# Context-Aware Data Quality Estimation in Mobile Crowdsensing

Shengzhong Liu[†*], Zhenzhe Zheng[†*], Fan Wu[†‡], Shaojie Tang[§], and Guihai Chen[†]
[†]Shanghai Key Laboratory of Scalable Computing and Systems, Shanghai Jiao Tong University, China
[§]Department of Information Systems, University of Texas at Dallas, USA
{liushengzhong1023, zhengzhenzhe220}@gmail.com; {fwu, gchen}@cs.sjtu.edu.cn; [§]shaojie.tang@utdallas.edu

***Abstract*—With the rapid growth of smart devices, mobile crowdsensing is becoming an important paradigm to acquire information from physical environments. Considering that the sensing data collected by mobile users are normally noisy and imprecise, one of the pressing problems in mobile crowdsensing is to evaluate the data quality in real time and to steer users to acquire data with high quality. However, it is challenging to estimate the data quality without the availability of ground truth data. In this paper, we observe that sensing context has a significant impact on data quality, which motivates us to propose a context-aware data quality estimation scheme. With historical sensing data, we train a context-quality classifier, which captures the relation between context information and data quality, to estimate data quality in an online manner. We apply such a context-aware data quality estimation scheme to guide user recruitment in mobile crowdsensing. We model the process of user recruitment as a stochastic submodular maximization problem, and design a random adaptive greedy algorithm to guarantee a constant approximation ratio. We evaluate our algorithm on a real-world temperature data set. The evaluation results show that our algorithm outperforms other existing techniques, in terms of prediction accuracy.**

## I. Introduction

In recent years, with the explosive increasing of smart devices embedded with various powerful sensors (*e.g.*, camera, microphone, accelerometer, digital compass, gyroscope, etc.), mobile crowdsensing (MCS) has been recognized as an innovative sensing data gathering paradigm [10], [18]. It has permeated many aspects of our daily life, including recording personal body indexes for health care [25], measuring environment phenomena like pollution level [10], monitoring traffic conditions (*e.g.*, availability of parking lot [19] and road congestion [20]), and sharing exercise data in social communities [11].

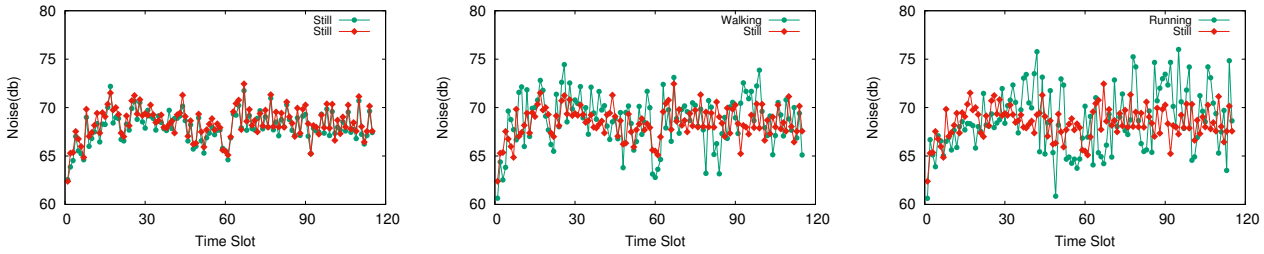
The main feature of mobile crowdsensing is the involvement of mobile users, which may be a double edged sword. On one hand, the service provider can leverage the intelligence of mobile users to improve the efficiency of data acquisition. For example, mobile users can easily identify the location of available parking lots and report them with pictures and comments, achieving much higher flexibility than the currently used ultrasound-based scanning system. On the other hand, compared with traditional wireless sensor networks [29], user's involvement may introduce even higher uncertainty in data quality, mainly due to various human activities during data collection. For example, in a noise measurement crowdsensing system, the collected sensing data would have poor quality if mobile users put their smartphones in pockets or even walk or run during data acquisition process.

Ensuring high data quality is a fundamental requirement to guarantee the success of mobile crowdsensing, which is the basic for other design components, such as user recruitment and incentive mechanism design. The data quality measures the degree of deviation to ground truth data, and is sometimes defined as data noise. There are many factors, captured by *sensing context* in this paper, that can influence the sensing data quality. We further divide these factors into two categories: hardware factors (*e.g.*, phone brand, sensor models, sensor calibration level, and etc) and human behaviour factors (*e.g.*, holding position of smartphone, human movement during sensing, and etc). Unfortunately, most of existing works in mobile crowdsensing did not investigate the impact of sensing context on data quality, or simply use a constant parameter to describe data quality [15]. However, different user activities in diverse context environments will lead to significantly different data qualities, indicating that a single constant parameter is not sufficient enough to describe the quality of data from mobile users. Considering that data quality may change over time (human behavior is always dynamic and variable), we have to determine the data quality in a real time manner, which is not a easy job without knowing the ground truth data. Peng *et al.*, [24] used unsupervised learning technique to estimate data quality, but this can only be done after collecting the historical data from all users. Our work, on the contrary, aims to develop a framework that can estimate the data quality on-the-fly.

There exist many challenges in the design of data quality estimation scheme. We list the major ones as follows:

- **Lack of Ground Truth**: It is trivial to estimate the data quality with the availability of ground truth. The sensing data consists of two parts: ground truth and data noise. With the ground truth data, we can extract the data noises from the sensing data directly, and further calculate its corresponding data quality in real time. However, in most crowdsensing applications, it is hard or even impossible to obtain the ground truth data, leading to the failure of ground truth-based data quality estimation schemes.
- **Lack of Historical Data**: Even without knowing the ground truth data, we can still determine the data quality in an offline manner as long as there is enough historical data. Given the historical data for a certain task, we can first estimate its corresponding ground truth, and then apply the ground truth-based scheme to calculate the data quality. However, in many scenarios, such as user recruitment, we have to determine the

[*]S. Liu and Z. Zheng make the same contribution to this work.
[‡]F. Wu is the corresponding author.
This work was supported in part by the State Key Development Program for Basic Research of China (973 project 2014CB340303), in part by China NSF grant 61672348, 61672353, 61422208, 61472252, 61272443 and 61133006, in part by Shanghai Science and Technology fund 15220721300, in part by CCF-Tencent Open Fund, and in part by the Scientific Research Foundation for the Returned Overseas Chinese Scholars. The opinions, findings, conclusions, and recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the funding agencies.Z. Zheng was also supported by Google PhD Fellowship and Microsoft Asia PhD Fellowship.




(a) Still vs Still (b) Still vs Walking (c) Still vs Running

Figure 1: An example to illustrate the influence of different phone sensing contexts on the quality of sensing data.

data quality for a new task in an online manner, such that no historical data could be used to calculate the ground truth and estimate the data quality.

Considering the challenges above, it is difficult to estimate data quality directly. In this paper, we seek to build a connection between contextual information and data quality, based on which we infer the data quality through the real-time contextual information. Specifically, we design this context-aware data quality estimation scheme by using supervised learning algorithms to train a context-quality classifier. In order to prepare training data set, we have to obtain data quality and contextual information for each piece of historical data. For data quality, we first build a Gaussian Mixture Model to describe sensing data, and propose an Expectation Maximization (EM)-based algorithm to estimate the ground truth of each task. Based on the ground truth data, we then calculate the data noises of the historical data, and derive the data quality distribution for each mobile user by applying the maximum likelihood estimation (MLE) method. With the data quality distribution as the prior distribution, we determine the data quality for each piece of historical data using the maximum a posterior probability (MAP) approach. To obtain contextual information, we directly read the hardware information from smart devices, and apply activity recognition technique to detect human behaviour based on the sensing data from idle sensors.

With the context-quality classifier, we are able to determine the data quality with the aid of contextual information, in a real-time manner. We integrate our context-aware data quality estimation scheme into a specific application: user recruitment, and model it as a stochastic submodular maximization problem. We cannot directly adopt the classical greedy algorithm from submodular optimization, because the data quality of each user is not determined in advance. Taking advantage of adaptive submodularity property, we design a random adaptive greedy algorithm, achieving a constant approximation ratio of $1/e$.

We summarize the main contributions of this paper:

• First, we make an in-depth study on real-time data quality estimation in mobile crowdsensing without the knowledge of ground truth data.

• Second, we investigate the relation between contextual information and data quality, and design a context-aware data quality estimation scheme. We calculate data quality and detect contextual information for each piece of historical data. With the training data set (contextual information and data quality pairs), we build a context-data quality classifier, which is used to estimate the data quality in real time.

• Third, we use the context-aware data quality estimation scheme to guide user recruitment process. We model it as an adaptive non-monotone submodular maximization problem, and propose a random adaptive greedy algorithm with a constant approximation ratio of $1/e$.

• Finally, we show the performance of our algorithm through simulations on a real-world sensing data set. The simulation results show that our algorithm outperforms the previous techniques in terms of prediction accuracy.

The rest of this paper is organized as follows: In Section II, we use a simple experiment to demonstrate the influence of sensing context on data quality. The context-quality model and problem formulation of user recruitment are presented in Section III. In Section IV, we illustrate the design details of the context-aware data quality estimation scheme, and further apply it to guide user recruitment in Section V. The evaluation results are shown in Section VI, followed by related work in Section VII. Finally, we conclude the paper in Section VIII.

## II. A Motivating Example

In order to capture the influence of different sensing contexts on sensing data quality, we carry out a simple experiment about noise pollution profile description in campus. In this experiment, we mainly focus on the human factor, *i.e.*, the influence of user activity on the data quality.

We install Noisetube [23] app on four Apple iPhone 6, and use the embedded acoustic sensor to measure the noise levels in our campus. Four volunteers holding these devices measure the noise level at the same location, but may be in different contexts, simultaneously. Two volunteers keep still during sensing, one volunteer keeps walking, and the remaining one keeps running. The whole process lasts for about 115 time slots. One slot is set to be two seconds.

We first compare the data quality collected by different volunteers under the same activity. The noise measurements from the two standing volunteers is presented in Figure 1(a). We can see that the two lines almost overlap with each other, implying that the impact of hardware on the data quality is negligible. We then compare the measurements from one standing volunteer and one walking volunteer. As shown in Figure 1(b), there exists obvious difference between the two lines, where the variance of the line corresponding to the walking volunteer is apparently greater than that of the standing volunteer. Finally, the measurements from one standing volunteer and one running volunteer are reported in Figure 1(c). We can observe that the variance of the line corresponding to the running volunteer is even greater than that of the walking volunteer.

The above experiment results indicate that different sensing contexts indeed have a significant impact on the quality of sensing data. It motivates us to explore the relation between context and data quality, and leverage such relation to estimate the data quality in the scenario that the ground truth data is unavailable while the contextual information is easy to collect.

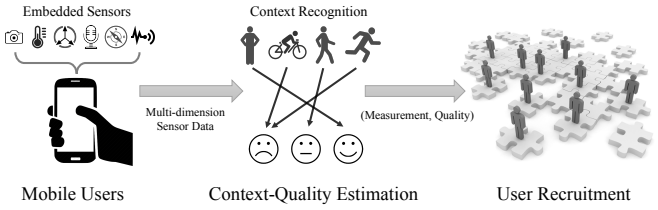



Figure 2: System Overview

## III. Preliminaries and Problem Formulation

In this section, we first present a system overview. We then describe the context-aware data quality estimation model. After that, we formally formulate the problem of user recruitment.

### A. System Overview

A typical mobile crowdsensing system contains three major components: a service provider, a set of clients, and a set of mobile users. The service provider is the central platform connecting clients and mobile users. The service provider receives queries about information in specific locations from clients, and announces a set of points of interest (PoIs) $\mathbb{L} \triangleq \{l_1, l_2, \ldots, l_M\}$. The PoIs are the physical locations, at which the service provider intends to acquire sensing data to answer the queries of clients. Based on the consideration of their current locations and available resources, the mobile users choose the preferred PoIs, indicating that they are willing to carry out the corresponding sensing tasks at these PoIs. We denote the $M$ mobile users by a set $\mathbb{V} = \{v_1, v_2, \ldots, v_M\}$. For convenience of discussion, we assume there is exact one mobile user at each of PoIs. Our results can be easily extended to the scenario that multiple mobile users stay at one PoI.

Due to the unreliable sensors and dynamic environment, the sensing data collected by mobile users are normally noisy and imprecise. Considering the uncertainty of sensing data, the service provider faces two fundamental problems in mobile crowdsensing: *how to evaluate the quality of collected sensing data in real time and how to recruit the reliable mobile users to maximize the service utility?* As shown in Figure 2, the service provider leverages the historical sensing data to estimate the relation between contextual information and data quality, and to measure the reliability of mobile users, which are two critical steps to tackle the above two problems. For the context-quality relation, we construct a classifier to map contextual information to data quality, such that we can evaluate the data quality through the real-time contextual information. For the reliability measurement, the service provider uses the historical sensing data of a mobile user to calculate her data quality distribution, indicating the probability of data quality of this user in future data acquisition. Such data quality distribution is considered as the reliability of the mobile user. We adopt the reliability of mobile users as selection criteria in user recruitment, leading to a high service utility.

### B. Data Quality and Context

We assume that the data quality can only be chosen from $N$ different levels, each of which corresponds to an independent gaussian distribution. The gaussian distribution has been widely used to describe sensing data [9], [14]. Given a sensing task, the sensing data from the $\phi$th data quality level is regarded as a random sample from the Gaussian distribution $\mathcal{N}(\mu, \sigma_\phi^2)$, where the mean $\mu$ is the ground truth of the task and the variance $\sigma_\phi^2$ is a predefined constant number. We denote the total possible data quality levels by a set $\mathbb{N} = (1, 2, \cdots, N)$.

Mobile users may involve in various and dynamic contexts during data acquisition process, leading to the collected data in diverse data quality levels. It is not sufficient enough to use a single fixed constant parameter to describe the data quality level of each user over the time. Therefore, for each mobile user $v_i \in \mathbb{V}$, we use a random variable $\Phi_i$ to represent her possible data quality level. The random variable $\Phi_i$ follows a multinomial distribution with the parameters $\{\pi_i(\phi), \phi \in \mathbb{N}\}$, where $\pi_i(\phi) = \mathbb{P}(\Phi_i = \phi)$, $0 \leq \pi_i(\phi) \leq 1$ and $\sum_{\phi=1}^{N} \pi_i(\phi) = 1$. We call such multinomial distribution $\{\pi_i(\phi)\}$ as *quality distribution* for the mobile user $v_i \in \mathbb{V}$. We use the vector $\boldsymbol{\Phi} = (\Phi_1, \Phi_2, \cdots, \Phi_M)$ to denote the random variables of all the mobile users.

As for contextual information, we represent it using a feature vector $\mathbf{c} = (c_1, c_2, \ldots, c_Q)$, in which each element denotes activity information or hardware information. For example, we can use $c_i$ to denote whether the user is walking or not. For hardware information, we can use one feature element to represent mobile phone brand and another for calibration level of sensors.

### C. User Recruitment

We model user recruitment in the scenario of unknown data quality level as a stochastic submodular maximization problem. The ground set is all the mobile users $\mathbb{V}$. Mobile users may have different possible data quality levels in different context situations. For a specific context, we can realize the random variable $\Phi_i$ to be a certain value $\phi_i$, and denote the realizations of all the mobile users as $\boldsymbol{\phi} = (\phi_1, \phi_2, \cdots, \phi_M)$. Before selecting mobile users, the service provider only knows the quality distributions of mobile users, and can calculate the probability distribution $\mathbb{P}(\boldsymbol{\phi})$ over a possible realization $\boldsymbol{\phi}$, *i.e.*, $\mathbb{P}(\boldsymbol{\phi}) = \prod_{i=1}^{M} \pi_i(\phi_i)$. Once the service provider selects a mobile user, she can collect the contextual information and exploit the context-quality classifier to determine the mobile user's specific data quality level. Therefore, for a selected subset of mobile users, the service provider can observe their partial realization, denoted by $\psi \subseteq \mathbb{V} \times \mathbb{N}$, which is a collection of mobile users-data quality level pairs $(v, \phi)$. For a partial realization $\psi$, we use $dom(\psi)$ to represent the contained users, *i.e.*, $dom(\psi) = \{v \in \mathbb{V} \mid \exists \phi \in \mathbb{N} : (v, \phi) \in \psi\}$. We write $\psi(v) = \phi$, when $(v, \phi) \in \psi$. Additionally, we call a partial realization $\psi$ consistent with a full realization $\boldsymbol{\phi}$, denoted by $\boldsymbol{\phi} \sim \psi$ when $\psi(v) = \boldsymbol{\phi}(v)$, for all $v \in dom(\psi)$, meaning that the realized quality levels of a user subset according to $\psi$ agree with that of the ground set according to $\boldsymbol{\phi}$.

The utility function of the service provider is $F : 2^{\mathbb{V}} \times N^{\mathbb{V}} \to \mathbb{R}$, which assigns a value to every subset of mobile users and the corresponding data quality level realization. The primal goal of the service provider is to select a user subset $\mathcal{V} \subseteq \mathbb{V}$ to maximize the resulting expected utility, subjecting to the cardinality constraint $|\mathcal{V}| \leq K$. We can formulate the user recruitment problem in unknown data quality scenario as:

$$\mathcal{V}^* = \underset{\mathcal{V} \in \mathbb{V}, |\mathcal{V}| \leq K}{\arg\max} \ \mathbb{E}[F(\mathcal{V}, \boldsymbol{\phi})], \tag{1}$$

where $\mathbb{E}[F(\mathcal{V}, \boldsymbol{\phi})]$ is the expected utility with respect to the realization probability distribution $\mathbb{P}(\boldsymbol{\phi})$:

$$\mathbb{E}[F(\mathcal{V}, \boldsymbol{\phi})] \triangleq \sum_{\boldsymbol{\phi}} \mathbb{P}(\boldsymbol{\phi}) \times F(\mathcal{V}, \boldsymbol{\phi}). \tag{2}$$

We now define an important property: *adaptive submodularity*, that the utility function should satisfy to achieve good performance guarantee. Adaptive submodularity is an extension of submodularity to adapt to the scenarios of unknown quality realization. We first give the formal definition of submodular function.

**Definition 1** (Submodularity)**.** *Given the ground set $\mathbb{V}$, a set function $f: 2^{\mathbb{V}} \to \mathbb{R}$ is called submodular if, for any $A \subseteq B \subseteq \mathbb{V}$ and $v \in \mathbb{V} \backslash B$, it satisfies that: $f(A \cup \{v\}) - f(A) \geq f(B \cup \{v\}) - f(B)$.*

Although many submodular maximization problems are NP-hard, we can apply the simple greedy algorithms to derive near-optimal performance [4], [5], [14]. However, the greedy algorithms cannot guarantee performance when the realization is not available before running the algorithms. Golovin and Krause introduced the concept of adaptive submodularity [13] to deal with the unknown realization scenario. We present the definition of conditional expected marginal utility.

**Definition 2** (Conditional Expected Marginal Utility)**.** *Given a partial realization $\psi$ and a user $v$, the conditional expected marginal utility of $v$ conditioned on $\psi$ is:*

$$\Delta(v|\psi) = \mathbb{E}[F(dom(\psi) \cup \{v\}, \boldsymbol{\phi}) - F(dom(\psi), \boldsymbol{\phi})|\boldsymbol{\phi} \sim \psi]. \quad (3)$$

*$\psi$ is consistent with $\boldsymbol{\phi}$, which is the realization of ground set.*

**Definition 3** (Adaptive Submodularity [13])**.** *Given the ground set $\mathbb{V}$, a set function $F$: $2^{\mathbb{V}} \times \mathbb{N}^{\mathbb{V}} \to \mathbb{R}$ is called adaptive submodular with respect to the realization distribution $\mathbb{P}(\Phi)$ if, for any partial realizations $\psi$ and $\psi'$, where $\psi$ is a subrealization of $\psi'$* i.e.*, $dom(\psi) \subseteq dom(\psi')$, and for any $v \in \mathbb{V} \backslash dom(\psi')$, it satisfies that: $\Delta(v|\psi) \geq \Delta(v|\psi')$.*

**Definition 4** (Adaptive Monotonicity [13])**.** *Given the ground set $\mathbb{V}$, a set function $F$: $2^{\mathbb{V}} \times \mathbb{N}^{\mathbb{V}} \to \mathbb{R}$ is called adaptive monotone with respect to the distribution $\mathbb{P}(\Phi)$ if, for any partial realization $\psi$ and user $v$, it satisfies that: $\Delta(v|\psi) \geq 0$.*

## IV. Context-aware Data Quality Estimation

In this section, we describe the design details of our context-aware data quality estimation scheme. The basic idea is to exploit contextual information to infer the data quality of mobile users, without knowing the ground truth of the sensing tasks. To fulfill this goal, we build a connection between contextual information and data quality by training a context-data quality classifier based on the historical sensing data.

### A. Quality Estimation

Expectation Maximization (EM) algorithm is a classical iterative method for finding maximum likelihood or maximum posteriori estimation of parameters in statistical models [8]. The key parts of adopting the EM algorithm are the choice of unobserved latent variables and the design of likelihood function. Different from the previous work about data quality management in mobile crowdsensing [24], [27], which choose the ground truth as the latent variable, we regard the latent variable as the data quality level $\boldsymbol{\Phi}$ of mobile users. We assume that there are $T$ tasks in historical data set, and $u_j$ is the ground truth of the $j$th task. We use a $M \times T$ matrix $\mathbf{X}$ to denote the historical data set, where $x_{ij}$ is the data that the mobile user $i$ collects for the $j$th task. Without loss of generality, we assume that each mobile user has carried out all the $T$ tasks, *i.e.*, $x_{ij} > 0$. We represent the collected data of the mobile user $i$ and the observations for the $j$th task as $x_{i,*} = (x_{i1}, x_{i2}, \cdots, x_{iT})$ and $x_{*,j} = (x_{1j}, x_{2j}, \cdots, x_{Mj})$, respectively.

Before estimating the data quality distribution, we first calculate the ground truth for each task by introducing a Gaussian Mixture Model (GMM). We can consider the $M$ data $X_{*,j}$ for the $j$th task are i.i.d samples from a probability density distribution (pdf) $p_\theta(x)$, which is the mixture of $N$ univariate Gaussian distributions:

$$p_\theta(x) = \sum_{i=1}^{M} g_i \sum_{\phi=1}^{N} \pi_i(\phi) N(x; \mu_j, \sigma_\phi^2) = \sum_{\phi=1}^{N} \bar{\pi}(\phi) N(x; \mu_j, \sigma_\phi^2),$$

where $N(x, \mu_j, \sigma_\phi^2)$ denotes the Gaussian pdf with mean $\mu_j$ and variance $\sigma_\phi^2$:

$$N(x; \mu_j, \sigma_\phi^2) \triangleq \frac{1}{\sqrt{2\pi}\sigma_\phi} exp\left(-\frac{x - \mu_j}{2\sigma_\phi^2}\right).$$

The value $g_i = 1/M$ is the probability that the data comes from the mobile user $i$. The probability $\bar{\pi}(\phi)$ can be considered as the average of all the $\pi_i(\phi)$, *i.e.*, $\bar{\pi}(\phi) = \sum_{i=1}^{M} g_i \pi_i(\phi)$. We note that the Gaussian distributions share the same mean $u_j$, which is the ground truth of the $j$th sensing task. To simply the notation, we omit the index $j$ in the following discussion. We can interpret such GMM model for the generalization of the observation data: we first draw a data quality level that takes value $\phi$ with probability $\bar{\pi}(\phi)$, and then generate the observation data $X_i \sim N(\mu, \sigma_\phi^2)$.

In GMM model, we consider mean $u$ and average probabilities $\{\bar{\pi}(\phi)\}$ as the parameters: $\theta \triangleq \{u, \bar{\pi}(\phi), 1 \leq \phi \leq N\}$, while the variance $\{\sigma_\phi^2\}$ is the given constant numbers. The latent variable is the average data quality level $\bar{\Phi}$, which follows the multinomial distribution with $\{\bar{\pi}(\phi)\}$. We introduce the incomplete-data log likelihood function for $\theta$ as

$$l_{id}(\theta) = \sum_{i=1}^{M} \log p_\theta(x_i) = \sum_{i=1}^{M} \log \sum_{\phi=1}^{N} \bar{\pi}(\phi) N(x_i; \mu, \sigma_\phi^2).$$

Maximizing this likelihood function with respect to the parameters $\theta$ is a nonconcave maximization problem, leading to the intractability to derive closed form solutions. Therefore, we adopt EM algorithm to iteratively estimate the parameters.

We first introduce the complete data log likelihood function. Let the complete data be $(X_i, \bar{\Phi}_i)$, $1 \leq i \leq M$, where $\bar{\Phi}_i$ is the random variable selected to produce the data $X_i$.

$$\begin{aligned} l_{cd}(\theta) &= \sum_{i=1}^{M} \log r_\theta(x_i, \bar{\phi}_i) = \sum_{i=1}^{M} \left[\log \bar{\pi}(\bar{\phi}_i) + \log p_\theta(x_i|\bar{\phi}_i)\right] \\ &= \sum_{i=1}^{M} \left[\log \bar{\pi}(\bar{\phi}_i) + \log N(x_i; \mu, \sigma_{\bar{\phi}_i}^2)\right] \\ &= -\frac{M}{2}\log(2\pi) + \sum_{i=1}^{M} \left[\log \bar{\pi}(\bar{\phi}_i) - \log \sigma_{\bar{\phi}_i} - \frac{(x_i - \mu)^2}{2\sigma_{\bar{\phi}_i}^2}\right]. \end{aligned}$$

EM algorithm consists of two major steps: the expectation (E) step and the maximization (M) step.

- **E-step:** Compute the expectation of the complete data log likelihood function $l_{cd}(\theta)$, with respect to the conditional

distribution of latent variables $\bar{\Phi}$ under the current estimate of parameters $\theta^{(k)}$:

$$\begin{aligned} & Q(\theta|\theta^{(k)}) \triangleq \mathbb{E}_{\theta^{(k)}}\left[l_{cd}(\theta)|X_{*,j}=x_{*,j}\right] \\ \overset{(a)}{=} & -\frac{M}{2}\log(2\pi) \\ & +\sum_{i=1}^{M}\mathbb{E}_{\theta^{(k)}}\left\{\left[\log\bar{\pi}(\bar{\Phi}_i)-\log\sigma_{\bar{\Phi}_i}-\frac{(x_i-\mu)^2}{2\sigma_{\bar{\Phi}_i}^2}\right]\middle| X_i=x_i\right\} \\ \overset{(b)}{=} & -\frac{M}{2}\log(2\pi) \\ & +\sum_{i=1}^{M}\sum_{\phi=1}^{N}\left[\log\bar{\pi}(\phi)-\log\sigma_\phi-\frac{(x_i-\mu)^2}{2\sigma_\phi^2}\right]\bar{\pi}_{\theta^{(k)}}(\phi|x_i), \end{aligned} \quad (4)$$

where the (a) holds because the random variable $\bar{\Phi}_i$ is conditionally independent of $\{X_k\}_{k\neq i}$ given $X_i$. In (b), we have introduced the conditional probability distribution

$$\bar{\pi}_{\theta^{(k)}}(\phi|x) \triangleq \frac{\bar{\pi}^{(k)}(\phi)N(x;\mu^{(k)},\sigma_\phi^2)}{\sum_{\phi=1}^{N}\bar{\pi}^{(k)}(\phi)N(x;\mu^{(k)},\sigma_\phi^2)},\ 1\leq\phi\leq N. \quad (5)$$

• **M-step:** Calculate the updated parameters $\theta^{(k+1)}$ that maximize the $Q(\theta|\theta^{(k)})$ in Equation (4), *i.e.*,

$$\theta^{(k+1)} = \arg\max_{\theta} Q(\theta|\theta^{(k)}). \quad (6)$$

The function $Q(\theta|\theta^{(k)})$ is concave quadratic in $\mu$. Setting the partial derivatives of $Q(\theta|\theta^{(k)})$ with respect to $\mu$ to zero, we obtain

$$0=\frac{Q(\theta|\theta^{(k)})}{\partial\mu}=\sum_{i=1}^{M}\sum_{\phi=1}^{N}\left[\frac{x_i-\mu}{\sigma_\phi^2}\pi_{\theta^{(k)}}(\phi|x_i)\right].$$

Hence, the updated mean $\mu^{(k+1)}$ is:

$$\mu^{(k+1)}=\frac{\sum_{i=1}^{M}x_i\sum_{\phi=1}^{N}\frac{\bar{\pi}_{\theta^{(k)}}(\phi|x_i)}{\sigma_\phi^2}}{\sum_{i=1}^{M}\sum_{\phi=1}^{N}\frac{\bar{\pi}_{\theta^{(k)}}(\phi|x_i)}{\sigma_\phi^2}},$$

which is a weight average of the observations $x_{*,j}$.

To derive the update of the mixture probabilities, we maximize the $Q(\theta|\theta^{(k)})$ with respect to the $\bar{\pi}(\phi)$. Here, we must take account of the constraint that the mixture probabilities sum to one, *i.e.*, $\sum_{\phi=1}^{N}\bar{\pi}(\phi)=1$. This can be achieved using a Lagrange multiplier and maximizing the following quantity:

$$Q(\theta|\theta^{(k)})+\lambda\left(\sum_{\phi=1}^{N}\bar{\pi}(\phi)-1\right).$$

Setting the partial derivatives of the above equality with respect to $\bar{\pi}_\phi$, and we have

$$0=\sum_{i=1}^{M}\frac{1}{\bar{\pi}(\phi)}\bar{\pi}_{\theta^{(k)}}(\phi|x_i)+\lambda.$$

If we now multiple both side by $\bar{\pi}(\phi)$ and sum over $N$, we can derive $\lambda=-M$. Using this to eliminate $\lambda$ and rearranging, we obtain the updated mixture probability $\bar{\pi}^{(k+1)}(\phi)$:

$$\bar{\pi}^{(k+1)}(\phi)=\frac{1}{M}\sum_{i=1}^{M}\bar{\pi}_{\theta^{(k)}}(\phi|x_i),\quad \forall 1\leq\phi\leq N.$$

We iteratively execute the E-step and the M-step until the converge condition holds. We can manually set the convergence condition, *e.g.*, the difference of log likelihood function between two iterations goes below a predefined threshold. The final ground truth for the task $j$ is denoted by $\mu_j^*$.

We now turn to calculate the data quality distribution, *i.e.*, the mixture probabilities $\{\pi_i(\phi)\}$, for each mobile user. We emphasize that the data quality levels are not relevant to the ground truths of the tasks, but only capture the noise of the collected data. We define the noise of the data collected by the mobile user $i$ for the task $j$ as $y_{ij}\triangleq x_{ij}-\mu_j^*$. For each mobile user $i$, we use the vector $y_{i,*}$ to denote the noises for all the collected data, *i.e.*, $y_{i,*}=(y_{i1},y_{i2},\cdots,y_{iT})$. We can consider that the data noises are drawn i.i.d from a probability density function $p_\theta(y)$, which is the mixture of $N$ univariate Gaussian distributions with respective probability $\pi_i(\phi)$, means 0, and variances $\sigma_\phi^2$, for $1\leq\phi\leq N$:

$$p_\theta(y)=\sum_{\phi=1}^{N}\pi_i(\phi)N(y;0,\sigma_\phi^2),$$

where $N(y;0,\sigma_\phi^2)$ is the Gaussian distribution with mean 0 and variance $\sigma_\phi^2$. In this Gaussian Mixture Model, the parameters $\theta$ are the mixture probabilities $\{\pi_i(\phi)\}, 1\leq\phi\leq N$. We try to derive the parameters $\theta$ to maximize the log likelihood function, subjecting to the constraint that the sum of the mixture probabilities is equal to 1. We formulate this log likelihood maximization problem as *MAX-ML*:

$$\begin{aligned} \text{Maximize} \quad & l_{id}(\theta) \\ \text{Subject to:} \quad & \sum_{\phi=1}^{N}\pi_i(\phi)=1, \end{aligned} \quad (7)$$

where the log likelihood function is defined as:

$$l_{id}(\theta)\triangleq\sum_{j=1}^{T}\log p_\theta(y_j)=\sum_{j=1}^{T}\log\sum_{\phi=1}^{N}\pi_i(\phi)N(y_j;0,\sigma_\phi^2).$$

Since the quantities $y_j$ and $\sigma_\phi^2$ have been given, $N(y_j;0,\sigma_\phi^2)$ is a constant value, such that *MAX-ML* is a concave maximization problem. We can derive the optimal results, denoted as $\pi_i^*(\phi)$, by applying the classical optimization technique [3].

We describe the detailed steps of the quality estimation scheme in Algorithm 1. The input of the algorithm is the historical sensing data $\mathbf{X}$ over $T$ tasks from $M$ mobile users, and the output of the algorithm is the data quality distribution of each mobile user. We first estimate the ground truth $\mu_j^*$ of each task $j$ using the EM algorithm with the input of data $x_{*,j}$ from the $M$ mobile users, and then derive the quality distribution $\{\pi_i^*(\phi)\}$ for each mobile user $i$ by solving a concave maximization problem over the data $x_{i,*}$ of $T$ tasks. In the EM algorithm, we set the initial ground truth of the $j$th task to be the average of all the data for the $j$th task, and the average quality distribution to be a uniform distribution (Lines 3 to 6). After the initialization stage, we iteratively execute the E-step (Lines 9 to 10) and the M-step (Lines 12 to 15) until the converge condition holds. To derive the data quality distribution $\{\pi_i^*(\phi)\}$ for the mobile user $i$, we calculate the noise vector $y_{i,*}$ for her collected data based on the estimated ground truths for the tasks, and then solve the *MAX-ML* problem using the optimization technique (Lines 20 to 22).

**Algorithm 1:** EM-based Quality Estimation Algorithm

**Input**: Historical data set $\mathbf{X}$; A set of variances $\{\sigma_\phi^2\}$
**Output**: Data quality distribution $\{\pi_i^*(\phi)\}$ of each user $i \in \mathbb{V}$.

// Estimate the ground truth
**for** $j = 1$ *to* $T$ **do**
  $\mu_j = \frac{\sum_{i=1}^{M} x_{ij}}{M}$;
  **for** $\phi = 1$ *to* $N$ **do**
    $\bar{\pi}(\phi) = \frac{1}{N}$;
  $k \leftarrow 0$; $\theta^{(k)} \leftarrow (\{\bar{\pi}(\phi)\}, \mu_j)$;
  **while** *not converged* **do**
    // E-step:
    Calculate $\bar{\pi}_{\theta^{(k)}}(\phi|x_{ij})$ using Equation (5);
    $Q(\theta|\theta^{(k)}) \leftarrow -\frac{M}{2}\log(2\pi) + \sum_{i=1}^{M}\sum_{\phi=1}^{N}\left[\log\bar{\pi}(\phi) - \log\sigma_\phi - \frac{(x_{ij}-\mu_j)^2}{2\sigma_\phi^2}\right]\bar{\pi}_{\theta^{(k)}}(\phi|x_{ij})$;
    // M-step:
    $\mu_j^{(k+1)} \leftarrow \frac{\sum_{i=1}^{M} x_{ij}\sum_{\phi=1}^{N}\frac{\bar{\pi}_{\theta^{(k)}}(\phi|x_{ij})}{\sigma_\phi^2}}{\sum_{i=1}^{M}\sum_{\phi=1}^{N}\frac{\bar{\pi}_{\theta^{(k)}}(\phi|x_{ij})}{\sigma_\phi^2}}$;
    **for** $\phi = 1$ *to* $N$ **do**
      $\bar{\pi}^{(k+1)}(\phi) \leftarrow \frac{1}{M}\sum_{i=1}^{M}\bar{\pi}_{\theta^{(k)}}(\phi|x_{ij})$;
    $\theta^{(k+1)} \leftarrow (\{\bar{\pi}^{k+1}(\phi)\}, \mu_j^{k+1})$;
    $k \leftarrow k+1$;
  $\mu_j^* \leftarrow \mu_j^k$;
// Estimate the data quality distribution
**for** $i = 1$ *to* $M$ **do**
  **for** $j = 1$ *to* $T$ **do**
    $y_{ij} \leftarrow x_{ij} - \mu_j^*$;
  $\{\pi_i^*(\phi)\} \leftarrow$ Solving the MAX-ML problem in (7);
**return** $\{\pi_i^*(\phi)\}$ *for each user* $i \in \mathbb{V}$;

We can use the obtained data quality distribution to determine the data quality level of each historical sensing data, which will be regarded as the training data set in the context-quality classifier. Here, we follow the idea of maximum a posteriori estimation: selecting the parameter $\phi_{ij}$ that maximizes the posterior distribution to be the quality level of the data $x_{ij}$.

$$\phi_{ij} = \arg\max_{\phi} \pi_i^*(\phi) N(y_{ij}; 0, \sigma_\phi^2), \quad \forall 1 \le i \le M, 1 \le j \le T.$$

### B. Context Recognition

We now discuss the context recognition, which is to collect hardware information and activity information. The hardware information, such as the accuracy of accelerometer and the resolution of camera, can be read directly from the devices. Thus, the remaining challenge lies in how to assess the user activity during data acquisition process.

Mobile phones are embedded with a bundle of powerful sensors, such as accelerometers, microphones, GPS, and etc. We can exploit the data collected from these sensors to recognize diverse activities in different environments. The activity recognition through analyzing the sensing data have been widely studied in the ubiquitous computing literature [2], [17], [21]. For example, Kwapisz *et al.* implemented a system that uses accelerometers to perform activity recognition [17]. Mun *et al.* designed a hybrid approach utilizing both Wi-Fi and GSM signals to infer the mobility patterns of users [21]. For specific mobile crowdsensing application, we can adopt selected approaches to conduct the activity recognition. Combing with the obtained hardware information, we can construct the complete context vector.

### C. Context-Quality Classifier

We rely on the previous two components: data quality estimation and context recognition, to train context-quality classifier. The historical sensing data consists of primary data (the data used to answer the queries of clients) and secondary data (the data used to obtain context information). For each piece of historical data, we derive data quality from the primary data by applying the data quality estimation scheme, and extract context vector from the secondary data using the activity recognition technique. Taking these context and data quality level pairs as the training set, the service provider construct a context-quality classifier using the classical supervised learning algorithm. We intend to train a multi-class classifier through the binary classification algorithm, such as support vector machine (SVM) [6]. We adopt a classical method that train a binary classifier between each data quality level and the rest. Once we build the context-quality classifier, we can determine the data quality for the collected data only through the contextual information, even without knowing the ground truth of the sensing task, in a real-time manner.

## V. Adaptive User Selection

In this section, we apply the context-quality classifier to guide the process of user recruitment. We first extend the classical Gaussian Process to model sensing data in the scenarios of diverse data quality levels, and define the specific utility function based on the concept of mutual information. Taking advantage of the adaptive submodularity property of the utility function, we design a random adaptive user selection algorithm, achieving a constant approximation ratio.

### A. Quality-related Gaussian Process

We introduce the classical Gaussian Process (GP) model for sensing data, and extend it to involve the consideration of data quality. In GP model, every point is associated with a random variable, following a univariate Gaussian distribution. The joint distribution over a set of random variables is a multivariate normal distribution. The parameters of GP model are a mean vector $\boldsymbol{\mu}$ and a covariance matrix $\boldsymbol{\Sigma}$, which is a symmetric positive-definite matrix. In mobile crowdsensing, we can use the Gaussian Process to model the sensing data collected at POIs. Specifically, we associate each PoI $l \in \mathbb{L}$ with a random variable $X_l$, which follows a one-dimension Gaussian distribution with mean $\boldsymbol{\mu}_l$ and variance $\boldsymbol{\Sigma}_{l,l}$. For each PoIs pair $l_1, l_2 \in \mathbb{L}$, their covariance is $\boldsymbol{\Sigma}_{l_1,l_2}$. The GP model is extremely powerful to represent sensing data [14]. If we observe sensing data on a set of PoIs, we can predict the data at unobserved PoIs, and provide the variance of such prediction. The classical GP model either does not consider the noises of the observed data, or simply assume that all the data noises are sampled from the same distribution. However, as we have discussed, different context environments may result in different data quality levels, implying that the observed data may have different levels of noises. Thus, the classical GP model fails to describe the diverse data quality in crowdsensing.

Suppose the set of PoIs selected to observe data is $A \subseteq \mathbb{L}$ and the rest unobserved PoIs are $B = \mathbb{L} \backslash A$.[1] We have the following expressions for the means and variances:

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_A \\ \boldsymbol{\mu}_B \end{bmatrix} \quad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{AA} & \boldsymbol{\Sigma}_{AB} \\ \boldsymbol{\Sigma}_{BA} & \boldsymbol{\Sigma}_{BB} \end{bmatrix}$$

In order to make GP model accordant with our data quality model, we introduce a noise matrix $\boldsymbol{\Gamma}$, where the diagonal entries represent the variances of the noise distributions at the corresponding PoIs and the others are zeros. In this case, we can capture the scenario that the data collected at different PoIs have different noise levels. We note that the vector of the diagonal entries from the matrix $\boldsymbol{\Gamma}$ is actually a realization $\boldsymbol{\Phi}$ of data quality levels from mobile users. Given the observations $x_A$ at the selected PoIs $A$, we can predict the values at the unobserved PoIs $B$, *i.e.*, the probability distribution $P(X_B|x_A)$, which is a conditional Gaussian distribution with mean $\boldsymbol{\mu}_{B|A}$ and variance $\boldsymbol{\Sigma}_{B|A}$:

$$\begin{cases} \boldsymbol{\mu}_{B|A} = \boldsymbol{\mu}_B + \boldsymbol{\Sigma}_{BA}(\boldsymbol{\Sigma}_{AA} + \boldsymbol{\Gamma}_{AA})^{-1}(x_A - \boldsymbol{\mu}_A) \\ \boldsymbol{\Sigma}_{B|A} = \boldsymbol{\Sigma}_{BB} - \boldsymbol{\Sigma}_{BA}(\boldsymbol{\Sigma}_{AA} + \boldsymbol{\Gamma}_{AA})^{-1}\boldsymbol{\Sigma}_{AB} \end{cases} \tag{8}$$

With the quality-related Gaussian Process model, now we can quantitatively measure the utility of the selected PoIs, and then define the detailed format of the utility function. Intuitively, the service provider always wants to select a set of PoIs $A$ that most significantly reduces the uncertainty about the prediction on the rest of PoIs $B$ [14]. A nature notion of uncertainty is *entropy*. Thus, for a set of selected PoIs $A$ and a data quality level realization $\Phi$, we define the utility function as the reduction of the entropy of the unselected PoIs $B$ before and after observing the random variables $X_A$:

$$F(A, \Phi) \triangleq H(X_B) - H(X_B|X_A). \tag{9}$$

We note that this reduction is also known as the *mutual information* between the selected PoIs $A$ and the rest of the PoIs $B$ [7]. According to the definition of entropy, we can calculate the entropy of the Gaussian random variable $X_B$ and the entropy of the random variable $X_B$ condition on the set of variables $X_A$:

$$H(X_B) = \frac{1}{2} \ln \left[ (2\pi e)^{|B|} \, |\boldsymbol{\Sigma}_{BB}| \right]$$
$$H(X_B|X_A) = \frac{1}{2} \ln \left[ (2\pi e)^{|B|} |\boldsymbol{\Sigma}_{B|A}| \right]$$

We integrate the above two equalities into Equation (9) to derive the specific format of the utility function. It is worth noting that the utility function depends on both the selected PoIs $A$ and the realization $\boldsymbol{\Phi}$. The realization $\boldsymbol{\Phi}$ determines the noise matrix $\boldsymbol{\Gamma}$ in calculating $H(X_B|X_A)$.

### B. Design Details

The challenges of designing efficient algorithms for stochastic submodular maximization problem lies in the unknown data quality levels of mobile users during the optimization. One trivial solution is to enumerate the possible realization of the data quality levels, and run the simple greedy algorithm, which guarantees the constant approximation ratio in non-stochastic submodular maximization setting [14], for each realization. There are $N$ possible quality levels for each mobile user. To compute the expectation of the utility over the user subset of size $K$, we have to take average of $O(N^M)$ instances of possible realizations. Thus, such solution leads to an exponential time complexity.

[1]As we have assumed there is exact one mobile user at one PoI, selecting PoIs is equal to selecting the corresponding mobile users at the PoIs.

**Algorithm 2:** Random Adaptive Greedy User Selection

**Input**: A set of mobile users $\mathbb{V}$; a context-quality classifier $Q$, data quality distribution $\{\pi_i(\phi)\}$ for each user $i \in \mathbb{V}$, cardinality constraint $K$.
**Output**: A set of selected mobile user $\mathcal{V}$.

1 $\mathcal{V} \leftarrow \varnothing$, $\psi \leftarrow \varnothing$;
2 **for** $i = 1$ ***to*** $K$ **do**
3 $\quad \mathcal{V}_i(\psi) \leftarrow \varnothing$;
4 $\quad$ Compute $\Delta(v|\psi)$ for all $v \in \mathbb{V} \backslash \mathcal{V}$;
5 $\quad$ **for** $k = 1$ ***to*** $K$ **do**
6 $\quad\quad \mathcal{V}_i(\psi) \leftarrow \mathcal{V}_i(\psi) \bigcup \underset{v \in (\mathbb{D} \cup \mathbb{V}) \backslash (\mathcal{V} \cup \mathcal{V}_i(\psi))}{\arg\max} \{\Delta(v|\psi)\}$;
7 $\quad$ Select $v_i$ randomly from $\mathcal{V}_i(\psi)$;
8 $\quad \mathcal{V} \leftarrow \mathcal{V} \cup \{v_i\}$;
9 $\quad$ Assign selected user to conduct the sensing task and obtain her context vector $\boldsymbol{c}_i$ and sensing data $x_i$;
10 $\quad \phi_i \leftarrow Q(\boldsymbol{c}_i)$;
11 $\quad \psi \leftarrow \psi \cup \{(v_i, \phi_i)\}$;
12 Remove all dummy users from $\mathcal{V}$;
13 **return** $\mathcal{V}$;

We turn to an adaptive selection policy. The service provider selects only one mobile user according to a certain criterion in each round. After that, the service provider assigns her a sensing task, receives her sensing data, and determines her data quality level. Based on the collected data and the realized data quality level, the service provider then updates the GP model according to Equations (8). The service provider iteratively executes these steps until selecting $K$ qualified mobile users. With such adaptive selection policy, we can reduce the time complexity to a polynomial order $O(KM)$.

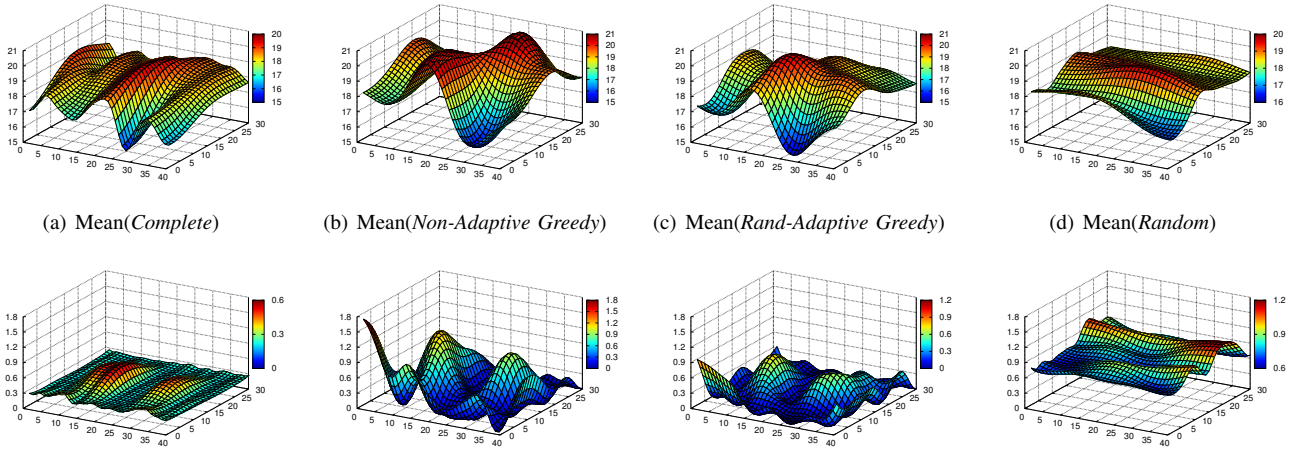
In order to design a good selection criterion, we take advantage of the adaptive submodularity of the utility function.

**Lemma 1.** *The utility function $F(\cdot, \cdot)$ defined in Equation (9) is adaptive submodular.*

Due to the limitation of space, we leave the detailed proof into our technical report [1].

We observe that $F(\mathbb{L}, \boldsymbol{\Phi}) = F(\varnothing, \boldsymbol{\Phi}) = 0$, thus the utility function is not adaptive monotone. When running the traditional adaptive greedy policy for the monotone objective function: selecting the user with the highest marginal utility $\Delta(v|\psi)$ in each round, the non-monotonicity would lead to the traps of low utility. We slightly modify the traditional adaptive greedy policy by introducing a random procedure, to deal with such traps. In order to avoid selecting the user with a negative marginal value, we introduce $2K-1$ additional dummy users, denoted by $\mathbb{D}$, to the ground set. The expectation of marginal utility for each dummy user $d \in \mathbb{D}$ is always 0, *i.e.*, $\Delta(d|\psi) = 0$. It is obvious that the dummy users do not affect the optimal policy, and can be removed from the solution of any selection policy, without affecting its expected utility.

We present the detailed steps of the random adaptive greedy user selection policy in Algorithm 2. We select $K$ candidate mobile users in an adaptive manner. In each iteration, given the current partial realization, we first calculate the expected marginal utility for each available users (Line 4). Then, we select $K$ candidate users with the maximum expected marginal utility, and randomly choose one of them as the selected user



(a) Mean(*Complete*) (b) Mean(*Non-Adaptive Greedy*) (c) Mean(*Rand-Adaptive Greedy*) (d) Mean(*Random*)

(e) Variance(*Complete*) (f) Variance(*Non-Adaptive Greedy*) (g) Variance(*Rand-Adaptive Greedy*) (h) Variance(*Random*)

Figure 3: Comparision between different gaussian process models.

in this iteration (Lines 5 to 8). When the selected user finishes the sensing task, we can obtain her contextual information and the sensing data. Taking the contextual information as input, the context-quality classifier outputs the data quality level of the user, which is used to update the partial realization. After selecting $K$ mobile users, we remove all the dummy users, and return the ultimate set as the result.

We now show that the random adaptive greedy algorithm achieves a constant approximation ratio of $1/e$.

**Theorem 1.** *For the adaptive submodular utility function $F(\cdot,\cdot)$, the user set $\mathcal{V}$ returned by the random adaptive greedy algorithm attains at least $1/e$ of the optimal value, that is:*

$$\mathbb{E}\left[F(\mathcal{V},\mathbf{\Phi})\right] \geq \frac{1}{e}\max_{|\mathcal{V}^*|\leq K}\mathbb{E}\left[F(\mathcal{V}^*,\mathbf{\Phi})\right]. \tag{10}$$

Due to the limitation of space, we supply the complete proof in our technical report [1].

## VI. Evaluation Results

In this section, we report the evaluation results on adaptive user selection process. We base on a real-world temperature data set to perform a sequence of simulations to emulate the behavior of mobile phone users and user selection process. 54 sensor nodes were deployed in a lab and kept collecting temperature information for several days. We select the samples between 1 am and 2 am to train a primal covariance matrix, which is used to compute the mutual information. Since the sensor readings have much smaller noises than user data, we regard the sensor readings as the ground truth here.

Then, we assume that there are 200 users willing to participate in this project. We generate a random quality distribution vector for each user, where there are 5 predefined qualities. They respectively map to the gaussian noise with variance 0.1, 0.5, 1, 2, and 5. User locations are randomly selected from the 54 sensor locations deployed in the lab.

In our first experiment, we construct four gaussian process models using data from (a)sensor readings from all of the 54 locations, (b)"noisy" readings from 20 users selected by non-adaptive simple greedy algorithm only based on their locations, (c)"noisy" readings from 20 users selected by our random adaptive greedy algorithm, and (d)"noisy" readings from 20 randomly selected users. The "noisy" reading of user is generated as follows: After determining the quality of the selected user, we sample a gaussian noise randomly from the corresponding normal distribution. Add the generated gaussian noise to the sensor reading, we then get the "noisy" user data. The recovering results of these four gaussian process models are presented in Figure 3. The predicted mean values are shown in (a), (b), (c), and (d), while the prediction variances are shown in (e), (f), (g), and (h). Besides, the relative deviations of these four recovered models are presented in Table I, which is defined as: Relative Deviation $= \frac{||\hat{y}-y||_2}{||y||_2}$ where $y$ denotes ground truth at those 54 locations, and $\hat{y}$ denotes the vector of predicted values.

Table I: Relative Deviation Comparision

| GP model | Relative Deviation |
|---|---|
| Complete Data Set | 0.0133 |
| Random Adaptive Greedy Algorithm | 0.0414 |
| Non-adaptive Greedy Algorithm | 0.0592 |
| Random Algorithm | 0.0876 |

As we can see, the GP model recovered from all of the 54 sensor readings are the best, with the lowest variance and relative deviation. The GP model recovered by our random adaptive greedy algorithm, although has relatively higher variance than complete model, shows better performance than the model recovered by random algorithm and non-adaptive greedy algorithm in both prediction variance and prediction accuracy. The non-adaptive greedy algorithm only considers the impact of locations, but not the impact of data quality. So it only outperforms the random algorithm, but is inferior to our random adaptive greedy algorithm.

In our second experiment, we compare the mutual information of the sets returned by our random adaptive greedy algorithm, random algorithm, non-adaptive greedy algorithm and the simple adaptive greedy algorithm. Each point in Figure 4 is the average of 10 runs of the algorithms. As shown in Figure 4, when the cardinality constraint is small, *i.e.*, less than 30, the simple adaptive greedy algorithm and

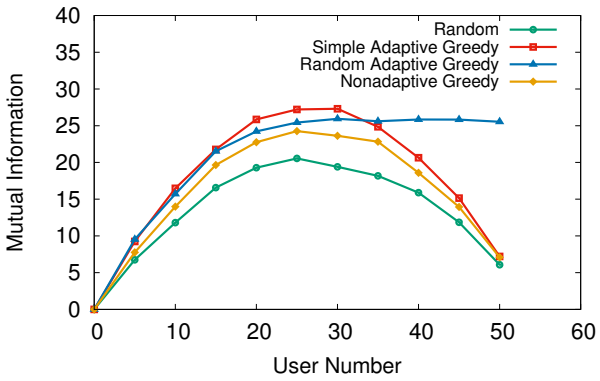



Figure 4: Mutual information of different algorithms.

the random adaptive greedy algorithm obtain similar mutual information. However, due to the non-monotonicity of mutual information, when the cardinality constraint gets approach to 50, the mutual information of sets returned by other three algorithm all decreases. Since we add enough dummy users and provide a random step in our random adaptive greedy algorithm, the mutual information gain will not decrease even when the cardinality constraint increases to 50. Moreover, since the non-adaptive greedy algorithm only considers the influence of location, mutual information of selected users is always lower than sets returned by simple adaptive greedy algorithm and random adaptive greedy algorithm. Here we can see the superiority of our random adaptive greedy algorithm.

## VII. Related Work

In this section, we briefly review the related works.

**Data Quality in crowdsensing.** In recent years, many researchers have paid their attention to the data quality problem in crowdsensing. However, they either ignored how to estimate the data quality and used it directly [15], [16], [26], [28], or only focused on estimating the data quality without further usage [24]. To the best of our knowledge, our work is the first to estimate the data quality and use it to guide user selection. Different from the labeling task in crowesourcing [30], sensor data in crosdsensing is always continuous. It is not suitable to describe its quality with confusion matrix model. We exploit the gaussian process model and use variance of gaussian noise to describe the quality of sensing data.

**Submodular maximization** is a well-studied mathematical problem. Nemhauser *et.al.* [22] firstly proved the approximation ratio of greedy algorithm in the maximization of monotone submodular functions. Then, in 2010, Golovin and Krause [13] proposed the concept of adaptive submodularity and proved the approximation ratio of simple adaptive greedy algorithm. There are some recent works on the maximization of non-monotone submodular function [4], [12].

## VIII. Conclusion

In this paper, we have studied real-time data quality estimation in mobile crowdesing. We have investigated the relation between sensing context and data quality, and proposed the context-aware data quality estimation scheme. We have integrated the data quality estimation scheme to guide user recruitment. We have modeled the user recruitment process as an adaptive non-monotone submodular maximization problem, and designed a random adaptive greedy algorithm to achieve a constant approximation ratio. Through simulation on a real-world temperature data set, we have shown the excellent performance of our algorithm when recovering the GP model in the whole target area with finite observed locations.

## References


[1] Context-aware data quality estimation in mobile crowdsensing. Technical report, https://www.dropbox.com/s/agj6qs2uva8b4av/mcs_liushengzhong.pdf?dl=0, 2017.

[2] Y. Arase, F. Ren, and X. Xie. User activity understanding from mobile phone sensors. In *UbiComp Adjunct*, 2010.

[3] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[4] N. Buchbinder, M. Feldman, J. Naor, and R. Schwartz. A tight linear time (1/2)-approximation for unconstrained submodular maximization. In *FOCS*, 2012.

[5] N. Buchbinder, M. Feldman, J. S. Naor, and R. Schwartz. Submodular maximization with cardinality constraints. In *SODA*, 2014.

[6] C.-C. Chang and C.-J. Lin. Libsvm: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27:1–27:27, 2011.

[7] T. M. Cover and J. A. Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[8] A. P. Dawid and A. M. Skene. Maximum likelihood estimation of observer error-rates using the em algorithm. *Applied statistics*, 28(1):20–28, 1979.

[9] A. Deshpande, C. Guestrin, S. R. Madden, J. M. Hellerstein, and W. Hong. Model-driven data acquisition in sensor networks. In *VLDB*, 2004.

[10] P. Dutta, P. M. Aoki, N. Kumar, A. Mainwaring, C. Myers, W. Willett, and A. Woodruff. Common sense: participatory urban sensing using a network of handheld air quality monitors. In *SenSys*, 2009.

[11] S. B. Eisenman, E. Miluzzo, N. D. Lane, R. A. Peterson, G.-S. Ahn, and A. T. Campbell. Bikenet: A mobile sensing system for cyclist experience mapping. *ACM Transactions on Sensor Networks*, 6(1):6, 2009.

[12] U. Feige, V. S. Mirrokni, and J. Vondrak. Maximizing non-monotone submodular functions. *SIAM Journal on Computing*, 40(4):1133–1153, 2011.

[13] D. Golovin and A. Krause. Adaptive submodularity: A new approach to active learning and stochastic optimization. In *COLT*, 2010.

[14] C. Guestrin, A. Krause, and A. P. Singh. Near-optimal sensor placements in gaussian processes. In *ICML*, 2005.

[15] H. Jin, L. Su, D. Chen, K. Nahrstedt, and J. Xu. Quality of information aware incentive mechanisms for mobile crowd sensing systems. In *MobiHoc*, 2015.

[16] R. Kawajiri, M. Shimosaka, and H. Kashima. Steered crowdsensing: Incentive design towards quality-oriented place-centric crowdsensing. In *UbiComp*, 2014.

[17] J. R. Kwapisz, G. M. Weiss, and S. A. Moore. Activity recognition using cell phone accelerometers. *SIGKDD Explorations Newsletter*, 12(2):74–82, 2011.

[18] N. D. Lane, E. Miluzzo, H. Lu, D. Peebles, T. Choudhury, and A. T. Campbell. A survey of mobile phone sensing. *IEEE Communications Magazine*, 48(9):140–150, 2010.

[19] S. Mathur, T. Jin, N. Kasturirangan, J. Chandrasekaran, W. Xue, M. Gruteser, and W. Trappe. Parknet: drive-by sensing of road-side parking statistics. In *MobiSys*, 2010.

[20] P. Mohan, V. N. Padmanabhan, and R. Ramjee. Nericell: rich monitoring of road and traffic conditions using mobile smartphones. In *SenSys*, 2008.

[21] M. Mun, D. Estrin, J. Burke, and M. Hansen. Parsimonious mobility classification using gsm and wifi traces. In *HotEmNets*, 2008.

[22] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher. An analysis of approximations for maximizing submodular set functions. *Mathematical Programming*, 14(1):265–294, 1978.

[23] Noisetube. http://www.noisetube.net/.

[24] D. Peng, F. Wu, and G. Chen. Pay as how well you do: A quality based incentive mechanism for crowdsensing. In *MobiHoc*, 2015.

[25] R. Pryss, M. Reichert, B. Langguth, and W. Schlee. Mobile crowd sensing services for tinnitus assessment, therapy, and research. In *MS*, 2015.

[26] Z. Song, B. Zhang, C. H. Liu, A. V. Vasilakos, J. Ma, and W. Wang. Qoi-aware energy-efficient participant selection. In *SECON*, 2014.

[27] D. Wang, L. Kaplan, H. Le, and T. Abdelzaher. On truth discovery in social sensing: A maximum likelihood estimation approach. In *IPSN*, 2012.

[28] Y. Wen, J. Shi, Q. Zhang, X. Tian, Z. Huang, H. Yu, Y. Cheng, and X. Shen. Quality-driven auction-based incentive mechanism for mobile crowd sensing. *IEEE Transactions on Vehicular Technology*, 64(9):4203–4214, 2015.

[29] J. Yick, B. Mukherjee, and D. Ghosal. Wireless sensor network survey. *Computer Networks*, 52(12):2292 – 2330, 2008.

[30] Y. Zhang, X. Chen, D. Zhou, and M. I. Jordan. Spectral methods meet em: A provably optimal algorithm for crowdsourcing. In *NIPS*, 2014.