Privacy Preserving Big Histogram Aggregation for Spatial Crowdsensing

Shaowei Wang[1], Liusheng Huang[2], Pengzhan Wang[3], Yao Shen[4], Hongli Xu[5], Wei Yang[6]
School of Computer Science and Technology, USTC, Hefei, 230027, China
Suzhou Institute for Advanced Study, USTC, Suzhou, 215123, China
Email: {wangsw[1], pzwang[3], shenyao[4]}@mail.ustc.edu.cn, {lshuang[2], xuhongli[5], qubit[6]}@ustc.edu.cn

***Abstract*—The popularity of mobile devices has far expanded the application scenarios of spatial crowdsensing, due to its ability to provide fine-grained multi dimensional sensor readings associated with location information. Privacy is one of the fundamental issues in crowdsensing, as these location-based sensor readings may reveal identities or activities of participants. In this paper, we adopts the state-of-art location privacy definition geo-indistinguishability, provide an efficient and effective privacy preserving histogram aggregation mechanism BFMM (Bit Flipping Matrix Mechanism) for fine-grained multi dimensional location-based data. Theoretical analyses and experimental results demonstrate the efficiency and effectiveness of our approach for fine-grained multidimensional location-based data. Specifically, the aggregation accuracy of our approach averagely outperforms existing methods by a factor of number of buckets in the histogram.**

I. Introduction

Stimulated by the popularity of personal mobile devices, spatial crowdsensing has recently been a fast growing paradigm for gathering and learning from data. With the engagements of a crowd of participants equipped with sensor-rich mobile devices, spatial crowdsensing enables a plenty of location-based collective applications ranging from people-centric scenarios (e.g. crowd mobility monitoring in [1]) to environmental-centric scenarios (e.g. urban noise mapping in [2], bus arrival time prediction in [3]). In many of these collective applications, histogram aggregates are the fundamental intermediate results between gathered location-based data and high-level distilled information. Such as, histograms representing multi dimensional joint distributions (e.g. counts of location-noise pairings for urban noise mapping), and histograms representing transition matrix (e.g. counts of location-location transitions for crowd mobility monitoring).

As the demand for more functionalities and better quality of services (QoS) grows in spatial crowdsensing applications, fine-grained multi dimensional sensing data is needed to feed these applications. As a consequence, *big histograms* (histograms with reasonably large number of buckets) are emerging in spatial crowdsensing. Specifically, to achieve better quality of services, spatial crowdsensing applications call for high-resolution location and other sensing data from participants. Moreover, usually a bunch of sensor readings (e.g. temperature, noise and ambient light) is being collected simultaneously to deliver rich functionalities in these applications.

On the other hand, privacy is a fundamental issue in crowdsensing, and collecting fine-grained multi dimensional sensing data from participants makes it even more severe. These high-resolution location data precisely reveals participants' whereabouts or even identities. What's worse, other readings of sensors may also implicitly leak sensitive information of participants [4].

Several pieces of methods have been proposed in the literature to preserve privacy in spatial crowdsensing. Many of them use spatial cloaking techniques (e.g. in [5]) by generalizing exact locations with coarse-grained ones, some (e.g. [6]) further impose k-anonymity or l-diversity constraints to adaptively choose the degree of granularity of locations. However, k-anonymity or l-diversity model along with data generalization techniques is volatile to privacy adversaries with background knowledge, and spatial cloaking sacrifices granularity of locations thus makes high-resolution location retrieval nearly impossible.

Recently, geo-indistinguishability [7] is proposed to achieve the state-of-art differential privacy for locations, and guides another piece of methods that use data perturbation techniques to preserving location privacy. In [7], Laplacian noises is added to a location's coordinates to preserving geo-indistinguishability, further in [8] and [9], randomized response techniques are used to obtain optimal location utility in location-based systems (LBSs) under geo-indistinguishability constraints. Unfortunately, these methods involve high computational overheads and achieve poor aggregation accuracy for fine-grained high dimensional data, where big histogram aggregates are needed. In concise, existing methods for privacy preserving spatial crowdsensing are not suitable for applications demanding fine-grained high dimensional sensing data, for their limited privacy protection or poor accuracy performance.

In this paper, we aim at designing an efficient and effective privacy preserving histogram aggregation mechanism for fine-grained and (or) high dimensional location-based data. As opposed to existing geo-indistinguishability mechanisms that randomized response with locations or location-based data in the original domain, bit flipping matrix mechanism (BFMM) proposed in this paper randomized response with a location set. As a result, occurrence rate of each location is less constrained by number of locations, and much better estimation accuracy is achieved for big histogram. In our mechanism, privacy of both location and other sensor readings is considered. Same as locations itself, sensor readings may potentially risk participants' privacy, thus distinguishability is also limited for sensor readings in our mechanism. The main contribution of this work is summarized as follows:

– We proposed a geo-indistinguishability histogram aggre-

978-1-4673-8590-9/15/$31.00 ©2015 IEEE

gation mechanism BFMM for multi dimensional data. Privacy for both location and multi dimensional sensing data is considered in the mechanism.
- We offer efficient algorithms implementing the BFMM mechanism along with analyses of their theoretical error bounds for histogram estimation.
- We conduct extensive experiments to evaluate proposed mechanism and algorithms. The experimental results demonstrate that the histogram estimation accuracy of our approach averagely outperforms existing approaches by a factor of $|S|$, which is the number of buckets of the histogram.

The remaining of the paper is organized as follows. Section II introduces our aggregation system model and privacy definition for multi dimensional location-based data. Section III proposes BFMM mechanism. Section IV provides efficient algorithmic implementations of BFMM and their theoretical error bounds. Section V gives experimental results of our approach on various spatial crowdsensing scenarios. Section VI reviews related privacy preserving approaches for spatial crowdsensing. Section VII concludes the whole paper.

II. Definitions

In this section, we introduce our privacy preserving aggregation system model for location-based data, and formal privacy definition for locations and multi dimensional sensing data.

A. System Model

There are N participants and one aggregator in the aggregation system, and the aggregator needn't to be trustable for participants. Let u_i denotes the i-th participant in the system, and l_i be the truly location of u_i, m_i be the spatial data u_i observed, where $l_i \in L$ and L is the set of points of interests (POIs), $m_i \in M$ and M is the set of all possible discrete sensor readings.

The spatial data m_i here may represents multi dimensional sensing data or even location, e.g. m_i is the pairing of air temperature and ambient light on location l_i or the next move of the participant since being location l_i. Without loss of generalization, we simply denote data associated with location l_i as m_i for participant u_i.

Since directly informing the aggregator the truly location-based data $s_i = (l_i, m_i)$ breaches privacy (as the aggregator might be a privacy adversary), the participant u_i releases only a sanitized random data s'_i probabilistically by applying a geo-indistinguishability mechanism $\tilde{K}$ on (l_i, m_i). As illustrated in Fig 1, sanitized data $s'_i = \tilde{K}(l_i, m_i)$ in our mechanism is a bit array representing a subset of domain of (l_i, m_i). After receiving all N sanitized reports from participants, the aggregator estimate statistical information (e.g. frequencies over domain of s_i) about locations and spatial data from $\{s'_1, s'_2, ..., s'_N\}$.

Note that the aggregator is not a trustable party for participants in our system model, and every participant takes responsibility for privacy protection of their own location-based data. The participants sanitize their location-based data locally and independently before sending it to the aggregator, thus privacy is protected without relying on the aggregator or other participants.

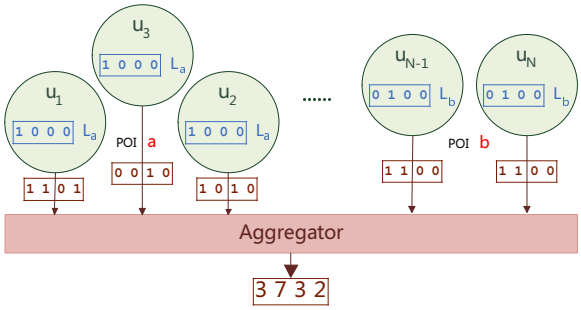

Fig. 1: Spatial crowdsensing aggregation model.

B. Privacy Definition

Geo-indistinguishability [7] is a notion for location privacy, it provides quantified indistinguishability of reported answers between any pairing of truly locations. Let metric $d(L_a, L_b)$ denotes the distance between location $L_a \in L$ and $L_b \in L$, $d(L_a, L_b)$ will be used to control the relative level of indistinguishability between sanitized random data $S'_a = \tilde{K}(L_a)$ and $S'_b = \tilde{K}(L_b)$. The $d(L_a, L_b)$ is usually set as Euclidean distance between location L_a and L_b though it can be any distance metric.

Definition 1 (ϵ-geo-indistinguishability [7]): A randomized mechanism $\tilde{K}$ gives ϵ–geo-indistinguishability iff for all possible pairings L_a and L_b belong to L, and all $T \subseteq range(\tilde{K})$,

$$\Pr[\tilde{K}(L_a) \in T] \leq \exp(\epsilon \cdot d(L_a, L_b)) \cdot \Pr[\tilde{K}(L_b) \in T].$$

Metric $d(L_a, L_b)$ in ϵ-geo-indistinguishability implies that we should report relatively similar answers when L_a and L_b is close due to privacy concern, and report distinguishable answers when L_a and L_b is far away from each other. By contrast, in the definition of differential privacy [10], indistinguishability metric $d(L_a, L_b)$ is identical for all possible pairings, thus isn't suitable for locations. Metric $d(L_a, L_b)$ defines relative distance or indistinguishable level between locations, meanwhile ϵ serves as a resilient multiplicative parameter to designate absolute indistinguishable levels. The choice of ϵ may differ in variety spatial crowdsensing scenarios, we can choose smaller ϵ for more rigid privacy preserving, and choose larger ϵ for potentially better data utility.

Intuitively, geo-indistinguishability is a generalization of differential privacy, and can be taken as privacy definition for universal data (as in [11], also known as data obfuscation [9]), including multi dimensional location-based data. For preserving privacy of other sensor readings besides location itself, here we slightly extends ϵ-geo-indistinguishability towards a unified and elastic privacy definition for multi dimensional location-based data $s = (l, m)$.

Similar as for location data l, indistinguishable levels (or distances) between multi dimensional location-based data s should firstly be defined. For a pair of location-based data $s_x = (L_a, M_c)$ and $s_y = (L_b, M_d)$, Euclidean distance or

Manhattan distance of s_x and s_y can be used as indistinguishability metric. Specially, since different sensor readings has different levels of sensitivity for participants' privacy, usually relative weights are assigned for different dimensions of location-based data in distance metric. A relatively small weights for a sensor reading specify that relatively small indistinguishability is allowed for different readings of the sensor, hence relatively rigid privacy constraints are applied. Therefore, a relatively small weights should be assigned for location dimension as it's highly sensitive information for a participant, and we can assign relatively large weights for air temperature or ambient light readings. Formally, we may define distance metric between multi dimensional location-based data $s_x = (L_a, M_c)$ and $s_y = (L_b, M_d)$ as follows:

$$d(s_x, s_y) = \|(w_l \cdot L_a - w_l \cdot L_b,\ w_m \cdot M_c - w_m \cdot M_d)\|_c\ ,$$

where c is dimensionality of metric, when $c = 1$, Manhattan distance is applied, and when $c = 2$, Euclidean distance is applied. The w_l is the relative weights for location data, w_m denotes the relative weights for other sensor readings. We now propose our privacy definition for multi dimensional location-base data in definition 2.

Definition 2 (ϵ-geo-indistinguishability for location-based data): A randomized mechanism $\tilde{K}$ gives ϵ-geo-indistinguishability iff for all possible pairings $s_x = (L_a, M_c)$ and $s_y = (L_b, M_d)$ belong to $L \times M$, and all $T \subseteq range(\tilde{K})$,

$$\Pr[\tilde{K}(L_a, M_c) \in T] \leq \exp(\epsilon \cdot d(s_x, s_y)) \cdot \Pr[\tilde{K}(L_b, M_d) \in T].$$

Despite the domain of data under consideration, there is no difference between ϵ-geo-indistinguishability in definition 2 and definition 1. The ϵ-geo-indistinguishability for location-based data in definition 2 simply replaces location data $l \in L$ with multi dimensional location-based data $s \in S$, where $S = L \times M$ denotes the set of all possible value of s. Therefore, in the following sections, only the more general location-based data $s \in S$ and its distance metric $d = \{d(s_x, s_y) \mid s_x, s_y \in S\}$ are discussed.

III. Bit Flipping Matrix Mechanism

Recall that in our system model, a participant u_i holds secret location-based data s_i of domain S, then releases a sanitized data $s'_i = \tilde{K}(s_i)$ that satisfying ϵ-geo-indistinguishability to the aggregator, the aggregator learns frequencies over domain S (namely histogram over S) from sanitized set $\{s'_1, s'_2, ..., s'_N\}$ contributed by N participants.

We now introduce Bit Flipping Matrix Mechanism (BFMM) as geo-indistinguishable mechanism $\tilde{K}$. We firstly characterize bit flipping matrix, then its constraints under ϵ-geo-indistinguishability.

A. Bit Flipping Matrix

In BFMM, taken as input the secret location-based data $s_i \in S$, a subset of S is randomly chosen as output with well-designed probability mass assignment. Denote the a-th element of S as S_a, we can represent S_a as a bit array, e.g., $S_a \in \{0,1\}^{|S|}$, where the a-th bit of S_a is set as 1, while other bits are set as 0. To satisfy constraints due to ϵ-geo-indistinguishability, BFMM flips every bit in the bit array of S_a. Specifically, for every bit array $S_a \in S$, BFMM flips the k-th bit in the array to 1 with a probability $F_{a,k} \in [0,1]$, where $a, k \in \{1, ..., |S|\}$.

Algorithm 1 Bit Flipping Matrix Mechanism

Input: Bit flipping matrix $F \in [0,1]^{|S|\times|S|}$, a participant's location-based data s_i, where $s_i = S_a \in S$
Output: A bit array $s'_i = S'_a \in \{0,1\}^{|S|}$ after bit flipping
1: **for** $k = 1$ **to** $|S|$ **do**
2: $\quad S'_{a,k} = 0$
3: $\quad$ Set $S'_{a,k} = 1$ with probability $F_{a,k}$
4: **end for**
5: **return** S'_a

Formally, we define bit flipping matrix $F \in [0,1]^{|S|\times|S|}$ as the matrix of probabilities to flip each secret bit in array S_a to 1, and $F_{a,k}$ is the probability flipping the k-th bit of array for location-based data S_a to 1. Conventionally, $F_{a,a} \geq 0.5$, and $F_{a,k} \leq 0.5$ when $a \neq k$, that is, the ones in bit array S_a remains one with a relatively high probability, and the zeros in bit array S_a remains zero with a relative high probability. Given bit flipping matrix F, as presented in Algorithm 1, BFMM flips k-th bit of S_a to 1 with probability $F_{a,k}$.

B. Privacy Constraints

In the former subsection, algorithm 1 establishes the basic technique in BFMM towards ϵ-geo-indistinguishability, further in theorem 1, we formalize constraints of bit flipping matrix F for achieving ϵ-geo-indistinguishability under matric d. Theorem 1 illustrates that for all pair of rows in F, the differences of flipping probability must be limited by exponential absolute distance between the corresponding location-based data $S_a, S_b \in S$.

Theorem 1: The bit flipping matrix F for domain S satisfies ϵ-geo-indistinguishability under indistinguishability metric $d = \{d(S_a, S_b) \mid S_a, S_b \in S\}$, iff for all possible pairings S_a and S_b belong to S,

$$\prod_{k=1..|S|} max\{\frac{F_{a,k}}{F_{b,k}}, \frac{1-F_{a,k}}{1-F_{b,k}}\} \leq \exp(\epsilon \cdot d(S_a, S_b)). \quad (1)$$

Proof: Let S'_a and S'_b denotes the sanitized version of S_a and S_b that randomized with bit flipping matrix F, where $S_a, S_b \in S$ and $a \neq b$, then we have:

$$\begin{aligned}
&max\{\frac{Pr[S'_a = t]}{Pr[S'_b = t]} \mid t \in \{0,1\}^{|S|}\} \\
&= max\{\prod_{k=1..|S|} (\frac{F_{a,k}}{F_{b,k}})^{[t_k=1]} \cdot (\frac{1-F_{a,k}}{1-F_{b,k}})^{[t_k=0]} \mid t \in \{0,1\}^{|S|}\} \\
&= \prod_{k=1..|S|} max\{(\frac{F_{a,k}}{F_{b,k}})^{[t_k=1]} \cdot (\frac{1-F_{a,k}}{1-F_{b,k}})^{[t_k=0]} \mid t_k \in \{0,1\}\} \\
&= \prod_{k=1..|S|} max\{\frac{F_{a,k}}{F_{b,k}}, \frac{1-F_{a,k}}{1-F_{b,k}}\}.
\end{aligned}$$

iff) We have $max\{\frac{Pr[S'_a=t]}{Pr[S'_b=t]} \mid t \in \{0,1\}^{|S|}\} \leq \exp(\epsilon \cdot d(S_a, S_b))$, hence $\prod_{k=1..|S|} max\{\frac{F_{a,k}}{F_{b,k}}, \frac{1-F_{a,k}}{1-F_{b,k}}\} \leq \exp(\epsilon \cdot d(S_a, S_b))$.
if) We have $\prod_{k=1..|S|} max\{\frac{F_{a,k}}{F_{b,k}}, \frac{1-F_{a,k}}{1-F_{b,k}}\} \leq \exp(\epsilon \cdot d(S_a, S_b))$, hence $max\{\frac{Pr[S'_a=t]}{Pr[S'_b=t]} \mid t \in \{0,1\}^{|S|}\} \leq \exp(\epsilon \cdot d(S_a, S_b))$. ■

IV. Implementation

In bit flipping matrix mechanism, there are unlimited potential choices of bit flipping matrix F satisfying ϵ-geo-indistinguishability. Among them all, we seek for best choices regarding error bounds for histogram estimation under computational constraints. Naturally, finding the optimal choices of bit flipping matrix F can be formalized as an optimization problem under privacy constraints in theorem 1 minimizing histogram estimation error. However, the original optimization problem involves with $|S|^2$ constraints on $|S|^2$ variables, and the objective function is non-linear, whereas, our goal is to provide an efficient and effective aggregation mechanism for big histogram where the location-based data domain $|S|$ is relatively large, thus it is not practical for implementation.

Actually, even finding a feasible bit flipping matrix F under ϵ-geo-indistinguishability constraints meanwhile with reasonable utility guarantee is not explicit. Though constructing the bit flipping matrix F under the definition of differential privacy where $(d(S_a, S_b) \equiv d_*)$ is kind of trivial as all $S_a \in S$ are symmetry. Under the notion of ϵ-geo-indistinguishability where distance $d(S_a, S_b)$ may differs for different pairings, such bit flipping matrix F is relatively hard to construct.

In this section, we focus on efficient approaches to concretely construct bit flipping matrix F with histogram estimation utility guarantee, two algorithms are presented, one has $|S|^2$ computational complexity, the other has $|S|^2 \log(|S|)$ computational complexity. We also provide detailed computational complexity analyses of our approaches and theoretical error bounds of constructed bit flipping matrix.

A. A Greedy Approach

There are $|S|^2$ ϵ-geo-indistinguishability constraints on bit flipping matrix F as implied in theorem 1, each limits $2 \cdot |S|$ flipping probabilities in two rows of F, this makes those constraints pretty complicate. Hence, in our greedy approach, we further inject symmetry constraints to bit flipping matrix F, to simplify the original ϵ-geo-indistinguishability constraints, formally, the symmetry constraints are as follows:

$$\begin{cases} F_{k,k} \geq 0.5, & for\ k \in 1, 2, ..., |S|; \\ F_{a,k} = 1 - F_{k,k}, & for\ a, k \in 1, 2, ..., |S|\ and\ a \neq k. \end{cases} \tag{2}$$

As a result, the constraints in equation (1) are reduced as follows:

$$\begin{aligned} &\prod_{k=1..|S|} max\{\frac{F_{a,k}}{F_{b,k}}, \frac{1-F_{a,k}}{1-F_{b,k}}\} \\ &= \frac{F_{a,a}}{F_{b,a}} \cdot \frac{1-F_{a,b}}{1-F_{b,b}} \\ &= \frac{F_{a,a}}{1-F_{a,a}} \cdot \frac{F_{b,b}}{1-F_{b,b}} \\ &\leq \exp(\epsilon \cdot d(S_a, S_b)). \end{aligned} \tag{3}$$

In addition, the number of independent variables in F is reduced to $|S|$.

Follow the reduced constraints in equation (3), we now proceed to present our greedy approach to construct bit flipping matrix F in algorithm 2. The algorithm greedily finding the closest elements of a element S_j in domain S, and apply the minimum distinguishability distance to its bit flipping probability $F_{j,j}$.

Algorithm 2 A Greedy Constructing Algorithm

Input: Privacy budget $\epsilon \in R^+$,
distance metric $d \in R^{|S| \times |S|}$.
Output: A bit flipping matrix F satisfying ϵ-geo-indistinguishability

1: **for** $j = 1$ **to** $|S|$ **do**
2: $d_{j,min} = min\{d_{j,1}, d_{j,2}, \ldots, d_{j,j-1}, d_{j,j+1}, \ldots, d_{j,|S|}\}$
3: $F_{j,j} = \frac{1}{\exp(-\epsilon \cdot d_{j,min}/2)+1}$
4: **for** $k = 1$ **to** $|S|$ **and** $k \neq j$ **do**
5: $F_{k,j} = 1 - F_{j,j}$
6: **end for**
7: **end for**
8: **return** F

***Theorem** 2:* The bit flipping matrix F constructed by Algorithm 2 for domain S satisfies ϵ-geo-indistinguishability under indistinguishability metric $d_{a,b} = d(S_a, S_b)$.

Proof:

$$\begin{aligned} &\prod_{k=1..|S|} max\{\frac{F_{a,k}}{F_{b,k}}, \frac{1-F_{a,k}}{1-F_{b,k}}\} \\ &= \frac{F_{a,a}}{1-F_{a,a}} \cdot \frac{F_{b,b}}{1-F_{b,b}} \\ &= \exp(\epsilon \cdot min\{d_{a,1}, d_{a,2}, \ldots, d_{a,a-1}, d_{a,a+1}, \ldots, d_{a,|S|}\}/2) \\ &\quad \cdot \exp(\epsilon \cdot min\{d_{b,1}, d_{b,2}, \ldots, d_{b,b-1}, d_{b,b+1}, \ldots, d_{b,|S|}\}/2) \\ &\leq \exp(\epsilon \cdot d_{a,b}/2) \cdot \exp(\epsilon \cdot d_{b,a}/2) \\ &\leq \exp(\epsilon \cdot d_{a,b}). \end{aligned}$$

■

The ϵ-geo-indistinguishability guarantee of constructed bit flipping matrix F is presented in theorem 2.

B. A Heuristic Approach

The proposed approach to construct bit flipping matrix F in algorithm 2 has the potential for further optimization with additional computational overhead. By relaxing distinguishability distance between settled elements and unsettled elements in domain S, we propose a heuristic approach in algorithm 3 based on algorithm 2.

Lets SC denotes the set of settled elements, the heuristic constructing algorithm greedily find the closest pair inside set SC or between set SC and $S - SC$ in line 4, after adding new element(s) in the pair into the settled set SC in line 6 and 16, the flipping probability of the new element(s) is settled in line 10 and 20, then, the temporary distance metric d' is

Algorithm 3 A Heuristic Constructing Algorithm

Input: Privacy budget $\epsilon \in R^+$,
distance metric $d \in R^{M \times M}$.

Output: A bit flipping matrix F satisfying ϵ-geo-indistinguishability

1: $SC = \Phi$
2: $d' = d$
3: **while** $SC \neq \{1, 2, ..., |S|\}$ **do**
4: $\quad d'_{a,b} = \min\{\ d'_{j,k}\ \mid\ j \neq k\ and\ j \notin SC\ \}$
5: $\quad$ **if** $a \notin SC$ **then**
6: $\quad\quad SC = SC \cup \{a\}$
7: $\quad\quad$ **for** $k = 1$ **to** M **and** $k \notin SC$ **do**
8: $\quad\quad\quad d'_{k,a} = 2 \cdot d'_{k,a} - d'_{a,b}$
9: $\quad\quad$ **end for**
10: $\quad\quad F_{a,a} = \frac{1}{\exp(-\epsilon \cdot d'_{a,b}/2)+1}$
11: $\quad\quad$ **for** $k = 1$ **to** $|S|$ **and** $k \neq a$ **do**
12: $\quad\quad\quad F_{k,a} = 1 - F_{a,a}$
13: $\quad\quad$ **end for**
14: $\quad$ **end if**
15: $\quad$ **if** $b \notin SC$ **then**
16: $\quad\quad SC = SC \cup \{b\}$
17: $\quad\quad$ **for** $k = 1$ **to** $|S|$ **and** $k \notin SC$ **do**
18: $\quad\quad\quad d'_{k,b} = 2 \cdot d'_{k,b} - d'_{a,b}$
19: $\quad\quad$ **end for**
20: $\quad\quad F_{b,b} = \frac{1}{\exp(-\epsilon \cdot d'_{a,b}/2)+1}$
21: $\quad\quad$ **for** $k = 1$ **to** $|S|$ **and** $k \neq b$ **do**
22: $\quad\quad\quad F_{k,b} = 1 - F_{b,b}$
23: $\quad\quad$ **end for**
24: $\quad$ **end if**
25: **end while**
26: **return** F

heuristically relaxed in line 7 and 17. The iterative greedy procedure stops when all elements in domain S is settled.

The ϵ-geo-indistinguishability guarantee of constructed bit flipping matrix F is presented in theorem 3.

***Theorem** 3:* The bit flipping matrix F constructed by Algorithm 3 for domain S satisfies ϵ-geo-indistinguishability under indistinguishability metric $d_{a,b} = d(S_a, S_b)$.

Proof: See appendix B. ∎

Apparently, the bit flipping matrix F^h constructed by heuristic approach in algorithm 3 exploits relaxations of F^g constructed by the greedy approach in algorithm 2, thus we have $F^h_{k,k} \geq F^g_{k,k}$, as a result, the expected histogram estimation accuracy of F^h is always no less favourable than F^g.

C. Histogram Estimator

Randomization with bit flipping matrix F may introduces bias, e.g. $E[S'_a] \neq S_a$ when F is not identical. In this subsection, we aim to eliminating bias introduced by bit flipping.

Let H denotes the histogram of secret set $\{s_1, s_2, ..., s_N\}$ of N participants over domain S, H' denotes the observed histogram, that is, the sum of sanitized set $\{s'_1, s'_2, ..., s'_N\}$. Actually, for bit flipping matrix F satisfying symmetry constraints as in equation (2), since the expectation of k-th bucket in H'_k is given as follows:

$$\begin{aligned} &E[H'_k] \\ &= \sum_{j=1..|S|} H_j \cdot F_{j,k} \\ &= H_k \cdot F_{k,k} + (N - H_k) \cdot (1 - F_{k,k}), \end{aligned}$$

there is an efficient unbiased histogram estimator as showed in algorithm 4.

Algorithm 4 Unbiased Histogram Estimator

Input: Sanitized set $\{s'_1, s'_2, ..., s'_N\}$ from N participants that randomized with bit flipping matrix F that satisfies equation (2).

Output: Unbiased estimation of H, where H is the histogram of secret set $\{s_1, s_2, ..., s_N\}$ over domain S.

1: $H' = \vec{\mathbf{0}}$
2: **for** $i = 1$ **to** N **do**
3: $\quad H' = H' + s'_i$
4: **end for**
5: **for** $k = 1$ **to** $|S|$ **do**
6: $\quad H''_k = \frac{H'_k - N \cdot (1 - F_{k,k})}{2 \cdot F_{k,k} - 1}$
7: **end for**
8: **return** H''

D. Error Bounds for Histogram Estimation

Let H denotes the truly histogram of participants' secret data set $\{s_1, s_2, ..., s_N\}$ over domain S, we have $H = sum\{s_1, s_2, ..., s_N\}$. Let H' denotes the observed histogram of sanitized data set $\{s'_1, s'_2, ..., s'_N\}$ over domain S, we have $H' = sum\{s'_1, s'_2, ..., s'_N\}$. The least square error bounds of constructed bit flipping matrix F in the greedy and heuristic approach is given in theorem 4.

***Theorem** 4:* For participants' secret data set $\{s_1, s_2, ..., s_N\}$ from domain S, each element in the set is independently randomized with bit flipping matrix F constructed by algorithm 2 or 3, and estimated histogram H'' given by algorithm 4, we have:

$$E[\|H'' - H\|_2^2] \leq \sum_{k=1..|S|} N \cdot \frac{\exp(\epsilon \cdot d_{k,min}/2)}{(\exp(\epsilon \cdot d_{k,min}/2) - 1)^2}.$$

Where $d_{k,min} = min\{d_{k,1}, d_{k,2}, \ldots, d_{k,k-1}, d_{k,k+1}, \ldots, d_{k,|S|}\}$.

Proof: See appendix A. ∎

E. Computational Overheads

To demonstrate the efficiency of our approaches, we now give detailed analyses of computational overheads of mechanism implementation, including overheads of constructing bit flipping matrix (denoted as *constructor*), overheads of randomizing bit array (denoted as *randomizer*) and overheads of estimating the histogram (denoted as *estimator*). We also

TABLE I: Average Computational Complexities

Approach	constructor	randomizer	estimator
LP mechanism in [8] [9]	$O(\|S\|^2 Poly(\|S\|^3))$	$O(\|S\|)$	$O(N+\|S\|^3)$
Exponential mechanism	$O(\|S\|^2)$	$O(\|S\|)$	$O(N+\|S\|^3)$
Greedy approach in algorithm 2	$O(\|S\|^2)$	$O(\|S\|)$	$O(N\cdot\|S\|+\|S\|)$
Heuristic approach in algorithm 3	$O(\|S\|^2\log\|S\|)$	$O(\|S\|)$	$O(N\cdot\|S\|+\|S\|)$

provide comparison with existing mechanisms (e.g. in [8] [9]) in table I.

Constructor: Apparently the one pass greedy approach in algorithm 2 constructing a bit flipping matrix in time $O(|S|^2)$, in contrast, the heuristic approach in algorithm 3 stops after most $|S|$ iterations, each iteration involving one min operation in distance oracle d and $O(|S|)$ distance updating operations, thus has overall average computational complexity $O|S|^2\log(|S|)$.

Randomizer: The randomizer in algorithm 1 flips each bit of location-base data $S_a \in \{0,1\}^{|S|}$, hence the computational overheads is $O(|S|)$.

Estimator: The histogram estimator in algorithm 4 simply estimate each H_k with H'_k, therefore the computational complexity is $O(N\cdot|S|+|S|)$.

Complexity Comparison: As comparison, the previous ϵ-geo-indistinguishability linear programming (LP) mechanism in [8] and [9] that output with a single location-based data instead of a set of location-based data in BFMM, are modeled as linear programming problem optimizing for single location reporting utility in Location-Based Services (LBSs),it involves with $|S|^3$ constraints on $|S|^2$ perturbing probabilities. Consequently, averagely $Poly(|S|^3)$ iterations and $O(|S|^2)$ arithmetic operation in each round is needed in LP mechanism, its average constructor computational complexity is $O(|S|^2)\cdot Poly(|S|^3)$. Additionally, in LP mechanism, perturbing a truly location-based data in domain costs $O(|S|)$, eliminating bias in estimator costs $O(|S|^3)$.

Another approach may achieve ϵ-geo-indistinguishability is the exponential mechanism [12] that is initially proposed for differential privacy preserving. By treating the distance of a pair of location-based data as loss function, the probability of perturbing the truly location-based data S_a to S_b in exponential mechanism is as follows:

$$Pr[S_b|S_a] = \frac{\exp(-\frac{d(S_a,S_b)}{2})}{\sum_{k=1..|S|}\exp(-\frac{d(S_a,S_k)}{2})}. \quad (4)$$

In exponential mechanism(EM), the computational complexity of computing perturbation probabilities is $O(|S|^2)$, the costs of randomizer and estimator is the same as LP mechanism.

In concise, the greedy approach in BFMM mechanism has much lower computational complexities than LP mechanism and EM mechanism respecting domain size $|S|$, and the heuristic in our mechanism achieves better histogram estimation accuracy with only slightly more overheads on bit flipping matrix constructing.

TABLE II: Summary of synthetic spatial crowdsensing scenarios, including the short name of the setup, the spatial map of the scenario, the discrete locations in the map, and the domain of data for histogram aggregation.

Code	Map	Locations L	Domain S
1DUS	a line of $[0,1]$	$\|L\|$ locations uniformly scattered in the line	L
1DUM			$L\times L$
1DRS		$\|L\|$ locations uniform randomly scattered in the line	L
1DRM			$L\times L$
2DUS	a square of $[0,1]\times[0,1]$	$\|L\|$ locations uniformly scattered in the square	L
2DRSX		$\|L\|$ locations uniform randomly scattered in the square	L
2DRSY			L
2DRSZ			L

V. EXPERIMENTS

In this section, we evaluate our BFMM mechanism on numerous spatial crowdsensing scenarios, mainly focus on histogram aggregation accuracy under ϵ-geo-indistinguishability privacy preserving, with comparison to the LP mechanism from [8] and [9] and exponential mechanism (EM) in equation (4).

The synthetic spatial crowdsensing scenarios include learning of histograms over locations scattered in 1-dimensional lines or 2-dimensional squares. As summarized in table II, in the 1-dimensional line scenario, two typical cases of are studied, in one case, $|L|$ locations are uniformly scattered in the line, and in the other case, $|L|$ locations are uniform randomly scattered in the line. We consider histogram aggregation over both the domain of locations and domain of location pairs (the transition matrix in section I) in the 1-dimensional line scenario; in the 2-dimensional square scenario, histogram aggregation over grid and uniform randomly scattered locations are studied.

All experiments simulate with $N = 100000$ participants with uniform distribution over domain S, and the Euclidean distance is used as distance metric d, the privacy budget ϵ is 5, which is an intermediate value as ϵ ranges from 0.01 to 10 in the literature [13]. The histogram estimation accuracy metric in our experiments is the square error $\|\frac{H''-H}{N}\|_2^2$ or its natural logarithm.

We implement the greedy approach (*Gre*) in algorithm 2 and the heuristic approach (*Heu*) in algorithm 3 for these experiments, and compare them with LP mechanism in [8] [9] and exponential mechanism (*EM*) in equation (4). However, in our 68 out of 92 experiments, the perturbing probability matrix constructed by LP mechanism is not invertible, due to fact that it optimizes for single report utility and tends to report central locations, thus can't eliminating bias of observed histogram, making the square error of its histogram estimation extremely large (e.g. $\geq 10^{10}$). Hence, the experimental results of LP mechanism is not showed in following figures.

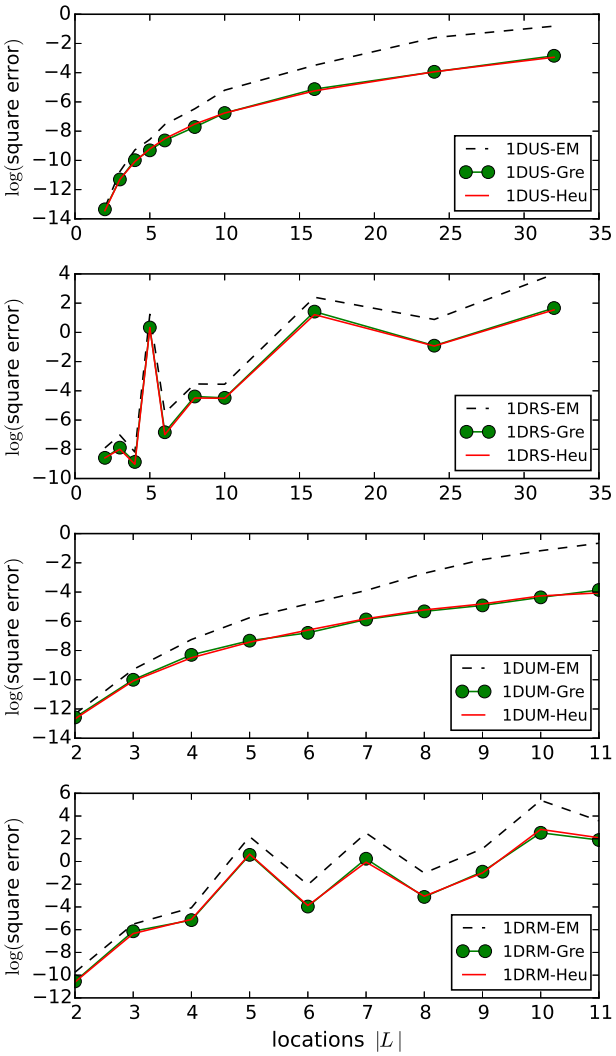

Fig. 2: Experiments of scenarios in line $[0, 1]$.

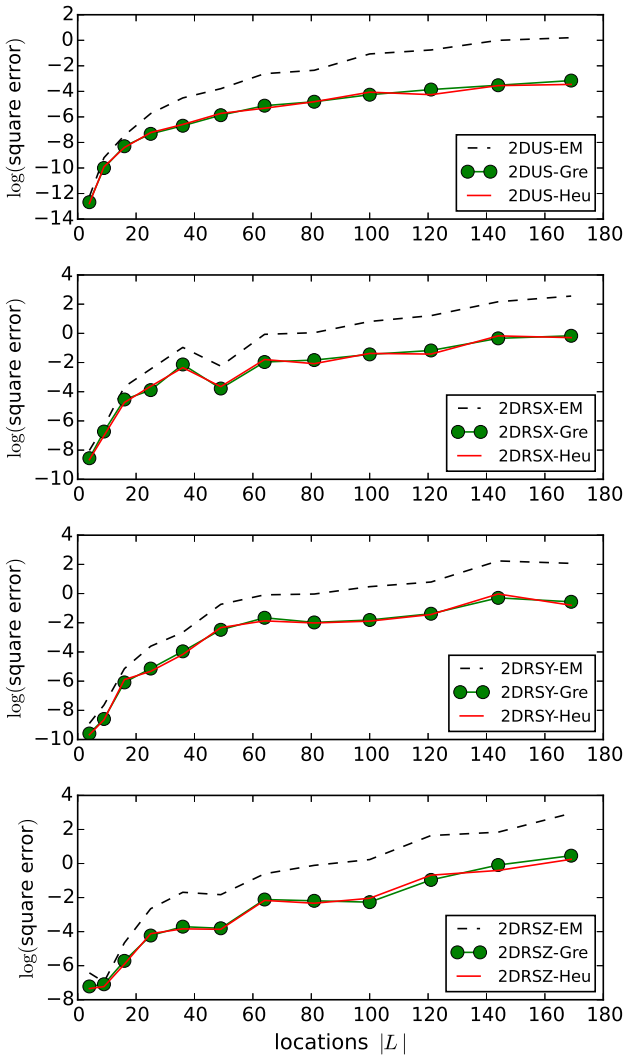

Fig. 3: Experiments of scenarios in square $[0, 1] \times [0, 1]$.

The scenarios of histogram estimation over locations scattered in line $[0, 1]$ is showed in Figure 2, which demonstrate that the greedy and heuristic approach in our mechanism significantly outperforms the EM mechanism, and usually by a factor of $\frac{|S|}{4}$ for reasonably large $|S|$ (e.g. $|S| \geq 10$).

The scenarios of histogram estimation over locations scattered in square territory $[0, 1] \times [0, 1]$ is showed in Figure 3, which demonstrate that the greedy and heuristic approach in our mechanism significantly outperforms the EM mechanism, and usually by a factor of $\frac{|S|}{5}$ for reasonably large $|S|$ (e.g. $|S| \geq 16$).

VI. Related Work

There are mainly three streams of methods for privacy preserving in crowdsensing: cloaking, perturbation and encryption.

Encryption: Encryption technique resort to cryptographic tools to keep participant's data secret, e.g. in [14], mechanisms in [15] [16] couple cryptographic with distributed noise generation to ensure differential privacy. However, these methods have high computational demanding, making them hard to implement in mobile devices, moreover, encrypted distributed noise is safe only for computational bounded adversaries and may be volatile to collusion between the aggregator and other participants. By contrast, the notion of geo-indistinguishability preserves participants' privacy information-theoretically and locally, without relying on any other party.

Cloaking: Cloaking technique protects participants' location or sensing data by generalizing location with a coarse one in [5], or further collaboratively imposing k-anonymity or l-diversity constraints on participants' coarse locations in [6]. However, its privacy guarantee is quite experimental, and is volatile to privacy adversaries with auxiliary information even with k-anonymity or l-diversity constraints [10].

Perturbation: Perturbation technique randomized response with participants' truly location or sensing data with limited probability. The approaches in [17] adopts negative survey for privacy preserving, it report truly data with zero probability to reduce computational overheads. However, it privacy guarantee is limited compared to rigorous differential privacy or

geo-indistinguishability definition. Recently, conforming the definition of differential privacy or geo-indistinguishability, [8] and [9] obfuscate truly data with optimized single data reporting utility.

Unfortunately, previous privacy preserving methods that perturbing data in the original domain achieves poor histogram estimation accuracies. In [18] and [13], randomized responding in the superset of original domain is proposed for histogram estimation under differential privacy constraints. However, differential privacy is not appropriate for location-based data as discussed in section II, as contrast, our work follows the generalized differential privacy notion geo-indistinguishability for privacy preserving.

VII. Conclusion

We propose a privacy preserving big histogram aggregation mechanism BFMM for spatial crowdsensing, it guarantees ϵ-geo-indistinguishability for multi dimensional location-based data. By representing the truly data and sanitized data as bit array of the domain size $|S|$, our mechanism output a subset of domain S. The efficiency and effectiveness of our mechanism is demonstrated by theoretical and experimental analyses, theoretical analyses illustrate that our mechanism implementation has much lower computational complexity than existing approaches, experimental results further illustrate that the aggregation accuracy outperforms existing approach averagely by a factor of $|S|$ for big histogram.

Acknowledgment

This paper is supported by the China Postdoctoral Science Foundation No.2015M570545, National Science Foundation of China under No. U1301256, 61170058, 61272133, Special Project on IoT of China NDRC (2012-2766), Research Fund for the Doctoral Program of Higher Education of China No. 20123402110019, and Jiangsu Planned Projects for Postdoctoral Research Funds No.1501085C.

References

[1] A. Stopczynski, J. E. Larsen, S. Lehmann, L. Dynowski, and M. Fuentes, "Participatory bluetooth sensing: A method for acquiring spatio-temporal data about participant mobility and interactions at large scale events," in *PerCom. IEEE*, 2013.

[2] E. DHondt, M. Stevens, and A. Jacobs, "Participatory noise mapping works! an evaluation of participatory sensing as an alternative to standard techniques for environmental monitoring," *PerCom. IEEE*, 2013.

[3] P. Zhou, Y. Zheng, and M. Li, "How long to wait?: predicting bus arrival time with mobile phone based participatory sensing," in *MobiCom. ACM*, 2012.

[4] D. Christin, "Privacy in mobile participatory sensing: Current trends and future challenges," *Journal of Systems and Software*, 2015.

[5] C.-Y. Chow, M. F. Mokbel, and X. Liu, "A peer-to-peer spatial cloaking algorithm for anonymous location-based service," in *SIGSPATIAL GIS. ACM*, 2006.

[6] H. Hu and J. Xu, "Non-exposure location anonymity," in *ICDE. IEEE*, 2009.

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

[8] N. E. Bordenabe, K. Chatzikokolakis, and C. Palamidessi, "Optimal geo-indistinguishable mechanisms for location privacy," in *CCS. ACM*, 2014.

[9] R. Shokri, "Privacy games: Optimal user-centric data obfuscation," *PETS. Springer*, 2015.

[10] C. Dwork, "Differential privacy," in *ICALP. Springer*, 2006.

[11] K. Chatzikokolakis, M. E. Andrés, N. E. Bordenabe, and C. Palamidessi, "Broadening the scope of differential privacy using metrics," in *PETS. Springer*, 2013.

[12] F. McSherry and K. Talwar, "Mechanism design via differential privacy," in *FOCS. IEEE*, 2007.

[13] Ú. Erlingsson, V. Pihur, and A. Korolova, "Rappor: Randomized aggregatable privacy-preserving ordinal response," in *CCS. ACM*, 2014.

[14] G. Ghinita, P. Kalnis, A. Khoshgozaran, C. Shahabi, and K.-L. Tan, "Private queries in location based services: anonymizers are not necessary," in *SIGMOD. ACM*, 2008.

[15] E. Shi, T.-H. H. Chan, E. G. Rieffel, R. Chow, and D. Song, "Privacy-preserving aggregation of time-series data." in *NDSS. ISOC*, 2011.

[16] J. Won, C. Y. Ma, D. K. Yau, and N. S. Rao, "Proactive fault-tolerant aggregation protocol for privacy-assured smart metering," in *INFOCOM. IEEE*, 2014.

[17] M. M. Groat, B. Edwards, J. Horey, W. He, and S. Forrest, "Enhancing privacy in participatory sensing applications with multidimensional data," in *PerCom. IEEE*, 2012.

[18] J. C. Duchi, M. I. Jordan, and M. J. Wainwright, "Local privacy and statistical minimax rates," in *FOCS. IEEE*, 2013.

Appendix

A. The proof of error bounds in theorem 4

Since $H''_k = \frac{H'_k - N\cdot(1-F_{k,k})}{2\cdot F_{k,k}-1}$ and H'_k is a bernoulli variable, we have:

$$
\begin{aligned}
&E[\|H''_k - H_k\|_2^2] = var[H''_k] \\
&= \frac{1}{(2\cdot F_{k,k}-1)^2}\cdot var[H'] \\
&= \frac{1}{(2\cdot F_{k,k}-1)^2}\cdot \sum_{j=1..|S|} H_j\cdot F_{k,j}\cdot(1-F_{k,j}) \\
&= \frac{1}{(2\cdot F_{k,k}-1)^2}\cdot N\cdot F_{k,k}\cdot(1-F_{k,k}) \\
&\le N\cdot \frac{\exp(\epsilon\cdot d_{k,min}/2)}{(\exp(\epsilon\cdot d_{k,min}/2)-1)^2}.
\end{aligned}
$$

B. The proof of privacy guarantee in theorem 3

Inductively consider the settled set SC, we firstly prove that, if the bit flipping probabilities of SC satisfy ϵ-geo-indistinguishability at the start of an iteration, then the settled set SC satisfies ϵ-geo-indistinguishability at the end of the iteration. Apparently $d'_{a,b} = 2\cdot d_{a,b} - \frac{2}{\epsilon}\log\frac{F_{b,b}}{1-F_{b,b}}$, thus for all possible $a \in S - SC$ and $b \in SC$, let $F_{a,a} = \frac{1}{\exp(-\epsilon\cdot d'_{a,min(SC)}/2)+1}$ as in line 10 and 20, where $d'_{a,min(SC)} = \min\{d'_{a,b} \mid b \in SC\}$, we have:

$$
\begin{aligned}
&\prod_{k=1..|S|} max\{\frac{F_{a,k}}{F_{b,k}}, \frac{1-F_{a,k}}{1-F_{b,k}}\} \\
&= \frac{F_{a,a}}{1-F_{a,a}}\cdot\frac{F_{b,b}}{1-F_{b,b}} \\
&= \exp(\frac{\epsilon\cdot d'_{a,min(SC)}}{2} + \log\frac{F_{b,b}}{1-F_{b,b}}) \\
&\le \exp(\frac{\epsilon\cdot d'_{a,b}}{2} + \log\frac{F_{b,b}}{1-F_{b,b}}) \\
&\le \exp(\epsilon\cdot d_{a,b}).
\end{aligned}
$$

Secondly, the empty set $SC = \Phi$ at starting satisfies ϵ-geo-indistinguishability, hence the final settled set $SC = S$ satisfies ϵ-geo-indistinguishability.