Incentivizing Crowdsensing With Location-Privacy Preserving

Xiong Wang, Zhe Liu, Xiaohua Tian, *Member, IEEE*, Xiaoying Gan, *Member, IEEE*, Yunfeng Guan, *Member, IEEE*, and Xinbing Wang, *Senior Member, IEEE*

***Abstract*— Crowd sensing systems enable a wide range of data collection, where the data are usually tagged with private locations. How to incentivize users to participate in such systems while preserving location-privacy is coming up as a critical issue. To this end, we consider location-privacy protection when motivating users to sense data instead of viewing them separately. Without loss of generality, k-anonymity is utilized to reduce the risk of location-privacy disclosure. Specifically, we propose a location aggregation method to cluster users into groups for k-anonymity preserving, and meanwhile mitigating the incurred information loss. After that, an incentive mechanism is carefully designed to select efficient users and calculate rational compensations based on clustered groups obtained in location aggregation, where the influences of both the information loss and k-anonymity in location-privacy preserving are captured into group values and sensing costs. Through theoretical analysis and extensive performances evaluated on real and synthetic data, we find out that the incentive payment increases sharply with more stringent privacy protection and the information loss can be further mitigated compared with conventional methods.**

***Index Terms*— Crowdsensing, incentive mechanism, location-privacy, reverse auction, k-anonymity.**

I. Introduction

THE past few years have witnessed the proliferation of smart mobile devices, like smartphones and tablets, which are embedded with powerful processors and plentiful sensors (e.g., GPS, thermometer, microphone, camera). Recently, the crowdsensing paradigm has been proposed to leverage the widely distributed mobile devices for sensing and collecting ubiquitous data, such as Ear-Phone to construct urban noise map [1], P-Sense to monitor air pollution [2], Nericell to sense road and traffic conditions [3], and Zee to build indoor fingerprint database [4]. Those sensory results are usually tagged with locations to form a database or map for information release and decision making.

Manuscript received January 14, 2017; revised June 2, 2017; accepted July 27, 2017. Date of publication August 4, 2017; date of current version October 9, 2017. This work was supported by the National Natural Science Foundation of China under Grant 61672342, Grant 61572319, Grant U1405251, Grant 61671478, Grant 61532012, Grant 61325012, and Grant 61521062. The associate editor coordinating the review of this paper and approving it for publication was Z. Sun. *(Corresponding author: Xinbing Wang.)*

X. Wang is with the Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai 200240, China. He is also with the State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an 710071, China (e-mail: wangxiongsjtu@sjtu.edu.cn).

X. Tian, X. Gan, Y. Guan, and X. Wang are with the Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: xtian@sjtu.edu.cn; ganxiaoying@sjtu.edu.cn; yfguan69@yahoo.com; xwang8@sjtu.edu.cn).

Z. Liu is with the State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an 710071, China (e-mail: zliuxidian@gmail.com).

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

Digital Object Identifier 10.1109/TWC.2017.2734758

Incentive reward is necessary to compensate the cost for user participation, because sensing data not only takes the time but also consumes resources like battery and wireless traffic. To motivate users, game and auction based incentive mechanism is designed to calculate the monetary payment [5]. Since sensing tasks are usually related to locations, incentivizing users to complete the tasks within their vicinities in location aware crowdsensing system becomes increasingly popular [6], [7]. The sensory data are meaningful only when they are associated to the corresponding locations in those crowdsensing systems, such as constructing a noise map or wireless signal fingerprint database.

Users' locations are vulnerable to malicious attacks. Even if users are protected by fake identities like pseudonym, an adversary can utilize their locations to infer the private information, such as political affiliations, alternative lifestyles, or medical problems [14]. Therefore, it is especially essential to achieve location-privacy protection. Location-privacy has been vastly studied in Location Based Service (LBS) and crowdsensing system, where the servers or platforms are often regarded untrusted. To preserve location-privacy, various methods are proposed including information caching [15], spatial cloaking [16], and data perturbation [17]. The goal is to prevent the servers or platforms inferring users' exact locations. A framework for protecting location-privacy in spatial crowdsourcing is proposed [18], where a cellular service provider is considered as a trusted third party to perform spatial cloaking and interact with the crowdsourcing server. However, the privacy preserving methods, such as spatial cloaking, need to process the reported locations, which frequently incurs large information loss. Since the information loss will lead to location deviations from true values which can reduce the data precision, it is vitally important to mitigate the inevitable information loss in location-privacy preserving.

Due to the concerns of compensating sensing costs and preventing personal information leakage, there is a need to consider both incentive mechanism design and location-privacy protection simultaneously. However, on the one hand, stimulating users to join is complicated because strategic users have different sensing costs and require different rewards. On the other hand, it is non-trial to achieve low information loss since location processing is essential to preserve privacy. Besides, integrating the incentive mechanism and location-privacy protection organically is no longer addressing them

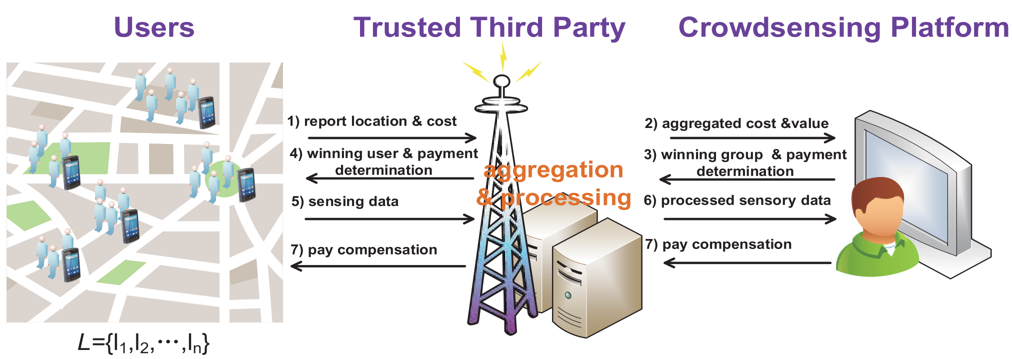


Fig. 1. Snapshot of crowdsensing system.

dividedly, which is also much more difficult. With these issues in mind, we aim to address the two problems in the same time.

In this paper, we consider incorporating location-privacy protection into incentive mechanism. Particularly, we focus on k-anonymity location-privacy in the reverse auction design, and study how privacy protection affects the auction process. A snapshot of the crowdsensing system is drawn in Fig. 1, which is composed of *Users, Trusted Third Party* and *Crowdsensing Platform*. The objective of the crowdsensing platform is to publish tasks and collect location aware data from the users. However, the crowdsensing platform is untrusted, while the users are vulnerable. To preserve user's location-privacy in uploaded sensory data, a trusted third party, such as an independent server or cellular service provider, will aggregate users into groups according to their locations and process sensory results within a group as well. The processed data, like the mean value, in a group with corresponding location centroid will be uploaded to the crowdsensing platform. After that, an incentive calculated based on data value and sensing cost will be paid to users. We promote a location aggregation method for location-privacy protection and information loss mitigation. Besides, we design a reverse auction, which combines the location aggregation process, to stimulate users and compensate sensing costs. Theoretical analysis and performance evaluations based on real and synthetic data are provided to demonstrate the system efficiency. We summarize the main contributions of this paper as follows:

- We propose a novel incentive mechanism with k-anonymity location-privacy preserving, where the incentive mechanism also captures the requirement and consequence of privacy preserving. Such an integrated framework can not only stimulate users but also protect private locations.
- We present a location aggregation scheme to preserve k-anonymity location-privacy, and meanwhile further mitigate the information loss. The scheme can prevent location information disclosure and obtain more accurate results.
- We design a reverse auction based incentive mechanism to select efficient users and compensate sensing costs, where the minimum total cost is approximated with a guaranteed ratio. The incentive mechanism also satisfies truthfulness, individual rationality and computational efficiency.

The rest of the paper is organized as follows. Section II presents related work including incentive mechanism field and privacy preserving field. Section III formulates location-privacy protection and the reverse auction, in which k-anonymity and data collection model are introduced. In Section IV, we propose the location aggregation algorithm to cluster users into groups while mitigating information loss. Section V provides the details of the auction design where winner selection and payment calculation are explained. Performances are evaluated in Section VI, and the conclusion is drawn in Section VII.

II. Related Work

The rise of crowdsensing paradigm has proliferated a broad range of mobile applications, among which, the requirements of sensory data tagged with locations become increasingly common, like outdoor sensory data map [1], [2] and indoor localization fingerprint construction [4]. The fundamental issue of crowdsensing system is to efficiently recruit and motivate user participation. In [8], Gao *et al.* leverage reverse VCG auction and Lyapunov optimization to incentivize users to upload location aware sensory data under the long term participation constraint. Wen *et al.* propose a quality driven auction to construct Wi-Fi fingerprint in indoor localization system [9]. However, a significant problem frequently neglected in these works is location-privacy protection of users. Since users' locations are correlated to their identities, it is necessary to protect the location privacy in crowdsensing system.

Some works, like [10], also take into account the privacy issue when incentivizing users, but often they stress incentive mechanism design and lack implementation of privacy protection. Considering differential privacy in incentive mechanism, Jin *et al.* add correlated noise to sensory data when aggregated to database, in which the crowdsensing platform is assumed trusted [11]. Similar framework is designed when the prices of users are homogeneous [12]. Actually the sensory data, such as Wi-Fi RSS and noise, are often not sensitive, while the tagged locations are sensitive instead. Besides, adding noise will destroy the data characteristic and may incur large information loss. Analogously, differential location-privacy is combined into winner selection rule of incentive mechanism [13], but the platform is regarded trusted and economic properties are guaranteed only with probabilities. Users behind the crowdsensing platform who publish the sensing tasks are often anonymous, thereby it is more appreciated to consider them potentially malicious. In this paper, we incorporate k-anonymity location-privacy into incentive mechanism in the data collection phase, where the crowdsensing platform is regarded untrusted.

As for location-privacy, Gruteser *et al.* reveal that an adversary can infer users' private information through locations, unless location-privacy protection like k-anonymity is implemented [14]. Niu *et al.* propose a spatial cloaking method to prevent the LBS server recovering users' actual locations [19]. They utilize information cache technique to store historical queries sent to the LBS sever which can reduce the location exposure risks. In crowdsensing system, Vu *et al.* utilize voronoi diagram to cluster users into different groups and preserve k-anonymity as well [20]. Besides, an overview of four location-privacy protection algorithms is concluded where real dataset is used to conduct the comparison experiments [21]. However, these methods often suffer from large information loss, which can damage the usability of sensory data. Microaggregation is identified as a good trade-off technique between privacy protection and information loss reduction [22]–[24]. We propose a location aggregation algorithm to protect the k-anonymity privacy and further mitigate the information loss.

III. Preliminary and System Model

A. Preliminary

1) k-Anonymity: k-anonymity is a metric for privacy preserving [25]. To protect user's privacy, k-anonymity requires that at least k reports are combined together before being released. Regarding k-anonymity of location-privacy, we are motivated to mix at least k users' locations into a group, in which an adversary cannot distinguish one user's location from the rest of others'. Besides, we are supposed to further carry out data fusion, including locations and sensory data, within a group such as calculating the mean values.

2) Information Loss: Suppose there is a dataset $\boldsymbol{x}$ which is aggregated into N_k groups. The within-group sum of squares for group i is defined as:

$$sse_i = \sum_{j=1}^{n_i} (x_{ij} - \overline{x}_i)^T (x_{ij} - \overline{x}_i), \tag{1}$$

where n_i is the count of data in group i, "T" is transposition and $\overline{x}_i$ is the mean value of group i. sse_i is the variance of group i, which can measure the aggregation property of data samples in the group. If sse_i is large, the data are "far" from each other, namely the group is not homogeneous. The total within-groups sum of squares SSE is the sum of sse_i:

$$SSE = \sum_{i=1}^{N_k} sse_i = \sum_{i=1}^{N_k} \sum_{j=1}^{n_i} (x_{ij} - \overline{x}_i)^T (x_{ij} - \overline{x}_i), \tag{2}$$

here SSE describes the overall group homogeneity after aggregation. When nearby data are grouped together, SSE will be small and the groups are homogeneous. The total sum of squares SST is:

$$SST = \sum_{i=1}^{N_k} \sum_{j=1}^{n_i} (x_{ij} - \overline{x})^T (x_{ij} - \overline{x}), \tag{3}$$

where $\overline{x}$ is the mean value of the whole dataset. SST is determined when the dataset is given and is also independent from group partition results. The information loss IL is the ratio between SSE and SST:

$$IL = \frac{SSE}{SST}. \tag{4}$$

Since IL is only related to SSE, thus minimizing IL is equivalent to minimizing SSE. Besides, if groups are homogeneous, the information loss will be low. Therefore, one of our goals is to keep the homogeneity after group aggregation.

B. System Model

In the crowdsensing system shown in Fig. 1, the crowdsensing platform solicits N users $\mathbb{U} = \{1, 2, \ldots, N\}$ to complete location aware sensing tasks, like sensing environmental data. However, the crowdsensing platform can be malicious, thus directly exposing users' raw locations is harmful to the location-privacy. A trusted third party, which is a cellular service provider in [18], is supposed to protect the location-privacy and process the sensory data. The workflow of the system is as follows:

1) Users report their locations $\mathbb{L} = \{l_1, l_2, \ldots, l_N\}$, outdoor or indoor, to the trusted third party for location-privacy protection. Here for user i, its location l_i is expressed by a tuple (x_i, y_i) indicating the two coordinates. Along with those locations, the users will also claim costs $\mathbb{B} = \{c_1, c_2, \ldots, c_N\}$ for taking the sensing tasks.
2) The trusted third party is supposed to perform aggregation on the locations and interact with the untrusted crowdsensing platform. In view of k-anonymity privacy, the trusted third party constructs N_k groups $\mathbb{G} = \{g_1, g_2, \ldots, g_{N_k}\}$ with each group size no less than k. Based on the group aggregation results, group values $\mathbb{V} = \{v_1, v_2, \ldots, v_{N_k}\}$ and group costs $\mathbb{B}^g = \{c_1^g, c_2^g, \ldots, c_{N_k}^g\}$ will be computed.
3) According to group values and costs, the crowdsensing platform utilizes the reverse auction to select the winning groups W_g and calculate corresponding group payments p_j^g, $\forall g_j \in W_g$.
4) The users in winning groups are winning users, and their payments will be computed based on group payments.
5) Winning users will undertake the sensing tasks, and the trusted third party will process the sensory results within the same groups, such as computing the mean values.
6) The processed sensory data tagged with groups centroids will be uploaded to the crowdsensing platform.
7) The crowdsensing platform pays the winning users the determined values through the trusted third party according to 3) and 4).

However, location aggregation for privacy protection will lead to the information loss, in that the original locations instead of the aggregated locations are corresponding to the sensory results. One of the main concentrations is to mitigate the information loss and select those users who are near to each other to form a group. Similar to Eq. (1), the sse_i of g_i is:

$$sse_i = \sum_{j=1}^{n_i}[(x_{ij} - \overline{x}_i)^2 + (y_{ij} - \overline{y}_i)^2], \tag{5}$$

where n_i is the number of users in g_i satisfying $n_i \geq k$, (x_{ij}, y_{ij}) is the location of the jth user with $(\overline{x}_i, \overline{y}_i)$ the mean location coordinates of g_i. Therefore, the total within-groups sum of squares is calculated:

$$SSE = \sum_{i=1}^{N_k} sse_i = \sum_{i=1}^{N_k}\sum_{j=1}^{n_i}[(x_{ij} - \overline{x}_i)^2 + (y_{ij} - \overline{y}_i)^2]. \tag{6}$$

One aim is to minimize SSE for information loss mitigation. With a little abuse of notations, we will utilize SSE to denote the total within-groups sum of squares, while sse to denote within-group sum of squares for a group. The accuracy and value of the data for each group are highly related to its sse. Specifically, larger sse means higher information loss and smaller group value, while smaller sse implies lower information loss and larger group value. In Section IV-B, we will step into details about the design scheme of calculating group value v_i when obtaining sse_i. Furthermore, the reason behind will also be explained.

The users are selfish and strategic to maximize their own utilities. The crowdsensing system aims to minimize the total cost for the recruited users, while collecting a certain amount of sensory data with acceptable quality. Considering the case of noise map construction, a relatively dense noise sample with low deviation can efficiently deliver the noise message to people. Therefore, assume that N_Q number of data are required with quality no less than Q, then we formulate this Cost Minimization with Quality and Number constraints problem (CMQN) as follows:

$$\begin{aligned} &min \sum_{g_i \in W_g} c_i^g \\ &s.t.\ |W_g| \geq N_Q \text{ (cardinality constraint)} \\ &\quad f(\sum_{g_i \in W_g} v_i) \geq Q \text{ (quality constraint)}, \end{aligned} \tag{7}$$

where W_g is the winning group set and $f(\cdot)$ maps group values to quality, which is non-decreasing and submodular:

$$\begin{aligned} f(S) \leq f(T) \quad \forall S \subseteq T \subseteq G \\ f(S \cup \{x\}) - f(S) \geq f(T \cup \{x\}) - f(T) \\ \forall S \subseteq T \subseteq \mathbb{G}, \quad x \in \mathbb{G} \setminus T. \end{aligned}$$

$f(.)$ reflects the diminishing marginal utility of sensory data, as formulated in [5], [7], and [9]. A typical form for f is log function, also instantiated in Eq. (25), which captures the subadditive feature of sensory data value. Another common instantiation is weighted sum dedicated to the case where the total value of sensory data is additive. N_Q and Q represent the demands of the crowdsensing platform for sensory data. Specifically, if the crowdsensing platform requires to collect a large data volume and high data quality, N_Q and Q should be set to large values, and vice versa.

Denote the payment to user i as p_i, then its utility u_i is:

$$u_i = \begin{cases} p_i - c_i & \text{if } i \text{ wins} \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

According to the workflow of the system, Eq. (7) and (8), the results of location aggregation are captured in the reverse auction. Particularly, group costs, group values and group aggregations will influence how winning users are selected and how users' payments are determined. In addition to solving CMQN, the reverse auction is also expected to satisfy the favored economic properties:

Truthfulness: For every user i, it cannot gain a larger utility by claiming a higher or lower price than its cost c_i:

$$u_i(\widetilde{c_i}) \leq u_i(c_i), \tag{9}$$

where $\widetilde{c_i}$ is the untruthful claimed cost.

Individual Rationality: For any user i, the corresponding utility is greater than 0 when claiming the truthful cost: $u_i(c_i) \geq 0$.

Computational Efficiency: The reverse auction should be completed in polynomial time.

IV. Location Aggregation for Privacy Protection

We utilize microaggregation method to cluster the users into groups according to their location coordinates in consideration of SSE minimization. Since k-anonymity privacy is to be preserved, [22] has proved that the multivariate microaggregation problem was NP-hard, and an optimal solution should output groups with size greater than or equal to k and less than $2k$. Therefore, we propose a heuristic location aggregation algorithm to obtain a suboptimal result where the incurred information loss turns out to be further mitigated. Based on Eq. (4), information loss mitigation is equivalent to SSE minimization, thereby our goal is to acquire a near optimal SSE.

A. Location Aggregation

Since users' locations are utilized to perform the aggregation, we call this procedure location aggregation which needs to process the two x and y location coordinates. Previous works have also investigated the multivariate aggregation problem, therefore before explicating our proposed scheme we will introduce the related algorithms for better comparison.

1) Maximum Distance to Average Vector (MDAV) [22] produces a group with fixed cardinality k each time. To generate a k size group, MDAV will iteratively add a new data which is closest to the first selected data. If some data are left, they will form a new group. However, MDAV still introduces relatively high SSE due to fixed group size and simple clustering technique
2) V-MDAV utilizes variable group size instead of the fixed ones [23] to further reduce SSE, in which the size of group is not necessary to be k. But V-MDAV still uses the first selected data to determine the next added data.

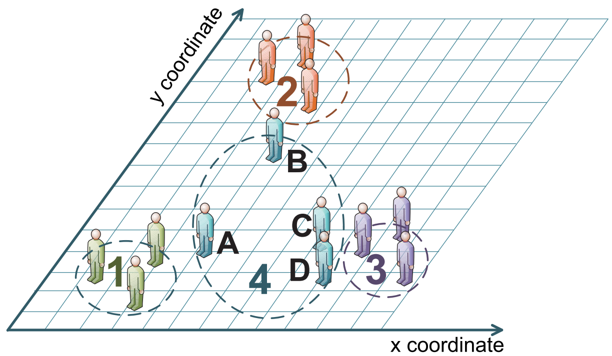


Fig. 2. Example of microaggregation with $k = 3$.

3) Importance Partitioning (IP) leverages the group centroid to add next data, but the size of generated group is still fixed [24].

Take Fig. 2 as an example where $k = 3$. MDAV first outputs groups $1, 2, 3$ and then clusters the rest into group 4. While V-MDAV will count A, B into group 1 and 2 respectively, and count C, D into group 3, that is V-MDAV outputs three groups. As for IP, it adds the rest of users into a group by comparing the distances between their locations and the group centroid. In this example, IP will add D into group 3 instead of group 4 since D is closer to the former group.

In this paper, we propose the Variable Centroid Location Aggregation (VCLA) scheme with an eye to variable size and distance to group centroid for location-privacy protection. Assume that we already have n_j users in g_j, then its sse_j is:

$$\begin{aligned} sse_j &= \sum_{m=1}^{n_j}[(x_{jm} - \frac{\sum_{l=1}^{n_j} x_{jl}}{n_j})^2 + (y_{jm} - \frac{\sum_{l=1}^{n_j} y_{jl}}{n_j})^2] \\ &= \sum_{m=1}^{n_j}[x_{jm}^2 - 2x_{jm}\frac{\sum_{l=1}^{n_j} x_{jl}}{n_j} + \frac{(\sum_{l=1}^{n_j} x_{jl})^2}{n_j^2} \\ &\quad + y_{jm}^2 - 2y_{jm}\frac{\sum_{l=1}^{n_j} y_{jl}}{n_j} + \frac{(\sum_{l=1}^{n_j} y_{jl})^2}{n_j^2}] \\ &= \sum_{m=1}^{n_j} x_{jm}^2 - \frac{(\sum_{l=1}^{n_j} x_{jl})^2}{n_j} + \sum_{m=1}^{n_j} y_{jm}^2 - \frac{(\sum_{l=1}^{n_j} y_{jl})^2}{n_j}. \end{aligned}$$

If a new user with $l' = (x', y')$ is added to g_j, the updated sse'_j is obtained:

$$sse'_j = \sum_{m=1}^{n_j+1} x_{jm}^2 - \frac{(\sum_{l=1}^{n_j+1} x_{jl})^2}{n_j+1} + \sum_{m=1}^{n_j+1} y_{jm}^2 - \frac{(\sum_{l=1}^{n_j+1} y_{jl})^2}{n_j+1},$$

notice that we count the new user as the $(n_j + 1)$th member in g_j. The incremental within-group sum of squares Δ_j is the difference between sse'_j and sse_j:

$$\begin{aligned} \Delta_j &= sse'_j - sse_j \\ &= x'^2 - \frac{(\sum_{l=1}^{n_j} x_{jl} + x')^2}{n_j+1} + \frac{(\sum_{l=1}^{n_j} x'_{jl})^2}{n_j} \\ &\quad + y'^2 - \frac{(\sum_{l=1}^{n_j} y_{jl} + y')^2}{n_j+1} + \frac{(\sum_{l=1}^{n_j} y'_{jl})^2}{n_j}. \end{aligned}$$

Denote $\overline{x}_j = \frac{\sum_{l=1}^{n_j} x_{jl}}{n_j}$, $\overline{y}_j = \frac{\sum_{l=1}^{n_j} y_{jl}}{n_j}$, and we obtain:

$$\begin{aligned} \Delta_j &= x'^2 - \frac{(n_j\overline{x}_j + x')^2}{n_j+1} + n_j\overline{x}_j^2 \\ &\quad + y'^2 - \frac{(n_j\overline{y}_j + y')^2}{n_j+1} + n_j\overline{y}_j^2 \\ &= \frac{n_j x'^2 - 2n_j\overline{x}_j + n_j\overline{x}_j^2}{n_j+1} + \frac{n_j y'^2 - 2n_j\overline{y}_j + n_j\overline{y}_j^2}{n_j+1} \\ &= \frac{n_j}{n_j+1}[(x' - \overline{x}_j)^2 + (y' - \overline{y}_j)^2]. \end{aligned} \tag{10}$$

Eq. (10) shows that we should add the user which is closest to the group centroid, instead of to the first selected user as in [22] and [23], for lower sse increase. We describe the location aggregation process in Algorithm 1, where $|\cdot|$ denotes the cardinality and $d(l_i, l_j)$ is the distance between l_i of user i and l_j of user j.

Algorithm 1 Variable Centroid Location Aggregation (VCLA)

Input: User set $\mathbb{U}$, location set $\mathbb{L}$ and k
Output: Aggregation results $\mathbb{G}$
1 Compute the location distance matrix D between users;
2 Compute the global centroid $\bar{l}$;
3 $\mathbb{G} = \emptyset$, $count = 1$;
4 **while** $|\mathbb{U}| \geq k$ **do**
5 $\quad i_{max} = \arg\max_{i\in\mathbb{U}} d(l_i, \bar{l})$;
6 $\quad g_{count} = \{i_{max}\}$, $\mathbb{U} = \mathbb{U} \setminus \{i_{max}\}$;
7 $\quad$ **for** $j = 1 : k-1$ **do**
8 $\quad\quad$ Update the centroid $\bar{l}_{count}$ of g_{count};
9 $\quad\quad i_{min} = \arg\min_{i\in\mathbb{U}} d(l_i, \bar{l}_{count})$;
10 $\quad\quad g_{count} = g_{count} \cup \{i_{min}\}$, $\mathbb{U} = \mathbb{U} \setminus \{i_{min}\}$;
11 $\quad g_{count} = extend(g_{count}, \mathbb{U})$;
12 $\quad \mathbb{G} = \mathbb{G} \cup g_{count}$, $count = count + 1$;
13 **if** $|\mathbb{U}| > 0$ **then**
14 $\quad$ **for** $i \in |\mathbb{U}|$ **do**
15 $\quad\quad i_g = \arg\min_{j\in\mathbb{G}} \frac{|g_j|}{|g_j|+1} d(l_i, \bar{l}_j)$;
16 $\quad\quad g_{i_g} = g_{i_g} \cup \{i\}$;
17 $\quad\quad$ Update $\mathbb{G}$ and $\bar{l}_{i_g}$;
18 **return** $\mathbb{G}$;

First, we compute the distance matrix and the global centroid in Lines 1-2 which will be utilized in the next group clustering process. In the *while* loop, the farthest user to the global centroid is picked each time in Line 5, then a group with size k is formed using the centroid of already selected users' locations where new users are aggregated one by one in Lines 6-10. Considering the variable size, each group is extended in Line 11 which will be analyzed later. The *while* loop guarantees that the group size is between k and $2k - 1$ with minimum sse increase every time. If there are still some users left, we will add each of them to the group which incurs the lowest sse increase in Lines 14-17. Note that the term $\frac{|g_j|}{|g_j|+1}$ comes from Eq. (10). Now recall the $extend()$ function which allows adaptive group size. For g_i with centroid $\bar{l}_i$, suppose that the nearest user's location in the unassigned set to $\bar{l}_i$ is l_{out}:

$$l_{out} = \arg\min_{j \in \mathbb{U}_{un}} d(\bar{l}_i, l_j),$$

where $\mathbb{U}_{un}$ is the unassigned user set. Let d_{out} represent the shortest distance between l_{out} and the rest unassigned users' locations:

$$d_{out} = \min_{j \in \mathbb{U}_{un} \setminus \{i_{out}\}} d(l_j, l_{out}),$$

and if the following condition holds:

$$d(\bar{l}_i, l_{out}) < \beta d_{out}, \tag{11}$$

corresponding user i_{out} will be counted into g_i. Here β is a constant related to the sparse characteristics of locations. We show the *extend*() function in Algorithm 2, which stops until group size grows to $2k - 1$ or Eq. (11) is not satisfied. In the *do-while* loop, we extend the input k cardinality group g_i to a larger set bounded by $2k - 1$.

Algorithm 2 Extend Groups

```
Input: Group g_i, unassigned set U and location set L
Output: Extended set g_i
Compute l̄_i;
do
    i_out = arg min_{j∈U} d(l_j, l̄_i);
    if d(l_out, l̄_i) > β d_out then
        break;
    g_i = g_i ∪ {i_out};
    Update l̄_i;
while |g_i| < 2k − 1;
return g_i;
```

If users have different k values, we can take the maximum k value to aggregate them, which still satisfies each user's requirement of location-privacy protection. To further mitigate information loss, we subdivide the users into several types with users of similar k values in one type and aggregate them separately using the maximum k value of each type, or just discard users who have very large k values but only constitute a small proportion.

B. Aggregation Analysis

Lemma 1: VCLA preserves k-anonymity of user's location-privacy.

Proof: According to Algorithm 1, the size of each group is at least k. Since the crowdsensing platform can only get access to the aggregated location, it cannot distinguish one's location from the rest of others'. That is k-anonymity of user's location-privacy is preserved. ■

Lemma 2: The computation complexity of VCLA is $O(N^2)$.

Proof: The computation complexity of distance matrix obtainment is $O(N^2)$. As for computing the centroid, it costs $O(N)$. The iteration times of the *while* is bounded by $\lfloor \frac{N}{k} \rfloor$, and computing farthest user to the centroid costs at most $O(N)$. The time complexity of clustering k size group is bounded by $(k-1)O(N)$. Note that acquiring group centroid only costs $O(1)$:

$$\bar{l}_{count} = \frac{(|g_{count}| - 1)\bar{l}_{count} + l_{i_{min}}}{|g_{count}|}.$$

The time complexity of *extend* function is bounded by $(k - 1)O(N)$. And adding the rest of users to assigned groups costs at most $(k-1)\lfloor \frac{N}{k} \rfloor$. Therefore, the overall time complexity is bounded by:

$$O(N^2) + O(N) + \left\lfloor \frac{N}{k} \right\rfloor \times [O(N) + (2k-2)O(N) + k - 1] = O(N^2)$$

■

Lemma 1 and 2 imply that our proposed VCLA can preserve the location-privacy in an efficient way. Moreover, in the computation complexity analysis, we find that the time consumption mainly lies in the distance computation and group clustering which are $O(N^2)$. Due to location aggregation, the sensory data within each group will be processed and further associated to the group centroid. In VCLA, users with adjacent locations are clustered. Follow the fact that sensory data tend to be stable in nearby locations, thus the processed sensory data can well represent the true values in group centroids owing to the aggregated adjacent locations and sensory data stability. We will also validate the time complexity and sensory data processing in the performance evaluation.

After location aggregation, we can obtain N_k groups and the corresponding sse. Furthermore, we are required to calculate the cost and value for each group which are specified as group cost and group value. For g_i, the group cost c_i^g is:

$$c_i^g = n_i \times max_{j \in g_i} c_j. \tag{12}$$

The reason behind Eq. (12) is to satisfy truthfulness and individual rationality, which will be discussed in Section V. In the meantime, the value v_i of g_i is obtained:

$$v_i = \alpha \frac{n_i^{1/\gamma}}{sse_i + 1}. \tag{13}$$

Next, we explain why we design the scheme of computing group values. Since larger sse_i means higher information loss within a group which will decrease the data quality, v_i is inversely proportional to sse_i, and the term $+1$ in the denominator is to prevent infinite v_i. Besides, more users mean more accurate aggregated results in that more sensory data mean lower random error in data fusion, for example when calculating the mean value of the sensory data within a group the result will be less impacted by the data randomness if more sensory data are collected. α and γ are the adjusting parameters depending on real situations.

Because the crowdsensing platform is regarded potentially malicious, thus location aggregation is carried out by the trusted third parity. Since the crowdsensing platform aims to gather sensory data, it will incentivize winning groups to complete sensing tasks after the location aggregation, where tasks are actually undertaken by the corresponding winning users. In this paper, we propose a reverse auction to stimulate users, in which the calculation of the auction process is based on the group values $\mathbb{V} = \{v_1, v_2, \ldots, v_{N_k}\}$ and the group costs $\mathbb{B}^g = \{c_1^g, c_2^g, \ldots, c_{N_k}^g\}$.

V. Reverse Auction for User Recruitment

Sensing data will consume resources and users are reluctant to participate for free, thus the crowdsensing platform needs to compensate the winning users for sensing costs. The crowdsensing platform is supposed to collect the sensory data from participatory users, which is formulated as CMQN in Eq. (7). Considering the sensory data procurement, we propose a reverse auction to model the winner determination process for selecting winners and the payment calculation process for compensating consumptions. The reverse auction is also expected to satisfy truthfulness, individual rationality and computational efficiency.

A. Winner Selection and Payment Calculation

Before stepping into the detailed description for reverse auction design, we first demonstrate that CMQN is NP-hard.

Theorem 1: The CMQN problem defined in Eq. (7) *is NP-hard.*

Proof: In order to prove that CMQN is NP-hard, we show that a Special case of CMQN (SCMQN) is NP-hard, which is formulated in Eq. (14):

$$\begin{aligned} &min \sum_{g_i \in W_g} c_i^g \\ &s.t.\ |W_g| = K \\ &\qquad \sum_{g_i \in W_g} w_i v_i \geq Q, \end{aligned} \tag{14}$$

where K is a constant and w_i is a factor associated to v_i. Next we will demonstrate the NP-hardness of SCMQN by reduction from the Minimum Size-constrained Weighted Set Cover (MSWSC) problem [26].

As for MSWSC, there are a set $\mathbb{W}$, and a set of subsets in $\mathbb{W}$ denoted as $W_g = \{g_1, g_2, \ldots, g_{N_k}\}$. Each $g_i \in W_g$ is assigned with a weight c_i^g. The objective of MSWSC is to find a subset of W_g of size K with the minimum total weight whose union can cover $\mathbb{W}$. To link the MSWSC and SCMQN, we carry out the following reduction:

- For each element $g_i \in W_g$, it has w_i copies.
- Each element in $\mathbb{W}$ should be covered Q times.

Since the reduction can be completed in polynomial time and every instance of MSWSC can be reducible to SCMQN, the NP-hardness of SCMQN is proved. Furthermore, CMQN generalizes SCMQN, thus CMQN is NP-hard. ■

Theorem 1 shows that CMNQ is NP-hard, thereby directly solving it is computationally intractable. Alternately, we implement an efficient greedy algorithm to obtain a guaranteed suboptimal result. Given set W, define the marginal value of g_i as:

$$\rho_{g_i}(W) = f(W \cup \{g_i\}) - f(W). \tag{15}$$

According to the submodularity of $f(.)$, $\rho_{g_i}(W)$ is diminishing over set W. The ratio between $\rho_{g_i}(W)$ and c_i^g is the marginal efficiency. We myopically select a group with the highest marginal efficiency, and if there is a tie, choose the group with a lower index. We depict the selection of winning groups in Algorithm 3.

Algorithm 3 Reverse Auction Winner Selection: CMQN

Input: Group set $\mathbb{G}$, value set $\mathbb{V}$, cost set $\mathbb{B}^g$, constraint parameters Q, N_Q
Output: Winning group set W_g

1 $W_g = \emptyset, Q' = Q, N_Q' = N_Q, \mathbb{G}' = \mathbb{G}$;
2 **while** $Q' > 0 \,||\, N_Q' > 0$ **do**
3 $\quad \rho_{g_i}(W_g) = f(W_g \cup \{g_i\}) - f(W_g)$;
4 $\quad g' = \arg\max_{g_i \in \mathbb{G}'} \frac{\rho_{g_i}(W_g)}{c_i^g}$;
5 $\quad W_g = W_g \cup \{g'\}, \mathbb{G}' = \mathbb{G}' \setminus \{g'\}$;
6 $\quad Q' = Q - f(W_g), N_Q' = N_Q' - 1$;
7 **return** W_g;

Conditions of *while* assure that the winning group set W_g satisfies the cardinality constraint and the quality constraint. Lines 3-4 choose the highest marginal efficiency group at each iteration. Lines 5-6 update winning group set, unselected group set and constraints. The output of algorithm 3 is also a suboptimal solution to CMQN, where an approximation ratio to the optimal result in the worst case will also be illustrated. After determining the winners, the crowdsensing platform needs to compute payments for winners. The basic idea behind payment calculation is to eliminate the influence of the winning group by excluding it from $\mathbb{G}$, which is illustrated in Algorithm 4.

In Algorithm 4, Line 4 describes the winner exclusion process, and then each time a group g' with the highest marginal efficiency is picked in Line 8. After that, a temporal payment p' is obtained which is the ratio between the marginal efficiency of g_j and the marginal efficiency of g' multiplying claimed cost c' in Lines 9-10. Parameters are updated and constraints are checked in Lines 11-13. Finally, the payment is the maximum value of all temporal payments in Line 14. Suppose that user j is in g_l, j will share the group payment evenly with the other group members:

$$p_j = \frac{p_l^g}{n_l}, \tag{16}$$

where p_l^g is the group payment obtained from Algorithm 4 and n_l is the number of users in g_l. Eq. (16) also implies that the payment to each winning user depends on both the location aggregation results and the reverse auction process.

B. Reverse Auction Analysis

In what follows, we will analyze the performances of the proposed reverse auction. In addition to the preferred economic properties, we also present the performance bound of the result for CMQN in Algorithm 3.

Algorithm 4 Reverse Auction Payment Calculation

Input: Winning group set W_g, group set $\mathbb{G}$, value set $\mathbb{V}$, cost set $\mathbb{B}^g$, constraint parameters Q, N_Q

Output: Group payment set $\mathbb{P}^g = \{p_1^g, p_2^g, \dots, p_{N_k}^g\}$

for $g_i \in \mathbb{G}$ **do**
  $p_i^g = 0$;
for $g_j \in W_g$ **do**
  $\mathbb{G}_{-j} = \mathbb{G} \setminus g_j, W_{-j} = \emptyset, \mathbb{P}_{-j} = \emptyset$;
  $Q' = Q, N'_Q = N_Q$;
  do
    $\rho_{g_l}(W_{-j}) = f(W_{-j} \cup \{g_l\}) - f(W_{-j})$;
    $g' = \arg\max_{g_l \in \mathbb{G}_{-j}} \frac{\rho_{g_l}(W_{-j})}{c_l^g}$;
    $\rho_{g_j}(W_{-j}) = f(W_{-j} \cup \{g_j\}) - f(W_{-j})$;
    $p' = \frac{\rho_{g_j}(W_{-j})}{\rho_{g'}(W_{-j})} c'$;
    $W_{-j} = W_{-j} \cup \{g'\}, \mathbb{P}_{-j} = \mathbb{P}_{-j} \cup \{p'\}$;
    $\mathbb{G}_{-j} = \mathbb{G}_{-j} \setminus \{g'\}$;
    $Q' = Q - f(W_g), N'_Q = N'_Q - 1$;
  while $Q' < 0$ & $N'_Q <= 0$;
  $p_j^g = \max \mathbb{P}_{-j}$;
return $\mathbb{P}^g$;

Lemma 3: A solution of CMQN is feasible if and only if the solution is feasible to the following problem:

$$
\begin{aligned}
&min \sum_{g_i \in \mathbb{G}} I'_i c_i^g \\
&s.t. \ |W'| \geq N_Q \\
&\quad \sum_{g_i \in \mathbb{G}} \rho_{g_i}(S) I'_i \geq Q - f(S) \quad \forall S \subseteq \mathbb{G} \\
&\quad I'_i \in \{0, 1\},
\end{aligned} \tag{17}
$$

where $\rho_{g_i}(S)$ is defined in Eq. (15) *and I'_i is an indicator:*

$$
I'_i = \begin{cases} 1 & \text{if } g_i \in W' \\ 0 & \text{otherwise.} \end{cases} \tag{18}
$$

Proof: Suppose W is a feasible solution to CMQN, and I_i is the corresponding indicator of g_i. Since W is feasible, we have $|W| \geq N_Q$ and $f(W) \geq Q$. Because $f(\cdot)$ is a non-decreasing and submodular function, according to [27], $f(\cdot)$ has the following property:

$$
f(T) \leq f(S) + \sum_{g_j \in T \setminus S} \rho_{g_j}(S) \ \forall S, T \subseteq \mathbb{G}.
$$

As for the second constraint:

$$
\begin{aligned}
\sum_{g_i \in \mathbb{G}} \rho_{g_i}(S) I_i &= \sum_{g_i \in W \setminus S} \rho_{g_i}(S) I_i \\
&\geq f(W) - f(S) \\
&\geq Q - f(S).
\end{aligned}
$$

On the contrary, assume that W' is a feasible solution to the problem described in Eq. (17), and I'_i is the indicator. Since W' is feasible, its cardinality is no less than N_Q, that is the first constraint in CMQN holds. Let $S = W'$, we obtain:

$$
0 = \sum_{g_i \in \mathbb{G}} \rho_{g_i}(W') I'_i \geq Q - f(W').
$$

Therefore $f(W') \geq Q$, which means the second constraint in CMQN is satisfied.

Overall, the two problems are equivalent. ■

Lemma 3 implies that we can view the original CMQN from the perspective of the marginal value.

Lemma 4: Given $0 < x_1 \leq x_2 \leq \dots \leq x_n$, and $y_1 \geq y_2 \geq \dots \geq y_n > 0$. If $Z = \sum_{i=1}^{n-1} x_i(y_i - y_{i+1}) + x_n y_n = x_1 y_1 + \sum_{i=1}^{n-1}(x_{i+1} - x_i) y_{i+1}$, we have:

$$
Z \leq (\max_i x_i y_i)[1 + \ln \min (\frac{x_n}{x_1}, \frac{y_1}{y_n})]. \tag{19}
$$

Proof: Since $Z = \sum_{i=1}^{n-1} x_i(y_i - y_{i+1}) + x_n y_n$ and $x_i \leq (\max_i x_i y_i)/y_i$, then $Z \leq (\max_i x_i y_i)[\sum_{i=1}^{n-1}(1 - \frac{y_{i+1}}{y_i}) + 1]$. Due to the fact that:

$$
\ln x \geq 1 - \frac{1}{x} \ \forall x \geq 1,
$$

we have:

$$
\begin{aligned}
Z &\leq (\max_i x_i y_i)[\sum_{i=1}^{n-1}(1 - \frac{y_{i+1}}{y_i}) + 1] \\
&\leq (\max_i x_i y_i)[\sum_{i=1}^{n-1} \ln \frac{y_i}{y_{i+1}} + 1] \\
&= (\max_i x_i y_i)[\sum_{i=1}^{n-1} \ln \frac{y_1}{y_n} + 1].
\end{aligned}
$$

Similarly, we can prove that $Z \leq (\max_i x_i y_i)[\sum_{i=1}^{n-1} \ln \frac{x_n}{x_1} + 1]$. Therefore, Eq. (19) is acquired. ■

As for CMQN , Algorithm 3 outputs a suboptimal solution W_g with total cost $\sum_{g_i \in W_g} c_i^g$ in T_Q iterations satisfying $T_Q \geq N_Q$. Denote the optimal result of CMQN as W_{opt} with total cost $\sum_{g_j \in W_{opt}} c_j^g$. Assume that a winning group g_i is added into W_g at the r_ith round, and the corresponding winning group set is $W_g^{r_i}$. Let $\theta^{r_i} = \frac{c_i^g}{\rho_{g_i}(W_g^{r_i - 1})}$ which represents the reciprocal of the marginal efficiency at the r_ith iteration, and $\delta_1 = \frac{\theta^{T_Q}}{\theta^1}$, $\delta_2 = \max_{g_i \in W_g} \frac{\rho_{g_i}(W_g^0)}{\rho_{g_i}(W_g^{r_i - 1})}$. We give the bound of the ratio between the suboptimal total cost to the optimal total cost in Theorem 2, in which we extend the result in [28] to the constrained size submodular set cover problem and provide detailed proofs.

Theorem 2: The suboptimal cost of Algorithm 3 satisfies $(1 + \ln \min\{\delta_1, \delta_2\})$-approximation to the optimal cost, that is:

$$
\frac{\sum_{g_i \in W_g} c_i^g}{\sum_{g_j \in W_{opt}} c_j^g} \leq 1 + \ln \min\{\delta_1, \delta_2\}. \tag{20}
$$

Proof: In Algorithm 3, we select a group with the highest marginal efficiency, namely the minimum θ^t, at each

tth iteration. According to Lemma 3, the problem in Eq. (17) is equivalent to CMQN. Consider its relaxation problem:

$$\begin{aligned} &min \sum_{g_i \in \mathbb{G}} I_i c_i^g \\ &s.t. \sum_{g_i \in \mathbb{G}} \rho_{g_i}(W_g^t) I_i' \geq Q - f(W_g^t) \quad t = 0, 1, 2, \ldots, T_Q - 1 \\ &\quad I_i \geq 0. \end{aligned} \tag{21}$$

Note that $T_Q \geq N_Q$ guarantees the final generated W_g satisfies the cardinality constraint, and I_i is non-negative instead of 0 or 1. Suppose C_g^l is the optimal value of Eq. (21), thus $C_g^l \leq \sum_{g_j \in W_{opt}} c_j^g$. Let $\Theta = (\theta^1, \theta^2 - \theta^1, \ldots, \theta^{T_Q} - \theta^{T_Q - 1})$, according to the group selection process and the non-decreasing submodular function $f(\cdot)$, $\theta^{t+1} \geq \theta^t$ and $\Theta \succeq \mathbf{0}$. For a given winning group g_i, the marginal value satisfies:

$$\rho_{g_i}(W_g^0) \geq \rho_{g_i}(W_g^1) \geq \ldots \geq \rho_{g_i}(W_g^{r_i - 1}) \geq \rho_{g_i}(W_g^{r_i}) = 0,$$

where "$\geq$" holds due to submodularity of $f(\cdot)$. In line with Lemma 4, we obtain:

$$\begin{aligned} &\theta^1 \rho_{g_i}(W_g^0) + (\theta^2 - \theta^1)\rho_{g_i}(W_g^1) + \ldots \\ &\quad + (\theta^{r_i} - \theta^{r_i - 1})\rho_{g_i}(W_g^{r_i - 1}) \\ &\leq \max_{t=1,2,\ldots,r_i} \{\theta^t \rho_{g_i}(W_g^{t-1})\}[1 + \ln \min\{\frac{\theta^{r_i}}{\theta^1}, \frac{\rho_{g_i}(W_g^0)}{\rho_{g_i}(W_g^{r_i - 1})}\}] \\ &\leq c_i^g [1 + \ln \min\{\delta_1, \delta_2\}], \end{aligned} \tag{22}$$

here $\max_{t=1,2,\ldots,r_i}\{\theta^t \rho_{g_i}(W_g^{t-1})\} \leq c_i^g$ because Algorithm 3 selects a group with the highest marginal efficiency (minimum θ^t) each time, and g_i is added at the r_ith round.

The dual problem of the relaxation problem is:

$$\begin{aligned} &max \sum_{t=0}^{T_Q - 1} [Q - f(W_g^t)] x_t \\ &s.t. \sum_{j=0}^{T_Q - 1} \rho_{g_i}(W_g^t) x_t \leq c_i^g \quad i = 0, 1, 2, \ldots, N_k \\ &\quad x_t \geq 0. \end{aligned} \tag{23}$$

According to Eq. (22), $(1 + \ln \min\{\delta_1, \delta_2\})^{-1}\Theta$ is a feasible solution to the dual problem. Therefore:

$$\begin{aligned} &(1 + \ln \min\{\delta_1, \delta_2\})^{-1}[\theta^1(Q - f(W_g^0)) + (\theta^2 - \theta^1) \\ &\quad \times (Q - f(W_g^1)) + \ldots + (\theta^{T_Q} - \theta^{T_Q - 1})(Q - f(W_g^{T_Q - 1}))] \\ &\leq C_g^l \leq \sum_{g_j \in W_{opt}} c_j^g. \end{aligned}$$

Besides:

$$\begin{aligned} \sum_{g_i \in W_g} c_i^g &= \sum_{t=1}^{T_Q} \theta^t \ (f(W_g^t) - f(W_g^{t-1})) \\ &= \theta^1(Q - f(W_g^0)) + \sum_{t=2}^{T_Q} (\theta^t - \theta^{t-1})(Q - f(W_g^{t-1})). \end{aligned}$$

Finally, we obtain:

$$\sum_{g_i \in W_g} c_i^g \leq \sum_{g_j \in W_{opt}} c_j^g (1 + \ln \min\{\delta_1, \delta_2\}), \tag{24}$$

and the theorem is proved. ■

Theorem 2 indicates that the worst performance of Algorithm 3 still maintains a guaranteed approximation ratio compared with the optimal result.

Theorem 3: The reverse auction is truthful.

Proof: To prove truthfulness, we need to demonstrate that the reverse auction is monotone and pays threshold values [29].

The reverse auction is monotone. Suppose a group g_j with claimed cost c_j^g wins the auction, and if the group claims another cost $\widetilde{c_j^g} \leq c_j^g$, according to the winner selection rule, the group will still be selected as a winner. As for the users in the group g_j, if claiming a lower cost, the group cost will not increase according to Eq. (12), and the user will still win the auction. Therefore, the reverse auction is monotone.

The reverse auction pays the threshold values. First, we prove that for $g_j \in W_g$, p_j^g is the threshold payment. Assume that the maximum value in $\mathbb{P}_j$ is the rth value. If the cost c_j^g is higher than p_j^g, the group will be sorted after r, and g_j will not be selected as a winner any more. Second, we demonstrate that the payment to any winning user is the threshold value. From Eq. (16), users in the same winning group share the group payment equally. If any user claims a cost higher than the payment, the group cost will be larger than the group payment, and the group will not be selected, namely the user will lose the auction. Thus, the reverse auction pays a threshold value to each winning user. ■

Theorem 4: The reverse auction is individual rational.

Proof: For any user losing the auction the utility is 0. Therefore, we focus on the winning user case. For any winning group g_j, the payment p_j^g is the maximum value in $\mathbb{P}_j$, which is assumed to appear in the rth place. Then, we obtain:

$$p_j^g = \frac{\rho_{g_j}(W_g^{r-1})}{\rho_{g'}(W_g^{r-1})} c',$$

where g' is the highest marginal efficiency group with regard to the already selected group W_g^{r-1} at rth round. If $p_j^g < c_j^g$, the group g' will be selected according to Algorithm 3, and g_j will be not selected any more, which is a contradiction. Thus $p_j^g \geq c_j^g$ and the payment of any user in group g_j satisfies:

$$\frac{p_j^g}{n_j} \geq \frac{c_j^g}{n_j} = \max_{i \in g_j}\{c_i\},$$

which means the payment is no less than the cost. That is the reverse auction is individual rational. ■

Theorem 5: The reverse auction is computational efficiency.

Proof: First, we discuss the winner selection process described in Algorithm 3. The maximum iteration times is bounded by $\max\{\left\lfloor \frac{Q}{\min\{v_j\}} \right\rfloor, N_Q\}$. The computation complexity of finding the highest marginal efficiency group is bounded by $O(N_k)$, where N_k is much larger than N_Q and $\frac{Q}{\min\{v_j\}}$ in the crowdsensing system. Therefore, the computation complexity is bounded by:

$$\max\{\left\lfloor \frac{Q}{\min\{v_j\}} \right\rfloor, N_Q\} O(N_k) \leq O(N_k^2).$$

Second, we provide the computation complexity of the payment calculation illustrated in Algorithm 4. The cardinality of W_g is bounded by N_k, and the *for* loop iterates at most N_k runs. In each iteration, the process is similar to the winner selection, and the complexity is bounded by $O(N_k^2)$. Computation complexity of getting the maximum value of all temporary payments is bounded by $O(N_k)$. Thus we have:

$$O(N_k) + N_k(O(N_k^2) + O(N_k)) = O(N_k^3).$$

Overall, the reverse auction can be completed in polynomial time, that is computational efficiency. ■

Theorem 5 gives the upper bound of the time complexity. However, the computation complexity also relates to the group values and group costs, which can lead to large variance in the computation complexity. After all, the computation complexity will not exceed $O(N_k^3)$.

VI. Performance Evaluation

In this section, we evaluate the performances with comparisons to baseline methods. We use the dataset from [30] which includes indoor locations and corresponding magnetometer measurements. The dataset provides Round-trip time-of-flight and magnetometer measurements at 1589 mobile node locations from 30 deployed stationary anchors in New Wing Yuan supermarket in Sunnyvale, CA. The data were collected during a period of 15 days and contained about 30000 sensing entries. To retain the repetitive locations in the dataset, we add small random noises to those locations. Due to the clustered locations in the dataset, β in Eq. (11) is assigned to 1.1 which prevents seriously deviated locations from being extended into a group, thus mitigating the information loss in location aggregation. α and γ are set to 2 and 3 in Eq. (13), in which γ represents that the group size is positively correlated to the data value but not the key factor, while α is an adaptive parameter which balances the data value and quality constraint to make the defined data value more appropriate for practical situation. Quality constraint Q and cardinality constraint N_Q are assigned 18 and 180, while users' costs are uniformly distributed in $(0, 3)$. We also utilize the diminishing return form as non-decreasing submodular function $f(\cdot)$:

$$f(W) = \lambda \ln(1 + \sum_{g_i \in W} v_i), \tag{25}$$

where λ is an adaption parameter and set to 3, whose reason is similar to α in terms of log function $f(.)$.

A. Location Aggregation

The information loss is only dependent on SSE, hence we use SSE to evaluate the location aggregation performance. For better comparison, we will show the SSE of our proposed VCLA and the other three baseline methods MDAV, V-MDAV, IP, using both real data from [30] and synthetic data. Regarding the synthetic data, users' locations are randomly distributed in a 50×50 square. The results are shown in TABLE I and II respectively, where the numbers of users are 10000, 20000, and 30000. Besides, we make the lowest SSE result bold in each case to better illustrate the algorithm comparisons.

TABLE I
Comparison of SSE With Real Data

user number	algorithm	$k=3$	$k=4$	$k=5$
10000	MDAV	2642.946	3505.837	4351.867
	V-MDAV	2439.105	3271.476	4356.011
	IP	2377.463	3134.386	3816.827
	VCLA	**2137.220**	**2975.861**	**3762.134**
20000	MDAV	4388.172	6119.494	7454.037
	V-MDAV	4117.844	5747.614	7439.348
	IP	4064.828	5420.741	6678.289
	VCLA	**3669.261**	**5063.391**	**6403.464**
30000	MDAV	5664.729	7699.375	9755.941
	V-MDAV	5261.037	7298.893	9444.971
	IP	5165.393	6888.234	8551.523
	VCLA	**4656.001**	**6348.007**	**8171.065**

TABLE II
Comparison of SSE With Synthetic Data

user number	algorithm	$k=3$	$k=4$	$k=5$
10000	MDAV	1409.491	1913.299	2416.289
	V-MDAV	1294.984	1865.387	2370.814
	IP	1303.121	1717.626	2178.939
	VCLA	**1142.731**	**1606.757**	**2064.143**
20000	MDAV	1417.705	1937.127	2405.056
	V-MDAV	1326.805	1877.956	2390.155
	IP	1307.064	1718.656	2161.395
	VCLA	**1148.575**	**1605.567**	**2039.887**
30000	MDAV	1395.474	1930.748	2454.773
	V-MDAV	1311.977	1850.126	2363.219
	IP	1290.863	1710.938	2152.349
	VCLA	**1129.970**	**1580.683**	**2042.002**

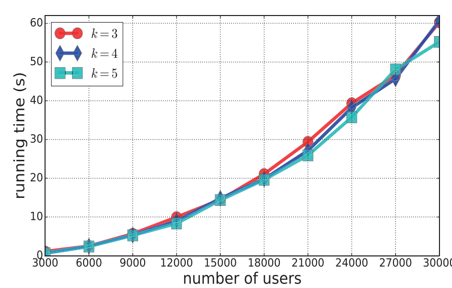


Fig. 3. Running time of VLCA.

Both results indicate that VCLA can always introduce lower information loss, which is very essential to obtain accurate sensory data. Besides, the information loss increases with k increasing, that is more stringent privacy protection will lead to more information loss. We also show the running time of VCLA using the real dataset when the number of users varies from 3000 to 30000 with an interval 3000 in Fig. 3, which reflects that the computation complexity is in polynomial time. Moreover, since the computation complexity is dominated by the location distance computation and k size group aggregation, the running time is mainly decided by the number of users, which is not much relevant to k.

Next we discuss the influence of location aggregation on sensory data processing. In the location aggregation process, we cluster the users when there are more likely near to each

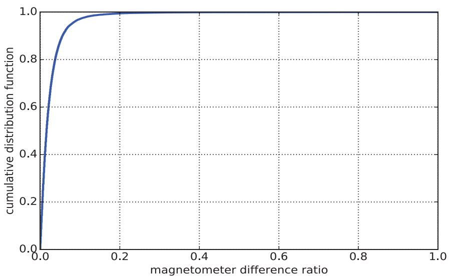


Fig. 4. Demonstration of spatial stability.

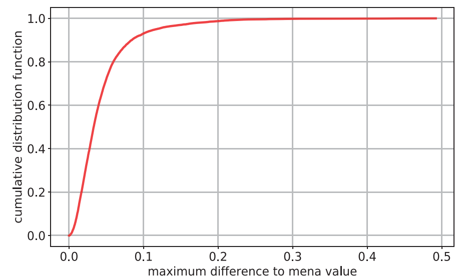


Fig. 5. Demonstration of magnetometer ratio.

other into a group. Therefore, if the sensory data of adjacent locations have little changes, which we call it spatial stability, the processed sensory data like calculating the mean values can well represent the real sensory data value of the group centroid. To this end, we use the magnetometer measurement at each location in the real dataset, where the missing magnetometer values are complemented based on those in adjacent locations. Define the magnetometer difference between two adjacent locations as:

$$diff(i, i-1) = |m(i) - m(i-1)|,$$

where $m(i)$ and $m(i-1)$ are the magnetometer measurements of two adjacent locations. Through computing the normalized magnetometer difference, we can measure the spatial stability of the sensory data:

$$\Delta diff(i, i-1) = \frac{diff(i, i-1)}{\max\{diff(i, i-1)\}}.$$

We show the cumulative distribution function of $\Delta diff(i, i-1)$ in Fig. 4. According to the figure, we obtain that the magnetometer data has good spatial stability. Furthermore, we study the magnetometer difference within a group compared to the processed magnetometer result which is the mean value in this case. For group g_i, its maximum magnetometer difference is computed:

$$maxdiff_i = \max_{j,k \in g_i} |m(j) - m(k)|.$$

Accordingly, the mean magnetometer measurement $\overline{m(i)}$ of g_i is:

$$\overline{m(i)} = \frac{\sum_{j \in g_i} m(j)}{n_i}.$$

Denote relative variation Δ_i as the ratio between $maxdiff_i$ and $\overline{m(i)}$. If Δ_i is small, the processed result has a little deviation from the real value, that is the processed result is accurate. Specifically, we consider the case $N = 30000$ and $k = 3$. We aggregate locations and calculate the Δ_i for each group. The average value of Δ_i is 0.0437 and the cumulative distribution Δ_i is displayed in Fig. 5. The results indicate that most Δ_i is less than 0.1. Beyond that, we also find that the large Δ_i is mainly caused by singular magnetometer measurements in the original dataset. Even so, the average Δ_i, equivalent to 0.0437, is still very small. Overall, the processing procedure has little influence on the final accuracy of sensory data.

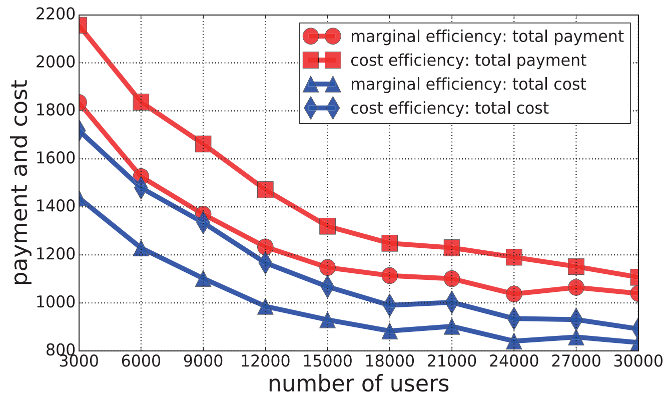


Fig. 6. Influence of user numbers.

B. Reverse Auction

We will show the results of our marginal efficiency method, and the cost efficiency method which selects winning users according to the descend cost and computes the payments similar to that of our marginal efficiency method. Analogously, the cost efficiency method also satisfies the favored economic properties. All the experiment results are obtained by using real dataset. First, we present the cost and payment influenced by the number of users and privacy protection respectively. The total costs and payments are shown in Fig. 6 when $k = 4$, which illustrates that with the increase in the number of users, the total costs and payments tend to decrease since the crowd-sensing platform has more options to select from. Besides, the cost and payment are in the same level for our marginal efficiency method which means the total payment also satisfies guaranteed approximation to the optimal total cost. Moreover, the total cost and payment of our marginal efficiency method are obviously lower than those of the cost efficiency method. Then we discuss the impact of privacy protection, namely k value. We set the number of users to 30000, and vary the k value from 2 to 6. The results are shown in Fig. 7, which indicates that the total costs and payments are growing hugely for more stringent privacy protection. The reason behind is that when k becomes large, more users are needed to form a group. According to the group bid calculation in Eq. (12), the bid will increase sharply with k, therefore the total cost and payment will grow dramatically under the same data cardinality Q.

Finally, we illustrate that the reverse auction satisfies truthfulness, individual rationality and computation efficiency. For better demonstration, we will give the results of the case where $N = 2000$ and $k = 4$. We run the algorithm for 100 times and

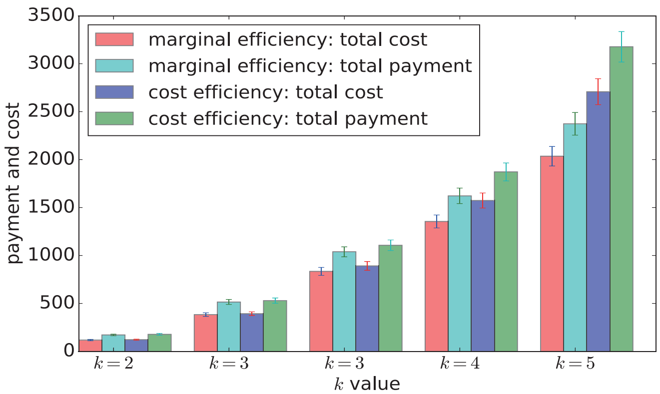


Fig. 7. Influence of privacy level.

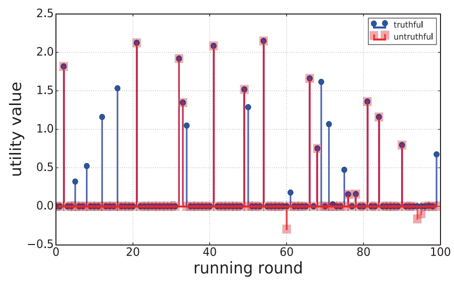


Fig. 8. Truthfulness.

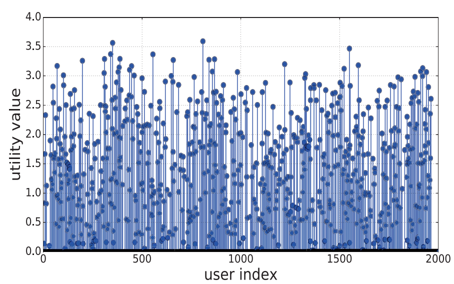


Fig. 9. Individual rationality.

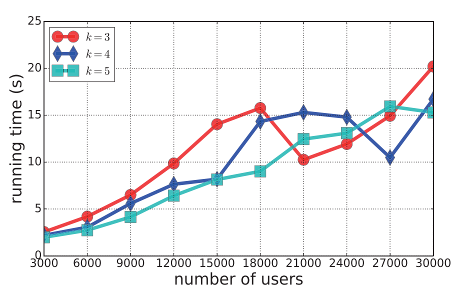


Fig. 10. Running time of reverse auction.

randomly choose an untruthful cost of a user, then compare the utilities between truthfully and untruthfully claiming cost, which are shown in Fig. 8. The figure proves that the user's utility of truthfully claiming is no less than that of untruthfully claiming, that is the reverse auction is truthful. All users' utilities are shown in Fig. 9, which reveals that users' utilities are non-negative, namely the reverse auction is individual rational. The running time of the reverse auction is drawn in Fig. 10, and we can see that the reverse auction is in polynomial computation time. Because the running time of winner selection and payment calculation are dependent on the group values, costs as well as the number of users, therefore the running time will fluctuate instead of monotonically increasing over the number of users.

VII. Conclusion

In this paper, we consider incorporating location-privacy protection into incentive mechanism design for crowdsensing system. We propose VCLA to aggregate locations for k-anonymity privacy preserving, and meanwhile mitigate the information loss. A reverse auction is designed to incentivize user participation, where the features of location-privacy preserving are also captured. The reverse auction meets a guaranteed approximation ratio and satisfies truthfulness, individual rationality, computational efficiency as well. Performances are evaluated to demonstrate the efficiency of the crowdsensing system.

References

[1] R. K. Rana, C. T. Chou, S. S. Kanhere, N. Bulusu, and W. Hu, "Ear-phone: An end-to-end participatory urban noise mapping system," in *Proc. ACM IPSN*, 2010, pp. 105–116.

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

[3] P. Mohan, V. N. Padmanabhan, and R. Ramjee, "Nericell: Rich monitoring of road and traffic conditions using mobile smartphones," in *Proc. ACM SenSys*, 2008, pp. 323–336.

[4] A. Rai, K. K. Chintalapudi, V. N. Padmanabhan, and R. Sen, "Zee: Zero-effort crowdsourcing for indoor localization," in *Proc. ACM Mobicom*, 2012, pp. 293–304.

[5] D. Yang, G. Xue, X. Fang, and J. Tang, "Crowdsourcing to smartphones: Incentive mechanism design for mobile phone sensing," in *Proc. ACM Mobicom*, 2012, pp. 173–184.

[6] Z. Feng, Y. Zhu, Q. Zhang, L. M. Ni, and A. V. Vasilakos, "TRAC: Truthful auction for location-aware collaborative sensing in mobile crowdsourcing," in *Proc. IEEE INFOCOM*, May 2014, pp. 1231–1239.

[7] X. Wang *et al.*, "A picture is worth a thousand words: Share your real-time view on the road," *IEEE Trans. Veh. Technol.*, vol. 66, no. 4, pp. 2902–2914, Apr. 2017.

[8] L. Gao, F. Hou, and J. Huang, "Providing Long-Term Participation Incentive in Participatory Sensing," in *Proc. IEEE INFOCOM*, Apr. 2015, pp. 2803–2811.

[9] Y. Wen *et al.*, "Quality-driven auction based incentive mechanism for mobile crowd sensing," *IEEE Trans. Veh. Technol.*, vol. 64, no. 9, pp. 4203–4214, Sep. 2015.

[10] D. Yang, X. Fang, and G. Xue, "Truthful incentive mechanisms for k-anonymity location privacy," in *Proc. IEEE INFOCOM*, Apr. 2013, pp. 2994–3002.

[11] H. Jin, L. Su, H. Xiao, and K. Nahrstedt, "INCEPTION: Incentivizing privacy-preserving data aggregation for mobile crowd sensing systems," in *Proc. ACM MobiHoc*, 2016, pp. 341–350.

[12] H. Jin, L. Su, B. Ding, K. Nahrstedt, and N. Borisov, "Enabling privacy-preserving incentives for mobile crowd sensing systems," in *Proc. IEEE ICDCS*, Jun. 2016, pp. 344–353.

[13] X. Jin and Y. Zhang, "Privacy-preserving crowdsourced spectrum sensing," in *Proc. IEEE INFOCOM*, Apr. 2016, pp. 2322–2330.

[14] M. Gruteser and D. Grunwald, "Anonymous usage of location-based services through spatial and temporal cloaking," in *Proc. USENIX MobiSys*, 2003, pp. 31–42.

[15] R. Shokri, G. Theodorakopoulos, P. Papadimitratos, E. Kazemi, and J. Hubaux, "hiding in the mobile crowd: Location privacy through collaboration," *IEEE Trans. Dependable Secure Comput.*, vol. 11, no. 3, pp. 266–279, Mar. 2013.

[16] I. J. Vergara-Laurens and M. A. Labrador, "Preserving privacy while reducing power consumption and information loss in LBS and participatory sensing applications," in *Proc. IEEE GLOBECOM Workshops*, Dec. 2011, pp. 1247–1252.

[17] F. Zhang, L. He, W. He, and X. Liu, "Data perturbation with state-dependent noise for participatory sensing," in *Proc. IEEE INFOCOM*, Mar. 2012, pp. 2246–2254.

[18] H. To, G. Ghinita, and C. Shahabi, "A framework for protecting worker location privacy in spatial crowdsourcing," *Proc. VLDB Endowment*, vol. 7, no. 10, pp. 919–930, Jun. 2014.

[19] B. Niu, X. Zhu, W. Li, and H. Li, "EPcloak: An efficient and privacy-preserving spatial cloaking scheme for LBSs," in *Proc. IEEE MASS*, Oct. 2014, pp. 398–406.

[20] K. Vu, R. Zheng, and J. Gao, "Efficient algorithms for K-anonymous location privacy in participatory sensing," in *Proc. IEEE INFOCOM*, Mar. 2012, pp. 2399–2407.

[21] K. L. Huang, S. S. Kanhere, and W. Hu, "Preserving privacy in participatory sensing systems," *Comput. Commun.*, vol. 33, no. 11, pp. 1266–1280, 2010.

[22] J. Domingo-Ferrer and J. M. Mateo-Sanz, "Practical data-oriented microaggregation for statistical disclosure control," *IEEE Trans. Knowl. Data Eng.*, vol. 14, no. 1, pp. 189–201, Jan./Feb. 2002.

[23] A. Solanas and A. Martinez-Balleste, "V-MDAV: A multivariate microaggregation with variable group size," in *Proc. COMPSTAT Symp. IASC*, 2006, pp. 917–925.

[24] G. Kokolakis and D. Fouskakis, "Importance partitioning in micro-aggregation," *Comput. Statist. Data Anal.*, vol. 53, no. 7, pp. 2439–2445, 2009.

[25] L. Sweeney, "K-anonymity: A model for protecting privacy," *Int. J. Uncertainty, Fuzziness Knowl.-Based Syst.*, vol. 10, no. 5, pp. 557–570, 2002.

[26] L. Golab, F. Korn, B. Saha, D. Srivastava, and F. Li, "Size-constrained weighted set cover," in *Proc. IEEE ICDE*, Apr. 2015, pp. 879–890.

[27] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher, "An analysis of approximations for maximizing submodular set functions—I," *Math. Programm.*, vol. 14, no. 1, pp. 265–294, 1978.

[28] L. A. Wolsey, "An analysis of the greedy algorithm for the submodular set covering problem," *Combinatorica*, vol. 2, no. 4, pp. 385–393, 1982.

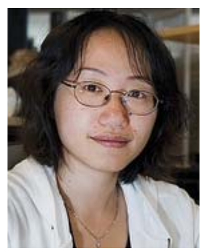
[29] R. Myerson, "Optimal auction design," *Math. Oper. Res.*, vol. 6, no. 1, pp. 58–73, 1981.

[30] A. Purohit, S. Pan, K. Chen, Z. Sun, and P. Zhang. (May 2014). *CRAWDAD Dataset CMU/Supermarket (v. 2014-05-27)*. [Online]. Available: http://crawdad.org/cmu/supermarket/20140527, doi: 10.15783/C7MW2Z.

Xiong Wang received the B.E. degree in electronic information engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2014. He is currently pursuing the Ph.D. degree in electronic engineering with Shanghai Jiao Tong University, Shanghai, China. His current research interests include crowdsourcing, network economics, and mobile computing.

Zhe Liu received the B.E. degree in telecommunication engineering from Hebei University, Baoding, China, in 2012, and the Ph.D. degree in telecommunication engineering, Xidian University, Xi'an, China, in 2017. His current research interests include the dynamic radio resource management in cognitive radio networks, big data, and information centric networks.

Xiaohua Tian received the B.E. and M.S. degrees in communication engineering from Northwestern Polytechnical University, Xian, China, in 2003 and 2006, respectively, the Ph.D. degree from the Department of Electrical and Computer Engineering, Illinois Institute of Technology, Chicago, IL, USA, in 2010. Since 2013, he has been with the Department of Electronic Engineering, Shanghai Jiao Tong University as an Assistant Professor with the title of SMC-B Scholar. He also served as a TPC Member for the IEEE INFOCOM 2014–2015, Best Demo/Poster Award Committee Member of the IEEE INFOCOM 2014, TPC Co-Chair for the IEEE ICCC 2014 International Workshop on Internet of Things, TPC Co-Chair for the 9th International Conference on Wireless Algorithms, Systems and Applications in 2014 on Next Generation Networking Symposium, Local Management Chair for the IEEE ICCC 2014, TPC member for the IEEE GLOBECOM 2011–2015 on Wireless Networking, Cognitive Radio Networks, and Ad Hoc and Sensor Networks Symposium, and TPC member for IEEE ICC 2013 and 2015 on Ad Hoc and Sensor Networks and Next Generation Networking Symposium. He serves as an Editorial Board Member on the computer science subsection of the journal *SpringerPlus*, and Guest Editor of the *International Journal of Sensor Networks*.



Xiaoying Gan received the Ph.D. degree in electronic engineering from Shanghai Jiao Tong University, China, in 2006. She is currently an Associate Professor with the Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai, China. From 2009 to 2010, she was a Visiting Researcher with the California Institute for Telecommunications and Information Technology, University of California San Diego, CA, USA. Her current research interests include network economics, social aware networks, heterogeneous cellular networks, multiuser multi-channel access, and dynamic resource management.

Yunfeng Guan received the Ph.D. degree in information engineering from Zhejiang University in 2003. He is currently an Associate Professor with the Institute of Wireless Communication Technology, Shanghai Jiao Tong University. He is mainly engaged in the study of digital TV systems and wireless access systems. He is also one of the main drafters of the China digital terrestrial television transmission standard GB20600-2006 (DTMB) and other important national DTV standards. He has developed a series of chips of DTMB demodulator, digital satellite television demodulator, HD-Decoder SoC. He has also published over 20 articles, obtained 28 patents for inventions, and finished a series of national and local major research projects. He received the First Prize of the Shanghai Science and Technology Development Award in 2005 and 2006, Second Prize of the National Science and Technology Development Award in 2008, the Shanghai Venus Award in 2009, and a Shanghai Technology Top Ten Youth Award nomination in 2010.

Xinbing Wang received the B.S. degree (Hons.) from the Department of Automation, Shanghai Jiao Tong University, Shanghai, China, in 1998, the M.S. degree from the Department of Computer Science and Technology, Tsinghua University, Beijing, China, in 2001 and the Ph.D. degree with a major in the Department of Electrical and Computer Engineering and a minor in the Department of Mathematics from North Carolina State University, Raleigh, NC, USA, in 2006. He is a currently a Professor with the Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai, China. He has been an Associate Editor for the IEEE/ACM Transactions on Networking and the IEEE Transactions on Mobile Computing, and a member of the technical program committees of several conferences including the ACM MobiCom 2012, the ACM MobiHoc 2012 and 2013, and the IEEE INFOCOM 2009–2014.