Accuracy-Guaranteed Event Detection via Collaborative Mobile Crowdsensing with Unreliable Users

Tong Liu[1,2(✉)], Wenbin Wu[1], Yanmin Zhu[3,4], and Weiqin Tong[1,2]

[1] School of Computer Engineering and Science, Shanghai University, Shanghai, China
{tong_liu,wenbinw,wqtong}@shu.edu.cn
[2] Shanghai Institute for Advanced Communication and Data Science, Shanghai University, Shanghai, China
[3] Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China
yzhu@sjtu.edu.cn
[4] Shanghai Key Lab of Scalable Computing and Systems, Shanghai, China

Abstract. Recently, mobile crowdsensing has become a promising paradigm to collect rich spatial sensing data, by taking advantage of widely distributed sensing devices like smartphones. Based on sensing data, event detection can be conducted in urban areas, to monitor abnormal incidents like traffic jam. However, how to guarantee the detection accuracy is still an open issue, especially when unreliable users who may report wrong observations are considered. In this work, we focus on the problem of user recruitment in collaborative mobile crowdsensing, aiming to optimize the fine-grained detection accuracy in a large urban area. Unfortunately, the problem is proved to be NP-hard, which means there is no polynomial-time algorithm to achieve the optimal solution unless P = NP. To meet the challenge, we first employ a probabilistic model to characterize the unreliability of users, and measure the uncertainty of inferring event occurrences given collected observations by Shannon entropy. Then, by leveraging the properties of adaptive monotonicity and adaptive submodularity, we propose an adaptive greedy algorithm for user recruitment, which is theoretically proved to achieve a constant approximation ratio guarantee. Extensive simulations are conducted, which show our proposed algorithm outperforms baselines under different settings.

Keywords: Collaborative mobile crowdsensing · Event detection · User recruitment · Adaptive greedy algorithm

1 Introduction

With the popularization of mobile devices equipped with rich embedded sensors and wireless communication modules, mobile crowdsensing [7,8,15] has emerged as a promising data sensing and collecting paradigm. Taking advantage of widely

X. Wang et al. (Eds.): CollaborateCom 2019, LNICST 292, pp. 729–744, 2019.
https://doi.org/10.1007/978-3-030-30146-0_49

distributed mobile devices, fine-grained event detection over an urban area can be conducted via collaborative sensing, which can support services such as traffic jam monitoring [18]. In this work, we consider a typical mobile crowdsensing system, consisting of a central *platform* built in cloud and a set of collaborative *users* equipped with sensing devices. The platform is responsible for recruiting some users to participate in event detecting, within a given budget. Then, the recruited users perform detecting during their movement. Finally, the platform infers all event occurrences in the monitored area, based on the observations reported by recruited users.

It is still an open issue that how to guarantee the detection accuracy, in terms of considering the unreliability of users and the budget constraint. Here, detection accuracy measures the deviation between ground truth and inference of event occurrences. On one hand, certain costs are incurred on users for detecting, such as energy consumption, bandwidth usage and interaction time. Thus, given a fixed budget, the number of recruited users is significantly limited. On the other hand, observations reported by different users have different accuracy levels, which may be influenced by device hardware or user experience. Which users are recruited and what observations are collected make a big difference on the fine-grained detection accuracy could be achieved. Considering these two aspects, users should be carefully selected by the platform, to satisfy the budget constraint and optimize the detection accuracy at the same time.

Recently, some works have paid efforts to figure out the problem of quality-aware data collection in mobile crowdsensing, which are the most related with our work. Different metrics are considered to measure data quality. A certain attained value of each user is considered to measure the quality of information in [13]. Data utility is measured by data granularity and quantity in [23] and prediction uncertainty and data density in [29], respectively. Yang *et al.* [30] consider the distance between measurements and true values estimated as the centroid of the measurements, to measure the quality of each user. Most other works [10,11,26–28,31] focus on designing truthful incentive mechanisms for quality-aware users, which model the quality of each user as a constant real value. Moreover, a few works [21,31] have noticed the unreliability of users. Zheng *et al.* [31] assume qualities of users follow certain multinomial distributions. A discrete probabilistic effort matrix is used in [21] to model the deviation between measurements and ground truth. Moreover, expectation maximization (EM) based algorithms are proposed to estimate the qualities of users. However, these works ignore digging how the unreliability of users influences the quality of data collection, e.g., fine-grained detection accuracy in our work, and accordingly proposing efficient user recruitment approaches.

In this work, we study the problem of user recruitment in collaborative mobile crowdsensing to provide event detection services, with the objective to optimize the detection accuracy under a fixed budget constraint. In addition, we consider unreliable users, whose observations may variously deviate from ground truth. Thus, multiple users should be recruited to detect one event, which can collaboratively improve the detection accuracy. Different from previous works,

we aim to propose an optimal user recruitment approach, by formally analyzing the relationship between the unreliability of recruited users and the fine-grained accuracy achieved given the observations collected from users.

However, the problem is particularly difficult due to the following challenges. *Firstly*, the more users are recruited, the higher detection accuracy could be achieved in intuition. However, the number of recruited users is significantly limited by the budget constraint. *Secondly*, observations of an unreliable user are nondeterministic before the user is recruited, which makes the platform hard to estimate the value of each user in terms of improving the overall detection accuracy. *Thirdly*, the improvement of detection accuracy made by the observations collected from a user is varying, which is also related with the observations have been collected from others. *Finally*, as the problem is formally proved to be NP-hard, there does not exist a polynomial-time algorithm to achieve the optimal solution if $P \neq NP$.

To meet the challenges, we propose an adaptive greedy algorithm for user recruitment in a budgeted mobile crowdsensing system in this paper. We first employ a probabilistic matrix to model the unreliability of each user, which consists of true-positive, false-positive, true-negative, and false-negative detection probabilities. Then, the probability of an event occurrence is estimated given the collected observations based on the Bayesian rule. Moreover, Shannon entropy is employed to measure the uncertainty of the estimation, which represents the detection accuracy. Next, by taking advantage of the properties of adaptive monotonicity and adaptive submodularity, we put forward an adaptive greedy algorithm, in which users are sequentially recruited according to the expectation of accumulated entropy reduction obtained by their observations. Our proposed algorithm is theoretically proved to achieve near-optimal performance with a constant approximation ratio guarantee. Extensive simulations are conducted to evaluate the performance of our proposed algorithm, compared with baselines. The comparative results show that our algorithm can achieve high detection accuracy under different settings.

The main contributions of this paper are summarized as follows:

- First, we employ a probabilistic model to characterize the unreliability of users, and Shannon entropy to measure the uncertainty of estimations on event occurrences. Moreover, the relationship between the detection accuracy achieved by collected observations and the unreliability of recruited users is formally established.
- Second, we propose an adaptive greedy user recruitment algorithm, which is proved to achieve approximately optimal performance with a constant approximation ratio guarantee.
- Third, we perform comprehensive simulations to evaluate our proposed algorithm, compared with baselines. The results show that our algorithm can achieve better performance under different settings.

The rest of the paper is organized as follows. Section 2 reviews related work. The system model, formal problem formulation and preliminary definitions are presented in Sect. 3. In Sect. 4, we illustrate the design details of our proposed

user recruitment algorithm, and its optimization analysis. Section 5 evaluates the performance of our algorithm compared with baselines via extensive simulations. Finally, we conclude our work in Sect. 6.

2 Related Work

With the proliferation of mobile sensing devices like smartphones, mobile crowdsensing has attracted a lot of attention from industry and academia. Many useful applications have been developed to collect various sensing data from crowds for environment monitoring [16,17,22], smart transportation [18,25], healthcare [5,9,20], and social interaction [1,3,4]. How to collect high-quality sensing data and extract accurate information is a fundamental issue for the success of mobile crowdsensing. Recently, quality-aware data collection has attracted some research efforts. In this section, we briefly review related works and point out the difference and contributions of our work.

The concept of Quality-of-Information (QoI) is first introduced into query-based mobile crowdsensing by [13]. QoI is formally formulated as a function of the required value of each query and the attained value of each user. An energy-efficient algorithm and a dynamic pricing scheme are proposed for deciding participants and allocating credits to participants respectively. Followed by [23], Song *et al.* propose a QoI-aware energy-efficient participant selection method, where QoI is measured by data granularity and quantity. Different from these two works, we use a probabilistic matrix to model the unreliability of users, and the data quality, i.e., fine-grained detection accuracy, is measured by the uncertainty of the estimations inferred based on observations of users.

Given the QoI of each user, reverse combinatorial auction-based incentive mechanisms are proposed for both single-minded and multi-minded users in [10], to maximize the profit defined on the accumulated QoI of selected users. Similarly, a few other works [11,26–28,31] have proposed quality-aware incentive mechanisms to encourage users to contribute high-quality data. Most of these works model qualities of users as known, certain and additive real values. They focus on how to guarantee the truthfulness of strategic users. In our work, we consider unreliable users who may contribute incorrect observations, and focus on discovering the truth with high certainty based on unreliable observations. We also propose an adaptive algorithm for greedily recruiting valuable users, while providing proper incentives to users is beyond the scope of our work.

Both Yang *et al.* [30] and Peng *et al.* [21] propose approaches to estimate qualities of users according to their measurements and then provide incentives based on their qualities. In [30], the quality of users is measured by the deviation of their measurements and ground truth, which is estimated as the centroid of the measurements. This truth discovery method is also employed in [11]. In [21], a probabilistic effort matrix is used for modeling the quality of each user, and an EM algorithm is proposed to estimate effort matrixes of users as well as ground truth. Developed by [14], data qualities of users, which are assumed following multinomial distributions, can also be estimated by an EM algorithm. Moreover,

a context-quality classifier is trained to discover the truth and a greedy-based algorithm is proposed for user selection. In [29], Xu *et al.* consider the platform can actively orchestrate queries for collecting annotation data, and data utility is measured by both prediction uncertainty and data density. A threshold-based method is proposed for online participant selection.

Similar with [21], we model the unreliability of users by a discrete probabilistic matrix. Different from the previous works, we focus on how to accurately discover ground truth with high certainty, given the observations collected from unreliable users under a fixed budget constraint.

3 Preliminaries and Problem Formulation

3.1 System Model

A typical mobile crowdsensing system is consisted of a central platform located in cloud and a universal set of mobile users equipped with smart devices, i.e., $\mathcal{U} = \{u_1, u_2, \cdots, u_K\}$, where K is the number of users. The platform is responsible to recruit some users to collect observations of event occurrences. For sake of describing the locations of events and users, we partition the whole detected area into fine-grained grids with equal size (e.g., a square of 200 m $\times$ 200 m). The set of all grids is denoted by $\mathcal{G} = \{g_1, g_2, \cdots, g_N\}$, where N is the number of grids. We use a Boolean variable X_n to denote whether there is an event occurring in grid g_n. Thus, the ground truth of event occurrences in the whole area can be expressed as $\mathbf{X} = \{X_n \in \{0, 1\}, \forall 1 \leq n \leq N\}$.

As users are mobile, we consider the trajectory of each user during the period of event detection is reported to the platform at the beginning. The trajectory of user $u_k, \forall 1 \leq k \leq K$ can be denoted by a set of grids, i.e., $\mathcal{G}_k \subseteq \mathcal{G}$, as shown in Fig. 1. If user u_k is selected as a participant by the platform, all the grids in $\mathcal{G}_k$ will be detected by u_k. The set of observations collected by u_k can be represented by $\mathcal{D}_k = \{D_{k,n} \in \{0, 1\}, \forall g_n \in \mathcal{G}_k\}$, where $D_{k,n} = 1$ means an event is detected by u_k in grid g_n. In addition, some costs are paid by participants, like power consumption and human-device interaction. We denote the cost of user u_k participating in event detection is c_k.

Unreliable Users. In our work, we consider unreliable users, who may report wrong observations to the platform, caused by device hardware, sensing contexts, or user experience. Thus, the observations of users are uncertain, even given the ground truth of event occurrences. To characterize the stochastic nature of detection results collected by users, we model each user is associated with a certain level of detection accuracy, denoted by a matrix of probabilities, i.e.,

$$\mathbf{P}_k = \begin{bmatrix} p_k^{\mathbb{T}} & 1 - p_k^{\mathbb{T}} \\ p_k^{\mathbb{F}} & 1 - p_k^{\mathbb{F}} \end{bmatrix}. \tag{1}$$

Here, $p_k^{\mathbb{T}}$ and $p_k^{\mathbb{F}}$ respectively represent the true-positive and false-positive detection probabilities, i.e.,

$$p_k^{\mathbb{T}} = \Pr(D_{k,n} = 1 | I_n = 1), \forall 1 \leq n \leq N,$$
$$p_k^{\mathbb{F}} = \Pr(D_{k,n} = 1 | I_n = 0), \forall 1 \leq n \leq N.$$

The detection probability matrix of each user is different and can be effectively estimated from the historical detecting records by the EM algorithm proposed in [21,31]. As it is not the focus of our work, we assume that the detection probabilities of all users are known by the platform for simplicity.

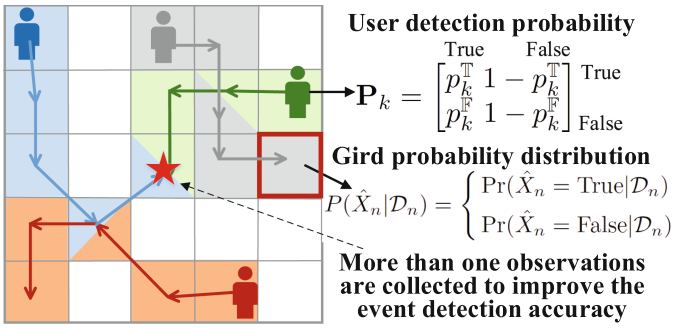

Fig. 1. Illustration of our collaborative mobile crowdsensing system model. Some users are recruited to collect observations along their trajectories for event detection. Different users have different accuracies represented by detection probabilities, while each grid is associated with a random variable to characterize the uncertainty of event occurring inference based on collected observations.

User Recruitment. In this work, we consider the platform sequentially recruits users, unless the total costs of recruited users exceed a given budget. Note that observations are collected once a user is recruited, and then the platform continues recruiting the next one. We use $I_k \in \{0, 1\}$ to indicate whether user u_k is recruited or not, and denote $\mathbf{I} = \{I_k \in \{0, 1\}, 1 \leq k \leq K\}$. Moreover, we denote $\mathcal{D}_n$ as the set of observations collected in grid g_n by all recruited users, i.e., $\mathcal{D}_n = \{D_{k,n} | I_k = 1 \text{ and } g_n \in \mathcal{G}_k, \forall 1 \leq k \leq K\}$, and we denote $\mathbf{D} = \{\mathcal{D}_n, \forall 1 \leq n \leq N\}$.

Given the set of recruited users indicated by $\mathbf{I}$ and their observations $\mathbf{D}$, we define a *realization* $\phi \triangleq \{(u_k, \mathcal{D}_k)\}$, indicating to what extent various users are recruited and their observations are collected. In addition, we use Φ to denote a random realization, in which the value of $\mathbf{I}$ is not determined. Then, the probability distribution over realization ϕ can be calculated as $\Pr(\Phi = \phi) = \prod_{I_k=1} \Pr(\Phi(u_k) = \phi(u_k))$, where $\phi(u_k) \triangleq \mathcal{D}_k$. After a user is recruited by the platform, the set of observations collected so far is updated, which is represented by a *partial realization* ψ. We define $\text{dom}(\psi)$ representing the recruited users given a partial realization ψ, i.e., $\text{dom}(\psi) = \{u_k | \exists (u_k, \mathcal{D}_k) \in \psi\}$. A partial

realization ψ is consistent with a full realization ϕ (denoted by $\phi \sim \psi$), if $\mathcal{D}_k$ is the same for all $u_k \in \text{dom}(\psi)$. Moreover, ψ is called a *subrealization* of $\psi^{'}$, if ψ and $\psi^{'}$ are both consistent with ϕ and $\text{dom}(\psi) \subseteq \text{dom}(\psi^{'})$.

Detection Accuracy. Given the observations collected by participants in grid g_n, whether there is an event occurring in g_n can be estimated. We use a random variable $\hat{X}_n$ to denote the estimation, which is associated with a probability distribution $P(\hat{X}_n|\mathcal{D}_n) = \Pr(\hat{X}_n = 1|\mathcal{D}_n)$. When a new observation $D_{k,n}$ in grid g_n is reported by user u_k, the probability can be updated according to the Bayesian rule [24] as

$$P(\hat{X}_n|\mathcal{D}_n \cup D_{k,n}) = \Pr(\hat{X}_n = 1|\mathcal{D}_n \cup D_{k,n}) \tag{2}$$
$$= \begin{cases} \dfrac{p_k^{\mathbb{T}} \cdot P(\hat{X}_n|\mathcal{D}_n)}{p_k^{\mathbb{T}} \cdot P(\hat{X}_n|\mathcal{D}_n) + p_k^{\mathbb{F}} \cdot (1 - P(\hat{X}_n|\mathcal{D}_n))}, & \text{if } D_{k,n} = 1, \\ \dfrac{(1 - p_k^{\mathbb{T}}) \cdot P(\hat{X}_n|\mathcal{D}_n)}{(1 - p_k^{\mathbb{T}}) \cdot P(\hat{X}_n|\mathcal{D}_n) + (1 - p_k^{\mathbb{F}}) \cdot (1 - P(\hat{X}_n|\mathcal{D}_n))}, & \text{if } D_{k,n} = 0. \end{cases}$$

Note that the Bayesian rule can be applied here because we assume the detection probabilities of users are independent from each other.

We denote the joint probability distribution over the discrete-valued random vector $\hat{\mathbf{X}} = [\hat{X}_1, \hat{X}_2, \cdots, \hat{X}_N]$ as $P(\hat{\mathbf{X}})$. With the observations collected in all grids, there exists

$$P(\mathbf{x}|\mathbf{D}) = \Pr(\hat{\mathbf{X}} = \mathbf{x}|\mathbf{D}) = \prod_{n=1}^{N} \Pr(\hat{X}_n = x_n|\mathcal{D}_n), \tag{3}$$

where $\mathbf{x} = \{x_1, x_2, \cdots, x_N\}$ and $x_n \in \{0, 1\}, \forall 1 \leq n \leq N$.

The fine-grained detection accuracy of the whole area can be measured by the reduction of the uncertainty of estimations $\hat{\mathbf{X}}$ in all grids, given all observations $\mathbf{D}$ collected by recruited users. Specially, *Shannon entropy* [21] is a commonly used criterion to measure the uncertainty of random variables. Given the joint probability distribution of $\hat{\mathbf{X}}$, its entropy can be calculated as

$$\begin{aligned} H(\hat{\mathbf{X}}|\mathbf{D}) &= -\sum_{\mathbf{x}} (P(\mathbf{x}|\mathbf{D}) \cdot \log P(\mathbf{x}|\mathbf{D})) \\ &= -\sum_{\mathbf{x}} \left[\prod_{n=1}^{N} \Pr(\hat{X}_n = x_n|\mathcal{D}_n) \cdot \sum_{n=1}^{N} \log \Pr(\hat{X}_n = x_n|\mathcal{D}_n) \right] \\ &= \sum_{n=1}^{N} \left[-\sum_{x_n \in \{0,1\}} \Pr(\hat{X}_n = x_n|\mathcal{D}_n) \cdot \log \Pr(\hat{X}_n = x_n|\mathcal{D}_n) \right] \\ &= \sum_{n=1}^{N} H(\hat{X}_n|\mathcal{D}_n). \end{aligned} \tag{4}$$

Then, the entropy reduction obtained by the observations $\mathbf{D}$ is $H(\hat{\mathbf{X}}) - H(\hat{\mathbf{X}}|\mathbf{D})$.

3.2 Problem Formulation

In this work, we consider the problem that how the platform recruits a proper subset of unreliable users given a fixed budget constraint, aiming to optimize the fine-grained detection accuracy of the whole area. Given the system model built in the last subsection, the problem can be formally formulated as follows,

$$\begin{aligned} &\max_{\mathbf{I}} f(\mathbf{I}, \Phi) \\ &s.t. \quad \sum_{u_k \in \mathcal{U}} c_k \cdot I_k \leq \eta, \\ &\qquad I_k \in \{0,1\}, \forall 1 \leq k \leq K, \end{aligned} \tag{5}$$

where $f(\mathbf{I}, \Phi) \triangleq \mathbb{E}[H(\hat{\mathbf{X}}) - H(\hat{\mathbf{X}}|\mathbf{D})]$, representing the expectation of the entropy reduction obtained by recruiting users indicated by $\mathbf{I}$, and η is the budget on the total costs of recruited users.

This problem is a stochastic 0-1 integer programming problem. We prove its NP-hardness in Theorem 1.

Theorem 1. *The user recruitment problem with a fixed budget is NP-hard.*

Proof. The decision version of this problem is that given a set of users and their trajectories, whether a subset of users can be found to achieve a given detection accuracy requirement, and the total costs of the users are no larger than η.

Then, we prove the NP-hardness of our problem by reducing a classical NP-hard problem, *vertex cover problem* [2], to our problem in polynomial time. An instance of the decision version of the vertex cover problem is, given an undirected graph $G = (V, E)$, whether a subset of n vertexes $V^{'} \subseteq V$ can be found, such that for $\forall uv \in E, u \in V^{'} \vee v \in V^{'}$ exists.

Next, we construct an instance of our problem, and show that the instance of the vertex cover problem can be transformed to the instance of our problem. We transform vertex set V and edge set E into the set of users $\mathcal{U}$ and grids $\mathcal{G}$, respectively. For $\forall uv \in E$, the corresponding grid is included in the trajectories of the users corresponding to u and v. For each user $u_k \in \mathcal{U}$, we set $p_k^{\mathbb{T}} = 1$, $p_k^{\mathbb{F}} = 0$, and $c_k = 1$. The detection accuracy requirement is set as $\sum_{g_n \in \mathcal{G}} H(\hat{X}_n|\mathcal{D}_n) = 0$. Thus, as long as grid g_n is included in the trajectory of a recruited user, there exists $H(\hat{X}_n|\mathcal{D}_n) = 0$. Also, we set $\eta = n$. Then, the instance of our problem is equal to select n users, the union of whose trajectories cover all grids.

Now, a solution of the instance of the vertex cover problem can be transformed to the solution of the instance of our problem. Specially, if the corresponding users in $\mathcal{U}$ for each $u \in V^{'}$ are recruited (denoted by $\mathcal{U}^{'}$), the detection accuracy requirement can be achieved, as any grid is included in the trajectories of at least one user in $\mathcal{U}^{'}$.

Thus, any instance of vertex cover problem is polynomial-time reducible to an instance of our problem. As vertex cover problem is NP-hard, we prove that our problem is NP-hard as well.

Unfortunately, there does not exist a polynomial-time algorithm to solve the user recruitment problem, unless P = NP. In the following, we design an adaptive greedy algorithm, which is proved to achieve a constant approximation ratio guarantee.

3.3 Preliminaries

In this subsection, we present the definitions of two important properties: *adaptive monotonicity* and *adaptive submodularity*, which are generalizations of monotonicity and submodularity [19] to adapt random realization. If the objective function of a stochastic 0-1 integer optimization satisfies these two properties, a good performance with a constant approximation ratio can be achieved by conducting an adaptive greedy algorithm.

Definition 1 (Conditional Expected Marginal Benefit [6]). *Given a partial realization ψ and an item e, the conditional expected marginal benefit of e conditioned on ψ is*

$$\Delta(e|\psi) = \mathbb{E}\left[F(dom(\psi) \cup \{e\}, \Phi) - F(dom(\psi), \Phi)|\Phi \sim \psi\right],$$

where the expectation is computed with respect to $p(\phi|\psi) = \Pr(\Phi = \phi|\Phi \sim \psi)$.

Definition 2 (Adaptive Monotonicity [6]). *A function $F : 2^E \times O^E \rightarrow \mathbb{R}$ is adaptive monotone with respect to distribution $p(\phi)$ if, for any partial realization ψ and for any $e \in E$, we have*

$$\Delta(e|\psi) \geq 0.$$

Definition 3 (Adaptive Submodularity [6]). *A function $F : 2^E \times O^E \rightarrow \mathbb{R}$ is adaptive submodular with respect to distribution $p(\phi)$ if, for any partial realization ψ and $\psi^{'}$, where ψ is a subrealization of $\psi^{'}$ (i.e., $dom(\psi) \subseteq dom(\psi^{'})$), and for any $e \in E \setminus dom(\psi^{'})$, we have*

$$\Delta(e|\psi) \geq \Delta(e|\psi^{'}).$$

4 Adaptive Greedy Algorithm

In this section, we first propose an adaptive greedy algorithm for the user recruitment problem, and then theoretically prove that the performance of our algorithm achieves a constant approximation ratio guarantee.

4.1 Algorithm Design

The basic idea of designing the algorithm is to greedily select the user, who achieves the most entropy reduction and has the least cost at the same time. Specially, users are sequentially recruited according to the following rule,

$$u_{k^*} = \arg\max_{u_k} \frac{\Delta(u_k|\psi)}{c_k}, \tag{6}$$

where $\Delta(u_k|\psi) = \mathbb{E}[H(\hat{\mathbf{X}}|\mathbf{D}) - H(\hat{\mathbf{X}}|\mathbf{D} \cup \mathcal{D}_k)]$, denoting the expectation of the entropy reduction obtained by observations collected from u_k, given partial realization ψ. As observations $\mathcal{D}_k$ are unknown before u_k is recruited, the expectation can be computed by considering any possible value of $\mathcal{D}_k$ and its probability respectively, i.e.,

$$\begin{aligned}\Delta(u_k|\psi) &= \mathbb{E}\left[\sum_{g_n \in \mathcal{G}_k} (H(\hat{X}_n|\mathcal{D}_n) - H(\hat{X}_n|\mathcal{D}_n \cup \{D_{k,n}\}))\right] \\ &= \sum_{g_n \in \mathcal{G}_k} \mathbb{E}[H(\hat{X}_n|\mathcal{D}_n) - H(\hat{X}_n|\mathcal{D}_n \cup \{D_{k,n}\})] \\ &= \sum_{g_n \in \mathcal{G}_k} [\Pr(D_{k,n} = 1) \cdot \Delta H(\hat{X}_n|D_{k,n} = 1) \\ &\qquad + \Pr(D_{k,n} = 0) \cdot \Delta H(\hat{X}_n|D_{k,n} = 0)]. \end{aligned} \tag{7}$$

Here, $\Pr(D_{k,n} = 1) = P_n \cdot p_k^{\mathbb{T}} + (1 - P_n) \cdot p_k^{\mathbb{F}}$, and $\Pr(D_{k,n} = 0) = P_n \cdot (1 - p_k^{\mathbb{T}}) + (1 - P_n) \cdot (1 - p_k^{\mathbb{F}})$.

The details of our proposed adaptive greedy algorithm for user recruitment are illustrated in Algorithm 1. We first initialize the probability distribution of each grid as uniform distribution without any priori information in line 1. Line 3 to 15 are repeatedly executed to sequentially recruit users, according to the rule in (6). Observations are collected once a user is recruited as shown in line 9, and then the probability distribution of each grid within the trajectory is updated according to (2). If the budget constraint cannot be satisfied by recruiting any user left, the algorithm ends. The time complexity of our algorithm is $O(K^2N)$.

4.2 Optimization Analysis

Theorem 2. *Let $\mathbf{I}^o$ indicate the set of recruited users returned by Algorithm 1, and $\mathbf{I}^*$ indicate the set of recruited users which achieves the maximal entropy reduction. Then, for any budget η, we have*

$$f(\mathbf{I}^o, \Phi) \geq (1 - 1/e) f(\mathbf{I}^*, \Phi) \tag{8}$$

Proof. First, we define function $\hat{f}(\{(u_k, \mathcal{D}_k)\}) = H(\mathbf{X}) - H(\hat{X}|\mathbf{D})$, which is monotone submodular as shown by Krause and Guestrin [12]. Obviously, there is $f(\mathbf{I}, \phi) = \hat{f}(\{(u_k, \mathcal{D}_k)|I_k = 1, \phi(u_k) = \mathcal{D}_k\})$ under realization ϕ.

Then, we prove this theorem, by proving f is adaptive monotone and adaptive submodular. Adaptive monotonicity is readily proved as $f(\cdot, \phi)$ is monotone for each ϕ. To prove adaptive submodularity, we aim to show $\Delta(u_k|\psi^{'} \leq \Delta(u_k|\psi)$ for any $\psi, \psi^{'}$ such that $\psi \subseteq \psi^{'}$ and any $u_k \notin \mathrm{dom}(\psi^{'})$. We define a coupled distribution p over pairs of realizations $\phi \sim \psi$ and $\phi^{'} \sim \psi^{'}$ such that $\phi(u_k) = \phi^{'}(u_k)$ for all $u_k \notin \mathrm{dom}(\psi^{'})$. Formally, $p(\phi, \phi^{'}) = \prod_{u_k \in \mathcal{U} \setminus \mathrm{dom}(\psi)} \Pr[\Phi(u_k) = \phi(u_k)]$ if $\phi \sim \psi, \phi^{'} \sim \psi^{'}$, and $\phi(u_k) = \phi^{'}(u_k)$; otherwise, $p(\phi, \phi^{'}) = 0$. Next, we calculate $\Delta(u_k|\psi^{'}$ and $\Delta(u_k|\psi)$ using p as follows,

Algorithm 1. Adaptive Greedy User Recruitment Algorithm

Input: A set of users $\mathcal{U}$, detection probabilities of each user $\mathbf{P}_k$, cost of each user c_k, budget η.
Output: A set of recruited users indicated by $\mathbf{I}$.

$I_k = 0$, $\mathcal{V} \leftarrow \emptyset$, $\psi \leftarrow \emptyset$, $P(\hat{X}_n) = 0.5$;
while $\sum_{u_k} c_k \cdot I_k < \eta$ and $\mathcal{U} \setminus \mathcal{V} \neq \emptyset$ **do**
  for each $u_k \in \mathcal{U} \setminus \mathcal{V}$ **do**
    Calculate $\Delta(u_k|\psi)$ according to (7);
  end for
  Select $u_{k^*} = \arg\max_{u_k \in \mathcal{U} \setminus \mathcal{V}} \frac{\Delta(u_k|\psi)}{c_k}$;
  if $\sum_{u_k} c_k \cdot I_k + c_{k^*} \leq \eta$ **then**
    $I_{k^*} = 1$;
    Collect observations $\mathcal{D}_{k^*}$;
    $\psi \leftarrow \psi \cup \{(u_{k^*}, \mathcal{D}_{k^*})\}$;
    for each $g_n \in \mathcal{G}_{k^*}$ **do**
      Update $P(\hat{X}_n)$ according to (2);
    end for
  end if
  $\mathcal{V} \leftarrow \mathcal{V} \cup \{u_{k^*}\}$;
end while
return $\mathbf{I}$;

$$\begin{aligned} f(\mathrm{dom}(\psi^{'}) \cup \{u_k\}, \phi^{'}) - f(\mathrm{dom}(\psi^{'}), \phi^{'}) &= \hat{f}(\psi^{'} \cup \{(u_k, \mathcal{D}_k)\}) - \hat{f}(\psi^{'}) \\ &\leq \hat{f}(\psi \cup \{(u_k, \mathcal{D}_k)\}) - \hat{f}(\psi) \\ &= f(\mathrm{dom}(\psi) \cup \{u_k\}, \phi) - f(\mathrm{dom}(\psi), \phi), \end{aligned}$$

where the inequality holds due to the submodularity of $\hat{f}$. Thus, we have

$$\begin{aligned} \Delta(u_k|\psi^{'}) &= \sum\nolimits_{(\phi,\phi^{'})} \Big[p(\phi, \phi^{'}) \cdot (f(\mathrm{dom}(\psi^{'}) \cup \{u_k\}, \phi^{'}) - f(\mathrm{dom}(\psi^{'}), \phi^{'})) \Big] \\ &\leq \sum\nolimits_{(\phi,\phi^{'})} \Big[p(\phi, \phi^{'}) \cdot (f(\mathrm{dom}(\psi) \cup \{u_k\}, \phi) - f(\mathrm{dom}(\psi), \phi)) \Big] \\ &= \Delta(u_k|\psi). \end{aligned}$$

According to Theorem 5.2 in [6], if a function f is adaptive monotone and adaptive submodular, and π is a greedy policy, then for any policy π^*, there exists $f(\pi) \geq (1 - 1/e) f(\pi^*)$. Thus, we can conclude that

$$f(\mathbf{I}^o, \Phi) \geq (1 - 1/e) f(\mathbf{I}^*, \Phi).$$

5 Performance Evaluation

In this section, we evaluate the performance of our proposed adaptive greedy algorithm (marked as AG in figures) by conducting comprehensive simulations.

5.1 Methodology and Setups

In our simulations, we compare our algorithm with three greedy-based baseline algorithms, which are illustrated in the following:

1. *Random Algorithm (RD).* Users are randomly selected by the platform, until the budget could not be satisfied if any one more user is recruited.
2. *User-Greedy Algorithm (UG).* This algorithm sequentially selects users with the most observations per cost, i.e., $|\mathcal{G}_k|/c_k$, under the budget constraint.
3. *Grid-Greedy Algorithm (GG).* This algorithm sequentially selects users with the highest accumulated entropy of all grids past through, i.e., $\sum_{g_n \in \mathcal{G}_k} H(\hat{X}_n)$.

Three metrics are employed to measure the performance achieved by our proposed algorithm and these three baselines. First, we compare the *entropy* achieved given the observations of recruited users selected by different algorithms, i.e., $H(\hat{\mathbf{X}}|\mathbf{D})$. Then, we employ two criterions, *precision* and *recall*, to measure the accuracy of event inference achieved by different algorithms. We infer there is an event occurring in grid g_n if $P(\hat{X}_n|\mathcal{D}_n) \geq 0.8$. If ground truth $X_n = 1$, then we consider the event is accurately inferred. Otherwise, an event is detected by mistake, or it is not found. Specifically, precision is calculated as the ratio between the number of accurately detected events and the number of estimations with $P(\hat{X}_n|\mathcal{D}_n) \geq 0.8$, while recall is calculated as the ratio between the number of accurately detected events and the number of events.

The default setting of all parameters in our simulations is illustrated as follows. All simulations are performed on a square area divided into $20 * 20$ grids (i.e., $N = 400$). Events randomly occur in 40 grids of them. There are 500 collaborative users in the system. For each user u_k, the true-positive and false-positive detection probabilities are randomly generated within $[0.5, 1]$ and $[0, 0.5]$ respectively, and cost c_k is uniformly distributed between \$0 and \$5. We limit the upper bound of the number of grids past by a user as 10. To generate the trajectory of a user, we first randomly choose a grid as the starting point. Next, the user may stay in the grid or move towards any direction[1]. For each grid, $P(\hat{X}_n)$ is initialized as 0.5. The default value of budget is set as 400. All simulation results are the average of 20 runs.

5.2 Performance Comparison

In this subsection, we evaluate the performance achieved by our adaptive greedy algorithm and the three baselines, by varying the number of users, the budget, and the number of events.

The performance achieved by different algorithms, when the number of users varies from 400 to 800, is plotted in Figs. 2, 3, and 4, respectively. We can find that generally the more users available in the system, the better performance

[1] We consider there are eight directions: northward, southward, westward, eastward, northwestward, northeastward, southwestward, southeastward.

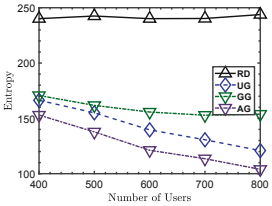

Fig. 2. Entropy vs. number of users.

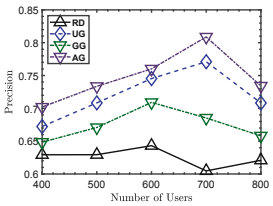

Fig. 3. Precision vs. number of users.

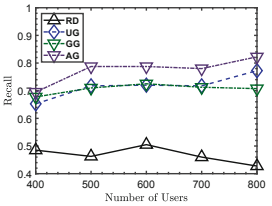

Fig. 4. Recall vs. number of users.

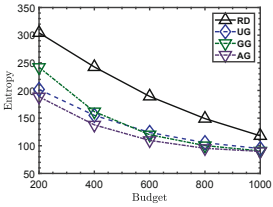

Fig. 5. Entropy vs. budget.

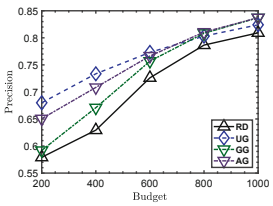

Fig. 6. Precision vs. budget.

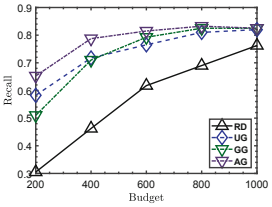

Fig. 7. Recall vs. budget.

achieved by each algorithm. Apparently, our algorithm outperforms the baselines in terms of the three metrics, no matter how many users there are. The user-greedy algorithm performs secondly, better than the other two baselines, because users with more observations per cost are recruited. When there are 700 users, entropy achieved by our algorithm is 15% lower than the user-greedy algorithm, and precision and recall are 5.2% and 8.3% higher than the user-greedy algorithm, respectively.

In Figs. 5, 6, and 7, we evaluate the performance achieved when the budget varies from 200 to 1000. Intuitively, the more budget is provided to recruited users, the higher detection accuracy can be achieved, while the marginal increment is reduced. We can find that our algorithm outperforms the baselines, except achieves a little lower precision than the user-greedy algorithm when the budget is less than 800. It may be caused by recruiting users with low detection accuracy, who report wrong observations in grids without events occurring. Specially, when budget is 200, our algorithm obtains 12% and 27% higher recall than the user-greedy algorithm and the grid-greedy algorithm, respectively.

We also vary the number of events from 10 to 50, to compare its impact on the performance of different algorithms, as shown in Figs. 8, 9, and 10. It can be found that precision achieved by the four algorithms increases when there are more events occurring, as the number of events detected by mistake deceases. On the other hand, entropy and recall have no obvious variation trend with the increase of number of events. Specially, when there are 40 events, entropy achieved by our algorithm is 12.5% and 17.3% lower than the user-greedy algorithm and the grid-greedy algorithm.

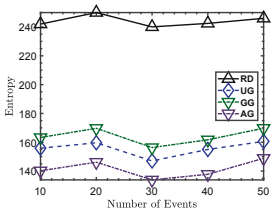

Fig. 8. Entropy vs. number of events.

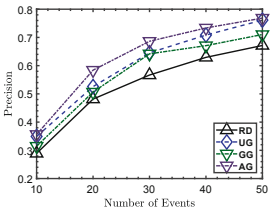

Fig. 9. Precision vs. number of events.

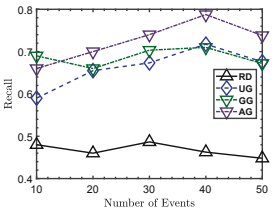

Fig. 10. Recall vs. number of events.

6 Conclusions

In this work, we have proposed a new approach for accuracy-guaranteed event detection via collaborative mobile crowdsensing with unreliable users. We first bridge the relationship between the uncertainty of event detection given observations of recruited users and the unreliability of recruited users, by building probabilistic models and applying the Bayesian rule. Then, leveraging the adaptive monotonicity and the adaptive submodularity, we propose an adaptive greedy algorithm for user recruitment, which is rigorously proved to achieve $(1 - 1/e)$-approximated performance. Extensive simulations are performed, whose results show that our algorithm outperforms the baselines in terms of achieving low entropy and high detection accuracy under different settings. When the budget is very limited, e.g., 200, our algorithm achieves at least 12% higher detection accuracy than the baselines.

Acknowledgements. This research is supported by NSFC (No. 61772341, 61472254, and 61802245), STSCM (No. 18511103002 and No. 16010500400), and KQJSCX20180329191021388. This work is also supported by the Program for Changjiang Young Scholars in University of China, the Program for China Top Young Talents, the Program for Shanghai Top Young Talents, Shanghai Engineering Research Center of Digital Education Equipment, SJTU Global Strategic Partnership Fund (2019 SJTU-HKUST), and the Shanghai Sailing Program (No. 18YF1408200).

References

1. Bao, X., Choudhury, R.R.: MoVi: mobile phone based video highlights via collaborative sensing. In: International Conference on Mobile Systems, Applications, and Services, pp. 357–370 (2010)
2. Cormen, T.T., Leiserson, C.E., Rivest, R.L.: Introduction to algorithms. Resonance **1**(9), 14–24 (2009)
3. Cox, L.P., Dalton, A., Marupadi, V.: SmokeScreen: flexible privacy controls for presence-sharing. In: International Conference on Mobile Systems, Applications, and Services, pp. 233–245 (2007)
4. Eagle, N., Pentland, A.: Social serendipity: mobilizing social software. IEEE Pervasive Comput. **4**(2), 28–34 (2005)

5. Gao, C., Kong, F., Tan, J.: HealthAware: tackling obesity with health aware smart phone systems. In: International Conference on Robotics and Biomimetics, pp. 1549–1554 (2009)
6. Golovin, D., Krause, A.: Adaptive submodularity: theory and applications in active learning and stochastic optimization. J. Artif. Intell. Res. **42**, 427–486 (2011)
7. Guo, B., et al.: Mobile crowd sensing and computing: the review of an emerging human-powered sensing paradigm. ACM Comput. Surv. (CSUR) **48**(1), 7 (2015)
8. Guo, B., Yu, Z., Zhou, X., Zhang, D.: From participatory sensing to mobile crowd sensing. In: 2014 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops), pp. 593–598. IEEE (2014)
9. Reddy, S., Parker, A., Hyman, J., Burke, J., Estrin, D., Hansen, M.: Image browsing, processing, and clustering for participatory sensing: lessons from a dietsense prototype. In: The Workshop on Embedded Networked Sensors, pp. 13–17 (2007)
10. Jin, H., Su, L., Chen, D., Nahrstedt, K., Xu, J.: Quality of information aware incentive mechanisms for mobile crowd sensing systems. In: ACM International Symposium on Mobile Ad Hoc Networking and Computing, pp. 167–176 (2015)
11. Jin, H., Su, L., Nahrstedt, K.: Theseus: incentivizing truth discovery in mobile crowd sensing systems. In: ACM International Symposium on Mobile Ad Hoc Networking and Computing, p. 1 (2017)
12. Krause, A., Guestrin, C.: Near-optimal observation selection using submodular functions. In: AAAI 2007, pp. 1650–1654 (2007)
13. Liu, C.H., Hui, P., Branch, J.W., Bisdikian, C., Yang, B.: Efficient network management for context-aware participatory sensing. In: 2011 8th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks (secon), pp. 116–124. IEEE (2011)
14. Liu, S., Zheng, Z., Wu, F., Tang, S., Chen, G.: Context-aware data quality estimation in mobile crowdsensing. In: INFOCOM 2017 - IEEE Conference on Computer Communications, pp. 1–9. IEEE (2017)
15. Ma, H., Zhao, D., Yuan, P.: Opportunities in mobile crowd sensing. IEEE Commun. Mag. **52**(8), 29–35 (2014)
16. Maisonneuve, N., Stevens, M., Niessen, M.E., Steels, L.: NoiseTube: measuring and mapping noise pollution with mobile phones. Environ. Sci. Eng. **2**(6), 215–228 (2009)
17. Mendez, D., Perez, A.J., Labrador, M.A., Marron, J.J.: P-sense: a participatory sensing system for air pollution monitoring and control. In: 2011 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops), pp. 344–347. IEEE (2011)
18. Mohan, P., Padmanabhan, V.N., Ramjee, R.: Nericell: rich monitoring of road and traffic conditions using mobile smartphones. In: Proceedings of the 6th ACM Conference on Embedded Network Sensor Systems, pp. 323–336. ACM (2008)
19. Nemhauser, G.L., Wolsey, L.A., Fisher, M.L.: An analysis of approximations for maximizing submodular set functions–I. Math. Program. **14**(1), 265–294 (1978)
20. Oliver, N., Floresmangas, F.: HealthGear: automatic sleep apnea detection and monitoring with a mobile phone. J. Commun. **2**(2)(2007)
21. Peng, D., Wu, F., Chen, G.: Pay as how well you do: a quality based incentive mechanism for crowdsensing. In: ACM International Symposium on Mobile Ad Hoc Networking and Computing, pp. 177–186 (2015)
22. Rana, R.K., Chou, C.T., Kanhere, S.S., Bulusu, N., Hu, W.: Ear-phone: an end-to-end participatory urban noise mapping system. In: Proceedings of the 9th ACM/IEEE International Conference on Information Processing in Sensor Networks, pp. 105–116. ACM (2010)

23. Song, Z., Zhang, B., Liu, C.H., Vasilakos, A.V.: QoI-aware energy-efficient participant selection. In: Eleventh IEEE International Conference on Sensing, Communication, and Networking, pp. 248–256 (2014)
24. Stone, J.V.: Bayes' Rule: A Tutorial Introduction to Bayesian Analysis. Sebtel Press, Sheffield (2013)
25. Thiagarajan, A., et al.: VTrack: accurate, energy-aware road traffic delay estimation using mobile phones. In: ACM Conference on Embedded Networked Sensor Systems, pp. 85–98 (2009)
26. Wang, H., Guo, S., Cao, J., Guo, M.: MeLoDy: a long-term dynamic quality-aware incentive mechanism for crowdsourcing. IEEE Trans. Parallel Distrib. Syst. **PP**(99), 1 (2018)
27. Wang, J., Tang, J., Yang, D., Wang, E., Xue, G.: Quality-aware and fine-grained incentive mechanisms for mobile crowdsensing. In: IEEE International Conference on Distributed Computing Systems, pp. 354–363 (2016)
28. Wen, Y., et al.: Quality-driven auction-based incentive mechanism for mobile crowd sensing. IEEE Trans. Veh. Technol. **64**(9), 4203–4214 (2015)
29. Xu, Q., Zheng, R.: When data acquisition meets data analytics: a distributed active learning framework for optimal budgeted mobile crowdsensing. In: INFOCOM (2017)
30. Yang, S., Wu, F., Tang, S., Gao, X., Yang, B., Chen, G.: Good work deserves good pay: a quality-based surplus sharing method for participatory sensing. In: IEEE International Conference on Parallel Processing, pp. 380–389 (2015)
31. Zheng, Z., Yang, Z., Wu, F., Chen, G.: Mechanism design for mobile crowdsensing with execution uncertainty. In: IEEE International Conference on Distributed Computing Systems, pp. 955–965 (2017)