QoI-Aware Multitask-Oriented Dynamic Participant Selection With Budget Constraints

Zheng Song, Chi Harold Liu, *Member, IEEE*, Jie Wu, *Fellow, IEEE*, Jian Ma, *Member, IEEE*, and Wendong Wang, *Member, IEEE*

***Abstract*—By using increasingly popular smartphones, participatory sensing systems can collect comprehensive sensory data to retrieve context-aware information for different applications (or sensing tasks). However, new challenges arise when selecting the most *appropriate* participants when considering their different incentive requirements, associated sensing capabilities, and uncontrollable mobility, to best satisfy the quality-of-information (QoI) requirements of multiple concurrent tasks with different budget constraints. This paper proposes a multitask-oriented participant selection strategy called "DPS," which is used to tackle the aforementioned challenges, where three key design elements are proposed. First is the QoI satisfaction metric, where the required QoI metrics of the collected data are quantified in terms of data granularity and quantity. Second is the multitask-orientated QoI optimization problem for participant selection, where task budgets are treated as the constraint, and the goal is to select a minimum subset of participants to best provide the QoI satisfaction metrics for all tasks. The optimization problem is then converted to a nonlinear knapsack problem and is solved by our proposed dynamic participant selection (DPS) strategy. Third is how to compute the expected amount of collected data by all (candidate) participants, where a probability-based movement model is proposed to facilitate such computation. Real and extensive trace-based simulations show that, given the same budget, the proposed participant selection strategy can achieve far better QoI satisfactions for all tasks than selecting participants randomly or through the reversed-auction-based approaches.**

***Index Terms*—Data collection, incentive schemes, participant selection, participatory sensing, quality-of-information (QoI).**

I. Introduction

PARTICIPATORY sensing was first proposed in [1], where the key idea was to have ordinary citizens collect and share sensory data from their surrounding environment by using their smartphones [2]. Early participatory sensing systems, such as PEIR [3] and SoundSense [4], were prototyped for a *single* sensing task or, simply, tasks. They did not explicitly consider the coexistence of multiple concurrent tasks or how to best motivate more users to contribute. Recent approaches such as Campaignr [5] and PRISM [6] can provide multidimensional sensory data simultaneously for multiple concurrent tasks. Meanwhile, the latest research system MEDUSA [7] points out that participatory sensing systems must support ways in which participants can be motivated by incentives to contribute sensory data, since participating in crowd sensing may incur real monetary costs (e.g., bandwidth usage). Therefore, support for multiple sensing tasks with rewards is critical for future participatory sensing systems and is our research path in its own right.

Our research is motivated by the application scenario shown in Fig. 1, which is also derived from the smartphone-based environmental monitoring system described in [8]. It shows a group of mobile users subscribing to a central server, which receives the sensing tasks from task publishers. Each task is associated with certain quality-of-information (QoI) requirements. Broadly speaking, QoI relates to the ability to judge whether information is *fit for use* for a particular purpose [9]–[11]. For the purposes of this paper, we will assume that QoI is characterized by a number of attributes, including the sensing region, sensing time period, data granularity, and quantity requirements, and the incentive budget it is willing to afford. For simplicity, we only consider a sensing region as a 2-D (longitude and latitude) area in the same plane. It is worth noting that the proposed strategy is also suitable for 3-D sensing space with altitude dimension. On the server's side, the targeted sensing region is divided into virtual areas of the same size, according to the task publisher's data granularity requirement. Each task lasts for a certain period of time, which is also divided into several discrete time slots according to the associated data granularity requirement. In each time slot on each area, a number of measurements (sensory data samplings) are required. According to [8], when there are adequate samples, the use of their average value per area, instead of the originals, does not affect the accuracy of the system. On the participant's side, they move around in the sensing region. When they receive the task requirements, they contact the central server to register their current location, required amount of incentive if participating in the data collection, and their sensing capabilities. In our considered scenario, the sensing capability of a participant is measured by how many data samples the user can collect in a time unit for all tasks, which is decided by both the sampling frequency and the type/amount of sensors equipped on his/her smartphone. For simplicity, we assume equal sampling frequency for all participants and, thus, use *the number of sensors*

Manuscript received November 6, 2013; revised February 24, 2014; accepted April 6, 2014. Date of publication April 15, 2014; date of current version November 6, 2014. This work was supported in part by the National Natural Science Foundation of China under Grant 61271041, Grant 61370197, and Grant 61300179. The review of this paper was coordinated by Prof. J. Tang. *(Corresponding author: C. H. Liu.)*

Z. Song, J. Ma, and W. Wang are with the State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: sonyyt@gmail.com; majian@mwsn.com.cn; wdwang@bupt.edu.cn).

C. H. Liu is with the School of Software, Beijing Institute of Technology, Beijing 100081, China (e-mail: chiliu@bit.edu.cn).

J. Wu is with the Department of Computer and Information Sciences, Temple University, Philadelphia, PA 19122 USA (e-mail: jiewu@temple.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TVT.2014.2317701

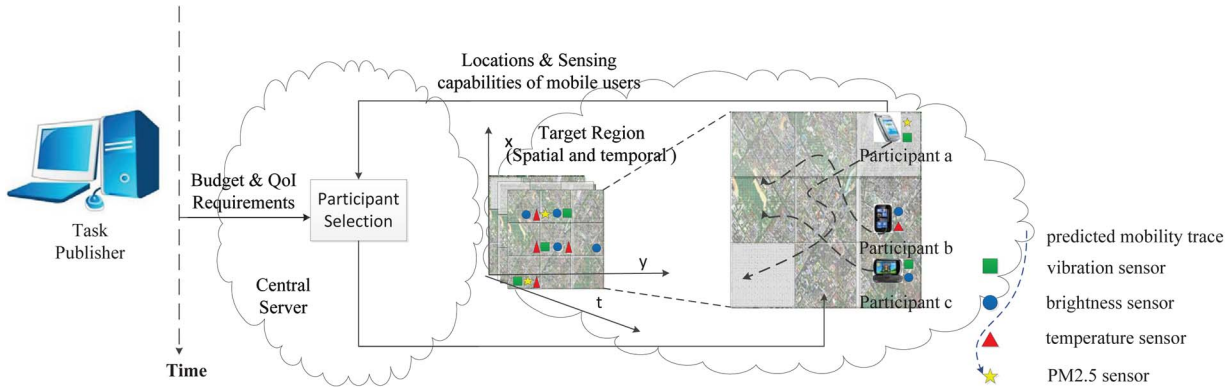

Fig. 1. Considered participatory sensing scenario, where the task publisher request for different types of sensory data is in a targeted region. A subset of all candidate participants is selected from all users with different sensing capabilities, initial locations, and incentive requirements to carry on sensing tasks.

to uniquely represent the sensing capability of a participant's smart device. When a particular participant is selected by the central server as the collector, his/her smart device measures the required environmental parameters periodically and uploads the collected data to the server. As a result, a reward is finally paid as a return. The environmental data collection phase stops either when all QoI requirements are fully satisfied or when the given task budget runs out.

In urban environments, the density of the available participants can be remarkably high. Thus, it is possible and necessary to use only part of all available participants to achieve multiple concurrent tasks. Different participant selection strategies lead to different system performance in terms of the achieved level of QoI satisfactions.

A highly simplified participant selection scene with four sensing tasks is shown in Fig. 1. For each task, samplings are required in (4, 4, 3, 2) areas, respectively. Three users named $\{a, b, c\}$ are walking in the region, where each carries two kinds of sensors. According to their future trajectory, selecting $\{a, b\}$ as participants can collect sensory data in (4, 3, 3, 2) areas for each task, respectively, whereas selecting $\{b, c\}$ can collect sensory data in (4, 4, 1, 0) areas for each task, respectively. Although $\{b, c\}$ can collect data for one more area for the second task, they collect much less data for the third and fourth tasks, compared with $\{a, b\}$. Therefore, in this highly simplified example, selecting $\{a, b\}$ as participants can obviously fit the tasks' requirements better than $\{b, c\}$ and achieve more accurate sensing results. However, in practice, there are far more tasks and participants; to evaluate how a subset of participants can satisfy the requirements of multiple tasks can be quite complicated. Moreover, it is not possible to assume that the trajectories of participants are known *a priori*. Therefore, selecting the most efficient participants to achieve the tasks' QoI requirements by minimum incentives and a constrained budget is the challenge that needs to be addressed.

Participant selection has always been a major challenge in participatory sensing systems, due to the diversity of users' sensing capabilities and incentive requirements. It is even more so when considering their unpredictable mobility pattern and the associated QoI requirements by various tasks. Many researchers proposed different participant selection strategies from different aspects, which either considered only one single sensing task, such as in [12]–[17], or ignored the budget constraint, such as in [18]–[20]. The aim of this paper is to find a subset of participants whose sensory data collection can best satisfy QoI requirements of multiple concurrent tasks in both temporal and spatial dimensions, with a constrained task budget. The contribution of this paper is fourfold as follows.

- We introduce a QoI satisfaction metric in terms of data granularity and quantity, which is used to quantify the degree of how collected sensory data can satisfy multidimensional task QoI requirements.
- We introduce a probability model to estimate the expected amount of data collected on a participant in the sensing region. From the historical trajectory information of a user, his/her probability of moving from one location to another can be calculated. Then, given the transition possibilities between quantized areas and their sensing capabilities, the expected amount of collected data by all participants can be predicted/computed.
- We propose a dynamic participant selection strategy called "DPS." The participants are selected based on a greedy algorithm that explicitly considers their expected amount of data collected, required QoI of multiple tasks, and users' incentive requirements, under the constraint of an aggregated task budget.
- The effectiveness and flexibility of the proposed strategy are extensively evaluated by real-trace-driven simulations.

The rest of this paper is organized as follows. Section II reviews the related research activities. Section III establishes a formal model of our system and Section IV describes the QoI satisfaction metric. Section V formulates the optimization problem of participant selection, describes our proposed DPS in detail, and gives the formal model to estimate the amount of data collected by mobile participants, as an input to DPS. Section VI extensively evaluates the performance of the proposed strategy by real-trace-driven simulations, and finally, Section VII concludes this paper.

II. Related Work

Since Burke *et al.* first proposed the concept of participatory sensing in [1], most early systems have been designed to support some specific task and are not suitable for multiple

tasks. For example, the Common Sense project [21] develops a participatory sensing system that allows individuals to measure their personal exposure to air pollution. Other applications of participatory sensing include the collection and sharing of information about noise pollution [22]. Moreover, vehicle sensing is one important aspect of participatory sensing, for vehicles exhibit better mobility than pedestrians and have no strict limits on processing power and sensing capabilities. Zhou *et al.* in [23] propose a probe-car-based traffic monitoring strategy, followed by an improvement in [24] by recruiting 4000 probe cars (taxis) to cooperatively work. Lee *et al.* in [25] take a further step forward by providing ordinary drivers a framework with which mobile users can participate in traffic monitoring.

As most of these early works do not have a participant selection scheme (i.e., they select participants randomly or just use laboratory workers as testers), they are not suitable for large-scale real-world deployment. Some recent research activities have proposed system models for multitask-oriented participatory sensing systems with reward. PRISM [22] first studies platforms for multiple sensing tasks and proposes a procedural programming language for collecting multiple kinds of sensor data from a large number of mobile phones. MEDUSA [7] synthesizes participatory sensing and crowdsourcing and puts forward a runtime system for multiple sensing tasks with the following stages: task submission, worker selection, and monetary incentives management.

The collaboration and scheduling of sensors is a problem similar to participant selection in wireless sensor networks (WSNs). Krause *et al.* in [26] consider the simultaneous placement and scheduling of sensors and propose an algorithm to decide where and when to place and activate the sensors using the submodularity of the utility function. He *et al.* in [27] consider the quality of sensing as the utility function and uses a greedy algorithm for sensor allocation. Joshi and Boyd in [28] consider minimizing the estimation error as the objective, involving convex optimization in solving it. However, WSNs are quite different from participatory sensing, since participants have uncontrolled mobility patterns and unpredictable incentive requirements.

Some researchers have noticed the lack of participant selection methods. Two representative works, i.e., [12] and [13], use the trajectories of participants in the participant selection phase. In [12], Reddy *et al.* develop a selection framework to enable organizers to identify well-suited participants for data collection, based on geographic and temporal availability as well as participation habits. In [13], Tuncay *et al.* exploit the stability of user behaviors and select participants based on the fitness of mobility history profiles. Similarly, Weinschrott *et al.* in [14] and Zhong and Cassandras in [15] discuss task assignment for opportunistic *in situ* sensing, and Lu *et al.* in [29] focus on initiating sampling around specific location "bubbles" (regions). Gaonkar *et al.* in [30] propose a coverage maximization algorithm that records participants' tracks and selects participants whose availability matches the campaign coverage constraints. Such methods rely heavily on the knowledge of participant trajectories and, thus, may lead to increased risk of mobile users' privacy leakage.

Participant selection in multitask systems is quite different from single-task systems. As far as we are concerned, the work of Duan *et al.* [19] is the first to propose the participant selection method for multiple tasks. Assuming that incentive requests of participants and the utility of sensory data on all locations are known, the proposed method selects a subset of participants, who have the maximum sensory data utility-deducting incentive requirements. In [18], Riahi *et al.* further improve the work in [19] by defining how to calculate sensory data utility on a certain location. Both works concentrate on selecting participants to maximize the difference between the value and the price of sensory data. Another approach [20] is aimed at minimizing the overall sensing cost of mobile devices with heterogeneous sensing capabilities while achieving the sensing tasks' requirements. The authors did not consider the scenario that in suburb areas with few people, the incentive requirement of noncompeting participants could be too high for the task publishers to afford, nor the scenario that sensing resources should lean toward satisfying those tasks affording higher budgets. It is more practical in real-world participatory sensing applications that task publishers offer the maximum incentive budget they are willing to afford and expect the central server (sensing platform) to provide them as high QoI satisfaction as possible.

As an incentive mechanism that also includes participant selection [16], [17], its aim is to obtain the maximum amount of sensory data by minimal payment. In [17], a reversed-auction-based approach is proposed, which gathers the bid price of all participants and selects those with the lowest bid price.

In comparison, we aim to best satisfy QoI requirements of multiple tasks by selecting participants with a constrained task budget. Moreover, we also consider the protection of user privacy, and only assume knowing their historical and current location, but not future trajectories.

III. System Model

This section presents a formal model for describing our participant selection system. We consider a scenario of multitask-oriented participatory sensing in a selected spatial region $\mathcal{L}$ during a particular time period $\mathcal{T}$, as shown in Fig. 1. The system is composed of a central server, a set of task publishers, and a set of M smartphone users moving in region $\mathcal{L}$ during time period $\mathcal{T}$ as candidate participants. They are denoted as $\mathcal{M} \triangleq \{m = 1, 2, \ldots, M\}$. Let q present, for example, an environmental monitoring task in the considered region within that time period and $\mathcal{Q}$ be the collection of all concurrent running tasks. The incentive budget of q is denoted as c_q, and the entire budget of all tasks can be calculated by $C_{\text{total}} = \sum c_q$, $\forall\, q \in \mathcal{Q}$. Moreover, according to the task's QoI requirement, or more precisely, the required data granularity of task q, the task publisher divides the entire region into a set of L areas, denoted as $\mathcal{L}_q \triangleq \{l = 1, 2, \ldots, L\}$, and also divides the entire sensing period into a set of time slots, denoted as $\mathcal{T}_q \triangleq \{t = 1, 2, \ldots, T\}$. On each virtual cube that is composed of a 2-D area and within a certain time slot, the required amount of data samples can be denoted as r_{lt}^q, $\forall\, l \in \mathcal{L}_q$, $\forall\, t \in \mathcal{T}_q$, which is also given by the task publisher as a requirement. Based

TABLE I
LIST OF NOTATIONS

Notation	Explanation
$\mathcal{Q}$	Sensing tasks
$\mathcal{M}$	Participants
c_q	Amount of incentive given by task q
d_m	Amount of incentive required by participant m
s_m^q	Sensing capability of a participant m for task q
$\mathcal{L}_q$	Sensing area division by task q
$\mathcal{T}_q$	Time slot division by task q
$\underline{R}^q$	Data requirement of task q
$\mathcal{X}$	A subset of all participants
$\underline{O}^q(\mathcal{X})$	Collected data by $\mathcal{X}$ for task q
$\underline{U}(\mathcal{X})$	QoI satisfaction vector achieved by $\mathcal{X}$
ω_q	Weight for task q
Δt	Participant's sample frequency
$\mathcal{H}$	Data samplings by a participant
$\underline{P}(\Delta t)$	Position transition matrix
$\underline{E}_m(t)$	Position matrix of participant m at time t
$\underline{E}_m(0)$	Initial location of participant m

on the given budget and QoI requirements of all tasks $\mathcal{Q}$, the central server aims to select part of all candidate participants for data collection at the beginning of $\mathcal{T}$, as the key algorithm we propose in this paper.

As for participants, each candidate demands some kind of monetary or virtual incentives, denoted as $d_m, \forall\, m \in \mathcal{M}$, during $\mathcal{T}$. When a particular participant is selected as the data contributor, his/her device samples the required environmental parameter periodically by the equipped sensor(s). To simplify the scenario and avoid loss of generality, we set an equal sampling frequency for all users, denoted as Δt. Thus, the number of data samples in $\mathcal{T}$ can be calculated by $\lceil \mathcal{T}/\Delta \mathrm{t} \rceil$. Finally, $s_m^q, \forall\, m \in \mathcal{M}, \forall\, q \in \mathcal{Q}$ is used to denote the sensing ability of a participant m to task q, where

$$s_m^q = \begin{cases} 0, & \text{if } m \text{ cannot collect data for } q \\ 1, & \text{otherwise.} \end{cases} \tag{1}$$

As mentioned earlier, we use the number of equipped sensors of a participant's device m to represent his/her sensing capability as S_m, as

$$S_m = \sum_{\forall\, q \in \mathcal{Q}} s_m^q, \quad \forall\, m \in \mathcal{M}. \tag{2}$$

It is worth noting that it is reasonably assumed that the initial locations of each participant m entering the sensory region is known and given, denoted as $E_m(0)$; the detailed definition in practice will be described in Section V-C. Table I shows the list of notations used in this paper.

IV. QUALITY OF INFORMATION SATISFACTION METRIC

As its name implies, the QoI satisfaction metric is used to describe the level of QoI satisfaction to the extent that the collected sensory data can satisfy the requirement of a task (data collection). Suppose that a subset $\mathcal{X}$ (of size $|\mathcal{X}|$) of all participants is selected for a task q and let $o_{lt}^q(\mathcal{X}), \forall\, l \in \mathcal{L}_q$, $\forall\, t \in \mathcal{T}_q$ denote the amount of samplings collected by $\mathcal{X}$ for task q on a certain area l, within a certain time slot t. The initial value of each $o_{lt}^q(\mathcal{X})$ is zero. When a new data sample on area l at time slot t is collected, if the amount of collected data $o_{lt}^q(\mathcal{X})$ is less than the amount of required data r_{lt}^q, then $o_{lt}^q(\mathcal{X})$ is increased by 1; otherwise, if the amount of collected data $o_{lt}^q(\mathcal{X})$ has reached the amount of required data, then r_{lt}^q, $o_{lt}^q(\mathcal{X})$ does not change. Such rules can also be applied to the condition of adding data collection of two different participants m_1 and m_2 together, as

$$o_{lt}^q(m_1 + m_2) = \begin{cases} o_{lt}^q(m_1) + o_{lt}^q(m_2), & \text{if } o_{lt}^q(m_1) + o_{lt}^q(m_2) \le r_{lt}^q \\ r_{lt}^q, & \text{if } o_{lt}^q(m_1) + o_{lt}^q(m_2) \ge r_{lt}^q. \end{cases} \tag{3}$$

Thus, two matrices, i.e., $\underline{R}^q$ and $\underline{O}(\mathcal{X})^q$, are used to denote the multidimensional QoI requirements of task q and the amount of sampling collected by $\mathcal{X}$, respectively. Thus

$$\underline{R}^q = \begin{bmatrix} r_{11}^q & r_{12}^q & \cdots & r_{1T}^q \\ r_{21}^q & r_{22}^q & & \\ \cdots & & \cdots & \\ r_{L1}^q & & & r_{LT}^q \end{bmatrix} \tag{4}$$

$$\underline{O}^q(\mathcal{X}) = \begin{bmatrix} o_{11}^q(\mathcal{X}) & o_{12}^q(\mathcal{X}) & \cdots & o_{1T}^q(\mathcal{X}) \\ o_{21}^q(\mathcal{X}) & o_{22}^q(\mathcal{X}) & & \\ \cdots & & \cdots & \\ o_{L1}^q(\mathcal{X}) & & & o_{LT}^q(\mathcal{X}) \end{bmatrix} \tag{5}$$

where $\forall\, l \in \mathcal{L}, \forall\, t \in \mathcal{T}$, we have

$$o_{lt}^q(\mathcal{X}) = o_{lt}^q\left(\sum_{\forall\, m \in \mathcal{X}} m\right). \tag{6}$$

Then, the following lemma immediately follows from the definition of $\underline{O}^q(\mathcal{X})$.

Lemma 1: Given $\mathcal{X}_1 \subset \mathcal{X}_2$, we have

$$\underline{O}^q(\mathcal{X}_1) \le \underline{O}^q(\mathcal{X}_2), \quad \forall\, \mathcal{X}_1, \mathcal{X}_2 \subset \mathcal{M}, \forall\, q \in \mathcal{Q}. \tag{7}$$

Lemma 2:

$$\underline{O}^q(\mathcal{X}) \le \underline{R}^q, \quad \forall\, q \in \mathcal{Q}, \forall\, \mathcal{X} \subset \mathcal{M}. \tag{8}$$

To best satisfy a sensing task's multidimensional QoI requirements, we aim to minimize the *difference* between the required and attained values (i.e., the defined two matrices $\underline{R}^q$ and $\underline{O}(\mathcal{X})^q$ in our case) by the Frobenius norm [31]. Since the Frobenius norm is mathematically used to measure the spatial length of a matrix, $\|\underline{R}^q - \underline{O}^q(\mathcal{X})\|$ can denote the vector difference for a particular task q. By using $(\|\underline{R}^q - \underline{O}^q(\mathcal{X})\|_F / \|\underline{R}^q\|_F)$, the difference is normalized with the value ranging from 0 to 1. Therefore, for a task q, its achieved QoI satisfaction metric can be computed as

$$u_q(\mathcal{X}) = 1 - \frac{\|\underline{R}^q - \underline{O}^q(\mathcal{X})\|_F}{\|\underline{R}^q\|_F}, \quad \forall\, q \in \mathcal{Q}, \quad \forall\, \mathcal{X} \subset \mathcal{M}. \tag{9}$$

In this way, the achieved QoI satisfaction metric of task q ranges from 0 to 1, where 0 indicates that no data is collected for task q, and 1 means that all considered QoI requirements at each area and within each time slot are fully satisfied. If the collected

data $\underline{O}(\mathcal{X})^q$ do not meet the requirement matrix $\underline{R}^q$, the QoI satisfaction metric can have room for increase when more data are collected. Moreover, the Frobenius norm can also denote the distribution of samplings. When the data collection amount is fixed, the Frobenius-norm-based QoI satisfaction metric shows better results when samplings are uniformly distributed among subregions than when samplings are gathered in a few subregions.

Proposition 1 below states an interesting fact—that given a random set of selected participants, adding a participant into the selected set can never decrease the QoI satisfaction of all tasks.

Proposition 1: Given $\mathcal{X}_1 \subset \mathcal{X}_2$, we have $u_q(\mathcal{X}_1) \leq u_q(\mathcal{X}_2)$.

Proof: Assume that

$$u_q(\mathcal{X}_1) > u_q(\mathcal{X}_2) \Longleftrightarrow u_q(\mathcal{X}_1) - u_q(\mathcal{X}_2) > 0,\ \exists q \in \mathcal{Q}. \tag{10}$$

According to our definition of the QoI satisfaction metric in (9), we have

$$u_q(\mathcal{X}_1) - u_q(\mathcal{X}_2) = \frac{\|\underline{R}^q - \underline{O}^q(\mathcal{X}_2)\|_F}{\|\underline{R}^q\|_F} - \frac{\|\underline{R}^q - \underline{O}^q(\mathcal{X}_1)\|_F}{\|\underline{R}^q\|_F} > 0,\ \exists q \in \mathcal{Q}. \tag{11}$$

According to the definition of Frobenius norm, we have

$$\|\underline{R}^q - \underline{O}^q(\mathcal{X})\|_F = \sqrt{\sum_{\forall l \in \mathcal{L}, \forall t \in \mathcal{T}} \left(r_{lt}^q - o_{lt}^q(\mathcal{X})\right)^2}. \tag{12}$$

According to Lemma 2, we have

$$\begin{aligned} &r_{lt}^q - o_{lt}^q(\mathcal{X}_1) \geq 0 \\ &r_{lt}^q - o_{lt}^q(\mathcal{X}_2) \geq 0 \\ &\exists q \in \mathcal{Q}, \quad \exists l \in \mathcal{L}, \quad \exists t \in \mathcal{T}. \end{aligned} \tag{13}$$

As r_{lt}^q can be taken as a constant here, when $o_{lt}^q(\mathcal{X})$ increases, $r_{lt}^q - o_{lt}^q(\mathcal{X})$ decreases, and $\|\underline{R}^q - \underline{O}^q(\mathcal{X})\|_F$ decreases. Thus

$$\begin{aligned} &\|\underline{R}^q - \underline{O}^q(\mathcal{X}_2)\|_F > \|\underline{R}^q - \underline{O}^q(\mathcal{X}_1)\|_F, \exists q \in \mathcal{Q} \Longleftrightarrow \\ &r_{lt}^q - o_{lt}^q(\mathcal{X}_2) > r_{lt}^q - o_{lt}^q(\mathcal{X}_1), \exists q \in \mathcal{Q}, \exists l \in \mathcal{L}, \exists t \in \mathcal{T} \Longleftrightarrow \\ &o_{lt}^q(\mathcal{X}_2) < o_{lt}^q(\mathcal{X}_1), \exists q \in \mathcal{Q}, \exists l \in \mathcal{L}, \exists t \in \mathcal{T}. \end{aligned} \tag{14}$$

However, according to Lemma 1, given $\mathcal{X}_1 \subset \mathcal{X}_2$, we have

$$o_{lt}^q(\mathcal{X}_2) \geq o_{lt}^q(\mathcal{X}_1), \quad \forall q \in \mathcal{Q}, \quad \forall l \in \mathcal{L}, \quad \forall t \in \mathcal{T}. \tag{15}$$

The contradiction shows that our assumption is fake. As a result, we have

$$u_q(\mathcal{X}_1) - u_q(\mathcal{X}_2) \leq 0 \Longleftrightarrow u_q(\mathcal{X}_1) \leq u_q(\mathcal{X}_2), \quad \forall q \in \mathcal{Q}. \tag{16}$$

Therefore, Proposition 1 is established. ■

V. Problem Formulation and Solutions

The goal of this paper is to find a set of participants whose collected amount of sensory data can best achieve the required QoI for all concurrent tasks being serviced. We denote the targeted set of selected participants as $\mathcal{X}^*$. Moreover, their total incentive requirement should be less than the total available budget C_{total} from all tasks. Hence, the optimization problem is formulated as

$$\begin{aligned} \text{Maximize}: \quad & \underline{U}(\mathcal{X}) = [u_1(\mathcal{X}), u_2(\mathcal{X}), \ldots, u_Q(\mathcal{X})]^T \\ \text{subject to}: \quad & \mathcal{X} \subseteq \mathcal{M}; \sum_{m \in \mathcal{X}} d_m \leq C_{\text{total}} \end{aligned} \tag{17}$$

where $\underline{U}(\mathcal{X})$ is a vector of objective functions. Each element of $\underline{U}(\mathcal{X})$ is the QoI satisfaction metric of the sensory data collected by $\mathcal{X}$.

Until now, the problem of participant selection remains an optimization problem. Different from the existing optimization problems formulated in [20] and [32] that aim at fully satisfying the data collection requirement of tasks, the novel optimization problem (17) treats the total budgets C_{total} of all tasks as constraint for selecting participants and aims at maximizing all tasks' QoI satisfaction, which is of more practical significance, for some data requirements cannot be fully satisfied due to the uncontrollable trajectories of participants.

A. Problem Transformation

Apparently, (17) is a multiobjective optimization (MOO) problem, whose optimal solution may not exist. Then, Pareto optimality can be used to describe solutions for MOO problems. A solution is Pareto optimal if it is not possible to move from that solution and improve at least one objective function, without detriment to any other objective function.

A simple but efficient problem transformation for MOO problems is the weighted-sum method [33]. By using it, one selects scalar weights ω_q for each task $\forall\, q \in \mathcal{Q}$ and minimizes the following composite objective function:

$$\sum_{q \in \mathcal{Q}} \omega_q u_q(\mathcal{X}). \tag{18}$$

If all weights are positive, as assumed in this paper, then minimizing the weighted sum provides a sufficient condition for Pareto optimality, which means the solution that can minimize the weighted transformation is always a Pareto optimal solution for (17).

Specifically, here we use the incentive budget of each task as the weight function to make sure that task publishers who pay more will eventually receive more data collection services, as

$$\omega_q = c_q \Big/ \sum_{q \in \mathcal{Q}} c_q, \quad \forall\, q \in \mathcal{Q}. \tag{19}$$

Hence, the optimization problem can be converted to a single-objective optimization problem to select a set of $\mathcal{X}^*$,

denoted as

$$\begin{aligned}\mathcal{X}^* &= \arg\max_x \sum_{q\in\mathcal{Q}} \omega_q u_q(\mathcal{X}) \\ &= \arg\max_x \sum_{q\in\mathcal{Q}} \omega_q \left(1 - \frac{\|\underline{R}^q - \underline{O}^q(\mathcal{X})\|_F}{\|\underline{R}^q\|_F}\right) \\ &\text{subject to} \quad \mathcal{X} \subseteq \mathcal{M}; \sum_{m\in\mathcal{X}} d_m \le C_{\text{total}}. \end{aligned} \tag{20}$$

B. Proposed DPS Strategy

The objective function of (20) fits the basic form of nonlinear knapsack problem [34]. The knapsack problem is a problem in combinatorial optimizations like the following: Given a set of items, each with mass and a value, determine the number of each item as to be included in a collection, so that the total weight is less than or equal to a given limit, and the total value is as large as possible. Several other knapsack-like problems exist, including the nonlinear knapsack problem [34].

The optimization target of the nonlinear knapsack problem is a function of set x, instead of the total value of x. The general statement of the nonlinear knapsack problem is given by

$$\begin{aligned} \min \quad & f(x) \\ \text{s.t.:} \quad & g(x) \leqslant b \\ & x \in \mathcal{S}. \end{aligned} \tag{21}$$

In our case, we have

$$f(\mathcal{X}) = \sum_{q\in\mathcal{Q}} \omega_q u_q(\mathcal{X}). \tag{22}$$

The decision problem form of the knapsack problem is NP-complete, and thus, greedy algorithms are frequently used to provide a suboptimal approximated solution. They first sort the items in decreasing order of value per unit of weight and then proceed to insert them into the sack, starting with as many of the front items as possible, until there is no longer any space in the sack for more.

The central part of our participant selection strategy is also in line with the heuristic greedy algorithm, which selects the most "efficient" participant in an iterative way. Here, the efficiency of a participant in each round of iteration is computed by the ratio between the increase in the optimization objective function by selecting him/her and the incentive paid to him/her. Specifically, let $\mathcal{X}'$ denote the set of participants that were selected in the previous round, then the efficiency $\vartheta(m, \mathcal{X}')$ of a participant m in this round can be calculated by

$$\vartheta(m, \mathcal{X}') = \left(\sum_{q\in\mathcal{Q}} \omega_q u_q(\mathcal{X}' + m) - \sum_{q\in\mathcal{Q}} \omega_q u_q(\mathcal{X}')\right) \Big/ d_m. \tag{23}$$

An example of how to calculate $\vartheta(m, \mathcal{X}')$ will be given in the detailed description of our proposed DPS, which runs at the beginning of the time period to select participants by rounds of iterations. The pseudocode of DPS is given in Algorithm 1, and a detailed description is given as follows.

Step 1: Initialization. At the beginning of the sensing time period, the participant selection strategy is initialized. All available participants are divided into two sets, i.e., the selected set $\mathcal{A}$ and the unselected set $\mathcal{B}$. In this step, all participants are put in $\mathcal{B}$, and $\mathcal{A}$ is set to $\emptyset$.

Step 2: Select one participant at a time from $\mathcal{B}$ to $\mathcal{A}$. A participant is selected from set $\mathcal{B}$ in each round of iteration.

Given the data collection expectation of each participant, the value of the optimization objective (20) by selecting $\mathcal{A}$ can be calculated by

$$\begin{aligned}&\sum_{q\in\mathcal{Q}} \omega_q \left(1 - \frac{\|\underline{R}^q - \underline{O}^q(\mathcal{A})\|_F}{\|\underline{R}^q\|_F}\right) \\ &\qquad = \sum_{q\in\mathcal{Q}} \omega_q \left(1 - \frac{\left\|\underline{R}^q - \sum_{m\in\mathcal{A}} \underline{O}^q(m)\right\|_F}{\|\underline{R}^q\|_F}\right). \end{aligned} \tag{24}$$

For each participant m in $\mathcal{B}$, if he/she is selected and moved from $\mathcal{B}$ to $\mathcal{A}$ to form a new set $\hat{\mathcal{A}}$, the change $\theta(m, \mathcal{A})$ of the optimization objective (20) is

$$\begin{aligned}\theta(m, \mathcal{A}) &= \sum_{q\in\mathcal{Q}} \omega_q \left(\frac{\|\underline{R}^q - \underline{O}^q(\mathcal{A})\|_F}{\|\underline{R}^q\|_F}\right) \\ &\quad - \sum_{q\in\mathcal{Q}} \omega_q \left(\frac{\|\underline{R}^q - \underline{O}^q(\mathcal{A}) - \underline{O}^q(i)\|_F}{\|\underline{R}^q\|_F}\right), \quad \forall\, m \in \mathcal{B}. \end{aligned} \tag{25}$$

According to Proposition 1, $\theta(m, \mathcal{A}) \ge 0, \forall\, m \in \mathcal{B}, \forall\, \mathcal{A} \subset \mathcal{M}$. The incentive m requests is d_m. For all tasks, a participant's efficiency $\vartheta(m, \mathcal{A})$ in this round can be calculated by the increase in optimization objective (20) divided by his incentive requests, as is denoted by

$$\begin{aligned}\vartheta(m, \mathcal{A}) = \frac{\theta(m, \mathcal{A})}{d_m} &= \left(\sum_{q\in\mathcal{Q}} \omega_q \left(\frac{\|\underline{R}^q - \underline{O}^q(\mathcal{A})\|_F}{\|\underline{R}^q\|_F}\right)\right. \\ &\quad \left. - \sum_{q\in\mathcal{Q}} \omega_q \left(\frac{\|\underline{R}^q - \underline{O}^q(\mathcal{A}) - \underline{O}^q(i)\|_F}{\|\underline{R}^q\|_F}\right)\right) \Big/ d_m. \end{aligned} \tag{26}$$

The selected participant of each round is denoted by (27) and is moved from $\mathcal{B}$ to $\mathcal{A}$, i.e.,

$$\arg\max_m \vartheta(m) = \frac{\theta(m, \mathcal{A})}{d_m}, \quad \forall\, m \in \mathcal{B}. \tag{27}$$

Step 3: Looping. Loop step 2, until the given budget for this unit sensing time period can afford no more participants or the QoI satisfaction metrics of all tasks reach 1 (fully satisfied).

It is worth noting that other optimization algorithms, e.g., dynamic programming and simulated annealing, are also applicable in solving optimization problem (20). However, these methods are more time consuming. In large-scale sensing applications that involve a huge amount of participants and tasks, dynamic programming and other algorithms may fail

to meet the time latency requirement of real-time participant selection.

Algorithm 1 DPS Algorithm

Require:
tasks $\mathcal{Q}$; incentive of tasks c_q,
area and time division of tasks $\mathcal{L}_q$, $\mathcal{T}_q$,
data requirement of each task $\underline{R}^q, \forall\, q \in \mathcal{Q}$;
participants $\mathcal{M}$;
incentive requirement of each participant d_m,
sensing capability of each participant s_m^q,
initial locations of participants $\underline{E}_m(0), \forall\, m \in \mathcal{M}$;
interval between samplings Δt;
transition matrix obtained from historic traces $\underline{P}(\Delta t)$.
Ensure:
Selected participants as set $\mathcal{X}^*$;
set of unselected participants $\mathcal{B} = \mathcal{M}$, set of selected participants $\mathcal{A} = \text{NULL}$
$\texttt{incentive\_left} = C_{\text{total}}$
while 1 **do**
  $\texttt{flag} \leftarrow 0$
  $\texttt{selected\_id} \leftarrow 0$
  $\texttt{max\_efficiency} \leftarrow 0$
  for mobile user $m \in \mathcal{B}$ **do**
    if $d_m > \texttt{incentive\_left}$ **then**
      continue
    end if
    compute m's efficiency $\vartheta(m, \mathcal{A})$ in (26)
    if $\vartheta(m, \mathcal{A}) > \texttt{max\_efficiency}$ **then**
      $\texttt{selected\_id} \leftarrow m$
      $\texttt{max\_efficiency} \leftarrow \vartheta(m)$
      $\texttt{flag} \leftarrow 1$
    end if
  end for
  if $\texttt{flag} = 0$ or $\texttt{selected\_id} = 0$ **then**
    break
  end if
  $\mathcal{A} \leftarrow \mathcal{A} + \texttt{selected\_id}$
  $\mathcal{B} \leftarrow \mathcal{B} - \texttt{selected\_id}$
  $\texttt{incentive\_left} \leftarrow \texttt{incentive\_left} - d_{\texttt{selected\_id}}$
end while
Return: final selected participant set $\mathcal{X}^* = \mathcal{B}$.

C. Expected Amount of Collected Data by Participants

When a particular participant is selected as the data contributor, he/she measures the required environmental data periodically by the sensor(s) embedded on his/her smartphone. Recall that the sampling frequencies of different users are set equal to Δt, and then, the time points when a participant m takes samples of environmental parameters can be denoted as $\mathcal{H} = \{h = 1, 2, \ldots, H\}$, $H = \lceil \mathcal{T}/\Delta \mathrm{t} \rceil$. The relationship between the time slots during a task's lifetime and the epochs when participants take samplings are demonstrated in Fig. 2, where $|\mathcal{H}|$ samplings are uniformly distributed in a total of $\mathcal{T}$ time slots.

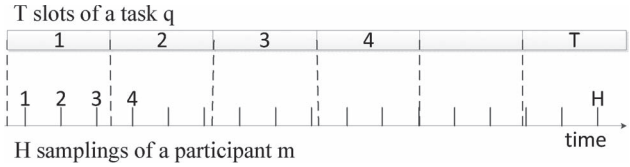

Fig. 2. Temporal relationship between time slots of tasks and samplings of participants.

To estimate how much data a participant can collect requires the knowledge of both his/her sensing capability and the areas he/she is in when the samplings are taken. Recall that all participants have registered their sensing capability and initial locations to the central server; we adopt a probability-based method [35] for estimating future locations of participants when they move around the sensing region, instead of trajectory prediction methods, for existing trajectory prediction methods suffer from rapid loss of accuracy when being applied to predict time-lasting movements [36], [37]. Specifically, their historical trajectories are used to calculate the probability to move from one location to another after a certain period of time. The assumption that participants' historical trajectories are known to the central server is reasonable, for most participatory sensing applications require the collected sensory data to be labeled with time and location information, as well as collector's ID [12], [22], some collaborative sensing application requires potential participants to periodically upload their GPS information [32], and some participatory sensing systems specifically collect the trajectories of participants [38].

Let p_{l_1,l_2} denote the probability that a user moves from area l_1 to l_2 within time interval Δt, and accordingly, let $P(\Delta t)$ denote the position transition matrix, computed as

$$\underline{P}(\Delta t) = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1L} \\ p_{21} & p_{22} & & \\ \cdots & & \cdots & \\ p_{L1} & & & p_{LL} \end{bmatrix}. \tag{28}$$

For convenience, let $\delta_l(t)$ denote the possibility that a participant appears in area l, $\forall\, l \in \mathcal{L}$ at time point t, and let $\underline{E}_m(t)$ denote the possibility that a participant m appears in $\mathcal{L}$ at t, where $\underline{E}_m(t) = [\delta_1(t), \delta_2(t), \ldots, \delta_L(t)], \forall\, m \in \mathcal{M}$. The initial location of the participant can be denoted as $\underline{E}_l(0)$, where

$$\delta_l(0) = \begin{cases} 0, & \text{if the participant is not in subarea } l \\ 1, & \text{if the participant is in subarea } l \end{cases} \qquad \forall\, l \in \mathcal{L}. \tag{29}$$

Based on [35], $\underline{E}_m(\Delta t) = \underline{E}_m(0) \times \underline{P}(\Delta t)$.

Theorem 1: Given the position matrix $\underline{P}(\Delta t)$, a participant m's data collection expectation $\underline{Q}^q(m)$ of task q can be calculated by his sensing abilities s_m^q and his initial location $\underline{E}_m(0)$.

Proof: For a particular participant m, the possibility that he/she takes a sample on each areas in $\mathcal{L}$ at sampling time $h \times \Delta t$ can be calculated by

$$\begin{aligned} \underline{E}_m(h \times \Delta t) &= \underline{E}_m\left((h-1) \times \Delta t\right) \times \underline{P}(\Delta t) \\ &= \cdots = \underline{E}_m(0) \times \underline{P}(h \times \Delta t), \quad \forall\, h \in \mathcal{H}. \end{aligned} \tag{30}$$

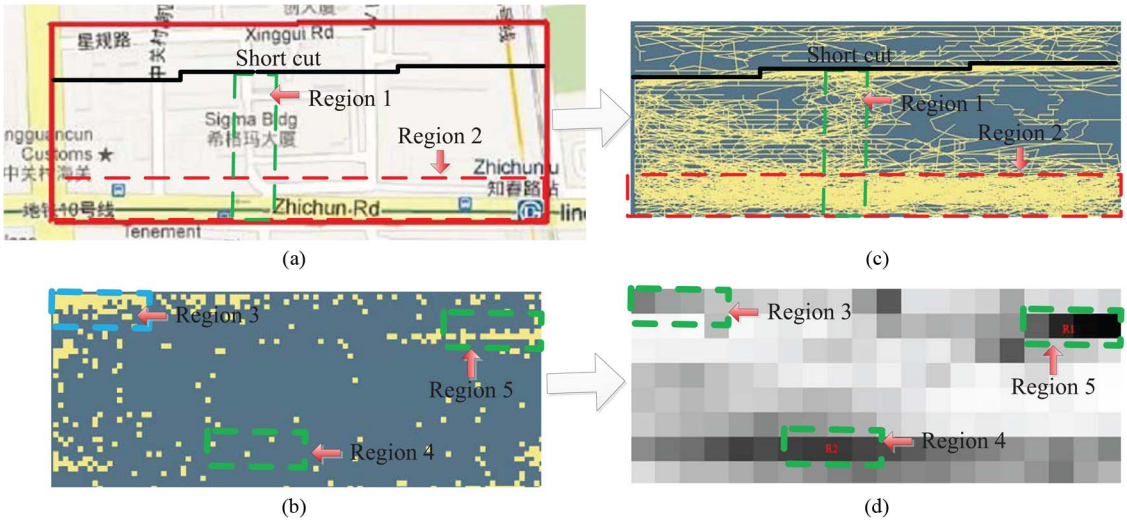

Fig. 3. Simulation setup and observations. (a) Simulation region (red rectangle). (b) Initial locations of mobile users. (c) User trajectories. (d) The possibility that a user stays in an area.

Knowing $\mathcal{T}$ and Δt, the expected number of samplings on each area within time slot t can be calculated as the sum of possibilities of all samplings taken during time slot t, as

$$[o_{1t}^q(m), o_{2t}^q(m), \ldots, o_{Lt}^q(m)] = \sum_{\Delta t \times h \geq \mathcal{T}_{t-1}}^{\Delta t \times h \leq \mathcal{T}_t} \underline{E}_m(h \times \Delta t) \times s_m^q. \tag{31}$$

Based on its definition, $\underline{O}^q(m)$ can be expressed by

$$\underline{O}^q(m) = [\underline{O}_1^q(m), \underline{O}_2^q(m), \ldots, \underline{O}_T^q(m)]^T \tag{32}$$

where

$$\underline{O}_t^q(m) = [o_{1t}^q(m), o_{2t}^q(m), \ldots, o_{Lt}^q(m)], \quad \forall\, t \in \mathcal{T} \tag{33}$$

and superscript "T" denotes the transposed set. Thus

$$\begin{aligned}
&\underline{O}^q(m) \\
&= \left[\sum_{h\Delta t \geq \mathcal{T}_0}^{h\Delta t \leq \mathcal{T}_1} \underline{E}_m(h \times \Delta t), \ldots, \sum_{h\Delta t \geq \mathcal{T}_{t-1}}^{h\Delta t \leq \mathcal{T}_t} \underline{E}_m(h \times \Delta t) \right]^T \times s_m^q \\
&= [\underline{E}_m(0) \times \kappa(1), \underline{E}_m(0) \times \kappa(2), \ldots, \underline{E}_m(0) \times \kappa(T)]^T \times s_m^q \\
&= \underline{E}_m(0) \times [\kappa(1), \kappa(2), \ldots, \kappa(T)]^T \times s_m^q
\end{aligned}$$

where

$$\kappa(\mathcal{T}_t) = \sum_{\forall h \in \{h | \mathcal{T}_{t-1} \leq h\Delta t \leq \mathcal{T}_t\}} \underline{P}(h \times \Delta t), \quad \forall\, t \in \mathcal{T}. \tag{34}$$

Thus, Theorem 1 is established. ■

Based on Theorem 1, when a participant's initial location and sensing capabilities are registered to the central server, his/her expected amount of collected data for all tasks can be calculated.

VI. Performance Evaluation

A. Setup

We assess the proposed DPS scheme by using the (Microsoft Research Asia) GeoLife data set [38], where real movement traces of ordinary citizens are used to represent mobile users in the considered scenario. The GeoLife project has collected 182 volunteers' trajectories in Beijing for three consecutive years. Each trajectory is marked by a sequence of time-stamped GPS points that contain users' latitude, longitude, and altitude at a given time. We adopt the following procedures to set up our simulation platform.

- As all traces were spread in different parts of Beijing, a specific rectangular region where the traces mostly appear is needed. We store all trajectories in a geographical MySQL database and find a 200 × 500 m^2 region that is of high movement density, as shown in Fig. 3(a), that happens to be around the area of the Microsoft Research Asia site. We use this region as the simulation area for the considered data collection application.
- If not specially mentioned in the following experiments, three tasks are simulated in the region, i.e., $|\mathcal{Q}| = 3$. For simplicity reasons, we consider the data granularity requirement of all tasks to be the same. Thus, for all tasks, the entire region is divided into 8 × 20 areas of 25 × 25 m^2, i.e., $|\mathcal{L}_q| = 160$, $\forall\, q \in \mathcal{Q}$. Moreover, by setting $|\mathcal{T}_q| = 10$, the lifetime of all tasks is composed of ten time slots. For each area, the required amount of data in a time slot is set to be 5 ($r_{lt}^q = 5, \forall\, q \in \mathcal{Q}, l \in \mathcal{L}, t \in \mathcal{T}$). Since a participant's incentive requirement could be realized in different formats in practice, such as real money or bonus points, we use dimensionless units to represent both the participants' incentive requests and the tasks' budget constraints. The default budget of each task is set to be 200 or $c_q = 200$, $\forall\, q \in \mathcal{Q}$.
- All 618 trajectories in the considered region are taken as potential (candidate) participants, i.e., $|\mathcal{M}| = 618$, as

shown in Fig. 3(c). Since these traces are recorded at different times, in our simulation, we simply neglect their time index and overlay them into the same time period. For each mobile user, given that the best GPS accuracy is about 5 m, we divide the entire region into 40×100 subregions, where each subregion covers an *area* of 5×5 m^2, and we identity them as the mobile users' locations instead of the original GPS coordinates. The first GPS record of each trajectory that falls into the aforementioned simulation region is used as the initial location of a mobile user. Fig. 3(b) shows initial locations of all 618 users. The total number of samplings $|\mathcal{H}|$ of each participant is set to 100, and the next 100 GPS records of each trajectory are used to represent the actual locations of samplings. In addition, users' sensing capabilities are randomly generated as a uniformly distributed random variable with which each user has 50% likelihood of carrying the sensor for each task, as $s_m^q = rand(0, 1)$, $\forall\, m \in \mathcal{M}$, $q \in \mathcal{Q}$. If not specifically mentioned in the following experiments, their incentive requirements d_m for participating in the entire sensing time period are also randomly generated from 1 to 20 units.

- To construct the location transition matrix $\underline{P}(\Delta t)$, we analyze the adjacent movement of all 618 trajectories from one point to another. Each square in the figure represents an area l, $\forall\, l \in \mathcal{L}$, and its gray value denotes the summed possibility for a participant to appear in this area from any initial location. It can also be regarded as the average amount of time that a participant spends in a specific area during the duration of simulation.

Collectively, it is interesting to put Fig. 3(a)–(d) together, and we obtain the following observations.

- The concentrated trajectories in "Region 1" [see Fig. 3(a) and (c)] are exactly the main road to the Sigma Building, as the office building of Microsoft Research Asia, where employees spend most of their day.
- The "yellow strip" in "Region 2" [see Fig. 3(a) and (c)] corresponds to the very busy Zhichun Road, where traffic is always high.
- As shown in Fig. 3(a), there is a shortcut from the west to the east, composed of three road sections between a couple of residential areas. Its west point connects another busy road, i.e., the 4th Road of South Zhongguancun. This observation is confirmed by trajectories in Fig. 3(c).
- "Region 3" [squared by the blue dashed line to the northwest corner; see Fig. 3(b)] has the highest population density if considering their initial locations. However, as shown in Fig. 3(d) for the average sojourn time, most of them are just pass-by users and will leave the simulation region soon after.
- Fig. 3(c) also shows that the density of participants' initial locations are quite high in "Region 5," where users spend most of their time [see Fig. 3(d)]. Through a field trip, we find that "Region 5" is a newly built residential community with outdoor fitness facilities (where the elderly like to be), and it is farthest from the main roads.
- Fig. 3(c) shows that "Region 4" has a relatively lower density, but Fig. 3(d) shows that many participants enter into that area and spend quite some time here. From the map, we see that "Region 4" is the exact spot of the Sigma Building.

All these features observed from the real-world trajectories further confirm the necessity to consider users' mobility patterns for participant selection.

B. Implementation

We refer to the proposed scheme as "DPS," and to compare the system performance, two other participant selection schemes are simulated, namely: 1) the random selection method (referred to as "RS") is considered as the benchmark, and 2) a reversed-auction-based method (referred to as "RA"), which is slightly modified from the existing algorithm RADP [17]. RS selects participants randomly until the total incentive budget runs out; RA is slightly modified from RADP [17] as to better fit our scenario. That is, the basic idea of RADP is to select participants who can provide higher sensing capabilities with a unit incentive request. According to our definition of participants' sensing capability in Section I, namely, the number of equipped sensors, RA used for comparison purposes select participants who have higher ratios between the number of equipped sensors and the incentive requirements, until the total incentive budget runs out. All three schemes and environmental settings are written by script files in the PHP programming language.

C. Results

We first show the accuracy of our probability-based data collection method. Ten randomly selected participants are taken as a test set, and their real-trajectory-based data collection in the first time slot is calculated. Meanwhile, their data collection expectations in the first time slot are also calculated using the generated location transmission matrix and their initial locations. In each round of the experiment, a location transmission matrix is generated by randomly selected 10, 50, 100, 200, . . ., 600 participants. The accuracy of each round is shown in Fig. 4, and we can observe that the average accuracy of data collection estimation rapidly increases when few trajectories are taken as the training set and finally reaches 77% when all participants are involved in the training set.

We show the running process of our proposed approach when the incentive budget of each task is given as 200. In each round of iteration, the efficiencies of all unselected participants are calculated, and the participant with the highest efficiency is selected and paid. Fig. 5 shows the efficiency of the selected participant in each round of iteration. We observe that, when 600 allowed incentives are given from three tasks, 107 participants are selected. Moreover, the predicted efficiency of the selected participant in each step sharply decreases in the first 20 steps, followed by a long tail after 40 steps, where 511 unselected participants' efficiencies are lower than 1. This implies that our approach can provide considerable QoI satisfaction for all sensing tasks, although the budget is quite limited.

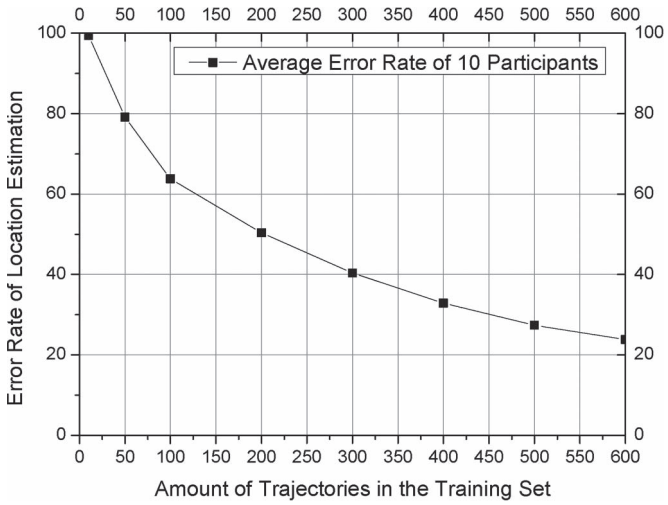

Fig. 4. Error rate of data collection estimation under different training sets.

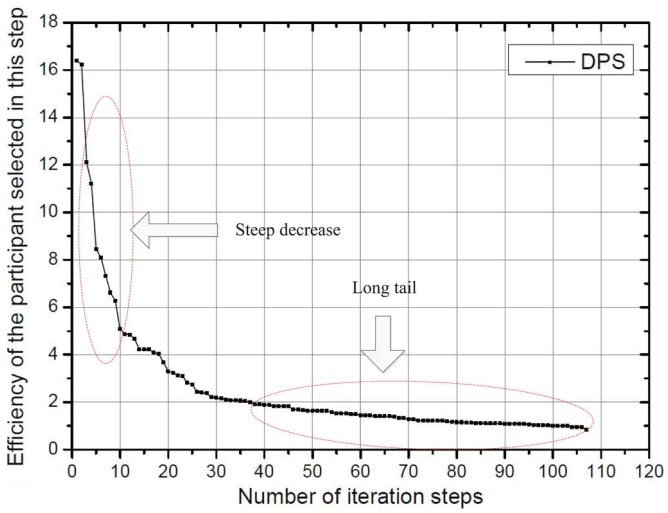

Fig. 5. Efficiency of the participant selected in each round of iteration.

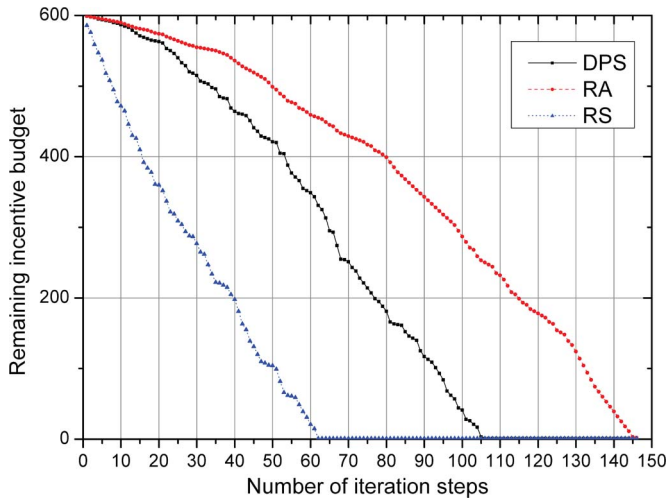

Fig. 6. Remaining incentive budget after each round of iteration.

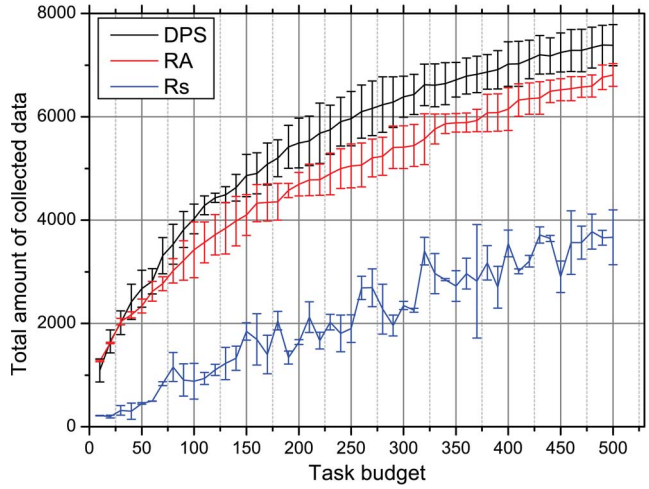

Fig. 7. Impact of task budget on the total amount of collected data.

We further show the remaining budget after a participant is selected and paid in each iteration (see Fig. 6). The remaining budget of the RS scheme decreases at an almost fixed rate, and the budget runs out after 62 steps. The remaining budget of our DPS scheme decreases at a much lower rate, since it considers both increasing the level of QoI satisfaction and satisfying the participant's incentive requirement, and it also runs out after 105 steps. The remaining budget of the RA scheme decreases at a lowest rate, since it only selects the cheapest participant without QoI guarantees, and it runs out after 145 steps.

In the experiment, when participants are selected, we are able to calculate their total collected data by their known trajectories. Next, we use the amount of collected data, the number of selected participants, and data uniformity as three indicators to evaluate their performance. Here, data uniformity is measured by the proposed method in [39]. That is, the entire region contains n areas ($n = |\mathcal{L}| = 160$ in our case), and let a denote the total amount of collected data. Then, n' ($n' = 50$ in our case) uniformly distributed areas are randomly generated, and the number of samplings on these areas is denoted as a' accordingly. If n' areas are randomly generated by enough times (e.g., 100 in our case), the average difference between a/n and $'/n'$ of each run will show the degree of inhomogeneity of sampling distribution. Thus, the disuniformity index of samplings can be denoted as

$$\sum_{i=1}^{100} \left| \frac{a'(i)}{n'} - \frac{a}{n} \right|. \tag{35}$$

The larger the disuniformity index is, the more inhomogeneous the distribution of collected samplings is.

First, we study the impact of tasks' total incentive budget on different approaches. We randomly generate 30 different combinations of incentive requests and sensing capabilities (i.e., the number of equipped sensors) for all participants. For each combination, we increase the allowed amount of incentives from each task every ten units, i.e., from 10, to 20, to 30, and so on, until it reaches 500 units.

Fig. 7 shows the total amount of collected data by three participant selection strategies. We observe that the amount of collected data by RA and DPS methods is significantly larger (>200%) than that of RS. Moreover, DPS exhibits 14.2% more data than that of RA. Furthermore, the overall trend for the amount of collected data, with respect to the task budget, is approximately linear for RS, whereas the amount of collected

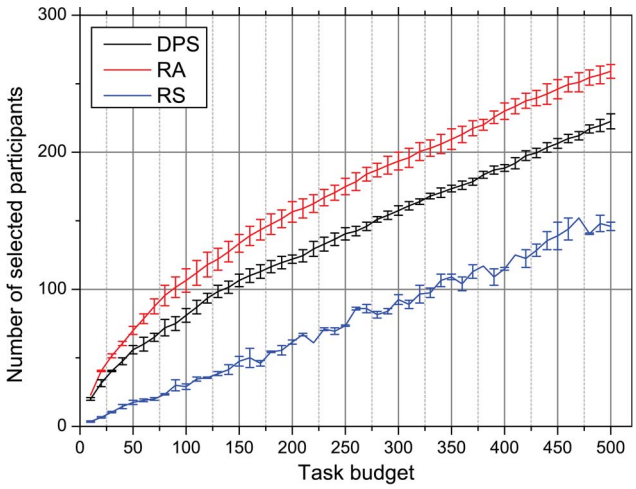

Fig. 8. Impact of task budget on the total number of selected participants.

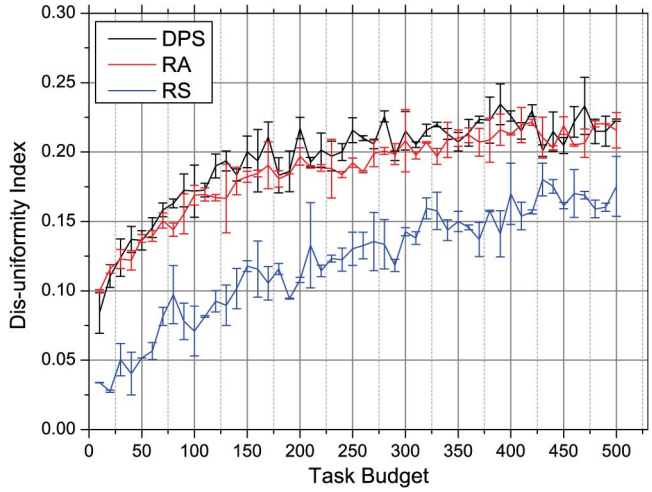

Fig. 9. Impact of task budget on the spatial distribution of the collected data when incentive requests are uniformly distributed among areas.

data rapidly increases for DPS and RA when the total task budget is inadequate. Thus, it is clear that DPS performs better in collecting more data (and, therefore, has better QoI experience), particularly under the condition that the total incentive is tight, consistent with what has been discussed in Section VI-B.

Fig. 8 shows the impact when changing the task budget on the total number of selected participants, where we observe that, for fixed task budget, our DPS method involves significantly fewer participants, i.e., 19.3% fewer than that for the RA scheme. However, compared with RS, DPS involves 118% more participants. Since it is fair to relate the number of participants with the total energy consumption, we can safely conclude that DPS uses nearly 20% more energy if compared with the RA scheme. This is because in each round of iteration, RA selects only the participants with lowest incentive requests, whereas DPS selects those considering both the incentive requests and data collection efficiencies, and thus, DPS recruits less participants, which is consistent with Fig. 6.

Fig. 9 shows the trend of the defined disuniformity of the collected data. It can be seen that RS achieves the best data uniformity, whereas DPS and RA behave closely. The indexes of DPS and RA rapidly increase when the budget is tight, which means that the collected data are more nonuniformly distributed in both temporal and spatial dimensions. However, when the budget reaches a certain threshold (200 from each task, as shown in the figure for the DPS scheme), the uniformity measurement stops increasing. This saturation is simply due to the spatially nonuniformity distributed participant trajectories—that when more budget is given, more participants are selected, thus leading to a higher degree of disuniformity. However, when all participants are given enough rewards, no more participants can be chosen, and thus, the defined disuniformity index stops increasing.

We further demonstrate the uniformity of data collection by DPS, RS, and RA approaches, when setting different incentive requests for different areas, i.e., incentive requirements are not uniformly distributed in the spatial dimension. Recall that in Fig. 3(d), many mobile users spend much time in residential Region 5 and office Region 4; we therefore set higher incentive requirements (ranging from 10 to 20) for the 157 mobile users

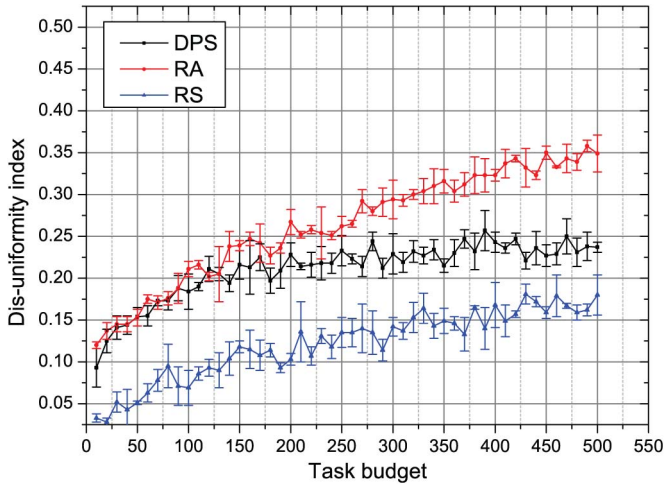

Fig. 10. Impact of task budget on the spatial distribution of the collected data when incentive requests are not uniformly distributed among areas.

near Region 4, and lower incentive requirements (ranging from 1 to 10) for the other 195 mobile users near Region 5. This is because users at home may have more time to contribute, whereas employees will have to be paid more to get them involved. Fig. 10 shows the change in the disuniformity index of the collected data. It can be seen that RA is greatly affected by spatially different requested incentives, because it tends to select incentive-efficient mobile users in Region 5. Meanwhile, DPS is also affected, but the average value of the disuniformity index just slightly increases by 0.023. When the overall task budget is quite tight, participants selected by DPS and RA are both randomly distributed in the spatial domain, since the incentive requests of most participants are randomly generated. However, when more budget is given, RA rarely selects those users from the regions in which their incentive requests are high, but our proposed DPS scheme is driven to select expensive participants in those regions by Frobenius norm in (9); as a result, the achieved QoI metric increases much faster when the required data collection resides in those regions having less data collected.

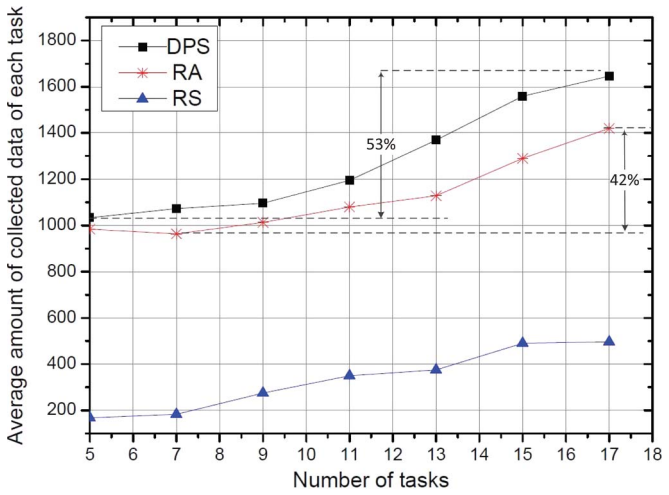

Fig. 11. Impact of the number of tasks on the average amount of collected data of each task.

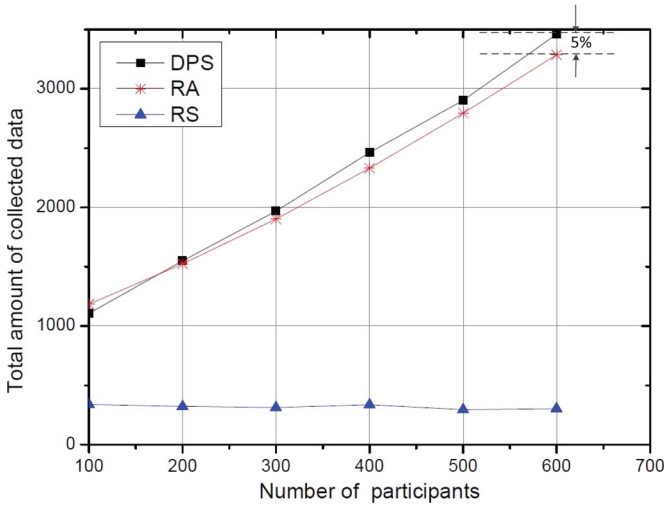

Fig. 12. Impact of the number of participants on the total amount of collected data.

Then, we verify our proposed DPS's performance while changing the number of tasks, as shown in Fig. 11. It can be seen that when 5–17 tasks are required, the average amount of collected data of all tasks obtained by DPS increases by 53%, whereas that of RA increases by only 42%, which indicates that DPS is more efficient when handling more tasks. This is because the total task budget increases when the number of task increases. With this extra budget, DPS can select more efficient participants to collect data, rather than randomly selecting users by RA, as consistent with Fig. 7.

Moreover, we conduct an experiment to verify the impact of the number of participants on DPS. We randomly select 100, 200, . . ., 600 trajectories as candidate participants. We decrease the budget of each task to 100, to motivate the selection of a subset of participants only. It is shown in Fig. 12 that the amount of collected data of RS does not increase with the increase in the number of participants, while the amount of collected data of DPS and RA increases. It indicates that both DPS and RA can select more efficient participants from the incremental extension of candidates, and thus, the achieved QoI satisfaction index will also increase.

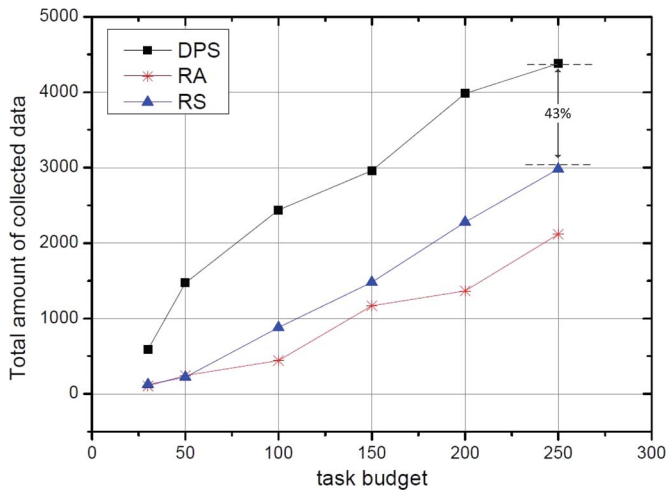

Fig. 13. Impact of task budget on the amount of collected data, when incentive request proportionally varies to participant's sensing capability.

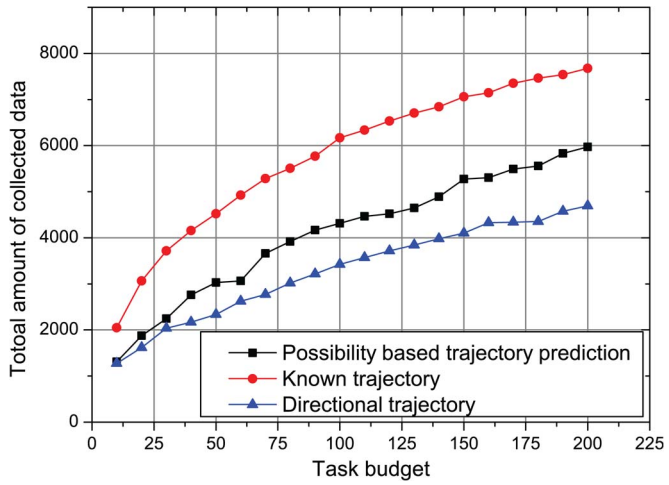

Fig. 14. Impact of different trajectory prediction methods on the amount of collected data.

Finally, we verify the impact of the participant's incentive requirement. In practice, most participants tend to request more rewards in return for contributing more sensory data. An extreme scenario is then simulated, where all incentive requirements are exactly proportional to their respective sensing capabilities or the number of equipped sensors on their smart devices. Fig. 13 shows that the amount of collected data from the RA scheme is lower than that from RS, whereas our proposed DPS still performs 43% better than RS. Higher QoI satisfaction is achieved by not only choosing participants whose incentive requests are relatively low, but more importantly, it fully considers how to best fit all participants' sensing capabilities to the task QoI requirements.

Finally, since our proposed DPS scheme selects participants based on their future trajectory estimations, this prediction accuracy eventually determines the overall system performance. To investigate this impact, we compare the amount of collected data and the number of selected participants achieved by DPS using the following three trajectories: 1) the probability-based trajectory prediction method we used in this paper; 2) trajectories known *a prori*; and 3) directional trajectory prediction. The directional trajectory prediction supposes that the first

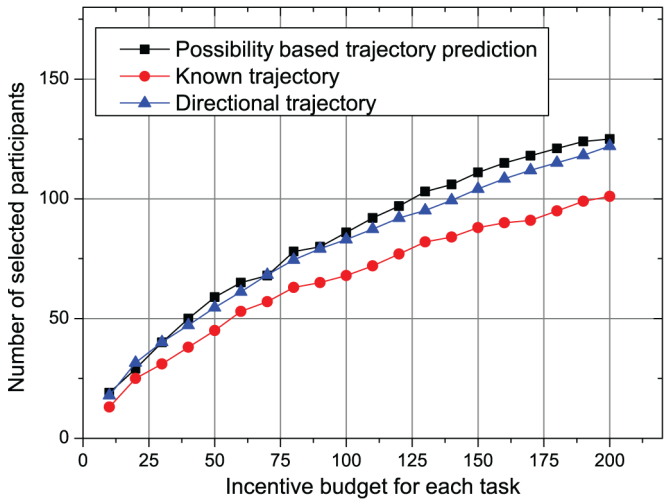

Fig. 15. Impact of different trajectory prediction methods on the number of selected participants.

two locations of each mobile user are known, and then users keep moving toward the same direction with the same speed. Figs. 14 and 15 show the result in terms of the amount of collected data and the number of selected participants among the aforementioned three trajectory prediction-powered DPS schemes. It can be seen that if the location prediction accuracy improves, the total amount of collected data can improve by 37% on average, while the energy consumption can be further reduced by 35%. On the other hand, probability-based trajectory prediction involves almost as many participants as that of the directional prediction scheme, but it collects 29% more data on average. This indicates that the probability-based trajectory prediction can achieve higher accuracy in predicting the data collection of participants compared with the directional prediction scheme.

VII. Conclusion and Future Directions

In this paper, DPS has been proposed to collect the maximum amount of sensory data for all sensing tasks in both temporal and spatial dimensions under budget constraints. A QoI satisfaction metric was introduced to quantify the degree of how collected sensory data can satisfy multidimensional task QoI requirements in terms of data granularity and quantity. The expected amount of collected data by each participant was then predicted by his/her sensing capability, initial location, and a probability model that together calculates his/her probability to move from one location to another, based on the historical trajectory information. Based on all these, DPS is thus established to select participants whose data collection expectations benefit the QoI satisfaction metrics of multiple tasks most compared with their incentive requirements. Extensive experimental results, based on a real trace in Beijing, show the effectiveness and robustness of DPS compared with other existing schemes.

As this paper mainly focuses on how to best fulfil multiple concurrent tasks' QoI requirements, incentive budgets of tasks and incentive requirements of participants are just taken as constraints to the optimization problem and are not expandingly discussed. However, to best fulfill the QoI requirements while minimizing the cost is another challenging research issue, particularly under the condition that participants' incentive requirements can be dynamically decided according to the amount of data (services) they provide or the energy status of their devices. In the future, we plan to extend our model to balance the incentive cost and the gain of QoI satisfaction, for as revealed by the experiment results in Fig. 5 in Section VI, the gain of QoI by providing extra incentive budget to those participants in the long tail section in Fig. 5 is very limited. We also plan to dig into the field of selecting the most energy efficient participants, to extend the overall lifetime of the participants' network as well as collecting satisfactory sensory data for tasks.

References

[1] J. A. Burke *et al.*, "Participatory sensing," in *Proc. ACM SenSys*, 2006, pp. 1–5.
[2] D. Christin, A. Reinhardt, S. S. Kanhere, and M. Hollick, "A survey on privacy in mobile participatory sensing applications," *J. Syst. Softw.*, vol. 84, no. 11, pp. 1928–1946, Nov. 2011.
[3] M. Mun *et al.*, "PEIR, the personal environmental impact report, as a platform for participatory sensing systems research," in *Proc. ACM MobiSys*, 2009, pp. 55–68.
[4] 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 *Proc. ACM MobiSys*, 2009, pp. 165–178.
[5] Campaignr. [Online]. Available: http://research.cens.ucla.edu/urban/
[6] T. Das, P. Mohan, V. N. Padmanabhan, R. Ramjee, and A. Sharma, "Prism: Platform for remote sensing using smartphones," in *Proc. ACM MobiSys*, 2010, pp. 63–76.
[7] M.-R. Ra, B. Liu, T. F. La Porta, and R. Govindan, "Demo: Medusa: A programming framework for crowd-sensing applications," in *Proc. ACM MobiSys*, 2012, pp. 481–482.
[8] D. Mendez, M. Labrador, and K. Ramachandran, "Data interpolation for participatory sensing systems," *Pervasive Mobile Comput.*, vol. 9, no. 1, pp. 132–148, Feb. 2013.
[9] C. Bisdikian *et al.*, "Building principles for a quality of information specification for sensor information," in *Proc. FUSION*, 2009, pp. 1370–1377.
[10] R. Y. Wang and D. M. Strong, "Beyond accuracy: What data quality means to data consumers," *J. Manag. Inf. Syst.*, vol. 12, no. 4, pp. 5–33, Mar. 1996.
[11] M. E. Johnson and K. C. Chang, "Quality of information for data fusion in net centric publish and subscribe architectures," in *Proc. FUSION*, Jul. 2005, pp. 1–5.
[12] S. Reddy, D. Estrin, and M. Srivastava, "Recruitment framework for participatory sensing data collections," in *Pervasive Computing*. Berlin, Germany: Springer-Verlag, 2010, pp. 138–155.
[13] G. S. Tuncay, G. Benincasa, and A. Helmy, "Autonomous and distributed recruitment and data collection framework for opportunistic sensing," in *Proc. ACM MobiCom*, 2012, pp. 407–410.
[14] H. Weinschrott, F. Durr, and K. Rothermel, "Streamshaper: Coordination algorithms for participatory mobile urban sensing," in *Proc. IEEE MASS*, 2010, pp. 195–204.
[15] M. Zhong and C. G. Cassandras, "Distributed coverage control and data collection with mobile sensor networks," in *Proc. IEEE CDC*, 2010, pp. 5604–5609.
[16] H. N. Pham, B. S. Sim, and H. Y. Youn, "A novel approach for selecting the participants to collect data in participatory sensing," in *Proc. IEEE SAINT*, 2011, pp. 50–55.
[17] J.-S. Lee and B. Hoh, "Sell your experiences: A market mechanism based incentive for participatory sensing," in *Proc. IEEE PerCom*, 2010, pp. 60–68.
[18] M. Riahi, T. G. Papaioannou, I. Trummer, and K. Aberer, "Utility-driven data acquisition in participatory sensing," in *Proc. EDBT*, 2013, pp. 251–262.
[19] L. Duan *et al.*, "Incentive mechanisms for smartphone collaboration in data acquisition and distributed computing," in *Proc. IEEE INFOCOM*, 2012, pp. 1701–1709.
[20] X. Lu, D. Li, B. Xu, W. Chen, and Z. Ding, "Minimum cost collaborative sensing network with mobile phones," in *Proc. IEEE ICC*, 2013, pp. 1816–1820.
[21] P. Dutta *et al.*, "Common sense: Participatory urban sensing using a network of handheld air quality monitors," in *Proc. 7th ACM Conf. Embedded Networked Sensor Systems*, 2009, pp. 349–350.

[22] E. Kanjo, "Noisespy: A real-time mobile phone platform for urban noise monitoring and mapping," *ACM/Springer MONET*, vol. 15, no. 4, pp. 562–574, Aug. 2010.
[23] J. Zhou, D. Gao, and D. Zhang, "Moving vehicle detection for automatic traffic monitoring," *IEEE Trans. Veh. Technol.*, vol. 56, no. 1, pp. 51–59, Jan. 2007.
[24] X. Li *et al.*, "Performance evaluation of vehicle-based mobile sensor networks for traffic monitoring," *IEEE Trans. Veh. Technol.*, vol. 58, no. 4, pp. 1647–1653, May 2009.
[25] U. Lee, E. Magistretti, M. Gerla, P. Bellavista, and A. Corradi, "Dissemination and harvesting of urban data using vehicular sensing platforms," *IEEE Trans. Veh. Tech.*, vol. 58, no. 2, pp. 882–901, Feb. 2009.
[26] A. Krause, R. Rajagopal, A. Gupta, and C. Guestrin, "Simultaneous optimization of sensor placements and balanced schedules," *IEEE Trans. Autom. Control*, vol. 56, no. 10, pp. 2390–2405, Oct. 2011.
[27] S. He *et al.*, "Maintaining quality of sensing with actors in wireless sensor networks," *IEEE Trans. Parallel Distrib. Syst.*, vol. 23, no. 9, pp. 1657–1667, Sep. 2012.
[28] S. Joshi and S. Boyd, "Sensor selection via convex optimization," *IEEE Trans. Signal Process.*, vol. 57, no. 2, pp. 451–462, Feb. 2009.
[29] H. Lu, N. D. Lane, S. B. Eisenman, and A. T. Campbell, "Bubble-sensing: Binding sensing tasks to the physical world," *Pervasive Mobile Comput.*, vol. 6, no. 1, pp. 58–71, Feb. 2010.
[30] S. Gaonkar, J. Li, R. R. Choudhury, L. Cox, and A. Schmidt, "Micro-blog: Sharing and querying content through mobile phones and social participation," in *Proc. ACM MobiSys*, 2008, pp. 174–186.
[31] A. L. Custódio, H. Rocha, and L. N. Vicente, "Incorporating minimum Frobenius norm models in direct search," *Comput. Optim. Appl.*, vol. 46, no. 2, pp. 265–278, Jun. 2010.
[32] X. Sheng, J. Tang, and W. Zhang, "Energy-efficient collaborative sensing with mobile phones," in *Proc. IEEE INFOCOM*, 2012, pp. 1916–1924.
[33] R. T. Marler and J. S. Arora, "The weighted sum method for multi-objective optimization: New insights," *Struct. Multidisciplinary Optim.*, vol. 41, no. 6, pp. 853–862, Jun. 2010.
[34] K. M. Bretthauer and B. Shetty, "The nonlinear knapsack problem—Algorithms and applications," *Eur. J. Oper. Res.*, vol. 138, no. 3, pp. 459–472, May 2002.
[35] X. Liu and H. A. Karimi, "Location awareness through trajectory prediction," *Comput., Environ. Urban Syst.*, vol. 30, no. 6, pp. 741–756, Nov. 2006.
[36] N. Schneider and D. M. Gavrila, "Pedestrian path prediction with recursive Bayesian filters: A comparative study," in *Pattern Recognition*. Berlin, Germany: Springer-Verlag, 2013, pp. 174–183.
[37] A. Asahara, K. Maruyama, A. Sato, and K. Seto, "Pedestrian-movement prediction based on mixed Markov-chain model," in *Proc. ACM SIGSPATIAL*, 2011, pp. 25–33.
[38] Y. Zheng, X. Xie, and W.-Y. Ma, "Geolife: A collaborative social networking service among user, location and trajectory," *IEEE Data Eng. Bull.*, vol. 33, no. 2, pp. 32–40, Jun. 2010.
[39] K.-T. Fang, D. K. Lin, P. Winker, and Y. Zhang, "Uniform design: Theory and application," *Technometrics*, vol. 42, no. 3, pp. 237–248, 2000.

Zheng Song received the Ph.D. degree from the State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications (BUPT), Beijing, China.

Prior to studying at BUPT, he was with the R&D Department of Sina Corporation. He is the holder of ten U.S. and China patents. His research interests include participatory sensing, indoor localization, and the Internet-of-Things.

Chi Harold Liu (S'05–M'10) received the B.Eng. degree from Tsinghua University, Beijing, China, and the Ph.D. degree from Imperial College, London, U.K.

He is an Associate Professor and the Department Head of Software Service Engineering, Beijing Institute of Technology. Before moving to academia, he was with IBM Research—China as a Staff Researcher and a Project Manager, after working as a Postdoctoral Researcher with Deutsche Telekom Laboratories, Berlin, Germany and as Visiting Scholar with the IBM T. J. Watson Research Center, Yorktown Heights, NY, USA. He has published more than 40 prestigious conference and journal papers. He is the holder of 11 EU/U.S./China patents. His current research interests include the Internet-of-Things (IoT), big data analytics, mobile computing, and wireless ad hoc, sensor, and mesh networks.

Dr. Liu received the Distinguished Young Scholar Award in 2013, the IBM First Plateau Invention Achievement Award in 2012, and the IBM First Patent Application Award in 2011. In 2011, he was also interviewed by EEWeb.com as the Featured Engineer. He serves as the Editor for the KSII *Transactions on Internet and Information Systems* and as a Book Editor for four books published by Taylor & Francis Group, USA. He has also served as the General Chair of the IEEE SECON'13 International Workshop on Internet-of-Things Networking and Control, the IEEE WCNC'12 Workshop on Internet-of-Things Enabling Technologies, and the ACM UbiComp'11 Workshop on Networking and Object Memories for Internet-of-Things. He is a member of the Association for Computing Machinery.

Jie Wu (M'89–SM'94–F'09) received the B.S. and M.S. degrees from Shanghai University of Science and Technology (currently Shanghai University), Shanghai, China, in 1982 and 1985, respectively, and the Ph.D. degree in computer engineering from Florida Atlantic University, Boca Raton, FL, USA, in 1989.

He is the Chair of and a Laura H. Carnell Professor with the Department of Computer and Information Sciences, Temple University, Philadelphia, PA, USA. Prior to joining Temple University, he was a Program Director with the National Science Foundation and a Distinguished Professor with Florida Atlantic University, Boca Raton, FL, USA. He has regularly published scholarly journals, conference proceedings, and books. His current research interests include mobile computing and wireless networks, routing protocols, cloud and green computing, network trust and security, and social network applications.

Mr. Wu serves on several editorial boards, including that of the IEEE TRANSACTIONS ON COMPUTERS, the IEEE TRANSACTIONS ON SERVICE COMPUTING, and the *Journal of Parallel and Distributed Computing*. He was the general Cochair/Chair for the IEEE International Conference on Mobile Ad Hoc and Sensor Systems (MASS 2006) and the IEEE International Symposium on Parallel and Distributed Processing (IPDPS 2008), the Program Cochair for the IEEE International Conference on Computer Communications (INFOCOM 2011), the General Chair for the IEEE International Conference on Distributed Computing Systems (ICDCS 2013) and the Program Chair for the China Computer Federation (CCF) China National Computer Congress (CNCC 2013). He is currently serving as the General Chair for the Association for Computing Machinery (ACM) International Symposium on Mobile Ad Hoc Networking and Computing (MobiHoc 2014). He was an IEEE Computer Society Distinguished Visitor, an ACM Distinguished Speaker, and the Chair for the IEEE Technical Committee on Distributed Processing. He is a CCF Distinguished Speaker. He received the 2011 CCF Overseas Outstanding Achievement Award.

Dr. Wu received the Ph.D. degree in computer engineering from Florida Atlantic University, Boca Raton, FL, USA, in 1989 and the B.S. and M.S. degrees, both from Shanghai University of Science and Technology (now Shanghai University), Shanghai, China, in 1982 and 1985, respectively.

Jian Ma (M'99) received the B.Sc. and M.Sc. degrees from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 1982 and 1987, respectively, and the Ph.D. degree from Helsinki University of Technology, Espoo, Finland, in 1994.

He is a Principal Member of the Research Staff with the Nokia Research Center, Beijing. He has also been a Guest Professor in computer science with the BUPT since 2002, Institute of Computing Technology, Chinese Academy of Sciences (ICT-CAS) since 2003, and College of Computer and Communication Engineering, Graduate University of Chinese Academy of Sciences (CCCE-GUCAS) since 2006. He is currently supervising over a dozen Ph.D. and M.Sc. students at the BUPT, CCCE-GUCAS, ICT-CAS, and Tsinghua University, Beijing. He has authored more than 160 conference and journal papers and four books and book chapters.

Dr. Ma has been involved in a few professional activities and is currently serving as a Council Member of the Beijing Communication Institute (a senior consultant club in communication industry and academia) and as Vice Chair of the Sensor Network Technical Committee, China Computer Federation. He has also served as a Cochair and a member of organization committees or program committees in several leading conferences, such as the ICC, AINA, PIMRC, ICDCS, WWW, and ChinaCom.

Wendong Wang (M'02) received the Bachelor's and Master's degrees from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 1985 and 1991, respectively.

He is currently a Full Professor with the BUPT. He is currently on the Assessment Panel of the National Natural Science Foundation Program and the National High Technology Research and Development Program of China. He has published more than 200 papers in various journals and conference proceedings. He is the holder of 14 U.S./China patents. His current research interests include next-generation network architecture; Internet-of-Things; participatory sensing; wireless ad hoc, sensor, and mesh networks; and mobile Internet.

Mr. Wang is a member of the Association for Computing Machinery.