Incentive Mechanism for Participatory Sensing under Budget Constraints

Zheng Song*, Edith Ngai†, Jian Ma*, Xiangyang Gong*, Yazhi Liu‡ and Wendong Wang*
*State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications,
Beijing 100876, China, Email: (sonyyt@gmail.com, majian@mwsn.com.cn, xygong@bupt.edu.cn, wdwang@bupt.edu.cn)
†Department of Information Technology, Uppsala University, Uppsala, Sweden, Email:edith.ngai@it.uu.se
‡Hebei United University, Tangshan 063000, China, Email:liuyazhi991@gmail.com

***Abstract*—Incentive strategy is important in participatory sensing, especially when the budget is limited, to decide how much and where the samples should be collected. Current auction-based incentive strategies purchase sensing data with lowest price requirements to maximize the amount of samples. However, such methods may lead to inaccurate sensing result after data interpolation, particularly for participants that are massing in certain subregions where the low-price sensing data are usually aggregated. In this paper, we introduce weighted entropy as a quantitative metric to evaluate the distribution of samples and find that the distribution of data samples is another important factor to the accuracy of sensing result. We further propose a greedy-based incentive strategy which considers both the amount and distribution of samples in data collection. Simulations with real datasets confirmed the impact of samples distribution to data accuracy and demonstrated the efficacy of our proposed incentive strategy.**

I. Introduction

Participatory sensing employs ordinary citizens to collect and share sensing data from their surroundings using their mobile phones [1]. It has been widely deployed for environmental sensing, such as noise [2] and air quality monitoring [3]. Its advantage is being able to gather widely-spread sensing data with reduced deployment cost.

Incentive mechanism serves as an important basis for environmental participatory sensing applications. The general procedure of incentive mechanism is outlined in Fig.1. The task publishers request to the central server for sensing data of certain geographic regions at specific time periods with incentive budgets. The entire region is divided into a set of subregions in both spatial and temporal dimensions according to the granularity requirement of the tasks [4]. In each subregion, mobile users are selected by the central server as sensing data collector according to the participant selection strategy. The incentive budget is distributed among the selected mobile users. When the incentive budget is not adequate for obtaining data from all the subregions, the missing data will be reconstructed by interpolation methods [5]. Data reconstruction is conducted considering the fact that the environmental data in nearby subregions are usually correlated.

Auction-based approach [6], [7] was firstly proposed as standard incentive mechanisms to maximize the amount of data samples collected by participants. Nevertheless, the auction-based approach suffers from high randomness of price claims in each subregion. The approach ignores the temporal and spatial distribution of the samples, which may cause missing data concentrated in certain subregions and lead to inaccurate data reconstruction results from interpolation.

Different from existing work, we propose a novel incentive mechanism that can achieve *high data quality* considering the *temporal and spatial distribution* of the sensing data. It allocates incentives targeting at minimizing mean error of the sensing result instead of maximizing the number of samples. We study the relations among the average data error, the amount and the distribution of data samples and discover that both larger amount and more even distribution of collected samples can improve the accuracy of sensing result. We formulate incentive allocation as an optimization problem and propose a greedy algorithm to solve it.

The main contributions of the paper are as follows:

1) We introduce an entropy-based metric to evaluate the uniformity of sample distribution. The mean error between the reconstructed data and the ground truth is captured by a function of the amount and distribution of data samples.

2) We formulate an optimization for incentive allocation and propose a greedy-based incentive allocation strategy to minimize the mean error of the sensing data.

3) We demonstrate the procedure and evaluate the performance of the proposed incentive allocation mechanism by simulations using real datasets. The results show that compared with the reversed auction approach, our approach can improve data accuracy by 32%, recruit 42% less participants and provide 72% more incentives to each participant in average.

The rest of the paper is organized as follows. Section 2 describes the related works. Section III describes the application scenario and formulates the mean error of the sensing data. Section IV introduces the entropy metric for measuring the distribution uniformity of data samples, and discusses the relations between the mean error and the amount and distribution of data samples. Sections V and VI present the proposed incentive allocation strategy and evaluate its performances by extensive simulations using real datasets. Finally, Section VII concludes this paper.

II. Related works

Lee et al. [6] firstly introduce reversed-auction-based incentive strategy, the basic idea of which is that the participants collects sensing data and sends data together with their bid

978-1-4799-3083-8/14/$31.00 ©2014IEEE

price to the central server, while the central server selects the lowest bid participants for data collection. Existing researches are mainly based on it, e.g. [7] further considers maximizing the overall coverage of collected samples. However, none of them have considered the distribution of samples.

Mendez et al. [5] first use data interpolation method in data complement, for the collected samples may not fully cover the entire targeted region. The paper further points out that since in a real participatory sensing system the density of the measurements are not uniform, the use of incentives is necessary in order to encourage the users to collect data in the required locations, and in some way, control the density of the measurements per area. This inspires us to dig into controlling the distribution of samples by incentives.

III. System model and notations

A typical participatory sensing system for environmental data collection is shown in Fig.1. It consists of a sensing task in a targeted region, a central server and a set of mobile users walking freely in that region. The aim of the sensing task is to gather sensing data of a specific environmental data type. The budget of the task is B. Table I shows the list of notations used in this paper.

According to [5], when the division of subregions are fine-grained, using one sample per subregion can provide enough sensing data accuracy for a sensing task. The entire sensing region and the entire lasting time of the sensing task are divided into a set $\mathcal{R} = \{r = 1, 2, ..., R\}$ of subregions according to granularity requirement of task publisher. In each subregion, only one sample of the targeted environmental parameter is required.

The procedure of incentive mechanism consists of two phases: First, in each subregion, all participants send the central server their bid price claims, and the lowest incentive requirement (bid claim) in each subregion is denoted as $\mathcal{I} = \{i_r | \forall r \in \mathcal{R}\}$. Second, a set $\mathcal{X} \subseteq \mathcal{R}$ of subregions are selected by the central server. The sum of their incentive requirements should to be less than or equal to the budget, i.e. $\sum_{\forall r \in \mathcal{X}} i_r \leq B$. For each selected subregion, the participant with the lowest bid price takes a sample, then uploads it to the central server and gets paid by the central server. In this paper, we assume that the samples taken by mobile users are close to the ground truth, which means that the measurement error of mobile devices is insignificant.

It is worth noting that in real-world scenarios, there may be no mobile users at all in some subregions, or the entire task budget is not adequate to cover all the subregions, only some of the subregions may have sensing results on them. We use $\mathcal{X}$, i.e. $\mathcal{X} \subset \mathcal{R}$ to represent the subregions that have data samples, and thus $\mathcal{R} \setminus \mathcal{X}$ represents the subregions without data samples. The sensing result in subregions without data samples has to be interpolated from data samples in $\mathcal{X}$. For simplicity, we use a popular and easy-to-implement interpolation method, inverse distance weighting(IDW) [8], as an example of the interpolation method in this paper. The missing data are calculated by a weighted average of the data available at the known points, while the inverse of the distance to each known point is used as weights.

TABLE I
List of notations

Notation	Explanation
$\mathcal{R}$	a set of subregions
i_r	the lowest price on each subregion
B	budget constraint
ε	mean error of sensing result
$\mathcal{X}$	a set of selected subregions. Incentive will be paid to participants on them
g_r	Ground truth of each subregion r
s_r	Sensing result of each subregion r by samples or interpolation
ε	average error of sensing result compared with ground truth
$\alpha(\mathcal{X})$	total amount of samples
$\beta(\mathcal{X})$	distribution of samples

Let $\mathcal{G} = \{g_r | \forall r \in \mathcal{R}\}$ denote the ground truth on all subregions $\mathcal{R}$. For those subregions with data samples, let $\mathcal{S} = \{s_r | \forall r \in \mathcal{X}\}$ denote the sensing results. Since the samples are close to the ground truth, we have $g_r \approx s_r, \forall r \in \mathcal{X}$. For those subregions without data samples, let $\mathcal{S}' = \{s'_r | \forall r \in \mathcal{R} \setminus \mathcal{X}\}$ denote the interpolated results, which can be denoted as Eq.(1) according to IDW.

$$s'_{r_1} = \sum_{\forall r_2 \in \mathcal{X}} d_{r_1 r_2} s_{r_2} \Big/ \sum_{\forall r_2 \in \mathcal{X}} d_{r_1 r_2}, \forall r_1 \in \mathcal{R} \setminus \mathcal{X}, \quad (1)$$

where, $d_{r_1 r_2}$ represent the distance between subregion r_1 and r_2.

We use ε to denote the average error of the sensing results to ground truth of all subregions $\mathcal{R}$, as the error mainly accumulates in subregions without data samples, we have:

$$\varepsilon \approx \sum_{\forall r \in \mathcal{R} \setminus \mathcal{X}} \frac{|g_r - s'_r|}{g_r} / |\mathcal{R}|, \quad (2)$$

where $|\mathcal{R}|$ represents the total number of subregions.

The aim of this paper is to achieve sensing results of high accuracy for the task publishers. It can be seen from Eq.(2) that the average error is related to the subregions with data samples, $\mathcal{S}$. We will discuss in detailed how the distribution of $\mathcal{S}$ generally influences the mean data error ε in the next section.

IV. Mean Error and Correlation with Samples

Intuitively, the accuracy of the interpolated result is related to the amount of samples collected. However, a large amount of data samples does not guarantee better interpolation results. For example, as shown in Fig.2, different distribution of the same amount of samples can lead to different mean errors. We generate 12×12 subregions and the ground truth in each subregion is continuous in the spatial domain. If all data samples are collected in the middle of the entire region, it will lead to a less accurate interpolation result ($\varepsilon = 0.41$), as shown in "Samples Set I". On the contrary, if the distribution of the samples is uniform, a better sensing result can be achieved ($\varepsilon = 0.17$), as shown in "Samples Set II".

Hence, we consider both the amount and the distribution of data samples $\mathcal{X}$, denoted by $\alpha(\mathcal{X})$ and $\beta(\mathcal{X})$ respectively. We introduce how to calculate $\alpha(\mathcal{X})$ and $\beta(\mathcal{X})$ from $\mathcal{X}$ in the following.

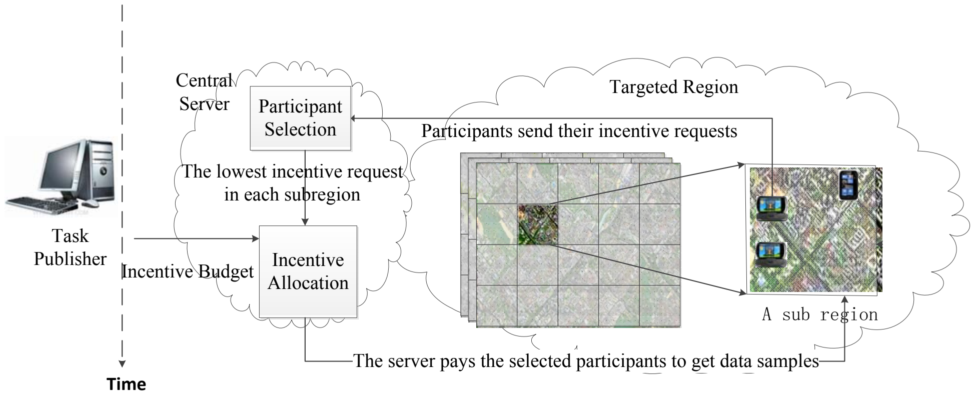

Central Server
Participant Selection
Targeted Region
Participants send their incentive requests
The lowest incentive request in each subregion
Task Publisher
Incentive Budget
Incentive Allocation
A sub region
The server pays the selected participants to get data samples
Time

Fig. 1. The considered participatory sensing scenario showing that the entire targeting area is divided into many subregions according to environmental data requirement of task publisher. A subset of participants in different subregions are selected to carry out the sensing tasks.

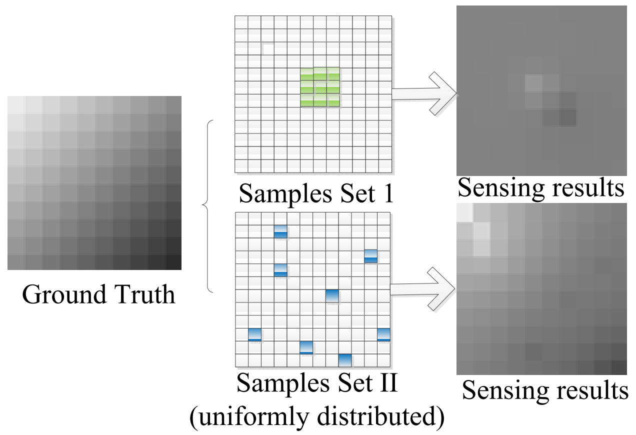


Fig. 2. The impact of distribution of collected data

A. Amount of Samples

As discussed earlier, a set of subregions $\mathcal{X}$ are supposed to be selected and paid. Only one data sample is selected and paid in each subregion, so that the amount of selected subregions with samples is equal to the amount of data samples in the entire region. The amount of samples $\alpha(\mathcal{X})$ can then be calculated by the cardicardal of $\mathcal{X}$ as:

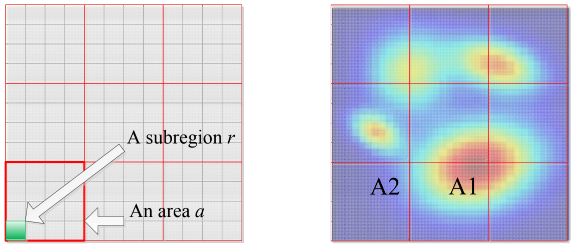
$$\alpha(\mathcal{X}) = |\mathcal{X}|. \tag{3}$$

B. Distribution of Samples

A metric $\beta(\mathcal{X})$ is needed here to indicate the distribution of samples, where higher value of $\beta(\mathcal{X})$ should be given to better sample distribution. To evaluate the distribution of samples, we divide the entire region into several discrete areas. Each area further consists of several subregions as shown in Fig.3(a). The metric $\beta(\mathcal{X})$ should have the following characteristics:

1) In general, if the collected samples are more uniformly distributed in the areas, $\beta(\mathcal{X})$ should indicate a better distribution;

2) The importance of data in different areas may vary with the environmental change. For example, Fig.3(b) shows the reading of noise level in an urban area [5]. The noise level changes tremendously in area “A1”, while changes steadily in area “A2”. If more data are collected in subregion A1, the results from interpolation will be more accurate. intuitively, higher weight should be given to areas with more tremendous environmental change, i.e. area “A1” in Fig.3(b).



(a) The entire region is divideds into several areas, each of which consists of many subregions

(b) The weight of each area is determined by how rapidly the data in this area changes

Fig. 3. The weight of areas could be different in calculating the sample distribution metric

Considering the above characteristics, we introduce weighted entropy to measure the distribution of samples, $\beta(\mathcal{X})$. The entropy of a source alphabet is widely taken as the measurement of uniformity [9], [10]: a source alphabet with n symbols has the highest possible entropy when the probability distribution of the alphabet is uniform. Based on the general form of entropy, weighted entropy is used in information theory when the importance of symbols are different [11].

In our case, the entire region is divided into a set $\mathcal{A} = \{a = 1, 2, ...A\}$ of areas. We use $X_a, \forall a \in \mathcal{A}$ to denote the set of samples that falls in each area. The probability of a sample falls in each area is calculated by:

$$p_a(\mathcal{X}) = \frac{|X_a|}{|X|}, \forall a \in \mathcal{A}. \tag{4}$$

Besides, the weight of each area is determined by how tremendously the environmental data change in different subregions. When the data in an area change steadily, less data samples are required to obtain accurate sensing results. Thus, the weight of each area is calculated by the standard deviation of data readings in each subregion:

$$w_a = \sqrt{\sum_{\forall r \in a} (g_r - \frac{\sum_{\forall r \in a} g_r}{|a|})^2}, \forall a \in \mathcal{A}. \quad (5)$$

Thus, $\beta(\mathcal{X})$ can be calculated by:

$$\beta(\mathcal{X}) = -\sum_{a \in \mathcal{A}} w_a p_a(\mathcal{X}) \log p_a(\mathcal{X}). \quad (6)$$

From Eq.6, the calculation of $\beta(\mathcal{X})$ consists of two parameters, p_a and w_a. However, the calculation procedure of w_a involves $g_r, \forall r \in \mathcal{R}$, which is supposed to be unknown till the incentive allocation procedure is over. In this paper, we consider the environmental data to be spatially and temporally correlated following certain pattern from time to time. Thus, $w_a, \forall a \in \mathcal{A}$ can be obtained from historic sensing results by using the former collected data as g_r. As a result, $w_a, \forall a \in \mathcal{A}$ can be taken as constant, and $\beta(\mathcal{X})$ is only related to the regional distribution $\mathcal{X}_a, \forall a \in \mathcal{A}$ of $\mathcal{X}$.

C. Relationship of ε, $\alpha(\mathcal{X})$ and $\beta(\mathcal{X})$

The change of environmental data depends on both the location and time, so it is hard to predict the precise interpolation error from the amount $\alpha(\mathcal{X})$ and distribution $\beta(\mathcal{X})$ of samples collected. For example, for the simulated data in Fig.2, the distribution of samples is more important than the amount of samples. For the data that have strong correlation in the spatial domain, small amount of uniformly distributed samples can accurately complete the sensing results. On the other hand, for the noise data in Fig.3, the amount of samples plays a more important role.

We observe that there is a stable correlation between the data accuracy and the amount $\alpha(\mathcal{X})$ and distribution $\beta(\mathcal{X})$ of samples. The interpolation accuracy increases with the amount of samples and the uniformity of their distribution. To express this relation, we use a function of $\varepsilon = f\big(\alpha(\mathcal{X}), \beta(\mathcal{X})\big)$ to correlate ε, $\alpha(\mathcal{X})$ and $\beta(\mathcal{X})$. We use Eq.7 as a basic form of function f.

$$\begin{aligned} \varepsilon &= f\big(\alpha(\mathcal{X}), \beta(\mathcal{X})\big) \\ &= c_1 - c_2 * \alpha(\mathcal{X}) - c_3 * \beta(\mathcal{X}), \end{aligned} \quad (7)$$

where c_1, c_2 and c_3 are constants depending on the environmental settings. c_2 and c_3 denote how $\alpha(\mathcal{X})$ and $\beta(\mathcal{X})$ affect ε respectively. We use Eq.7 for two reasons: First, in such form, the changes of ε, $\alpha(\mathcal{X})$ and $\beta(\mathcal{X})$ follow a general trend. Second, as mentioned above, there are still occasions that the sensing data change differently as time passes. Eq.7 can capture these variations with a simple form. Based on the environmental settings, the parameters c_1, c_2 and c_3 can be obtained by curve fitting using historic sensing results. We will demonstrate this procedure in Section V.

V. Incentive Mechanism

The mean data error in Eq.2 can be formulated as the objective function in the optimization problem related to the amount and the distribution of samples. According to Eq.7, we have:

$$\min(\varepsilon) = \min(c_1 - c_2 * \alpha(\mathcal{X}) - c_3 * \beta(\mathcal{X})). \quad (8)$$

To minimize the objective function, we have to find a set of subregions $\mathcal{X}$ that can maximize the following objective function and satisfy the budget constraint:

$$\begin{aligned} &\mathcal{X}^* = \arg\max_{\mathcal{X}} (\alpha(\mathcal{X}) + \frac{c_3}{c_2} * \beta(\mathcal{X})), \\ &s.t.: \sum_{\forall r \in \mathcal{X}} i_r \le B \end{aligned} . \quad (9)$$

We propose an incentive strategy to find the targeted set $\mathcal{X}$ with a greedy algorithm, in which the subregions are selected iteratively. The improvement of accuracy in each subregion is measured by the rate of increase in the objective function in Eq.9 and its incentive request i_r. In each iteration, the most cost-effective subregion will be selected. The proposed algorithm is given in pseudocode (see Algorithm 1) and the detailed descriptions are provided below.

Step 1: Initialization

At the beginning of incentive strategy, all price claims from participants are gathered to form $i_r, \forall r \in \mathcal{R}$. All subregions are divided into two sets, the selected set $\mathcal{S}_1$ and the unselected set $\mathcal{S}_2$. In this step, all subregions are put in the unselected set $\mathcal{S}_2$, and $\mathcal{S}_1$ is set to $\emptyset$.

Step 2: Selection of one subregion in each iteration step

For each subregion r in $\mathcal{S}_2$, if it is selected and moved from $\mathcal{S}_2$ to $\mathcal{S}_1$ to form a new set $\mathcal{S}_1$, the change of the objective function in Eq.9 will be:

$$\begin{aligned} &\alpha(\mathcal{S}_1 + r) + \frac{c_3}{c_2} * \beta(\mathcal{S}_1 + r) - \alpha(\mathcal{S}_1) - \frac{c_3}{c_2} * \beta(\mathcal{S}_1) \\ &= \frac{1}{|\mathcal{R}|} + \frac{c_3}{c_2} * (\beta(\mathcal{S}_1 + r) - \beta(\mathcal{S}_1)) \end{aligned} \quad (10)$$

In subregion r, the lowest incentive request is i_r. If r is selected, the change of the objective function in Eq.9 per unit incentive can be calculated by:

$$\frac{\frac{1}{|\mathcal{R}|} + \frac{c_3}{c_2} * (\beta(\mathcal{S}_1 + r) - \beta(\mathcal{S}_1))}{i_r}. \quad (11)$$

A subregion r that provides the highest efficiency is selected and moved from $\mathcal{S}_2$ to $\mathcal{S}_1$ in each step of iteration. Incentive i_r is allocated to the selected subregion r and paid to the participant with the lowest incentive request to get the data samples.

Step 3: Looping

Repeat step 2 until the given budget B runs out or all the subregions are selected.

VI. Simulations

We conduct extensive simulations using real datasets to evaluate the performance of the proposed incentive strategy.

Algorithm 1 The proposed incentive mechanism

```
Require:
    budget $B$; Subregions$\mathcal{R}$; Lowest incentive request on each subregion $i_r, \forall r \in \mathcal{R}$;
    Areas division$A$; weight of entropy $w_a \forall a \in \mathcal{A}$
Ensure:
    Selected subregion set $\mathcal{X}^*$;
set of unselected subregions $\mathcal{S}_2 = \mathcal{R}$, set of selected subregions $\mathcal{S}_1 =$ NULL
incentive_left $= B$
while 1 do
    flag $\leftarrow 0$
    selected_id $\leftarrow 0$
    max_eff $\leftarrow 0$
    for subregions $r \in \mathcal{S}_2$ do
        if $i_r >$ incentive_left then
            continue
        end if
        compute $r$'s efficiency Eff$_r$ in (11)
        if Eff$_r >$ max_eff then
            selected_id $\leftarrow r$
            max_eff $\leftarrow$ Eff$_r$
            flag $\leftarrow 1$
        end if
    end for
    if flag $= 0$ or selected_id $= 0$ then
        break
    end if
    $\mathcal{S}_1 \leftarrow \mathcal{S}_1 +$ selected_id
    $\mathcal{S}_2 \leftarrow \mathcal{S}_2 -$ selected_id
    incentive_left $\leftarrow$ incentive_left $- i_{\text{selected\_id}}$
end while
Return: final set of selected subregions $\mathcal{X}^* = \mathcal{S}_1$.
```

A. Setup Procedure

We use the environmental sensing data collected by stationary WSNs as ground truth. The dataset was collected by LUCE (Lausanne Urban Canopy Experiment) [12] which was a measurement campaign took place in the EPFL campus. Besides, we use the trajectory dataset [13] collected in a theme park to simulate the mobility of participants. The trajectory dataset contains 41 mobility traces on a particular day. We take the following steps to setup the experiment:

1) Setup of subregions $\mathcal{R}$ and incentive requests i_r: The rectangular sensing field is divided into 4×4 subregions, the entire 24 hours of lasting time is divided into 24 time slots, thus $|\mathcal{R}| = 16 \times 24 = 384$. We feed the mobility traces of the 41 participants into the sensing field. The incentive request of each participants is randomly generated between [1, 20]. We calculate $i_r, \forall r \in \mathcal{R}$ according to the trajectories of the participants and their incentive requests.

2) Calculation of w_a, c_2 and c_3: In the training phase, we use the temperature data collected on December 1st, 2006 as training data to get the ground truth $g_r, \forall r \in \mathcal{R}$. The sensing field is divided into $|\mathcal{A}| = 16$ area. Each area further contains $2 \times 2 \times 4$ subregions. The weight w_a of each area is calculated according to Eq.5. We then generate 40,000 sets of selected subregions of different amount $\alpha(\mathcal{X})$ and different distribution $\beta(\mathcal{X})$. For each set, IDW is used to reconstruct the sensing result for subregions without any samples. We then calculate the mean error ε using the sensing result after interpolation and the ground truth of each set according to Eq.2. Finally, we conduct curve fitting to obtain c_1, c_2, c_3 in Eq.7 using the 40,000 generated sets of α, β and ε.

B. Results

1) The impact of sample distribution to data accuracy: We calculate the mean data error $\varepsilon(\mathcal{X})$ considering different amount and distribution of the 40,000 generated sets of $\mathcal{X}$. Fig.4(a) shows that the distribution of samples significantly impacts the accuracy of the sensing result. Given the same amount of samples, the mean error of uniformly distributed samples can be 92% lower than samples with uneven distribution.

2) Our approach vs. reversed auction: We input all the 40,000 sets of $\varepsilon(\mathcal{X})$, $\alpha(\mathcal{X})$ and $\beta(\mathcal{X})$ into Matlab and use its curve fitting module to calculate c_1, c_2, and c_3 from Eq.7. The result is given by:

$$\varepsilon = 0.3777 - 0.0642\alpha - 0.1247\beta \tag{12}$$

.

From this expression we observe that the sample distribution plays a more important role than the amount of samples in this case. Fig.4(b) shows the predicted error based on Eq.12. Although Fig.4(a) and Fig.4(b) look different, the average distance between the actual error and the predicted error of all the 40,000 sets of samples is 0.021, which is rather small and acceptable.

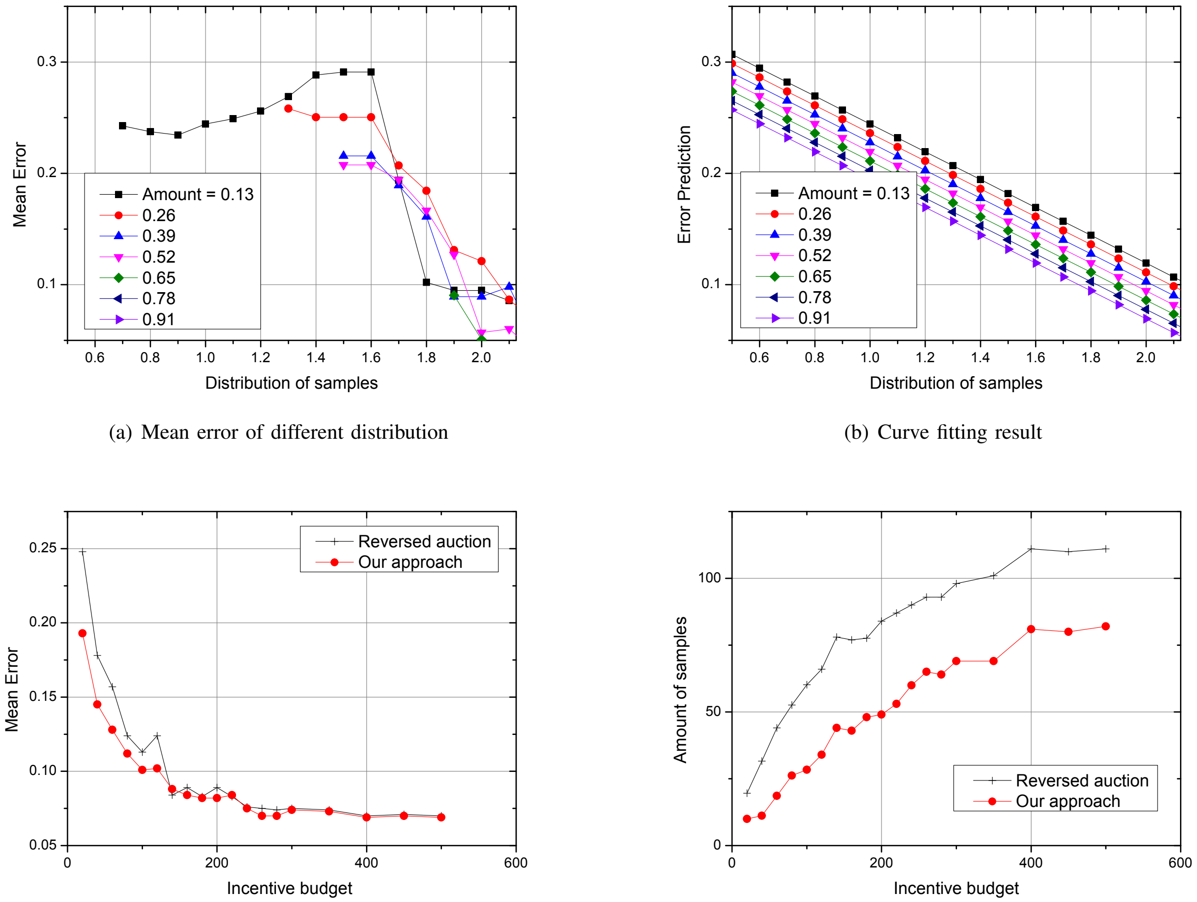
We use the temperature data collected on December 8st, 2006 as ground truth to compare the performance of our approach with the reversed auction approach [6]. The mean error and the amount of selected regions of the two approaches are plotted in Fig.4(c) and Fig.4(d) varying the incentive budget B.

From Fig.4(c) we observe that the difference of mean error between our approach and reversed auction is larger when the budget is inadequate. When the budget is 20, the error of reversed auction is 0.248 while the error of our approach is only 0.193 *(32% less)*. This is because our approach gives higher incentive to collect data in dispersive subregions, while the reversed auction approach only selects the subregions with the lowest incentive requests.

From Fig.4(d) we observe that the amount of selected subregions are much smaller in our approach compared with the reversed auction approach. Given a sufficient budget, the data accuracy of the two approaches are almost the same. Nevertheless, our approach requires 42% *less data samples*, which can reduce the energy consumption of the mobile devices. It also gives 72% more incentive to each participant on average, which can encourage mobile users with high quality data to participate in data collection.

VII. Conclusions

In this paper, we proposed a novel incentive mechanism that can achieve higher data accuracy with constrained budget for participatory sensing. Different from existing incentive strategies, our approach considers not only the total amount of samples but also the distribution of samples in data collection. We studied and formulated the relationship among the mean



(a) Mean error of different distribution

(b) Curve fitting result

(c) Mean error under different budget

(d) Amount of samples under different budget

Fig. 4. Simulation Results

data error, the amount of samples and the distribution of samples. The minimization of mean data error can be converted into an optimization problem for incentive allocation in each subregion. We proposed a greedy incentive allocation algorithm to solve the optimization problem. Extensive simulations with real datasets demonstrated the efficacy of our proposed strategy. Our incentive allocation strategy can increase the data accuracy and the benefits of participants significantly compared with the existing reversed auction approach.

VIII. Acknowledgement

Edith's work is partially supported by SSF ProFun project in Sweden. Others' work was supported in part by the National Natural Science Foundation of China (Grants no. 61271041 and 61202436).

References

[1] J.Burke, D.Estrin, M.Hansen, A.Parker, N. Ramanathan, S. Reddy, and M. Srivastava, "Participatory sensing," in *ACM SenSys'06*, 2006, pp. 1–5.

[2] H. Lu, W. Pan, N. D. Lane, T. Choudhury, and A. T. Campbell, "Soundsense: Scalable sound sensing for people-centric sensing applications on mobile phones," in *MobiSys'09*, 2009, pp. 165–178.

[3] S. Devarakonda, P. Sevusu, H. Liu, R. Liu, L. Iftode, and B. Nath, "Real-time air quality monitoring through mobile sensing in metropolitan areas," in *SIGKDD'13*. ACM, 2013, p. 15.

[4] S. Isaacman, R. Becker, R. Cáceres, S. Kobourov, J. Rowland, and A. Varshavsky, "A tale of two cities," in *HotMobile'10*. ACM, 2010, pp. 19–24.

[5] D. Mendez, M. Labrador, and K. Ramachandran, "Data interpolation for participatory sensing systems," *Pervasive and Mobile Computing*, 2012.

[6] J.-S. Lee and B. Hoh, "Sell your experiences: a market mechanism based incentive for participatory sensing," in *IEEE Percom'10*. IEEE, 2010, pp. 60–68.

[7] L. G. Jaimes, I. Vergara-Laurens, and M. A. Labrador, "A location-based incentive mechanism for participatory sensing systems with budget constraints," in *IEEE Percom'12*. IEEE, 2012, pp. 103–108.

[8] D. Shepard, "A two-dimensional interpolation function for irregularly-spaced data," in *Proceedings of the 1968 23rd ACM national conference*. ACM, 1968, pp. 517–524.

[9] E. J. Dudewicz and E. C. Van Der Meulen, "Entropy-based tests of uniformity," *Journal of the American Statistical Association*, vol. 76, no. 376, pp. 967–974, 1981.

[10] G. Judge and D. Miller, *Maximum entropy econometrics: Robust estimation with limited data*. John Wiley, 1996.

[11] B. X. Guan, B. Bhanu, N. S. Thakoor, P. Talbot, and S. Lin, "Automatic cell region detection by k-means with weighted entropy," in *Biomedical Imaging (ISBI), 2013 IEEE 10th International Symposium on*. IEEE, 2013, pp. 418–421.

[12] G. Barrenetxea, H. Dubois-Ferriere, R. Meier, and J. Selker, "A weather station for sensorscope," *IPSN'06*, 2006.

[13] G. Solmaz, M. I. Akbas, and D. Turgut, "Modeling visitor movement in theme parks," in *Local Computer Networks (LCN), 2012 IEEE 37th Conference on*. IEEE, 2012, pp. 36–43.