# *ALC*$^2$: When Active Learning Meets Compressive Crowdsensing for Urban Air Pollution Monitoring

Tong Liu, *Member, IEEE*, Yanmin Zhu, *Senior Member, IEEE*, Yuanyuan Yang, *Fellow, IEEE*, and Fan Ye

***Abstract*—As metropolises develop, air pollution has become a serious problem, especially in developing countries like China. Many governments and researchers have devoted themselves to tackling and solving this problem. With the proliferation of smartphones, mobile crowdsensing is becoming a promising paradigm for monitoring large-scale environmental phenomena. In a practical crowdsensing system, incentives should be provided to encourage the participation of rational smartphone users, because it incurs various costs on users to collect sensing data. However, monitoring fine-grained air pollution in a large urban area based on crowdsensing will lead to high payments, which makes designing an efficient incentive mechanism a challenging problem. Fortunately, compressive sensing (CS) has been proved as an effective technology to reduce the amount of collected data via exploiting the spatial correlations among sensing data. In this article, we employ CS in the air pollution monitoring application, in which only a sampled set of locations are selected to collect data and provide incentives to the participants, and air pollution concentrations in unselected locations are inferred via CS. We propose an active learning scheme, which iteratively selects valuable locations to collect sensing data. Moreover, an expectation maximization-based algorithm is designed to detect the contexts in which sensing data are collected, and an efficient incentive mechanism is provided to encourage users with low costs participating. Comprehensive simulations are conducted to demonstrate the performance of our proposed scheme.**

***Index Terms*—Active learning (AL), air pollution monitoring, compressive sensing (CS), crowdsensing, incentive.**

## I. Introduction

WITH the modernization of peoples' lives, air pollution has emerged as an acute problem in urban areas, especially developing nations (e.g., China and India). Long-term exposure to air pollutants (e.g., $NO_2$, $PM_{2.5}$, and CO) leads to a high risk of several health issues, such as respiratory infections, heart disease, and lung cancer. To track and solve this problem, some cities try to monitor air pollution by deploying stationary measuring stations. However, building such stations is significantly limited by the availability of land and large costs of maintenance (30 000 USD per year), which results in difficulties of obtaining measurements of fine-grained air pollution concentrations in a large urban area. For example, only 22 measuring stations have been built in Beijing, a city with an area of 16 400 km$^2$. Although several works [1], [2] have focused on inferring fine-grained air qualities by exploiting the correlations with other datasets, such as traffic flows and points-of-interest (POIs), direct measurements of fine-grained air pollution provide fidelity and accuracy unsurpassable by other ways.

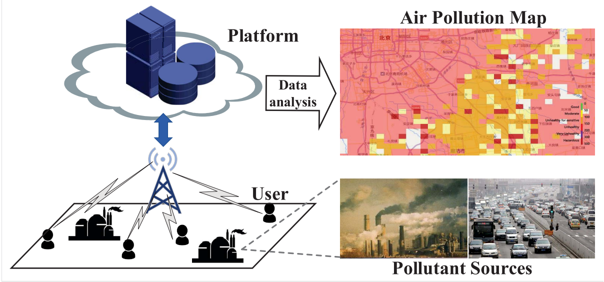


Fig. 1. Illustration of a crowdsensing system for air pollution monitoring.

Mobile crowdsensing provides an unprecedented opportunity for collecting sensing data on a large scale (e.g., community or city), which takes advantage of widely distributed modern mobile devices (e.g., smartphones) equipped with abundant sensors. Numerous environment-centric applications have been developed based on the paradigm of crowdsensing, such as traffic monitoring [3] and noise mapping [4]. In these applications, smartphone users report their location-based measurements to a central platform via wireless networks. After aggregating plenty of geographically distributed measurements, the platform can obtain a fine-grained overview of an environmental phenomenon. Similarly, air pollution monitoring can be conducted based on crowdsensing, as shown in Fig. 1. Although smartphones are not equipped with environmental sensors at present, fortunately, sensor-integrated portable external hardware [5] has been developed,

Manuscript received May 24, 2019; revised August 12, 2019; accepted August 27, 2019. Date of publication September 5, 2019; date of current version December 11, 2019. This work was supported in part by NSFC under Grant 61772341, Grant 61472254, and Grant 61802245, in part by Science and Technology Committee of Shanghai under Grant 18511103002 and Grant KQJSCX20180329191021388, in part by the Program for Changjiang Young Scholars in University of China, in part by the Program for China Top Young Talents, in part by the Program for Shanghai Top Young Talents, in part by the Shanghai Engineering Research Center of Digital Education Equipment, in part by the Shanghai Jiao Tong University (SJTU) Global Strategic Partnership Fund under Grant 2019 SJTU-HKUST, and in part by the Shanghai Sailing Program under Grant 18YF1408200. *(Corresponding author: Tong Liu.)*

T. Liu is with the School of Computer Engineering and Science, Shanghai University, Shanghai 200444, China (e-mail: tong_liu@shu.edu.cn).

Y. Zhu is with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China, and also with the Shanghai Key Laboratory of Scalable Computing and Systems, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: yzhu@sjtu.edu.cn).

Y. Yang and F. Ye are with the Department of Electrical and Computer Engineering, Stony Brook University, Stony Brook, NY 11794 USA (e-mail: yuanyuan.yang@stonybrook.edu; fan.ye@stonybrook.edu).

Digital Object Identifier 10.1109/JIOT.2019.2939552

and smartphones incorporating small low-cost environmental sensors are coming soon [6].

Recently, several efforts [7]–[10] have been put into developing crowdsensing systems for monitoring air pollution in an urban area. However, most of them focus on the implementation of portable sensing devices or smartphone applications. For example, a crowdsensing system, named P-Sense, is designed in [8], where external sensing devices can measure the concentrations of several gases like carbon dioxide, combustible gas, and carbon monoxide. Yang *et al.* [10] proposed an architecture of the people-centric Internet of Things system for urban environment measuring, in which an interaction mechanism between human and sensing devices is provided. Sensing data can be transferred from devices to smartphones via Bluetooth. Different from these works, we assume smartphones can measure air pollution concentrations already, and we focus on designing an efficient air pollution monitoring scheme based on compressive crowdsensing, which includes an incentive mechanism to encourage smartphone users with low costs participating and an algorithm to select a sampled set of locations to collect measurements to reduce payments.

An efficient and appropriate incentive mechanism is a key component in a practical crowdsensing system. On the one hand, certain monetary rewards are expected by smartphone users to participate in sensing, because numerous resources are consumed, such as power, bandwidth, and human efforts. As a rational individual, a smartphone user will not participate if the reward he/she earns is less than his/her cost. On the other hand, the platform aims at minimizing its total payment under the condition of guaranteeing the quality and quantity of the collected data. Although a number of incentive mechanisms [11]–[14] have been proposed based on game theory, they focus on ensuring the truthfulness of participants. They do not consider the various values of sensing data from an overall perspective and pay each measurement through balancing its value and cost.

To monitor fine-grained air pollution in an urban area, a large amount of measurements in different locations should be collected, which is still costly for the platform. Fortunately, there exists an inherent spatial correlation among sensing data in different locations, which has been observed in real datasets [1]. The correlation exists because air pollutants released by pollutant sources disperse in 3-D space according to a certain model (e.g., the Gaussian model [15]). By exploiting the correlation, *compressive sensing* (CS) [16] can employed to significantly reduce the amount of collected sensing data (i.e., only a sampled set of locations are selected to collect measurements). The air pollution concentrations in unselected locations can be accurately inferred based on collected measurements. A few existing works [17]–[19] have employed CS in crowdsensing systems. However, these works simply assume smartphone users are cooperative, who will do sensing tasks allocated to them without incentives. In contrast, we combine CS into our incentive mechanism design, providing location-dependent incentives to encourage valuable smartphone users with low costs participating.

In this article, we consider a practical crowdsensing system with rational smartphone users to monitor fine-grained air pollution in a large urban area, where the participation of users is strongly stimulated by the incentives provided to them. A central platform located in cloud aggregates all collected sensing data and recover the whole air pollution map via CS, with the aim of minimizing the total payment spent for collecting measurements. An efficient incentive mechanism is expected to dynamically adjust the incentives provided to users in different locations, according to the measurements already collected and the spatial distribution of users. Given such an incentive mechanism, only valuable and low-cost users will be encouraged to participate and paid. In addition, sufficient measurements in different locations should be collected, to guarantee the accuracy of recovering the whole air pollution map via CS.

This problem is highly difficult due to several challenges. First, the relationship between an arbitrary incentive and the participation of smartphone users is not clear. It is impractical to collect the cost information of all users and then choose the cheapest ones in a real crowdsensing system, as it takes a lot of time and energy of smartphone users, lowering their participating motivation. Second, the quality of each collected measurement is not guaranteed, due to measuring errors of hardware and sensing contexts (e.g., indoor versus outdoor), which can greatly impact the usability of measurements. Note that indoor and outdoor measurements have significantly different values in the same location and only outdoor measurements are useful. Third, the value of a measurement for detecting the whole air pollution map and the incentive provided to it are tightly coupled. The value needs to be estimated before deciding what incentive to provide, while incentives published to users influence which measurements can be collected.

To address these challenges, we propose an *active learning* (AL) scheme which iteratively collects measurements in selected locations, to obtain an accurate and fine-grained air pollution map and reduce the total payment as much as possible. We first employ a Gaussian air pollution dispersion model to formally analyze the relationship between the fine-grained pollution concentrations in different locations and the emission rates of all pollutant sources existing in the urban area. Then, we build a probabilistic model to characterize the participation of a crowd of rational smartphone users given a certain incentive. Next, we develop an expectation maximization (EM)-based algorithm, to distinguish indoor and outdoor measurements collected in the same location and estimate the pollution concentration of the location based on outdoor measurements. Given the estimated concentrations in a sampled set of locations, we employ CS to recover the whole air pollution map and detect pollutant sources. To collect sufficient and valuable measurements in different locations, we propose an iterative algorithm based on the idea of AL. In each iteration, a subset of locations without measurements are modestly selected, and proper incentives are provided in these locations.

The major contributions of this article are summarized as follows.

1) We combine CS into incentive mechanism design in a crowdsensing system, which can significantly reduce the amount of collected measurements, and therefore, decrease the total payment spent to monitor fine-grained air pollution in a large urban area.
2) We propose an AL-based scheme to collect sensing data, in which a sampled set of valuable locations without measurements are iteratively selected to query users by providing proper incentives. The incentives in different locations are updated adaptively in each iteration according to the collected measurements and the geographical distribution of users.
3) We also provide an EM-based algorithm to detect the contexts of collected measurements, to distinguish indoor and outdoor measurements.
4) We perform comprehensive simulations and the results confirm the superiority of our scheme in terms of the total payment and the accuracy of detection.

The remainder of this article proceeds as follows. The models and preliminaries are presented in Section II. Section III illustrates the problem formulation, as well as the workflow of our proposed scheme. In Section IV, the designs of our proposed scheme are described in detail, including the EM-based pollution concentration estimation algorithm, the CS-based pollutant source detection algorithm, the AL-based location selection algorithm, and the incentive mechanism. Section V shows the performance of our simulations. Finally, we discuss related work and conclude this article in Sections VI and VII, respectively.

## II. Models and Preliminaries

### A. Air Pollution Dispersion Model

In an urban area, air pollutants are always released by several natural or anthropogenic pollutant sources, such as power plants and wild fires. Several mathematical models [20]–[22] have been studied to characterize the nature of air pollution dispersion, considering the emission rates of pollutant sources, the direction and velocity of wind, atmospheric turbulence, and so on. These models can be employed to simulate the movement of pollutants in atmosphere and predict future concentrations in different scenarios. In this article, we employ a most widely used one, Gaussian model, in this article. Here, we emphasize that any other air dispersion model can be used in our proposed compressive crowdsensing-based urban air pollution monitoring system. Choosing a proper air pollution dispersion model is also a critical issue, specially in urban areas (due to the effects of skyscrapers, streets, temperatures, and so on), which is not within the scope of this article.

Basically, Gaussian model assumes pollution concentrations decay in 3-D space according to the Gaussian distribution as shown in Fig. 2. Accordingly, the *complete equation for Gaussian dispersion modeling* is formulated as follows:

$$C = Q \cdot \frac{1}{\pi \bar{u} \sigma_y \sigma_z} \cdot e^{\frac{-y^2}{2\sigma_y{}^2}} \cdot e^{\frac{-H^2}{2\sigma_z{}^2}} \tag{1}$$

where $C$ denotes the pollution concentration observed in a location caused by a pollutant source. $Q$ is the pollutant emission rate of the source, $\bar{u}$ represents the wind velocity, $y$ is the crosswind distance between the observed location and the source, and $H$ is the height of the source. $\sigma_y$ and $\sigma_z$ are two constant dispersion parameters, which measure the atmospheric turbulence. Fig. 2 plots an illustration of the Gaussian dispersion model, in which we can see pollutants diffuse quickly in the downwind direction, and the pollution concentrations decay as the Gaussian distribution in the crosswind and vertical directions.

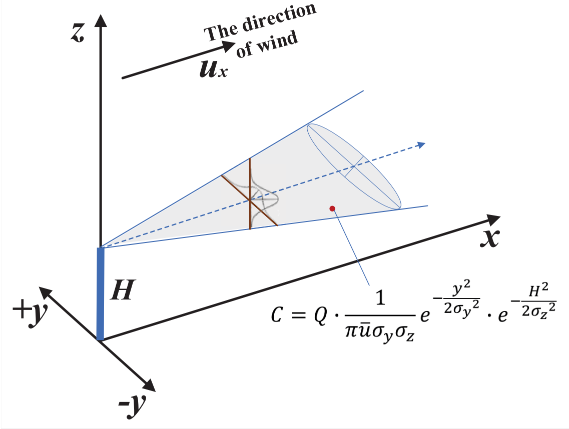

z
The direction of wind
$u_x$
x
H
+y
-y
$C = Q \cdot \frac{1}{\pi \bar{u} \sigma_y \sigma_z} e^{-\frac{y^2}{2\sigma_y{}^2}} \cdot e^{-\frac{H^2}{2\sigma_z{}^2}}$

Fig. 2. Illustration of the Gaussian air pollution dispersion model.

In this article, we only consider static pollutant sources and assume they release pollutants continuously at certain emission rates. Moreover, we suppose the wind velocity and direction can be known in prior from other datasets like meteorological data, and we set the height of pollutant sources as a fixed value (e.g., 50 m) according to common knowledge. Then, given the locations and emission rates of all pollutant sources in the whole urban area, the pollution concentration of any location can be calculated according to the dispersion model.

### B. Crowdsensing System Model

In this article, we aim to detect pollutant sources and their emission rates based on a crowdsensing system, as shown in Fig. 3, in which a plenty of mobile users equipped with monitoring sensors participate in collecting sensing data. All sensing data is aggregated and analyzed by a central platform resided in cloud.

For the convenience, we virtually divide the whole monitored urban area into $N$ small grids of the same size, e.g., 200 m × 200 m. The set of girds is denoted by $\mathcal{N} = \{1, 2, \ldots, N\}$. The air pollution concentration in each grid can be seen as uniform while different grids may have different values.

We suppose there are $k$[1] pollutant sources in the whole monitored urban area, and they locate in different grids. Thus, the location of a pollutant source can be denoted by a grid. The emission rates of the pollutant sources are denoted by $\mathbf{Q} = \{Q_1, Q_2, \ldots, Q_k\}$. Note that the number of pollutant sources $k$ and their emission rates $\mathbf{Q}$ are unknown by the platform in prior and need to be monitored by crowd users.

[1] Note that compared with the number of grids, the number of pollutant sources is sparse, i.e., $k \ll n$.

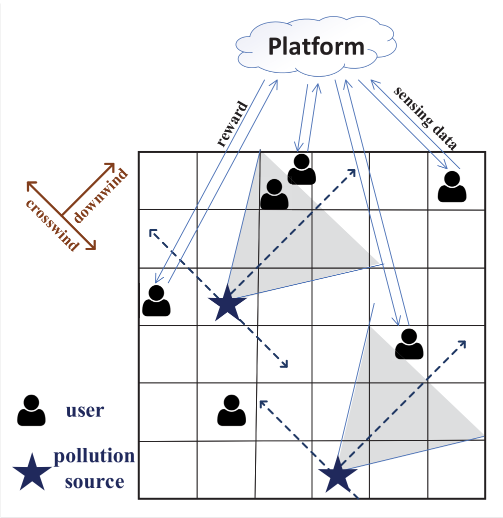


Fig. 3. Illustration of the urban air pollution monitoring system based on mobile crowdsensing.

We define an index vector $\mathbf{g} = \{g_1, g_2, \ldots, g_N\}^T$ to indicate whether a grid contains a pollutant source as (2). If a pollutant source is in grid $i$, then, $g_i$ equals to the emission rate of the pollutant source; Otherwise, $g_i = 0$

$$g_i = \begin{cases} Q_j, & \text{if pollutant source } j \text{ is in grid } i \\ 0, & \text{if there is no source in grid } i \end{cases}, \quad \forall 1 \leq i \leq n. \tag{2}$$

Air pollutants released by these sources disperse in space, vitally influenced by winds, which lead to different pollution concentrations in different locations. The pollutants in a grid are the accumulation of pollutants dispersed to the grid from all sources. We use $\mathbf{C} = \{C_1, C_2, \ldots, C_N\}^T$ to denote the pollution concentrations in all grids.

By employing the dispersion model introduced above, we can mathematically analyze the relationship between the pollution concentration of each grid and the emission rate of each pollutant source. Specially, we build a transfer matrix $\Omega \in \mathbb{R}^{N \times N}$, which satisfies the following equation:

$$\begin{bmatrix} C_1 \\ \vdots \\ C_N \end{bmatrix} = \begin{bmatrix} \Omega_{11} & \cdots & \Omega_{1N} \\ \vdots & \ddots & \vdots \\ \Omega_{N1} & \cdots & \Omega_{NN} \end{bmatrix} \times \begin{bmatrix} g_1 \\ \vdots \\ g_N \end{bmatrix}. \tag{3}$$

According to (1), we can derive that

$$\Omega_{ij} = \frac{1}{\pi \bar{u} \sigma_y \sigma_z} \cdot e^{\frac{-d_{ij}^2}{2\sigma_y{}^2}} \cdot e^{\frac{-H^2}{2\sigma_z{}^2}} \tag{4}$$

where $d_{ij}$ denotes the downwind distance from grid $g_j$ (the location of a pollutant source) to grid $g_i$ (the location of an influenced grid).

Given a certain incentive in grid $i$, a set of measurements sensed by different users can be collected by the platform, which is denoted by $\mathbf{M}_i = \{m_1^{(i)}, m_2^{(i)}, \ldots, m_\gamma^{(i)}\}$, where $\gamma$ represents the number of collected measurements. In this article, we consider the measurements may be sensed in different contexts, e.g., indoor and outdoor.[2] However, only outdoor measurements are useful to detect the pollution concentration. We use a latent vector $\mathbf{Z}_i = \{z_1^{(i)}, z_2^{(i)}, \ldots, z_\gamma^{(i)}\}$ to indicate whether a measurement is sensed indoor or outdoor, i.e.,

$$z_r^{(i)} = \begin{cases} 0, & m_r^{(i)} \text{ is sensed indoor} \\ 1, & m_r^{(i)} \text{ is sensed outdoor} \end{cases} \quad \forall r = 1, 2, \ldots, \gamma. \tag{5}$$

Note that the value of $\mathbf{Z}_i$ is unknown by the platform.

### C. User Participation Model

In our scheme, the platform provides the same payment for each measurement collected in the same grid for the sake of fairness, while payments for different grids can differ. We denote the payment for each measurement in grid $i$ as $P_i$, and define a payment vector as $\mathbf{P} = \{P_1, P_2, \ldots, P_N\}$. Given a certain payment $P_i$, rational smartphone users with lower costs in grid $i$ will actively participate in sensing for earning money. To understand the relationship between payment $P_i$ and participation behavior of crowded users, we build a probabilistic participation model for a crowd of rational smartphone users in the following.

As some resources (e.g., power and bandwidth) are consumed for collecting sensing data, costs are incurred on smartphone users. The cost of a specific user depends on many factors, such as the hardware of smartphones, the remaining power of batteries, the quality of wireless networks, and the impact of his/her participation. These factors lead to various costs on different users for collecting a measurement in the same grid, which is private information of each user. We denote the cost of user $s$ as $c_s \in [c_{\min}, c_{\max}]$, where $c_{\min}$ and $c_{\max}$ represent the lower bound and the upper bound, respectively.

First, we build a model to characterize the participation of one rational smartphone user given a certain payment. Apparently, a rational user will not participate in sensing if the payment he/she earns is less than his/her cost. We use a random variable $X_s$ to indicate whether user $s$ will participate in sensing given an arbitrary payment $P$, and thus, the value of $X_s$ can be defined as

$$X_s = \begin{cases} 0, & \text{if payment } P \text{ is less than cost } c_s \\ 1, & \text{otherwise.} \end{cases} \tag{6}$$

As it is unnecessary to collect measurements from all users, the platform tends to choose users with the lowest costs for the sake of saving money. However, it is impractical to collect the cost information of all users. First, collecting cost information from all users is costly in both latency and power. Second, the smartphone users, who submit their cost information without being chosen, will lose interest in participating in the future.

To avoid collecting cost information, we set a uniform payment $P \in [c_{\min}, c_{\max}]$ for each measurement collected in the same grid. Thus, only users with $c_s \leq P$ will participate in sensing. The extra money $(P - c_s)$ paid for each measurement can be seen as the overpay for avoiding collecting private

[2] Note that our EM-based pollution concentration estimation algorithm proposed in Section IV-A can be easily extended to the situation with more than two contexts.

TABLE I
NOTATIONS AND DESCRIPTIONS

| No. | Descriptions |
|---|---|
| $N$ | Number of grids |
| $k$ | Number of pollutant sources |
| $\mathbf{Q}$ | Emission rates of pollutant sources |
| $\mathbf{g}$ | Index vector indicating whether a grid contains a pollutant source |
| $\mathbf{C}$ | Pollution concentration in each grid |
| $\boldsymbol{\Omega}$ | Transfer matrix between pollutant sources and pollution concentrations |
| $\gamma$ | Number of collected measurements in each grid |
| $d_{ij}$ | Downwind distance from grid $g_i$ to grid $g_j$ |
| $\mathbf{M}_i$ | Set of measurements collected in grid $i$ |
| $\mathbf{Z}_i$ | Latent vector indicating whether measurements are sensed indoor or outdoor |
| $\mathbf{P}$ | Payments provided for each measurement collected in different grids |
| $c_s$ | Cost of user $s$ |
| $X_s$ | Random variable indicating whether user $s$ participates or not |
| $n_i$ | Number of smartphone users in grid $i$ |
| $Y_i$ | Number of participants in grid $i$ |
| $\widehat{\mathbf{g}}$ | Estimation of pollution sources |
| $m$ | Number of grids with collected measurements |
| $\boldsymbol{\pi}$ | Index vector indicating which grids are selected |
| $\widehat{\mathbf{C}}$ | Estimated pollution concentrations in selected grids |

information from users. The probability distribution of random variable $X_s$ can be represented as

$$f(X_s;\ p) = \begin{cases} p, & \text{if } X_s = 1 \\ 1-p, & \text{if } X_s = 0 \end{cases}$$

which is the Bernoulli distribution with success probability $p = \Pr(X_s = 1) = \Pr(c_s \le P)$.

Second, we analyze the participation behavior of a crowd of rational users in the same grid. If the population in grid $i$ is known as $n_i$, the number of participants in grid $i$, $Y_i$, can be represented as $Y_i = \sum_{s=1}^{n_i} X_s$. We assume the costs of a crowd of users follow a certain probabilistic model according to the law of large numbers. The model can be learned by counting the numbers of participants under different payments. Take the uniform distribution as an example, e.g., $c_s \sim \mathcal{U}(c_{\min}, c_{\max})$. We can deduce that given a payment $P_i$, $Y_i$ obeys the Binomial distribution as $Y_i \sim B(n_i, p_i)$, where $p_i = [(P_i - c_{\min})/(c_{\max} - c_{\min})]$. Thus, the probability of collecting $\gamma$ measurements equals to

$$\Pr(Y_i = \gamma) = \frac{n_i!}{\gamma!(n_i - \gamma)!} p_i^{\gamma}(1 - p_i)^{n_i - \gamma}. \tag{7}$$

For convenience, all main notations and their descriptions used in this article are summarized in Table I.

### D. EM Algorithm

In this section, we briefly introduce some basics of the EM algorithm [23], which will be applied in estimating the pollution concentration in a grid. The EM algorithm is an iterative method for finding the maximum likelihood estimate (MLE) of parameters in statistical models. Given observed data $\mathbf{M}$ generated by a statistical model, an unobserved latent data $\mathbf{Z}$ and a vector of unknown parameters $\boldsymbol{\theta}$, along with a likelihood function $L(\boldsymbol{\theta}; \mathbf{M}, \mathbf{Z}) = p(\mathbf{M}, \mathbf{Z}|\boldsymbol{\theta})$, the MLE of unknown parameters $\boldsymbol{\theta}$ is

$$L(\boldsymbol{\theta}; \mathbf{M}) = p(\mathbf{M}|\boldsymbol{\theta}) = \sum_{\mathbf{Z}} p(\mathbf{M}, \mathbf{Z}|\boldsymbol{\theta}).$$

The EM algorithm finds the MLE of $\boldsymbol{\theta}$ by iteratively performing an expectation (E) step and a maximization (M) step.

1) *E-Step:* Calculates the expectation of the log-likelihood function under the current estimate of $\boldsymbol{\theta}$, with respect to the conditional distribution of $\mathbf{Z}$ given $\mathbf{M}$

$$Q\Big(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}\Big) = \mathbb{E}_{\mathbf{Z}|\mathbf{M},\boldsymbol{\theta}^{(t)}}\big[\log L(\boldsymbol{\theta}; \mathbf{M}, \mathbf{Z})\big].$$

2) *M-Step:* Calculates the parameters which maximize the expectation of log-likelihood

$$\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} Q\Big(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}\Big).$$

The iterations stop upon convergence.

### E. CS Technology

Some basics of CS are given in this section, which is employed to recover the whole pollution map based on measurements collected from a sampled set of grids. CS is a promising technique for reducing the sample rate of data with a sparse structure. Consider a target data $\mathbf{y} = [y_1, \dots, y_n]^T$, which can be decomposed under a basis $\Psi = \{\Psi_i\}_{i=1}^n \in \mathbb{R}^{n\times n}$. Therefore, $\mathbf{y}$ can be represented as $\mathbf{y} = \Psi\mathbf{x} = \sum_{i=1}^n x_i\Psi_i$, in which $x_i$ is the coefficient of basis vector $\Psi_i$. $\mathbf{y}$ is called *k-sparse* if the coefficient vector $\mathbf{x}$ has only $k$ nonzero elements and $k \ll n$.

CS employs a linear encoder to compress an $n$-dimensional vector into an $m$-dimensional vector, where $m < n$. Assume matrix $\Phi = \{\Phi_i\}_{i=1}^m \in \mathbb{R}^{m\times n}$ is a collection of measuring vectors and vector $\mathbf{z} = [z_1, \dots, z_m]^T$ are measurements. A measurement $z_i$ is the inner products of $\mathbf{y}$ and $\Phi_i$, as

$$\mathbf{z} = \Phi\mathbf{y} = \Phi\Psi\mathbf{x} = \Omega\mathbf{x}$$

where $\Omega = \Phi\Psi \in \mathbb{R}^{m\times n}$, called sensing matrix.

A widely used reconstruction approach is the $\ell_1$-norm minimization, which can be solved in polynomial time by linear programming

$$\arg\min_{\widehat{\mathbf{x}}} \|\widehat{\mathbf{x}}\|_{\ell_1}, \text{ s.t. } \mathbf{z} = \Omega\widehat{\mathbf{x}}$$

where $\widehat{\mathbf{x}}$ is the estimate of $\mathbf{x}$. According to theory of CS, $\mathbf{y}$ can be accurately reconstructed if $\Omega$ satisfies the *restricted isometry property* (RIP) [24] of order $3k$. Moreover, RIP can be achieved with a high probability if $m = O(\text{poly}(k, \log n))$ [25].

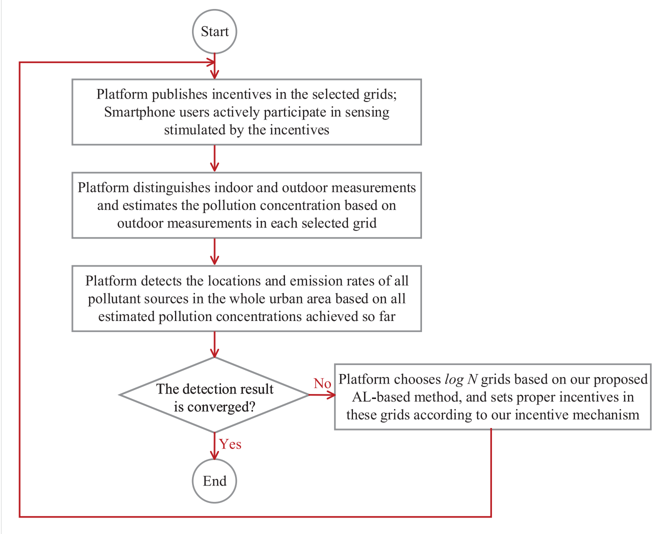

Start
Platform publishes incentives in the selected grids; Smartphone users actively participate in sensing stimulated by the incentives
Platform distinguishes indoor and outdoor measurements and estimates the pollution concentration based on outdoor measurements in each selected grid
Platform detects the locations and emission rates of all pollutant sources in the whole urban area based on all estimated pollution concentrations achieved so far
The detection result is converged?
No
Platform chooses log N grids based on our proposed AL-based method, and sets proper incentives in these grids according to our incentive mechanism
Yes
End

Fig. 4. Illustration of the workflow of our proposed iterative scheme.

## III. Problem and Overview

### A. Problem

In this article, we consider the problem of minimizing the total payment for accurately detecting pollutant sources in a large urban area, via designing an iterative AL-based scheme with incentives to collect sensing data. The problem can be formulated as

$$\arg\min_{\mathbf{P}} \sum_{i=1}^{N} (\gamma \cdot P_i), \text{ s.t. } \|\widehat{\mathbf{g}} - \mathbf{g}\|_{\ell_2} \le \delta \tag{8}$$

where $\widehat{\mathbf{g}}$ represents the estimation of $\mathbf{g}$, and $\delta$ is a predefined threshold to guarantee the accuracy of detection. To solve this problem, we iteratively select a sampled set of grids to collect sensing data, instead of collecting data from all grids. By providing an efficient incentive mechanism $\mathbf{P}$, $\gamma$ measurements can be collected in each selected grid. Then, pollution concentrations in these grids can be estimated, and further the locations and emission rates of pollutant sources can be detected.

This problem is very challenging because there exists a tradeoff between accurately detecting pollutant sources and minimizing the total payment. On the one hand, collecting measurements from the users with the lowest costs can achieve the minimum payment. However, these measurements may suffer a high data redundancy, missing important information for recovering the whole pollution map. On the other hand, the measurements which carry the most valuable information for detecting pollutant sources may come from users with high costs, incurring high payments. Additionally, how to determine the value of a measurement in detecting pollutant sources is still unclear.

### B. Overview

The workflow of our proposed iterative scheme is illustrated in Fig. 4. In each iteration, there are four major steps.

1) Smartphone users in the selected grids participate in sensing, stimulated by the incentives published by the platform. As a result, measurements in these grids are collected.
2) For each grid with collected measurements, the platform estimates the pollution concentration by distinguishing indoor and outdoor measurements.
3) Based on the estimated pollution concentrations obtained so far, the platform detects the locations and emission rates of pollutant sources. If the detection result converges to the ground truth, the iteration stops.
4) Otherwise, the platform continually selects $\lfloor \log N \rfloor$ new grids and sets proper incentives to them. Then, the next iteration starts.

To realize this scheme, there exist four key issues that need to be addressed.

1) *Quality of Measurements:* First, it is unknown to the platform whether a measurement is sensed indoor or outdoor, while only outdoor measurements are valid in estimating the pollution concentration. Second, a group of measurements from the same grid are needed to eliminate their measuring errors cooperatively.
2) *Unknown Valuable Grids for CS:* Choosing grids to collect measurements which are valuable for detecting pollutant sources via CS, can reduce the amount of collected measurements, as well as the total payment. However, how to measure the value of each grid in detecting pollutant sources remains unsolved.
3) *Unknown Number of Iterations:* It is difficult to determine how many iterations are sufficient to guarantee the accuracy of pollutant source detection (namely judging the convergence), because the gap between the estimations and the ground truth cannot be calculated directly, and the enough number of sampled grids depends on an unknown parameter $k$ according to RIP.
4) *Balance Between Accuracy and Payment:* In terms of choosing grids to collect measurements in each iteration, achieving high detection accuracy and low total payment should be balanced.

In response to these issues, we propose an EM-based algorithm for pollution concentration estimation, a CS-based algorithm for pollutant source detection, an AL-based algorithm for grid selection, and an incentive mechanism, which are described in the following sections, respectively.

## IV. Scheme for Air Pollution Monitoring

In this section, we describe the details of the iterative scheme.

### A. Pollution Concentration Estimation

At the beginning of an iteration, smartphone users in grids with positive incentives are stimulated to participate in sensing. Based on the measurements collected from the same grid, the platform needs to estimate the pollution concentration in each grid with measurements. In practice, a measurement may be sensed under different contexts, like indoor and outdoor. The values of indoor and outdoor measurements are significantly different, while only outdoor measurements are valid

**Algorithm 1** EM-Based Algorithm for Pollution Concentration Estimation

```
Input: A group of measurements $\mathbf{M}_i$ and $\epsilon$
Output: Estimated parameters $\boldsymbol{\theta}_i = \{I_i, \rho_i, C_i, \sigma_i\}$
// Initialization
$I_i = \min(\mathbf{M}_i)$, $C_i = \max(\mathbf{M}_i)$;
$\rho_i = \sigma_i = \mathrm{var}(\mathbf{M}_i)$;
$\lambda_i^I = \lambda_i^O = 0.5$;
$\log L^{(0)}(\boldsymbol{\theta}_i|\mathbf{M}_i) = -\infty$;
while $|\log L^{(t)}(\boldsymbol{\theta}_i|\mathbf{M}_i) - \log L^{(t-1)}(\boldsymbol{\theta}_i|\mathbf{M}_i)| > \epsilon$ do
    $\log L^{(t)}(\boldsymbol{\theta}_i|\mathbf{M}_i) = 0$;
    // E-step
    for $r = 1$ to $\gamma$ do
        $w_r^I = f(m_r^{(i)}|I_i, \rho_i) * \lambda_i^I$;
        $_r^O = f(m_r^{(i)}|C_i, \sigma_i) * \lambda_i^O$;
        $\log L^{(t)}(\boldsymbol{\theta}_i|\mathbf{M}_i) += \log(w_r^I + w_r^O)$;
        $w_r^I = \frac{w_r^I}{w_r^I + w_r^O}$ and $w_r^O = \frac{w_r^O}{w_r^I + w_r^O}$;
    end for
    // M-step
    $\lambda_i^I = \sum_{r=1}^{\gamma} w_r^I$ and $\lambda_i^O = \sum_{r=1}^{\gamma} w_r^O$;
    $I_i = \sum_{r=1}^{\gamma} m_r^{(i)} * w_r^I$ and $C_i = \sum_{r=1}^{\gamma} m_r^{(i)} * w_r^O$;
    $\rho_i = \sum_{r=1}^{\gamma} (m_r^{(i)} - I_i)^2 * w_r^I$ and $\sigma_i = \sum_{r=1}^{\gamma} (m_r^{(i)} - C_i)^2 * w_r^O$;
end while
return $\boldsymbol{\theta}_i$.
```

in the pollution concentration estimation. Therefore, the platform needs to distinguish indoor and outdoor measurements first. In this section, we show how to apply the EM algorithm, which can classify indoor and outdoor measurements and find the MLE of the pollution concentration in a grid.

We assume that either indoor or outdoor measurements collected in a grid obeys a normal distribution with the expectation equal to the ground truth of the grid. In grid $i$, the indoor and outdoor distributions are denoted by $\mathcal{N}(I_i, \rho_i)$ and $\mathcal{N}(C_i, \sigma_i)$, respectively.[3] Given measurements $\mathbf{M}_i$ collected in grid $i$, parameters $\boldsymbol{\theta}_i = \{I_i, \rho_i, C_i, \sigma_i\}$ can be estimated via the EM algorithm as shown in Algorithm 1.

In Algorithm 1, the E-step and the M-step are iteratively executed until the log-likelihood function converges (i.e., the gap between two successive iterations is less than a small threshold $\epsilon$). In the E-step, for each measurement $m_r^{(i)}$ collected in grid $i$, probabilities $w_r^I$ and $w_r^O$ are updated, respectively, based on current estimated parameters $\boldsymbol{\theta}_i$ and mixture proportions of indoor and outdoor normal distributions $\{\lambda_i^I, \lambda_i^O\}$. Note that $\{w_r^I, w_r^O\}$ is the probability distribution of latent variable $z_r^{(i)}$ to denote whether measurement $m_r^{(i)}$ is sensed indoor or outdoor. In the M-step, parameters of indoor and outdoor normal distributions are updated according to the new values of $\{w_r^I, w_r^O\}$.

Due to the existence of various contexts and measuring errors, a group of measurements should be collected to guarantee the accuracy of the pollution concentration estimation. To make sure the number of measurements needed, we do an extensive simulation as an example, to show its impact on the estimation accuracy. In the simulation, the values of measurements are randomly generated under parameters $I = 30$, $\rho = 15$, $C = 100$, and $\sigma = 10$. As shown in Fig. 5, we plot the estimation error with varying the number of measurements and the percent of indoor measurements. It is observed that the estimation error declines in general as the number of measurements increases no matter how many indoor measurements are mixed in. However, the decrease of estimation error becomes minimal after a certain number of measurements. In our case, 25 measurements are sufficient, which achieves 98% accuracy in average and higher than 90% accuracy in the worst case.

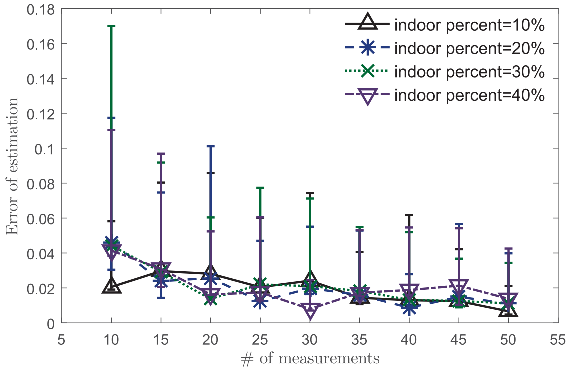


Fig. 5. Estimation error versus number of measurements under different percents of indoor measurements are mixed in.

## *B. Pollutant Source Detection*

In this section, we show how to employ CS to detect the locations and emission rates of pollutant sources, based on the obtained pollution concentration estimations so far. We denote the set of grids with collected measurements by $\boldsymbol{\pi} = \{\pi_1, \pi_2, \ldots, \pi_m\}$, where $m$ is the number of grids with collected measurements and $\pi_i \in \{1, 2, \ldots, N\}$. The estimated pollution concentrations in these grids are denoted by $\widehat{\mathbf{C}} = \{\widehat{C}_{\pi_1}, \widehat{C}_{\pi_2}, \ldots, \widehat{C}_{\pi_m}\}^T$.

As there is $\mathbf{C} = \Omega\mathbf{g}$, $\mathbf{C}$ can be seen as decomposed into $\mathbf{g}$ based on basis vectors $\Omega$, although $\Omega$ is not orthogonal. Due to the sparsity of $\mathbf{g}$, CS can be employed to recover the whole pollution map based on $\widehat{\mathbf{C}}$. Given $\widehat{\mathbf{C}}$, its corresponding transfer matrix is $\Omega' = \{\Omega_{\pi_i}\}_{i=1}^m$. Therefore, the value of $\mathbf{g}$ can be estimated as $\widehat{\mathbf{g}}$, by solving the following problem:

$$\arg\min_{\widehat{\mathbf{g}}} \|\widehat{\mathbf{g}}\|_{\ell_1}, \text{ s.t. } \widehat{\mathbf{C}} = \Omega'\widehat{\mathbf{g}}. \tag{9}$$

The nonzero elements in $\widehat{\mathbf{g}}$ point out the locations and emission rates of pollutant sources.

To obtain an accurate estimation $\widehat{\mathbf{g}}$, transfer matrix $\Omega'$ should be carefully chosen according to RIP. However, verifying an arbitrary matrix $\Omega'$ satisfies RIP or not is combinatorially complex in time, and there are $(2^n - 1)$ different choices of $\Omega'$. Thus, it is impossible to determine the sampled set of grids $\boldsymbol{\pi}$ at the start of the scheme. As illustrated in Fig. 4, we enlarge

[3] Note that we estimate the pollution concentration in a grid as the expectation of the outdoor distribution achieved by Algorithm 1. We do not distinguish the estimation and the ground truth in this section.

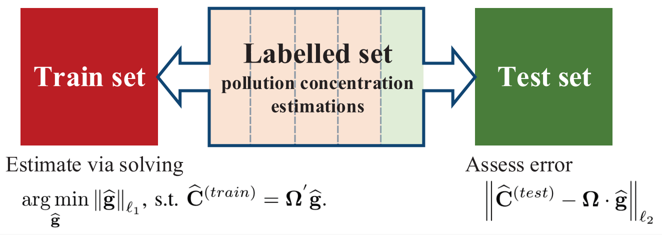

Train set
Labelled set
pollution concentration estimations
Test set
Estimate via solving
$\arg\min_{\widehat{\mathbf{g}}} \|\widehat{\mathbf{g}}\|_{\ell_1}$, s.t. $\widehat{\mathbf{C}}^{(train)} = \mathbf{\Omega}'\widehat{\mathbf{g}}$.
Assess error
$\left\|\widehat{\mathbf{C}}^{(test)} - \mathbf{\Omega}\cdot\widehat{\mathbf{g}}\right\|_{\ell_2}$

Fig. 6. Illustration of the fivefold cross validation method used to judge convergence.

the sampled set of grids step by step. In each iteration, $\lfloor \log n \rfloor$ new grids are selected to collect measurements, and proper incentives are set in these grids. As $k$ is unknown, the number of iterations cannot be decided in prior. The iteration should end when the estimation of pollutant sources $\widehat{\mathbf{g}}$ converges to the ground truth $\mathbf{g}$.

However, due to the unknown $\mathbf{g}$, it is nontrivial to judge when the convergence is achieved. To test the difference between our estimation $\widehat{\mathbf{g}}$ and ground truth $\mathbf{g}$, we adopt the *k-fold cross-validation* method. Cross-validation is a technique for assessing how accurately a predictive model will perform in practice, which is trained given a set of labeled samples. In a round of cross-validation, the labeled dataset is divided into two subsets: one is used to train the model (called *training set*); the other is used to validate (called *testing set*). The mean squared error (MSE) of the testing set is always used to assess the accuracy of the trained model. Multiple rounds are performed using different partitions to reduce the assessment error. In our scheme, we use fivefold cross validation, where the set of pollution concentration estimations $\widehat{\mathbf{C}}$ is randomly partitioned into five equal sized subsets, as shown in Fig. 6. In each round, a single subset acts as the testing set $\widehat{\mathbf{C}}^{(\text{test})}$ to assess error $\|\widehat{\mathbf{C}}^{(\text{test})} - \Omega \cdot \widehat{\mathbf{g}}\|_{\ell_2}$, while the other four subsets compose the training set $\widehat{\mathbf{C}}^{(\text{train})}$, which are used to achieve $\widehat{\mathbf{g}}$ by solving (9).

## C. Incentive Mechanism Design

To guarantee collecting sufficient measurements (e.g., $\gamma = 25$) with a high probability, a proper payment should be provided to stimulate low-cost smartphone users. The proper payment set in a certain grid can be deduced according to the user participation model described in Section II-C.

We take costs of users following uniform distribution (e.g., $c_s \sim \mathcal{U}(c_{\min}, c_{\max})$) as an instance, to illustrate how to deduce the payment set to a grid. Given the number of users in grid $i$, $n_i$, we have derived that the number of participants $Y_i$ follows the Binomial distribution, i.e., $Y_i \sim B(n_i, p_i)$, where $p_i = [(P_i - c_{\min})/(c_{\max} - c_{\min})]$. To stimulate at least $\gamma$ users participating with success probability no less than 99%, i.e.,

$$1 - \sum_{r=0}^{\gamma-1} \Pr(Y_i = r) \geq 99\%.$$

According to (7), $p_i$ can be computed by solving the above inequality. Then, payment $P_i$ can be calculated as $P_i = p_i \cdot (c_{\max} - c_{\min}) + c_{\min}$. Fig. 7 plots the payment when varying the number of users under two different cost distributions, where $c_{\min} = 1$ and $c_{\max} = 20$. We can observe that as the number of users grows, the required payment for each measurement declines, while the marginal decrease becomes smaller and smaller.

**Algorithm 2** AL-Based Algorithm for Grid Selection in an Iteration

```
Input: Set of grids $\mathcal{N}$, transfer matrix $\Omega$, payment $\mathbf{P}$, current
    sampled grids $\boldsymbol{\pi}^{(t)}$ and estimation of pollutant sources $\widehat{\mathbf{g}}^{(t)}$
    and $\widehat{\mathbf{g}}^{(t-1)}$
Output: Updated sampled grids $\boldsymbol{\pi}^{(t+1)}$
for each $i \in \mathcal{N} \setminus \boldsymbol{\pi}^{(t)}$ do
    Calculate $I_i$ and $R_i$ according to Eq. (10) and Eq. (11);
end for
for $l = 1$ to $\lfloor \log N \rfloor$ do
    $\boldsymbol{\pi}^{(t+1)} = \boldsymbol{\pi}^{(t)} \cup \{\arg\max_i \frac{I_i}{R_i \cdot P_i}\}$;
end for
return $\boldsymbol{\pi}^{(t+1)}$;
```

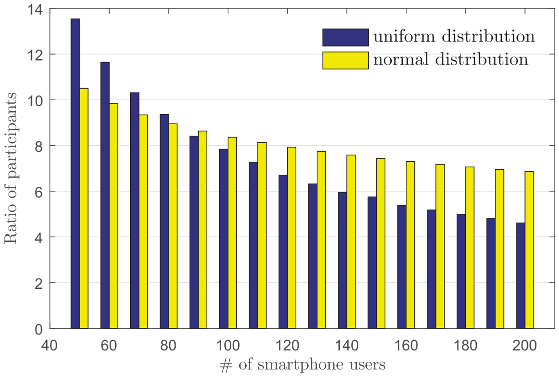


Fig. 7. Payment set for each measurement versus number of smartphone users in a grid.

## D. Grid Selection

According to the design of our scheme shown in Fig. 4, in each iteration $\lfloor \log N \rfloor$ new grids should be selected to set proper incentives by the platform if the detection of pollutant sources does not converge. By carefully selecting partial grids to collect measurements, the platform can obtain more labeled data (pollution concentration estimations) to solve the CS-based pollutant source detection problem, which coincides with the general framework of pool-based AL as shown in Fig. 8. AL [26] is a major solution to exploit unlabeled data in machine learning, where the learner can decide which unlabeled data to pose queries.

In this article, with the objective of enhancing the accuracy of pollutant source detection and maintaining a low payment, we consider employ a density-weighted AL method to decide which grid is chosen. The method considers not only the payment for collecting measurements in the grid but also the *informativeness* and *representativeness* of its pollution concentration estimation in pollutant source detection.

First, we show the mathematical definition of informativeness and representativeness of the labeled data of a grid, given

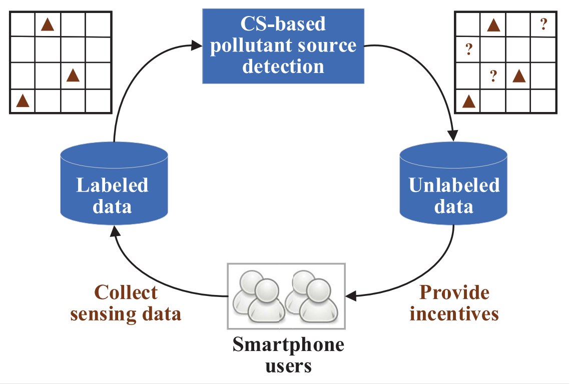


Fig. 8. Illustration of the pool-based AL framework for pollutant source detection.

the current sampled data. We denote the estimations of pollutant sources in iteration $(t-1)$ and iteration $t$ as $\widehat{\mathbf{g}}^{(t-1)}$ and $\widehat{\mathbf{g}}^{(t)}$, respectively. For an arbitrary grid $i \in \mathcal{N}\backslash\boldsymbol{\pi}$, its informativeness $I_i$ is defined as the difference between the two pollution concentrations computed based on $\widehat{\mathbf{g}}^{(t-1)}$ and $\widehat{\mathbf{g}}^{(t)}$, respectively, i.e.,

$$I_i = \Omega_i \times \left|\widehat{\mathbf{g}}^{(t)} - \widehat{\mathbf{g}}^{(t-1)}\right|. \tag{10}$$

Intuitively, the higher $I_i$, the more information contained by grid $i$. The representativeness of grid $i$ is defined as how the instance of grid $i$ distinguishes with the current sampled ones in data structure, i.e.,

$$R_i = \frac{1}{|\boldsymbol{\pi}^{(t)}|}\sum_{j\in\boldsymbol{\pi}^{(t)}} \mathrm{sim}(\Omega_i, \Omega_j) \tag{11}$$

where $\mathrm{sim}(\Omega_i, \Omega_j)$ calculates the cosine similarity of the two vectors. According to CS technology, the higher $R_i$, the lower value of gird $i$ to detect pollutant sources.

Then, we design a heuristic algorithm for selecting $\lfloor \log N \rfloor$ grids by balancing their payments, informativeness and representativeness at the same time. As shown in Algorithm 2, the informativeness and representativeness of each unsampled grid are calculated based on the set of sampled grids (lines 1–3). Then, $\lfloor \log N \rfloor$ grids with the largest value of metric $[I_i/(R_i \cdot P_i)]$ are selected in each iteration (lines 4–6), and proper incentives are set in these grids to collect measurements.

## V. Performance Evaluation

In this section, we evaluate the performance of our proposed scheme, especially Algorithm 2.

### A. Methodology and Setups

It has been proven by many previous works [19], [27] that CS performs well in recovering a sparse vector from a few samples, compared with common interpolation methods, such as linear interpolation and Kriging interpolation [28]. Thus, in our simulations, we concentrate on showing the performance of our proposed grid selection algorithm, compared with three baseline algorithms.

1) *Random:* This algorithm randomly chooses $\lfloor \log N \rfloor$ new grids in each iteration until convergence.
2) *Greedy-Payment:* The cheapest $\lfloor \log N \rfloor$ grids are selected in each iteration until convergence.
3) *Greedy-Informativeness:* In each iteration, $\lfloor \log N \rfloor$ grids with the most information are picked up.
4) *Greedy-VRR [29]:* This algorithm considered the ratio between payment and informativeness as the metric to choose grids.

Three metrics are used to evaluate the performance of the four algorithms from different aspects.

1) *Error of Estimation:* This metric is calculated as $[(\|\widehat{\mathbf{g}} - \mathbf{g}\|_{\ell_2})/n]$, where $\widehat{\mathbf{g}}$ is estimated according to (9) based on all measurements collected until convergence. It measures the accuracy of detecting pollutant sources based on our iterative scheme.
2) *Total Payment:* We sum up the incentives given to all participants collecting measurements in selected grids as $\sum_{i\in\boldsymbol{\pi}}(\gamma \cdot P_i)$ to represent the total payment.
3) *Number of Iterations:* This metric is proportional to the total number of sampled grids as well as the time consumed for the crowdsensing process.

The default setting of system parameters is as follows. All simulations are conducted on a square area divided into 50 × 50 grids ($N = 2500$), and the size of each grid is equal to 200 m×200 m. The wind blows from west to east at $v = 5$ m/s. Transfer matrix $\Omega \in \mathbb{R}^{2500\times2500}$ can be computed according to (4) given $\sigma_y = 200$ and $\sigma_z = 1000$. The locations and the emission rates of pollutant sources are randomly chosen from $\{1, \ldots, 2500\}$ and $\{1000 \text{ mg/s}, \ldots, 5000 \text{ mg/s}\}$, respectively. The population $n_i$ in each grid is randomly generated varying in $[50,\ 200]$. We conduct simulations considering both uniform distribution and normal distribution for the costs of smartphone users, with $c_{\min} = 1$ and $c_{\max} = 20$. Given $n_i$, the value of payment $P_i$ provided to each measurement can be known according to Fig. 7. For example, if there are 60 users in a grid, the payment is set as 11.64 and 9.83, respectively, considering uniform distribution and normal distribution of costs of users. We study the performance of Algorithm 2 and the four baseline algorithms by varying the number of pollutant sources $k$ from 5 to 25, respectively. The result of each setting is the average of ten runs. The simulations are implemented in MATLAB R2018a on a Dell server (PowerEdge T420, Intel E5-2400, 1.8-GHz CPU, 4-GB DDR3 memory, 300-GB Disk) with Windows 10 operation system.

### B. Simulation Results

Figs. 9–11 plot the performance of the five algorithms under uniformly distributed costs of users, while Figs. 12–14 plot the performance under normal distributed costs of users.

Figs. 9 and 12 show the error of estimated pollution sources achieved by different algorithms. We can find that our algorithm can accurately detect pollution sources via CS if sufficient measurements are collected. The estimation error achieved by our algorithm is less than 30%, no matter how the number of pollution sources varies. The result also demonstrates that the cross-validation method works well for judging the convergence of our iterative scheme. We can find that

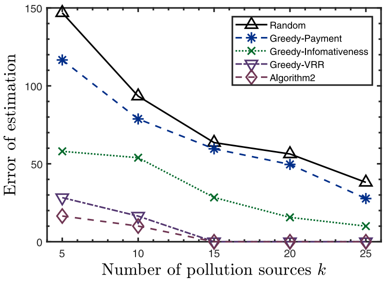


Fig. 9. Error of estimation versus number of pollutant sources under uniform distributed costs.

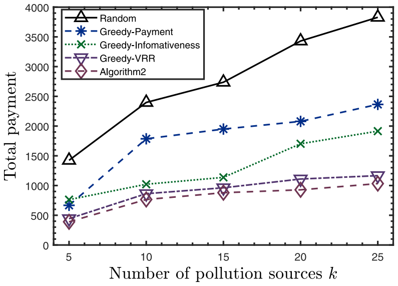


Fig. 10. Total payment versus number of pollutant sources under uniform distributed costs.

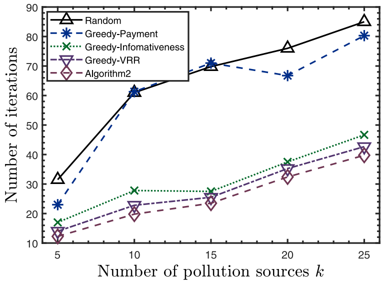


Fig. 11. Number of iterations versus number of pollutant sources under uniform distributed costs.

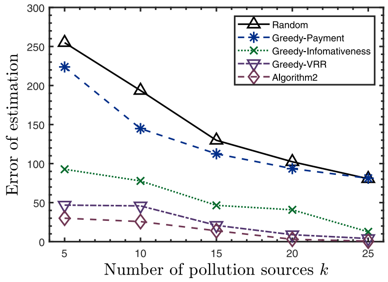


Fig. 12. Error of estimation versus number of pollutant sources under normal distributed costs.

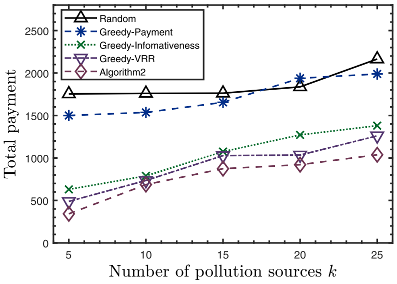


Fig. 13. Total payment versus number of pollutant sources under normal distributed costs.

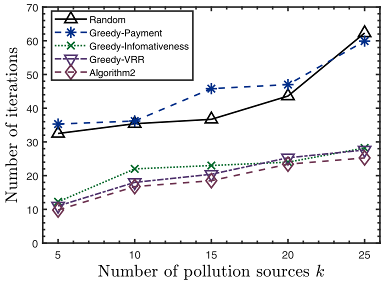


Fig. 14. Number of iterations versus number of pollutant sources under normal distributed costs.

greedy-VRR algorithm and Algorithm 2 perform better in different settings, compared with the other three algorithms, as both cost and informativeness of measurements are considered. Moreover, all algorithms perform better under uniform distributed costs than normal distributed costs because more measurements are collected as shown in Figs. 11 and 14. When there are ten pollution sources, the estimation errors achieved by our algorithm are 10.0 and 25.7 under different distributed costs, which are 39.0% and 43.9% lower than greedy-VRR algorithm, and 64.8% and 67.0% lower than greedy-informativeness algorithm, respectively.

Figs. 10 and 13 show that more payment is needed as $k$ increases. Although greedy-payment algorithm chooses the cheapest grids in each iteration, it consumes more money compared with greedy-informativeness algorithm, greedy-VRR algorithm, and Algorithm 2. This is because the cheapest measurements may suffer poor values in pollutant source detection, which leads to collecting more measurements as shown in Figs. 11 and 14, and thus, incurring a high payment. When there are 25 pollutant sources, Algorithm 2 can save 64.9% and 33.1% payment compared with greedy-informativeness algorithm, and save 13.0% and 21.5% payment compared with greedy-VRR under two different cost distributions.

As shown in Figs. 11 and 14, more iterations are needed by the four baseline algorithms, compared with our algorithm. In other words, more measurements are collected by the baselines to achieve accurate pollutant source estimations, which incurs higher payments. We can find that random algorithm and greedy-payment algorithm need significantly more measurements, as cheap but low-value measurements are chosen by them. Specifically, 57.4% and 51.5% more measurements are needed by random algorithm and greedy-payment under uniform distributed costs, compared with our algorithm, when there are 20 pollution sources.

## VI. Related Work

In environment-centric crowdsensing applications (e.g., pollution mapping and traffic monitoring), the platform needs to aggregate plenty of sensing data and pay to participants. In this section, we review related works from the following two aspects important for reducing the total payment.

### A. Compressive Crowdsensing

CS has been proved to be efficient in reducing the amount of sampled data [27], [30]. A few works [17]–[19] have applied it in crowdsensing. Xu *et al.* [19] solved a fundamental problem when applying CS in crowdsensing applications, which is the base of transforming sensing data into sparse representation that is unknown. In this article, the base can be derived according to the air pollution dispersion model. Xu *et al.* [17] considered cost diversity of sensing samples and design cost-aware CS to balance total cost and recovery accuracy. However, the cost of each sample should be known a prior and the cost minimization is achieved given a fixed amount of sampled data. In [18], an online task allocation algorithm is proposed to minimize the number of collected sensing data by leveraging the spatial and temporal correlation. Similarly, we repeat the crowdsensing process and choose specific grids to collect data in each cycle. However, instead of assuming cooperative smartphone users, we consider them as rational and provide incentives to them, which does not incur latencies and overheads in gathering their costs before each round.

### B. Incentive Mechanism Design

A series of studies [12]–[14] have been dedicated to designing incentive mechanisms for crowdsensing applications to stimulate smartphone users participating in sensing. In [12], a recurrent reverse auction is employed to select participants according to their locations given constraints in budget and coverage. Koutsopoulos [13] derived a mechanism that minimizes the total cost paid to participants by tracking the quality of their reported cost information and using it for determining participation level and payment. Zhao *et al.* [14] proposed online incentive mechanisms by considering smartphone users randomly arrive one by one. Under a budget, the value of services provided by participants is maximized before a given deadline. All these mechanisms collect private information from each smartphone user before the sensing process, and focus on making them truthful. Different from these studies, we analyze the participation model in terms of a group of users rather than an individual, which follows statistic laws. Therefore, incentives can be designed according to the population in interested areas.

## VII. Conclusion

This article has focused on reducing the total payment by exploiting the spatial correlations of sensing data for compressive crowdsensing-based urban air pollution monitoring. Specifically, we first employ a Gaussian air pollution dispersion model to characterize the relationship between the fine-grained pollution concentrations and the locations and emission rates of pollutant sources. Then, we propose an iterative scheme to recruit smartphone users collecting measurements. In the scheme, we provide an EM-based algorithm for detecting measurements collected in different contexts. Also, an incentive mechanism is designed to stimulate smartphone users participating. Finally, the framework of pool-based AL is employed to select the most informative and representative grids to set incentives in each iteration. Comprehensive simulations have been conducted to confirm the superiority of our proposed algorithms.

## References

[1] Y. Zheng, F. Liu, and H.-P. Hsieh, "U-Air: When urban air quality inference meets big data," in *Proc. 19th ACM SIGKDD Int. Conf. Knowl. Disc. Data Min.*, 2013, pp. 1436–1444.

[2] H.-P. Hsieh, S.-D. Lin, and Y. Zheng, "Inferring air quality for station location recommendation based on urban big data," in *Proc. KDD*, 2015, pp. 437–446.

[3] A. Thiagarajan *et al.*, "VTrack: Accurate, energy-aware road traffic delay estimation using mobile phones," in *Proc. 7th ACM Conf. Embedded Netw. Sensor Syst. (SenSys)*, 2009, pp. 85–98.

[4] M. Stevens and E. D'Hondt, "Crowdsourcing of pollution data using smartphones," in *Proc. Workshop Ubiquitous Crowdsourcing*, 2010, pp. 1–4.

[5] Accessed: May 2019. [Online]. Available: https://en.wikipedia.org/wiki/Air_Quality_Egg

[6] Accessed: May 2019. [Online]. Available: https://www.rmit.edu.au/news/all-news/media-releases/2015/october/revolutionary-new-weapon-in-air-pollution-fight/

[7] P. Dutta *et al.*, "Common sense: Participatory urban sensing using a network of handheld air quality monitors," in *Proc. 7th ACM Conf. Embedded Netw. Sensor Syst. (SenSys)*, 2009, pp. 349–350.

[8] D. Mendez, A. J. Perez, M. A. Labrador, and J. J. Marron, "P-sense: A participatory sensing system for air pollution monitoring and control," in *Proc. IEEE Int. Conf. Pervasive Comput. Commun. Workshops (PERCOM Workshops)*, 2011, pp. 344–347.

[9] D. Hasenfratz, O. Saukh, S. Sturzenegger, and L. Thiele, "Participatory air pollution monitoring using smartphones," in *Proc. Mobile Sens.*, 2012, pp. 1–5.

[10] L. Yang, W. Li, M. Ghandehari, and G. Fortino, "People-centric cognitive Internet of Things for the quantitative analysis of environmental exposure," *IEEE Internet Things J.*, vol. 5, no. 4, pp. 2353–2366, Aug. 2018.

[11] D. Yang, G. Xue, X. Fang, and J. Tang, "Crowdsourcing to smartphones: Incentive mechanism design for mobile phone sensing," in *Proc. 18th Annu. Int. Conf. Mobile Comput. Netw.*, 2012, pp. 173–184.

[12] L. G. Jaimes, I. J. Vergara-Laurens, and M. A. Labrador, "A location-based incentive mechanism for participatory sensing systems with budget constraints," in *Proc. IEEE Int. Conf. Pervasive Comput. Commun. (PerCom)*, 2012, pp. 103–108.

[13] I. Koutsopoulos, "Optimal incentive-driven design of participatory sensing systems," in *Proc. IEEE INFOCOM*, 2013, pp. 1402–1410.

[14] D. Zhao, X.-Y. Li, and H. Ma, "How to crowdsource tasks truthfully without sacrificing utility: Online incentive mechanisms with budget constraint," in *Proc. IEEE INFOCOM*, 2014, pp. 1213–1221.

[15] M. R. Beychok, *Fundamentals of Stack Gas Dispersion*. Irvine, CA, USA: M. R. Beychok, 1995.

[16] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[17] L. Xu, X. Hao, N. D. Lane, X. Liu, and T. Moscibroda, "Cost-aware compressive sensing for networked sensing systems," in *Proc. 14th Int. Conf. Inf. Process. Sensor Netw.*, 2015, pp. 130–141.

[18] L. Wang *et al.*, "CCS-TA: Quality-guaranteed online task allocation in compressive crowdsensing," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput. (UbiComp)*, 2015, pp. 683–694.

[19] L. Xu, X. Hao, N. D. Lane, X. Liu, and T. Moscibroda, "More with less: Lowering user burden in mobile crowdsourcing through compressive sensing," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput. (UbiComp)*, 2015, pp. 659–670.

[20] Accessed: May 2019. [Online]. Available: https://en.wikipedia.org/wiki/Outline_of_air_pollution_dispersion

[21] S. P. Arya *et al.*, *Air Pollution Meteorology and Dispersion*, vol. 6. New York, NY, USA: Oxford Univ. Press, 1999.

[22] S. Janhäll, "Review on urban vegetation and particle air pollution—Deposition and dispersion," *Atmos. Environ.*, vol. 105, pp. 130–137, Mar. 2015.

[23] G. McLachlan and T. Krishnan, *The EM Algorithm and Extensions*, vol. 382. New York, NY, USA: Wiley, 2007.

[24] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inf. Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006.

[25] W. Wang, M. N. Garofalakis, and K. Ramchandran, "Distributed sparse random projections for refinable approximation," in *Proc. ACM 6th Int. Conf. Inf. Process. Sensor Netw.*, 2007, pp. 331–339.

[26] B. Settles, "Active learning," in *Synthesis Lectures on Artificial Intelligence & Machine Learning*, vol. 6. San Rafael, CA, USA: Morgan & Claypool, 2012, p. 765.

[27] Y. Zhu, Z. Li, H. Zhu, M. Li, and Q. Zhang, "A compressive sensing approach to urban traffic estimation with probe vehicles," *IEEE Trans. Mobile Comput.*, vol. 12, no. 11, pp. 2289–2302, Nov. 2013.

[28] R. Woodard, "Interpolation of spatial data: Some theory for kriging," *Technometrics*, vol. 42, no. 4, pp. 436–437, 2000.

[29] T. Liu, Y. Zhu, Y. Yang, and F. Ye, "Incentive design for air pollution monitoring based on compressive crowdsensing," in *Proc. IEEE Glob. Commun. Conf. (GLOBECOM)*, 2016, pp. 1–6.

[30] L. Kong, M. Xia, X.-Y. Liu, M.-Y. Wu, and X. Liu, "Data loss and reconstruction in sensor networks," in *Proc. IEEE INFOCOM*, 2013, pp. 1654–1662.

**Tong Liu** (M'19) received the B.Eng. and Ph.D. degrees from the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China, in 2012 and 2017, respectively.

She is an Assistant Professor with the School of Computer Engineering and Science, Shanghai University, Shanghai. Her current research interests include mobile crowdsensing, edge computing, and urban computing.

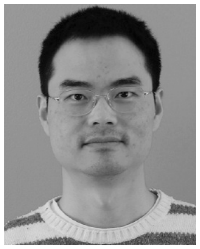
**Yanmin Zhu** (SM'17) received the B.Eng. degree in computer science from Xi'an Jiaotong University, Xi'an, China, in 2002, and the Ph.D. degree in computer science from the Hong Kong University of Science and Technology, Hong Kong, in 2007.

He is a Professor with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China. His current research interests include sensor network, vehicular ad hoc networks, and mobile computing.

**Yuanyuan Yang** (F'09) received the B.Eng. and M.S. degrees in computer science and engineering from Tsinghua University, Beijing, China, and the M.S.E. and Ph.D. degrees in computer science from Johns Hopkins University, Baltimore, MD, USA.

She is a Professor of computer engineering and computer science with Stony Brook University, Stony Brook, NY, USA, and the Director of Communications, and Devices Division New York State Center of Excellence in Wireless and Information Technology. Her current research interests include wireless networks, data center networks, optical networks, and high-speed networks. She has published over 270 papers in major journals and refereed conference proceedings and holds seven U.S. patents in the above areas.

Prof. Yang is currently an Associate Editor-in-Chief for the IEEE TRANSACTIONS ON COMPUTERS and an Associate Editor for the *Journal of Parallel and Distributed Computing*. She has served as an Associate Editor for the IEEE TRANSACTIONS ON COMPUTERS and the IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS. She has served as the general chair, the program chair, or the vice chair for several major conferences and a program committee member for numerous conferences.



**Fan Ye** received the B.E. and M.S. degrees in automation and computer science from Tsinghua University, Beijing, China, and the Ph.D. degree in computer science from the University of California at Los Angeles, Los Angeles, CA, USA.

He then joined IBM T. J. Watson Research, Yorktown Heights, NY, USA, as a Research Staff Member, researching on stream processing systems, cloud messaging, and mobile computing. He is currently an Associate Professor with the Department of Electrical and Computer Engineering, Stony Brook University, Stony Brook, NY, USA. He holds over a dozen U.S./international patents and patent applications. His current research interests include mobile computing, mobile cloud, wireless networks, sensor networks, and their applications.