Differential Location Privacy for Sparse Mobile Crowdsensing

Leye Wang*†, Daqing Zhang‡†, Dingqi Yang§, Brian Y. Lim¶ and Xiaojuan Ma*
*Hong Kong University of Science and Technology, China;†SAMOVAR, Institut Mines Telecom/Telecom SudParis, France;
‡Peking University, China; §University of Fribourg, Switzerland; ¶National University of Singapore, Singapore
wangleye@gmail.com, dqzhang@sei.pku.edu.cn, dingqi@exascale.info, limyl@comp.nus.edu.sg, mxj@cse.ust.hk

***Abstract*—Sparse Mobile Crowdsensing (MCS) has become a compelling approach to acquire and make inference on urban-scale sensing data. However, participants risk their location privacy when reporting data with their actual sensing positions. To address this issue, we adopt *ϵ-differential-privacy* in Sparse MCS to provide a theoretical guarantee for participants' location privacy regardless of an adversary's prior knowledge. Furthermore, to reduce the data quality loss caused by differential location obfuscation, we propose a 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, which aims to minimize the uncertainty in data adjustment. We also propose a fast *approximated* variant. Third, we propose an *uncertainty-aware inference* algorithm to improve the inference accuracy of 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 existing differential privacy methods.**

I. Introduction

Mobile Crowdsensing (MCS) [1] is an emerging paradigm that leverages the recent surge of sensor-equipped smartphones to collect urban-scale information, such as noise [2] and traffic [3]. 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* to impute information of the uncovered regions by combining historical records with available sensing data from nearby regions [4]. In Sparse MCS, participants report the sensing data with time stamps and geographical coordinates, which may introduce serious privacy risks. Therefore, ensuring location privacy is essential to attract participants.

A large body of work on location-based systems (LBS) studies location privacy, and has proposed two general protective mechanisms [5]: (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. This paper focuses on *obfuscation*.

One of the most popular obfuscation mechanisms is *cloaking* [5], [6]. It represents a user's location as a cloaked region containing multiple fine-grained cells instead of a specific place or cell. However, the effectiveness of cloaking can be greatly impaired if the adversary has prior knowledge about the target user's location distribution [7]. For example, if the cloaked region where a user appears consists of a school and a government office and it is known that the user is a student, the adversary can conclude rather confidently that the user would be at the school.

To address this problem, *differential privacy* [8], [9] 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 [7]. The more similar the probability for each region is, the harder it is to infer users' original positions, leading to better privacy protection.

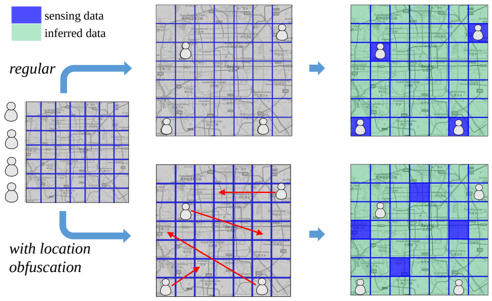
In conventional LBS, the data loss introduced by applying differential privacy is measured by the distance between the actual and the obfuscated locations. However, in Sparse MCS, the data quality loss is determined by the difference of sensing data between the actual and the obfuscated locations, instead 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 [7], [10], we need to redesign the obfuscation mechanisms for Sparse MCS applications.

In this paper, we explore how to balance three key elements in the location privacy-preserving mechanism for Space MCS applications: the *participant's privacy requirements*, the *adversary's prior knowledge about the participant's actual location distribution*, and the *data quality degradation* stemming from the location obfuscation. The main contributions of this work are:

1) To the best of our knowledge, this is the first work to apply differential location privacy to Sparse MCS while reducing the data quality loss.

2) A quality-assured privacy-preserving framework, which consists of three components: (i) a *data adjustment* function to fit the original sensing data to the obfuscated location; (ii) an *optimal* obfuscation function, *DUM-ϵe*, and its fast approximation, *FDUM-ϵe*, to minimize the uncertainty in data adjustment under the constraints of ϵ-differential-privacy and evenly-distributed obfuscation; and (iii) an *uncertainty-aware* inference algorithm to improve the inference accuracy for the obfuscated data.

3) Empirical evaluations with real temperature and traffic monitoring datasets, which validates that compared to the existing differential privacy mechanisms, our framework



(1) Before task starts: fine-grained regions & participant recruitment.

(2) Participants upload actual (or obfuscated) locations and data.

(3) After the participants upload data, the server infers the complete sensing map.

Figure 1: Regular data reporting for Sparse MCS (Top) and with location privacy protection using obfuscation (Bottom).

with *DUM-$\epsilon\epsilon$* can reduce the data quality loss by up to 42%; compared to *DUM-$\epsilon\epsilon$*, *FDUM-$\epsilon\epsilon$* increases the quality loss by <3%, but with only <1% of the computation time.

II. Preliminary: Sparse MCS

Sparse MCS Use Case. Suppose an organizer launches a temperature monitoring task in a target urban area, which is divided into fine-grained regions. The goal is to update the temperature map once every hour (sensing cycle) based on sensing results from selected participants. These participants need to upload the temperature sensed on their smartphones to the server, along with the actual sensing locations (see Figure 1 (2-Top)). Typically, with a limited budget, the selected participants cannot fully cover the area; hence, the server needs to infer the temperature values of the overlooked regions (see Figure 1 (3-Top)).

Collected Sensing Matrix. Data inference is modeled as a matrix completion problem: let *Collected Sensing Matrix* (C) be a matrix to record the data collected from participants, such that $C[r, t]$ represents the data of region r in 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 fill in the missing data.

Data Inference Algorithm. Recently, *compressive sensing* has been proven to be more accurate than most of the other methods in inferring urban sensing data such as temperature and traffic [3], [11], [12], so we use it as the inference method in this paper.

As a corollary from the compressive sensing theory, Candes et al. [13] postulate two critical assumptions for applying compressive sensing to matrix completion problem:

1) *Even Data Distribution*. To ensure effective data inference, 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 [13].

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 holds. When the sampled entries contain noise or uncertainty, the total inference error is proportional to the uncertainty level of the sampled entries [13].

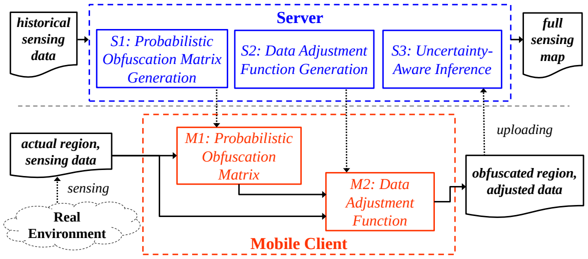

historical sensing data
Server
S1: Probabilistic Obfuscation Matrix Generation
S2: Data Adjustment Function Generation
S3: Uncertainty-Aware Inference
full sensing map
uploading
actual region, sensing data
M1: Probabilistic Obfuscation Matrix
M2: Data Adjustment Function
obfuscated region, adjusted data
sensing
Real Environment
Mobile Client

Figure 2: Location privacy-preserving framework for Sparse MCS.

We will revisit the two essential assumptions when discussing the design of *quality-optimized* location privacy-preserving mechanisms.

III. Location Privacy-Preserving Framework

Regular data collection in Sparse MCS needs participants to report their actual regions. Using obfuscation to add location privacy protection can allay participants' concerns, but may lead to data quality loss if the sensing data assigned to an obfuscated region were not representative of the actual situation. 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 a temperature monitoring use case, where participants report their obfuscated locations to the server rather than their actual locations. Figure 2 is an overview of our proposed location privacy-preserving framework for Sparse MCS. It consists of two tiers — server side and mobile client side. Before a Sparse MCS task starts, based on the historical sensing data, the server side generates a *probabilistic obfuscation matrix* (Step S1) and a *data adjustment function* (Step S2) in an offline manner. The matrix encodes the probabilities of obfuscating any one region to another. We can safeguard users' location privacy by carefully selecting the probabilities, which can make it impossible to accurately infer an actual region from its obfuscated counterpart, *even if the adversary knows the obfuscation matrix*. The data adjustment function is used to reduce data uncertainty due to region obfuscation. It is learned by analyzing the correlation between any two regions' sensing data in the historical log.

After pre-downloading both the obfuscation matrix and the data adjustment function to their mobile phones, participants can execute the sensing task on the mobile clients as follows. First, each mobile phone senses its actual location. Then, based on the probabilistic obfuscation matrix, it maps the associated region to another region (Step M1). Afterward, the data adjustment function alters the original sensing data to fit the properties of the obfuscated region (Step M2). The mobile client then uploads the modified region and data to the server. The server then infers the full sensing map from all the obfuscated regions collectively, which contains a certain degree of uncertainty compared to the actual data (Step S3). Next, we introduce how to apply differential location privacy to Sparse MCS.

IV. Differential Location Privacy

Adversary Model. In this paper, we focus on a common adversary model – Bayesian attack [7], [10]. Specifically, suppose an adversary has some *prior* knowledge about the probabilistic distribution of a user's actual region r, denoted as $\pi(r)$; 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, denoted as $\sigma(r)$, based on Bayes' rule:

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

Remark 1: The location obfuscation process is done on the smartphone side so that the MCS server has no knowledge of participants' true locations like an adversary.

Remark 2: This adversary model is a *snapshot* localization attack: the adversary can infer a participant's actual region using only the currently reported (obfuscated) position. The study on *trajectory* attack will be our future work.

Definition. Under this 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., $\sigma(r)/\pi(r)$. Intuitively, if two regions r and r' have *similar* probabilities of being mapped to r^*, then an adversary, if observing r^*, will be unable to distinguish whether the true region is r or r'.

Definition 1. *ϵ-differential-privacy. Suppose the target 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 is the parameter indicating the level of privacy, and $P[r, r^*]$ denotes the entry in P of obfuscating r to r^*.

The smaller ϵ is, the stronger privacy protection is. Note that ϵ-differential-privacy is just a constraint on the obfuscation matrix; hence, given a certain ϵ, multiple matrices may satisfy the ϵ-differential-privacy.

Privacy Guarantee. We can prove that ϵ-differential-privacy theoretically limits the knowledge gain in the previous adversary model, i.e., $\sigma(r)/\pi(r)$, whatever the adversary's prior knowledge $\pi(r)$ is (Theorem 3.2 in [7]).

Theorem 1. *If an obfuscation matrix satisfies ϵ-differential-privacy, then for an adversary with any prior knowledge π, his posterior knowledge σ satisfies:*

$$1/e^{\epsilon} \leq \sigma(r)/\pi(r) \leq e^{\epsilon}, \quad \forall r \in \mathcal{R} \tag{3}$$

V. Differential Location Privacy with Data Quality Loss Reduction

As illustrated above, there may exist many obfuscation matrices satisfying ϵ-differential privacy. Our goal is to select the one that can minimize the data quality loss brought in by location obfuscation.

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* [13]. However, introducing differential location privacy may compromise these two requirements:

1) *Even Obfuscated Region Distribution*. While the selected participants' actual location distribution is even, the distribution of the obfuscated regions may be unbalanced. Consider an extreme case of the obfuscation matrix where no region can be obfuscated to region i. Then in the collected sensing matrix, all the values of the ith row are unknown.

2) *Small Data Uncertainty in Obfuscated Regions*. The participant's actual sensing data corresponds to the original region. Although data adjustment can reduce the discrepancy between the reported data and the true data of the obfuscated region, inevitably there still exists some uncertainty after this process.

B. Optimal Obfuscation Matrix Generation

We seek to reduce the data uncertainty and control the distribution evenness of the obfuscated regions arose in the location obfuscation. To achieve the first goal, we optimally select an obfuscation matrix that can minimize the *expectation of data uncertainty* between the reported and true data in the obfuscated regions. Regarding the second aspect, we introduce an *evenness constraint* to the obfuscation matrix.

Objective: Data Uncertainty Minimization

The first step to reduce the data uncertainty is applying a *data adjustment* function to adapt the original sensing data to the obfuscated region. As environmental data usually have high spatial correlations [11], we learn a *linear regression* model for data adjustment, based on the historical sensing data of the original and obfuscated regions. In our framework (Figure 2), model learning is conducted on the server (Step S2), while the linear fit estimation is performed on the mobile clients (Step M2).

We define an *uncertainty matrix*, U, to represent the intrinsic error or uncertainty of the proposed data adjustment model. Note that $U[r, r^*]$, the data uncertainty incurred by obfuscating region r to r^*, can be computed by the *residual standard error* of the linear regression adjustment model.

Intuitively, smaller uncertainty leads to better quality. Hence, we can formulate the problem as finding an obfuscation matrix P that can minimize the overall *expectation* of data uncertainty in U, i.e.,

[1] The adversary could obtain P through spoofing to be a participant and receive P directly from the server.

$$\bar{U} = \sum\nolimits_{r \in \mathcal{R}} p(r) \cdot \sum\nolimits_{r^* \in \mathcal{R}} U[r, r^*] \cdot P[r, r^*] \tag{4}$$

where $p(r)$ is the overall probability of any one participant appearing in the region r ($\sum_{r \in \mathcal{R}} p(r) = 1$). Usually $p(r)$ is assumed to be a uniform distribution or modeled as overall human mobility pattern (e.g., via anonymous mobile phone call records [14]). For a well-designed Sparse MCS task, $p(r)$ should be roughly even-distributed to ensure the data quality. For simplicity, $p(r)$ is set to $1/|\mathcal{R}|$ (uniform). Then, to improve data quality for Sparse MCS, the objective is to minimize Eq. 4, with the following constraints.

Constraint 1: ϵ-Differential-Privacy

P must satisfy *ϵ-differential-privacy* (Eq. 2).

Constraint 2: Even Obfuscated Region Distribution

The obfuscated regions need to be evenly distributed to guarantee the inferred data quality, i.e.,

$$\psi(r^*) = \sum\nolimits_{r \in \mathcal{R}} p(r) \cdot P[r, r^*] = 1/|\mathcal{R}| \tag{5}$$

Linear Optimization: DUM-ϵe

With the objective of reducing data quality loss, we formulate a linear program, ***D**ata **U**ncertainty-**M**inimization under constraints of **ϵ**-differential-privacy and **e**venly-distributed obfuscation* (*DUM-ϵe*), to obtain the optimal P:

$$\arg\min_P \sum_{r \in \mathcal{R}} p(r) \cdot \sum_{r^* \in \mathcal{R}} U[r, r^*] \cdot P[r, r^*] \tag{6}$$

$$\text{s.t. } P[r, r^*] \le e^{\epsilon} \cdot P[r', r^*] \qquad \forall r, r', r^* \in \mathcal{R} \tag{7}$$

$$\sum_{r \in \mathcal{R}} p(r) \cdot P[r, r^*] = 1/|\mathcal{R}| \qquad \forall r^* \in \mathcal{R} \tag{8}$$

$$P[r, r^*] \ge 0 \qquad \forall r, r^* \in \mathcal{R} \tag{9}$$

$$\sum_{r^* \in \mathcal{R}} P[r, r^*] = 1 \qquad \forall r \in \mathcal{R} \tag{10}$$

where Eq. 7 is *ϵ-differential-privacy*; Eq. 8 is the *evenness* constraint; Eq. 9 and Eq. 10 are constraints for probabilities.

C. Approximation of Optimal Obfuscation Matrix

DUM-ϵe needs to make $O(|\mathcal{R}|^3)$ comparisons between different regions to ensure ϵ-differential-privacy, which makes it hard to scale. We thus approximate *DUM-ϵe* to reduce the number of comparisons, while still ensuring ϵ-differential-privacy.

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

Now, we introduce the definition of *diameter-2-critical* graph [15], whose diameter is 2 and the deletion of any edge increases its diameter. Then, the following theorem holds:

Theorem 2. *If $\mathcal{G}(\mathcal{R}, \mathcal{E})$ is a diameter-2-critical graph, an obfuscation matrix P satisfies ϵ-differential-privacy if it satisfies the following tighter constraint:*

$$P[r, r^*] \le e^{\frac{\epsilon}{2}} \cdot P[r', r^*], \quad \forall \langle r, r' \rangle \in \mathcal{E}, r^* \in \mathcal{R} \tag{11}$$

The number of comparisons in Eq. 11 is $O(|\mathcal{E}||\mathcal{R}|)$. To reduce $O(|\mathcal{E}||\mathcal{R}|)$, we identify the diameter-2-critical graph with the minimal number of edges [16]: one vertex is joined by an edge with all others; then, $O(|\mathcal{E}|) = O(|\mathcal{R}|)$, and the number of comparisons is $O(|\mathcal{R}|^2)$. We then approximate *DUM-ϵe* by replacing Eq. 7 with Eq. 11, calling it *Fast DUM-ϵe* (*FDUM-ϵe*). To create the minimal diameter-2-critical graph, any region can be chosen as the "central" vertex that connects to all others. Our experiments show that this selection has negligible effect on the data quality, and thus we randomly pick one as the central vertex.

D. Uncertainty-Aware Inference Algorithm

Ordinary compressive sensing inference for matrix completion treats all the collected data instances equally in the learning process [3]. 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* [17] learning process and give different sampling weights to different entries in the collected sensing matrix. The weight assigned is based on the overall uncertainty $\bar{u}(r^*)$ of the obfuscated region r^*:

$$\bar{u}(r^*) = \sum\nolimits_{r \in \mathcal{R}} p(r) \cdot P[r, r^*] \cdot U[r, r^*] \tag{12}$$

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

$$w(r^*) = w_0 + (1 - w_0) \cdot \frac{\bar{u}_{\max} - \bar{u}(r^*)}{\bar{u}_{\max} - \bar{u}_{\min}} \tag{13}$$

where $u_{\max}$ and $u_{\min}$ 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 as found in our empirical results (see Section VI).

VI. Evaluation

Baselines. We employ three privacy-preserving baseline mechanisms. The difference between them and *(F)DUM-ϵe* is the method for generating the obfuscation matrix P.

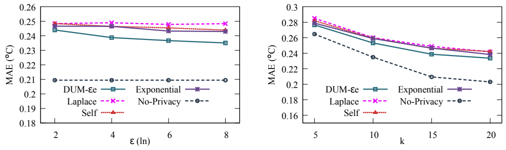
Self [18] assigns higher probability to self-obfuscation pairs: $P_{self}[r, r^*] \propto e^{\epsilon}$, if $r^* = r$; 1, otherwise.

Laplace [7] is to add Laplacian noise to the actual location data. Intuitively, *Laplace* tends to obfuscate a region to its nearby regions with high probability.

Exponential [9] is also a widely-used differential privacy mechanism. We set its scoring function as the uncertainty matrix; thus, the smaller $U[r, r^*]$ is, the higher $P[r, r^*]$ is.

Evaluation Scenarios. We evaluate our framework on two datasets from real environment and traffic monitoring.

Environment monitoring: We use the temperature sensing datasets from *SensorScope* [19]. We divided the target area, EPFL campus (300m×500m), into 100 equal-sized regions



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

Figure 3: MAE in temperature monitoring ($|\mathcal{R}| = 57$).

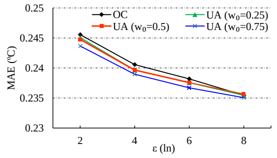
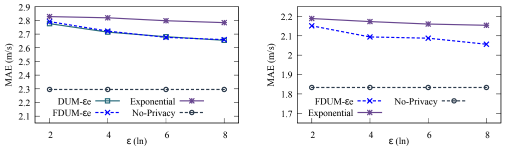


Figure 4: MAE of UA vs. OC (temperature, $k = 15$, *DUM-ϵe*).



(a) $|\mathcal{R}| = 100$ (b) $|\mathcal{R}| = 500$

Figure 5: MAE in traffic monitoring ($k = 0.3|\mathcal{R}|$, varying ϵ).

(30m×50m), and got 57 regions deployed with the sensors. These 57 regions are used in the evaluation. The sensing data spanned one week, with a sensing cycle of 30 minutes. We use the SWIM model with the 'Dartmouth' setting [20] to generate the moving traces of 1000 candidate participants.

Traffic monitoring: We use a four-day trajectory dataset of ∼30,000 taxis in Beijing [21]. The sensing cycle is set to 1 hour. Road segments (between two neighboring road intersections) are seen as 'regions'. We choose the top 100-500 road segments which have the most frequent taxi visits to construct the target sensing area.

For both scenarios, the data of the first day is used for learning the data adjustment function and region obfuscation matrix, while the rest days are for testing.

Experiment Parameters. We set up the experiment with the privacy level, ϵ, and the number of participants selected in each cycle, k, as independent variables.

Data Quality Metric. To measure data quality, we calculate the *Mean Absolute Error (MAE)*. For each experiment setting of k and ϵ, *MAE* is computed over five repeated trials. We focus on data quality loss and define $Loss_{MAE}$:

$$Loss_{MAE}(DUM\text{-}\epsilon e) = MAE(DUM\text{-}\epsilon e) - MAE(No\text{-}Privacy)$$

Evaluation Results: Computation Resources. The most time-consuming part of our framework is learning the obfuscation matrix. The time needed is proportional to the number of regions in the target sensing area. We use *IBM CPLEX* to solve *DUM-ϵe* or *FDUM-ϵe*. In our test computer (CPU: Intel Core i7-3612QM@2.10GHz, RAM: 8 GB, OS: Windows 7), ***DUM-ϵe* can only process at most 150 regions**, which consumes 1119 seconds; otherwise, CPLEX terminates exceptionally with 'out of memory'. For *FDUM-ϵe*, computation time is less than 1% of *DUM-ϵe*, and **could handle up to 500 regions** within just 121 seconds. As learning obfuscation matrix is an offline process, the above computation time is acceptable.

Evaluation Results: Data Quality. In general, our results show that *DUM-ϵe* can reduce data quality loss by up to 42% compared to the three baselines, and *FDUM-ϵe* can achieve similar performance as *DUM-ϵe* (<3% additional quality loss).

Environment Monitoring. Figure 3a shows the temperature MAE under varying privacy level ϵ, with a fixed number of participants (k = 15). As expected, *No-Privacy* achieves the best data quality. Among the privacy-preserving mechanisms, *DUM-ϵe* incurs the smallest $Loss_{MAE}$. When varying ϵ in $[\ln(2), \ln(8)]$, *DUM-ϵe* can reduce $Loss_{MAE}$ by 11.2-31.6%, 11.3-24.3% and 7.1-20.9% compared to *Laplace*, *Self* and *Exponential*, respectively. Also, MAE of *DUM-ϵe* decreases the most sharply. Hence, relaxing privacy level leads to more improvements in data quality for *DUM-ϵe*.

Figure 3b shows that MAE decreases for all the mechanisms with more participants. When $\epsilon = \ln(4)$, by varying k from 5 to 20, $Loss_{MAE}$ of *DUM-ϵe* is always the smallest among all the privacy-preserving mechanisms. Specifically, $Loss_{MAE}$ of *DUM-ϵe* is smaller than the three baselines by 13.5-42.1%. Furthermore, we can see that to ensure a certain data quality level, organizers have the trade-offs between smaller, more manageable recruitment populations (k) and the participants' privacy level (ϵ). For example, to achieve MAE $\leq$ 0.235, organizers can recruit 10 participants with *No-Privacy*, or 20 participants with *DUM-ϵe* when $\epsilon = \ln(4)$. Relaxing to $\ln(8)$ allows for a smaller recruitment size of 15.

To evaluate our *Uncertainty-Aware inference algorithm* (UA), we compare it with *Ordinary Compressive sensing* (OC), as shown in Figure 4. We experimented with w_0 = 0.25, 0.5 and 0.75 for UA, and found that 0.75 performs the best. Overall, UA achieves a smaller MAE than OC when ϵ is low. For a larger $\epsilon = \ln(8)$, UA does not appreciably improve accuracy, probably because at higher ϵ (lower privacy), the obfuscation leads to less uncertainty, and thus there is not much quality loss for UA to recover.

Traffic Monitoring. Figure 5 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. For clarity, we only show the best baseline, *Exponential*. Generally, *DUM-ϵe* and *FDUM-ϵe* achieve similar data quality, much better than *Exponential*. For example, *FDUM-ϵe* degrades data quality only by <3% compared to *DUM-ϵe* when 30 taxis are randomly selected on 100 road segments with different privacy levels.

VII. Related Work

Location privacy has been widely studied in recent years because of the growing popularity of location-based applications [5]. Popular locaiton privacy protection mechanisms include *cloaking* [6] and *dummy points* [22]. However, both of them are sensitive to the adversary's prior knowledge about the target user's location distribution [7]. The same drawback exists when applying cloaking-based location privacy protection in MCS, e.g., [23], which is still a common practice in this area [24].

Recently, researchers introduced *differential privacy* [8] to location-based services (LBS) [7], [10], [25] to alleviate the impact of an adversary's prior knowledge. Traditionally, differential location privacy alters a user's actual location to an obfuscated location by applying appropriately chosen Laplacian noise [7], [25]. Our work builds on this idea, while considering the data uncertainty incurred by obfuscation. We propose to achieve the optimal obfuscation via linear programming, which can outperform the Laplace noise. Note that [10] also uses linear programming to obtain an optimal obfuscation function, but they sought to optimize LBS services by minimizing the expected distance between actual and obfuscated locations, which can not be directly applicable to MCS.

To et al. have developed differential privacy in the participant recruitment process of MCS tasks [26]. Since their objective is to achieve high task assignment rate with short travel distance, differential location privacy is preserved by perturbing the total participant count in a certain region, rather than obfuscating any specific user's actual location as discussed in this work.

VIII. Conclusion

This paper presents a differential 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. The proposed framework includes: (1) a data adjustment function, (2) a linear program and its fast approximation to obtain the optimal location obfuscation matrix satisfying ϵ-differential-privacy, and (3) an uncertainty-aware data inference algorithm. Empirical evaluation with real-world datasets shows that our framework can provide adequate privacy protection with reduced data quality loss. In the future, we will extend this framework by adding other privacy protection guarantees and relaxing the requirement of accurate data for learning the obfuscation matrix.

Acknowledgement. This paper is partially supported by NSFC under Grant No. 61572048, Swiss National Science Foundation under grant No. PP00P2_153023.

References

[1] D. Zhang, L. Wang, H. Xiong, and B. Guo, "4w1h in mobile crowd sensing," *IEEE Commun. Mag.*, vol. 52, no. 8, pp. 42–48, 2014.

[2] 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 *IPSN*, 2010, pp. 105–116.

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

[4] L. Wang, D. Zhang, Y. Wang, C. Chen, X. Han, and A. M'hamed, "Sparse mobile crowdsensing: challenges and opportunities," *IEEE Commun. Mag.*, 2016.

[5] J. Krumm, "A survey of computational location privacy," *PUC*, vol. 13, no. 6, pp. 391–399, 2009.

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

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

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

[9] F. McSherry and K. Talwar, "Mechanism design via differential privacy," in *FOCS*, 2007, pp. 94–103.

[10] N. E. Bordenabe, K. Chatzikokolakis, and C. Palamidessi, "Optimal geo-indistinguishable mechanisms for location privacy," in *CCS*, 2014, pp. 251–262.

[11] L. Kong, M. Xia, X. Liu, G. Chen, Y. Gu, and M. Wu, "Data loss and reconstruction in wireless sensor networks," *IEEE TPDS*, vol. 25, no. 11, pp. 2818–2828, 2014.

[12] L. Wang, D. Zhang, A. Pathak, C. Chen, H. Xiong, D. Yang, and Y. Wang, "CCS-TA: quality-guaranteed online task allocation in compressive crowdsensing," in *UbiComp*, 2015, pp. 683–694.

[13] E. J. Candes and Y. Plan, "Matrix completion with noise," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 925–936, 2010.

[14] H. Xiong, D. Zhang, L. Wang, and H. Chaouchi, "Emc3: Energy-efficient data transfer in mobile crowdsensing under full coverage constraint," *IEEE TMC*, vol. 14, no. 7, pp. 1355–1368, 2015.

[15] L. Caccetta and R. Häggkvist, "On diameter critical graphs," *Discrete Mathematics*, vol. 28, no. 3, pp. 223–229, 1979.

[16] P. Erdős, A. Rényi, and V. Sós, "On a problem of graph theory," *Studia Sci. Math. Hungar*, vol. 1, pp. 215–235, 1966.

[17] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, no. 8, pp. 30–37, 2009.

[18] S. Agrawal and J. R. Haritsa, "A framework for high-accuracy privacy-preserving mining," in *ICDE*, 2005, pp. 193–204.

[19] F. Ingelrest, G. Barrenetxea, G. Schaefer, M. Vetterli, O. Couach, and M. Parlange, "Sensorscope: Application-specific sensor network for environmental monitoring," *ToSN*, vol. 6, no. 2, pp. 17:1–17:32, 2010.

[20] S. Kosta, A. Mei, and J. Stefa, "Large-scale synthetic social mobile networks with swim," *IEEE TMC*, vol. 13, no. 1, pp. 116–129, 2014.

[21] J. Shang, Y. Zheng, W. Tong, E. Chang, and Y. Yu, "Inferring gas consumption and pollution emission of vehicles throughout a city," in *KDD*, 2014, pp. 1027–1036.

[22] H. Kido, Y. Yanagisawa, and T. Satoh, "Protection of location privacy using dummies for location-based services," in *ICDE Workshops*, 2005, pp. 1248–1248.

[23] A. Kapadia, N. Triandopoulos, C. Cornelius, D. Peebles, and D. Kotz, "Anonysense: Opportunistic and privacy-preserving context collection," in *Pervasive Computing*. Springer, 2008, pp. 280–297.

[24] 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 Record*, vol. 44, no. 4, pp. 23–34, 2016.

[25] R. Dewri, "Local differential perturbations: Location privacy under approximate knowledge attackers," *IEEE TMC*, vol. 12, no. 12, pp. 2360–2372, 2013.

[26] H. To, G. Ghinita, and C. Shahabi, "A framework for protecting worker location privacy in spatial crowdsourcing," *Proceedings of the VLDB Endowment*, vol. 7, no. 10, pp. 919–930, 2014.