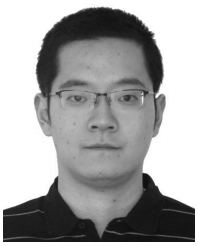
[26] T. Schneider, "Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values," *J. Climate*, vol. 14, no. 5, pp. 853–871, Mar. 2001.

[27] L. Kong *et al.*, "Data loss and reconstruction in wireless sensor networks," *IEEE Trans. Parallel Distrib. Syst.*, vol. 25, no. 11, pp. 2818–2828, Nov. 2014.
[28] B. Bamba, L. Liu, P. Pesti, and T. Wang, "Supporting anonymous location queries in mobile environments with privacygrid," in *Proc. 17th Int. Conf. World Wide Web (WWW)*, 2008, pp. 237–246.
[29] M. F. Mokbel, C.-Y. Chow, and W. G. Aref, "The new casper: Query processing for location services without compromising privacy," in *Proc. VLDB*, Sep. 2006, pp. 763–774.
[30] R. Shokri, G. Theodorakopoulos, C. Troncoso, J.-P. Hubaux, and J.-Y. Le Boudec, "Protecting location privacy: Optimal strategy against localization attacks," in *Proc. ACM Conf. Comput. Commun. Secur. (CCS)*, 2012, pp. 617–627.
[31] R. Dewri, "Local differential perturbations: Location privacy under approximate knowledge attackers," *IEEE Trans. Mobile Comput.*, vol. 12, no. 12, pp. 2360–2372, Dec. 2013.
[32] L. Wang, G. Qin, D. Yang, X. Han, and X. Ma, "Geographic differential privacy for mobile crowd coverage maximization," in *Proc. AAAI*, Apr. 2018, pp,. 1–8.
[33] L. Pournajaf, D. A. Garcia-Ulloa, L. Xiong, and V. Sunderam, "Participant privacy in mobile crowd sensing task management: A survey of methods and challenges," *ACM SIGMOD Rec.*, vol. 44, no. 4, pp. 23–34, May 2016.
[34] C. Cornelius, A. Kapadia, D. Kotz, D. Peebles, M. Shin, and N. Triandopoulos, "Anonysense: Privacy-aware people-centric sensing," in *Proc. 6th Int. Conf. Mobile Syst., Appl., Services (MobiSys)*, 2008, pp. 211–224.
[35] I. Krontiris and T. Dimitriou, "Privacy-respecting discovery of data providers in crowd-sensing applications," in *Proc. IEEE Int. Conf. Distrib. Comput. Sensor Syst.*, May 2013, pp. 249–257.
[36] L. Pournajaf, L. Xiong, V. Sunderam, and S. Goryczka, "Spatial task assignment for crowd sensing with cloaked locations," in *Proc. IEEE 15th Int. Conf. Mobile Data Manage.*, Jul. 2014, pp. 73–82.
[37] I. J. Vergara-Laurens, D. Mendez, and M. A. Labrador, "Privacy, quality of information, and energy consumption in participatory sensing systems," in *Proc. IEEE Int. Conf. Pervas. Comput. Commun. (PerCom)*, Mar. 2014, pp. 199–207.
[38] H. To, G. Ghinita, and C. Shahabi, "A framework for protecting worker location privacy in spatial crowdsourcing," *Proc. VLDB Endowment*, vol. 7, no. 10, pp. 919–930, Jun. 2014.
[39] H. To, C. Shahabi, and L. Xiong, "Privacy-preserving online task assignment in spatial crowdsourcing with untrusted server," in *Proc. IEEE 34th Int. Conf. Data Eng. (ICDE)*, Apr. 2018, pp. 833–844.
[40] L. Wang, D. Yang, X. Han, T. Wang, D. Zhang, and X. Ma, "Location privacy-preserving task allocation for mobile crowdsensing with differential geo-obfuscation," in *Proc. 26th Int. Conf. World Wide Web (WWW)*, 2017, pp. 627–636.
[41] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.
[42] E. J. Candes and Y. Plan, "Matrix completion with noise," *Proc. IEEE*, vol. 98, no. 6, pp. 925–936, Jun. 2010.
[43] J. Shang, Y. Zheng, W. Tong, E. Chang, and Y. Yu, "Inferring gas consumption and pollution emission of vehicles throughout a city," in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2014, pp. 1027–1036.
[44] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*. New York, NY, USA: Springer, 2013.
[45] D. A. Spielman and S.-H. Teng, "Smoothed analysis of algorithms: Why the simplex algorithm usually takes polynomial time," *J. ACM*, vol. 51, no. 3, pp. 385–463, 2004.
[46] L. Caccetta and R. Häggkvist, "On diameter critical graphs," *Discrete Math.*, vol. 28, no. 3, pp. 223–229, 1979.
[47] P. Erdős, A. Rényi, and V. Sós, "On a problem of graph theory," *Studia Sci. Math. Hungar*, vol. 1, pp. 215–235, Dec. 1966.
[48] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, no. 8, pp. 30–37, Aug. 2009.
[49] S. Agrawal and J. R. Haritsa, "A framework for high-accuracy privacy-preserving mining," in *Proc. ICDE*, 2005, pp. 193–204.
[50] W. Eltarjaman, R. Dewri, and R. Thurimella, "Location privacy for rank-based geo-query systems," *Proc. Privacy Enhancing Technol.*, vol. 2017, no. 4, pp. 77–96, Oct. 2017.
[51] F. Ingelrest, G. Barrenetxea, G. Schaefer, M. Vetterli, O. Couach, and M. Parlange, "SensorScope: Application-specific sensor network for environmental monitoring," *ACM Trans. Sensor Netw.*, vol. 6, no. 2, pp. 1–32, Feb. 2010.
[52] S. Kosta, A. Mei, and J. Stefa, "Large-scale synthetic social mobile networks with SWIM," *IEEE Trans. Mobile Comput.*, vol. 13, no. 1, pp. 116–129, Jan. 2014.
[53] Y. Zheng, F. Liu, and H.-P. Hsieh, "U-air: When urban air quality inference meets big data," in *Proc. KDD*, 2013, pp. 1436–1444.
[54] D. Yang, D. Zhang, and B. Qu, "Participatory cultural mapping based on collective behavior data in location-based social networks," *ACM Trans. Intell. Syst. Technol.*, vol. 7, no. 3, pp. 1–23, Jan. 2016.
[55] G. Danezis, S. Lewis, and R. J. Anderson, "How much is location privacy worth?" in *Proc. WEIS*, vol. 5, Jun. 2005, pp. 1–13.



Leye Wang received the Ph.D. degree in computer science from Pierre and Marie Curie University and Telecom SudParis, France, in 2016. He was a Postdoctoral Researcher with The Hong Kong University of Science and Technology. He is currently an Assistant Professor with the Key Laboratory of High Confidence Software Technologies and the Department of Computer Science and Technology, EECS, Peking University, China. His research interests include ubiquitous computing, mobile crowdsensing, and urban computing.

Daqing Zhang (Fellow, IEEE) received the Ph.D. degree from the University of Rome "La Sapienza" in 1996. He is currently a Professor with the CS Department, School of EECS, Peking University, China, and Telecom SudParis, IP Paris, France. He has published more than 260 technical papers in leading conferences and journals. His research interests include context-aware computing, urban computing, mobile computing, big data analytics, and pervasive elderly care. He received the Ten Years CoMoRea Impact Paper Award at IEEE PerCom 2013 and IEEE UIC 2019, the Honorable Mention Award at ACM UbiComp 2015 and 2016, and the Best Paper Award at IEEE UIC 2015 and 2012. He served as the general or program chair for more than ten international conferences, giving keynote talks at more than 20 international conferences. He is the Associate Editor for *ACM Transactions on Intelligent Systems and Technology* and the IEEE PERVASIVE COMPUTING.



Dingqi Yang received the Ph.D. degree in computer science from Pierre and Marie Curie University and the Institut Mines-TELECOM/TELECOM SudParis, where he received both the CNRS SAMOVAR Doctorate Award and the Institut Mines-TELECOM Press Mention in 2015. He is currently a Senior Researcher with the University of Fribourg, Switzerland. His research interests include big social media data analytics, ubiquitous computing, and smart city applications.

Brian Y. Lim received the B.S. degree in engineering physics from Cornell University, Ithaca, NY, USA, in 2006, and the Ph.D. degree in human–computer interaction from Carnegie Mellon University, Pittsburgh, USA, in 2012. He is currently an Assistant Professor with the Department of Computer Science, National University of Singapore (NUS). His research interests include explainable artificial intelligence, ubiquitous computing, human–computer interaction, and applications for urban data analytics and smart healthcare.

Xiao Han received the Ph.D. degree in computer science from Pierre and Marie Curie University and the Institut Mines-TELECOM/TELECOM SudParis in 2015. She is currently an Assistant Professor with the Shanghai University of Finance and Economics, China. Her research interests include social network analysis, fintech, and privacy protection.

Xiaojuan Ma received the Ph.D. degree in computer science from Princeton University. She was a Researcher of human–computer interaction with Noah's Ark Lab, Huawei Tech. Investment Co., Ltd., Hong Kong. She was a Post-Doctoral Researcher with the Human-Computer Interaction Institute (HCII), Carnegie Mellon University (CMU), and before that a Research Fellow with the Information Systems Department, National University of Singapore (NUS). She is currently an Assistant Professor of human–computer interaction (HCI) with the Department of Computer Science and Engineering (CSE), The Hong Kong University of Science and Technology (HKUST). Her background is in human–computer interaction. She is particularly interested in data-driven human-engaged computing in the domain of ubiquitous, social, and crowd computing and human–robot interaction. She was a recipient of Computing Innovation Fellows by Computing Research Association in 2010, and named Outstanding Chinese Young Leaders in HCI by the International Chinese Association of Computer Human Interaction in 2016.