IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 28, 2018, accepted August 27, 2018, date of publication September 3, 2018, date of current version October 19, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2868250

An Effective Crowdsourcing Data Reporting Scheme to Compose Cloud-Based Services in Mobile Robotic Systems

YINGYING REN[1], WEI LIU[2], YUXIN LIU[1], NEAL N. XIONG[3], ANFENG LIU[1,4], AND XUXUN LIU[5], (Member, IEEE)

[1]School of Information Science and Engineering, Central South University, Changsha 410083, China
[2]School of Informatics, Hunan University of Chinese Medicine, Changsha 410208, China
[3]Department of Mathematics and Computer Science, Northeastern State University, Tahlequah, OK 74464, USA
[4]State Key Laboratory of Industrial Control Technology, Zhejiang University, Hangzhou 310027, China
[5]College of Electronic and Information Engineering, South China University of Technology, Guangzhou 510641, China

Corresponding author: Yuxin Liu (yuxinliu@csu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61772554, in part by the Open Research Project of the State Key Laboratory of Industrial Control Technology, Zhejiang University, China, under Grant ICT1800391, and in part by the National Basic Research Program of China (973 Program) under Grant 2014CB046345.

ABSTRACT The smart device combined with artificial intelligence can act as robot system to perform data collection task. To minimize the data collection cost and to guarantee the quality of service (QoS) of tasks are two vital issue in such mobile robot system. Data collection platform and data reporter often needs to negotiate with each other before start of data collection which will generate a certain cost. Once the platform and the data reporter agree to the cooperation, data reporter will collect and report data for a period. However, in previous researches, it was often considered that data reporters can report data at any time without considering the cost of interaction and negotiation, which is not suitable for the practice. In this paper, we propose an efficiency cost data collection scheme (ECDCS) in which the data reporter is selected according to the contribution that all the data it collects have on the whole system rather than a single data samples. Because there exists correlation in data, matrix completion technology can be adopted to recover the missing data samples with partial data while guarantee the QoS of the task. So, a data reporter selection scheme ECDCS based on the matrix completion technology is proposed in which the selection is in terms of the cooperation effect of the reporters rather than a single data sample. The main goal is to select the reporter set with low cost and high QoS which has the best cooperative effect. By doing so, in the proposed data collection scheme, the missing of partial data can be tolerated which can reduce data collection cost while guarantee the QoS. The extensive experiments results indicate that the proposed scheme can effectively reduce the data cost while maintain the QoS of application.

INDEX TERMS Mobile crowdsourcing, data collection, matrix completion technique, low cost, data samples.

I. INTRODUCTION

With the development of microelectronics technology, the processing capability of modern sensor-based devices has been developed rapidly [1]–[4]. And volume and cost of these devices have decreased significantly. These changes have led to the widespread application of sensor-based devices which greatly promoted the development of Internet of Things (IoTs) [5]–[9]. Since 2011, the number of devices (such as smartphones, mobile vehicles, industrial-aware devices, etc.) connected to the Internet of Things on the Earth has exceeded the population, reaching 9 billion. And it is estimated that by 2020, the number of devices connected to the network will reach 24 billion [10], [11]. Furthermore, recent advances in sensor have made it can be embedded into various devices of Internet of Things, which forms the infrastructure of data sensing and acquisition [12]–[15]. And a large number of mobile devices embedded in smart sensor devices such as mobile phone, mobile vehicles, autonomous automobiles and autonomous unmanned aerial vehicles combined with artificial intelligence (AI) technologies act as mobile robots

to perform massive tasks in varied environments [15]–[23]. With wireless communications, these mobile robot's system can be connected to Internet to exchange information and cooperate to perform extensive tasks [17], [18], [23]–[27]. In such tasks, these mobile smart sensor-based devices take a participatory way to sense data and report data samples to data centers in cloud through wireless communications at low cost [17], [18]. Data centers in cloud process the received data and provide the composition intelligent services to users [17], [18]. This scheme of data collection is promising for many applications. For example, the study of certain migratory birds requires data on their migration routes, time, population scale or amount, habits and rules. If only rely on the observation of the researchers to obtain these data, the system will need many researchers and must establish a large number of observation stations which will take a long time and a large number of human and material resources. And the observed data obtained in this way is often incomplete and unsystematic [28]–[32]. In the current IoT network, participatory data sensing scheme can be applied to data collection well. The researchers published the task of bird observation in cloud. The information in the published tasks includes the name of bird, the time, the scope, the report forms and the important elements and so on. In this way, a large number of mobile sensor-based devices can sense the time, location, quantity and other information of birds. And the reports can be in a variety of forms: text, sound, pictures, videos, and so on. The trajectory data of bird activity collected in this way can have longer period than the data collected by the researchers themselves. And the data is more detailed in the content. It can be considered as no boundary in the observation because the distribution of mobile devices is wide. And the distribution is flexible. In this way, the cost of data acquisition can be decreased a lot [33]–[34]. For example: the VTrack project is a typical application of big data network [10], [17], [18]. Vtrack is an application which can provide real-time traffic information. Users can obtain real-time traffic information of the city so that they can optimize the transportation to save fuel, time and other resources. NoiseTube project is an application which can provide the service that demonstrates the distribution of urban noise [10], [17], [18]. In such applications, crowdsourcing is generally used for data collection. The task publisher publishes the task of data collection so that the crowd with smart phones in the city can sense data through the embedded sensors and report the corresponding data to VTrack or NoiseTube.

The advantages of participatory sensing scheme with mobile sensor devices have aroused widespread concern of researchers. Many related studies have been proposed. The most challenge issue in these studies is how to reduce the amount of data collected while maintaining high QoS. Reducing the amount of data has different meanings in different applications, but the key point of that is to minimize the cost of data collection. One of the effective ways to reduce the cost of data collection is to reduce the data mount and select the reporters with high quality and low cost. To ensure high QoS is to guarantee that the data collected contains rich and comprehensive information. For example, data samples are needed for every monitoring object and monitoring site, which means that data should cover the whole monitoring area. This QoS dimension is called data coverage. Another QoS dimension is low redundancy in data. Redundant data cannot add new information but increases the cost of the system. The QoS dimension also includes many other aspects such as collection time of data and so on. The researchers have done some studies on these QoS dimension. For example, some researchers mainly focus on how to use participatory sensing scheme to complete the task. Research [35]–[36] discussed how to use incentive mechanism to stimulate participants to participate actively in the data collection tasks. In addition, some studies have considered the cost of data collection, the selection and optimization of reporters and QoS.

Despite these studies, there are still some issues to be studied. (1) In many studies, the basic unit of data collection is data sample. Therefore, these studies usually use QoS as a measure to select data samples to optimize the performance of the data collection scheme. Such strategies often use incentives and promote the reward of data samples in specific areas to stimulate data collection. For example: to ensure that data collected can cover the entire monitoring area, high reward is given to those areas which are difficult to obtain data. High reward will stimulate data collectors to participate in the data collection actively. The data collection strategy of multi-dimensional goal is similar. These strategies are solid and have good theoretical basis. It can be proved theoretically that the Pareto Optimality can be achieved. But in practice, once a data collector is selected to report data sample, more than one data sample will be collected. Because it takes much time and cost to determine which data sample should be collected. But the data negotiation even need more cost and time than data report. Therefore, it is not practical to take data sample as the basic unit of data collection. And this scheme is also not suitable for the mobile robots' system in this paper. Therefore, once a mobile device is selected, the mobile device will autonomously collect data for a period, and report data samples during the movement. The above practical situation shows that in the design of the data collection strategy, the data reporter should be considered as a basic unit of data collection to meet the practical needs and reduce the cost. (2) The main way to reduce the amount of data collected in many previous studies is to reduce the redundancy in data collection. But the problem in this method is: (a) Redundancy of data is difficult to avoid in practice. And the removal of redundant data requires additional cost. Thus, the effect of reducing redundant data is not satisfying. For example, in the incentive mechanism for data collection, the amount of collected data samples is large in the area where the data collector is dense even if the participatory radio of data collector is small and the reward for the data collectors is low. And the data redundancy is large. But in the remote

area, because there are few data collectors, the data collected is still not enough even at high reward. (b) The amount of data reduced by this kind of method is limited. If the monitored area is divided into $n \times n$ grids, each grid needs to collect at least one data sample. In theory, even if there is no redundancy in the data collection which is the most ideal situation, the system needs to collect n^2 data samples at least. From the view of the above situation, this paper proposes a data collection strategy utilizing the matrix completion technology. With the matrix completion technology, if the sampling matrix has a low rank character, the missing data can be recovered even if only part of data is collected. It is noted that according to the method proposed in this paper, the data samples that needed to be collected is less than n^2 in theory. In this way, the amount of data collected can be effectively reduced.

Based on the above analysis, this paper takes matrix completion technology as the theoretical foundation and takes reporter as the basic data collection unit. The optimized data reporter group which has low cost and high QoS is constructed through the combination of the trajectory data. Therefore, the proposed scheme is better than the previous data collection strategies in terms of data amount, cost and other QoS indicators. The main contributions of this paper are summarized as follows:

(1) For the application whose sampling matrix has the feature of low-rank in the crowdsourcing, this paper proposed an Efficiency Cost Data Collection Scheme (ECDCS) which takes the advantage of matrix completion technology. By using the correlation between data, it is possible to collect only a part of the data to construct a complete application. The matrix completion technology can recover the missing data through partial data, thereby effectively reducing the amount of data to be collected. ECDCS can select a set of data collectors that can meet the requirements of matrix completion technology and maximize the collaboration effect of data collectors.

(2) ECDCS can select a group of data collector with low cost which can satisfy the requirements for constructing the application. Although the application of matrix completion technology to data collection can reduce the redundancy, the selection of data collector still matters. Because the trajectories of different data collectors are different, the covered area and the expected reward is also different. Therefore, this paper proposes a reporter selection algorithm which optimize the combination of multiple reporters. Under the conditions of application construction, the proposed algorithm selects the reporter combination with lower cost and better collaboration effect. The proposed algorithm selects the data reporter according to the EC value. The EC value of a data reporter is the ratio of the reward it receives to its efficiency. After each selection, the EC value of the remaining data reporters will be updated. Thus, the combination of data reporters with better collaboration effect will be selected.

(3) Finally, we validated the effectiveness of the proposed ECDCS through extensive experiments. Compared to the traditional No Matrix Completion Data Collection Scheme (NMCDCS) which does not apply matrix completion technology, ECDCS reduces the cost by 97.53%. ECDCS reduced the cost by 56.76% and 81.94% respectively, compared to Price First Data Collection Scheme (PFDCS) and Random Selection Data Collection Scheme (RSDCS), which also used matrix completion technology.

The rest of the paper is organized as follows. We review related work in Section 2. In Section 3, we describe the sys-tem model and formulate the problem of our data collection strategy. Sections 4 present the details of Efficiency Cost Data Collection Scheme (ECDCS). We evaluate the proposed ECDCS scheme via simulations in Section 5. We conclude the paper in Section 6.

II. PRELIMINARY KNOWLEDGE AND RELATED WORK

With the development of intelligent devices, more and more data-based applications can be developed, making it possible for the implementation of smart grids, smart homes, and smart cities [37], [38]. In the reality, such big data-based application like smart city need to collet a large number of data. Sensing tasks are often very large which can't be done with a single smart device or a small number of smart devices. Although the tasks of such applications are usually large and complex, it is easy to collet a single data. Therefore, the tasks of constructing big data-based applications can be completed by the cooperation of multiple smart devices.

Crowdsourcing leverages the cooperation of multiple smart devices to accomplish complex sensing tasks that a single device cannot perform. Crowdsourcing can use incentives like money or virtual rewards to recruit a large number of mobile smart devices to perform the crowdsourcing tasks. Complex sensing tasks can be accomplished through the cooperation of a large number of smart devices.

The crowdsourcing system utilizes smart devices distributed in the city to accomplish large-scale, wide-scale sensing tasks. The system framework of crowdsourcing as shown in Figure 1 mainly consists of three parts:

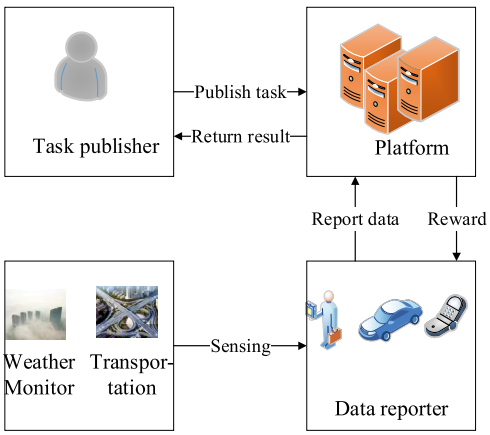


FIGURE 1. The framework of crowdsourcing.

(1) Task publisher. Task publisher can also be called application publisher or data requester. Task publisher publish the

sensing task according to the construction requirements of application. After the task publisher obtains the data through the platform of the crowdsourcing, the data will be filtered, refined and processed. The collected data will be processed to construct application that meet the market needs or specific requirements. The task publisher needs to undertake the overhead during the data collection.

(2) Platform. The task publisher only publishes the requirements of the application but don't care how to accomplish these tasks. The platform determines how to allocate sensing tasks, select the suitable data reporters to perform the tasks, and how to pay these reporters. The data reporters are paid by the platform rather than task publisher directly.

(3) Data reporter. Data reporter is the holder of the smart devices which can be also called data collector. These smart devices can be vehicles embedded with sensors, mobile phones, IPADs or other intelligent devices.

For example, to monitor air quality in different parts of the city, it is necessary to collect meteorological data from different locations in the city. However, for a single data reporter, the location that it can reach is highly restricted. A single smart device can only collect a portion of the data needed by the application. Different data reporters can reach different regions. Therefore, a high-quality application can be constructed by the collaboration of multiple data reporters. And tasks that cannot be completed by a single data reporter can be accomplished through the cooperation of multiple data reporters. Although the locations that each reporter can reach are different, the applications can be successfully built as long as the selected reporter combination can cover the areas that the application needs. Therefore, crowdsourcing can accomplish the tasks through the collaboration between multiple intelligent devices, which cannot be accomplished by a single intelligent device. Such tasks can be water pollution monitoring, noise monitoring, traffic flow detection, etc. According to the requirements of the application, different data reporter combination can be selected to perform different sensing tasks.

Although the collaboration of multiple reporters can perform large-scale sensing tasks and build high quality application, different combination of the reporters has different collaboration effects. Therefore, it is a problem worthy of study in the crowdsourcing to employ appropriate reporters to achieve the best cooperative effect according to the requirements of the application. A single smart device, which is data collector, incurs some overhead when collecting data. In order to motivate more reporters to accomplish the sensing tasks with high quality and provide the appropriate data for application stably, a certain reward will be given to the data reporters. The reward can be money, virtual credits or other forms. Through the incentive mechanism, the application can obtain stable and reliable data from the data reporters. But the incentive mechanism will also bring additional cost to the system, how to use the minimum cost to achieve maximum cooperative effect is the key content of crowdsourcing research.

Research [39] proposed two modes of crowdsourcing. The first sensing mode is platform-centric. The system first sets the reward amount, and the data reporters compete to participate in the crowdsourcing task. The Nash equilibrium is achieved by Stackelberg method to maximize the overall utility of the system. The second sensing mode is reporter-centric. The reporter-centric mode selects appropriate data reporters through auctions. The measure of cooperative effect in research [40] is whether the system can recruit sufficient number of data reporters. In the crowdsourcing model in [40], the total reward is given. Whether the publisher of the task knows that cost of the data reporter is a factor that affects the total compensation. If task publishers can obtain the cost of reporters, they can pay lower prices to recruit low cost data reporters. But in practice, it is difficult to obtain the cost of data reporters because the cost of different data reporters is different and even the devices may be different. So, it is very difficult to get the true cost of data reporters. Therefore, in practice, the reward is often presented by the data reporter rather than the application. The reward presented by the data reporter is often bigger than the true cost of the data reporter. Because different data reporters have different expectations for the reward, the proposed rewards of the reporters are different.

The scheme of crowdsourcing can also be divided into online-scheme and offline-scheme. The candidate group of data reporters in online-scheme is dynamically changing. The candidate group of data reporters in offline-scheme is unchanged. Azzam *et al.* [41] proposed a dynamic online-scheme crowdsourcing model. To meet the different requirements of the application for sensing tasks in different periods, the model continuously increases or decreases the number of data reporters. The main purpose of this model is to select a group of data reporters with higher stability. To motivate data reporters to perform the data sensing tasks better, the model utilizes Cooperative Game Theory. In this model, stability is an important factor to measure the effect of data reporters' collaboration. But, dynamic models can also bring large much additional computation. If the data reporter group is changed frequently, it will bring much unnecessary overhead. Wang *et al.* [42] proposed a combination of online mode and offline mode. In their proposed model, the candidate reporter set is first statically determined, and then the data reporter who ultimately performs the sensing tasks is dynamically determined by auction. Although online-scheme can improve the cooperation flexibility of data reporters, it will also bring a lot of additional computation cost. And the stability of the sensing task is difficult to maintain. Therefore, in the actual sensing task, the offline-scheme is widely used.

If there are multiple sensing tasks to be completed and the number of reporters is not enough, it will be impossible to accomplish all the sensing tasks. Therefore, the problem of task allocation appears [43]. Research [44] proposed the LRBA algorithm to determine the allocation of sensing tasks according to the location. At the same time, by adjusting the price of the tasks, a certain number of data reporters are

maintained in the system. Research [45] proposed an online task allocation model QASCA. QASCA takes into account the time and cost of task assignment to maximize system performance. Research [46] considers how to maximize individual benefits from the perspective of data reporters. The system can motivate data collectors to actively participate in data collection by maximizing their personal profits. In the crowdsourcing system, the problem that task assignment cannot solve is the shortage of data reporters. To solve the problem that the number of sensing tasks does not match the number of data reporters, Zhang *et al.* [47] proposed a CAPR algorithm to balance the number of sensing tasks and the number of data reporters. CAPR adjusts the system's incentives based on the ratio of sensing tasks and the number of data reporters.

Insufficient data reporters and uneven distribution of data reporters in urban areas are challenges in crowdsourcing systems. Some researches motivate more data reporters to collect data in remote areas by increasing the reward. Pu *et al.* [48] proposed a crowdsourcing model for predicting data reporters' trajectories. In their proposed crowdsourcing model, vehicles are used to collect data. This model can guarantee the cooperative effect of the data reporters for a period. However, it is still impossible to completely solve the problem of insufficient data reporters [49]. Currently, few researchers apply data recovery technology to crowdsourcing. Data recovery technology is an effective solution to reduce data redundancy and data missing in crowdsourcing. Matrix completion is a relatively mature data recovery technology. There are many mature matrix completion algorithms such as SET [50], OptSpace [51], and SVT [52]. The matrix completion technology can solve the problem of insufficient data reporters and data redundancy in the crowdsourcing. How to apply matrix completion technology to crowdsourcing is a very worthwhile research.

III. SYSTEM MODEL AND PROBLEM STATEMENT

A. SYSTEM MODEL

Some data-based applications need to collect data at specific location and time to perform computational work and data analysis. For example, in order to measure the air pollution at different location in a city during a day, relevant indicators need to be measured at different time and different location. A single mobile intelligent device has limited range of activity, and it can only collect data in a certain location within a certain period. Fortunately, the task that a single mobile intelligent device cannot accomplish can be completed by multiple mobile intelligent devices. Different mobile intelligent devices have different data sampling locations due to their different range of activities. Therefore, crowdsourcing system can accomplish a complex task through the collaboration of multiple mobile devices. When the number of devices that volunteer to participate in data collection reaches a certain amount, it can make up for the shortage of data collected by a single device. The set of data reporter can be expressed as:

$$U = \{U_1, U_2, \ldots, U_k, \ldots, U_q\}, \quad k \in [1, q]$$

Using smart devices distributed in various regions of the city, the crowdsourcing system can select appropriate data reporters from the set of candidate data reporters to complete the corresponding crowdsourcing tasks collaboratively.

To record data collected at different locations, the crowdsourcing system numbers different regions. Divide the sensing area into n grids, and number each grid, then different sampling areas can be expressed as:

$$\mathcal{A} = \{a_1, a_2, \ldots, a_i, \ldots, a_m\}$$

In addition to the location of the data, the data acquisition time is also recorded to meet the requirements of the applications like meteorological data monitoring. For example, to detect the meteorological data of the entire city, it is necessary to collect data at different times in different areas of the city to obtain meteorological data of a certain day or a certain moment. The time granularity of different application in data collection is different. Assume that the application needs to collect data within the time period $\mathcal{T}$. during this period $\mathcal{T}$, the application needs to collect data of the corresponding position at regular intervals. According to the sampling interval, the time period $\mathcal{T}$ can be divided into smaller time slots:

$$\mathcal{T} = \left\{t_1, t_2, \ldots, t_j, \ldots, t_T\right\}, \quad j \in [1, T]$$

t_i denotes the time data collected. The application expects to collect data e_i^j at location a_i time t_j. Then the data sampling matrix that the application expects to collect can be denoted as:

$$E^{m\times T} = \begin{bmatrix} e_1^1 & e_1^2 & \cdots & e_1^j & \cdots & e_1^T \\ e_2^1 & e_2^2 & \cdots & e_2^j & \cdots & e_2^T \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ e_i^1 & e_i^2 & \cdots & e_i^j & \cdots & e_i^T \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ e_m^1 & e_m^2 & \cdots & e_m^j & \cdots & e_m^T \end{bmatrix} \quad i \in [1, m], \quad j \in [1, T] \tag{1}$$

However, due to the limitations in the number and distribution of data reporters, it is difficult for applications to obtain the expected sampled data matrix. At some sampling points, there may be no data; at other sampling points, multiple duplicate data may be collected. Therefore, the amount of data that the platform actually collects at each sampling point can be expressed as:

$$N^{m\times T} = \begin{bmatrix} n_1^1 & n_1^2 & \cdots & n_1^j & \cdots & n_1^T \\ n_2^1 & n_2^2 & \cdots & n_2^j & \cdots & n_2^T \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ n_i^1 & n_i^2 & \cdots & n_i^j & \cdots & n_i^T \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ n_m^1 & n_m^2 & \cdots & n_m^j & \cdots & n_m^T \end{bmatrix} \quad i \in [1, m], \quad j \in [1, T] \tag{2}$$

In the sampling matrix, only the sampling points where the data is acquired are valid sampling points. Then the valid sampling point matrix of the sampling data matrix is:

$$X^{m\times T}=\begin{bmatrix} x_1^1 & x_1^2 & \cdots & x_1^j & \cdots & x_1^T \\ x_2^1 & x_2^2 & \cdots & x_2^j & \cdots & x_2^T \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_i^1 & x_i^2 & \cdots & x_i^j & \cdots & x_i^T \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_m^1 & x_m^2 & \cdots & x_x^j & \cdots & x_m^T \end{bmatrix} \quad i\in[1,m],\quad j\in[1,T] \tag{3}$$

in which:

$$x_i^j=\begin{cases}0, & n_i^j=0\\ 1, & n_i^j\geq 1\end{cases}$$

In the sampling data matrix, if there is no data at the sampling point, that is $n_i^j=0$, the sampling point is an invalid sampling point, which is represented as $x_i^j=0$. If there is at least one data at the sampling point, that is $n_i^j\geq 1$,then the sampling point is an valid sampling point, which is represented as $x_i^j=1$. Although data redundancy may occur at the valid sampling point, the sampling point is a valid sampling point as long as there is data.

Data reporters generate a certain amount of cost when collecting data. To increase the enthusiasm of data reporters to collect data, a certain amount of reward is given to the data reporter to increase their enthusiasm. The expected reward proposed by the data reporter is:

$$R=\{r_1,r_2,\ldots,r_q\}$$

The platform pays data reporters according to R to motivate them to participate in data collection.

But even if there is certain reward to motivate the data reporters, individual data reporters are not always able to submit data at sampling time due to activity time and environment limitation such as signal strength, power and other resources. The data submitted by the data reporter U_k can be expressed as:

$$D_k=\begin{bmatrix} d_1^1 & d_1^2 & \cdots & d_1^j & \cdots & d_1^T \\ d_2^1 & d_2^2 & \cdots & d_2^j & \cdots & d_2^T \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ d_i^1 & d_i^2 & \cdots & d_i^j & \cdots & d_i^T \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ d_m^1 & d_m^2 & \cdots & d_m^j & \cdots & d_m^T \end{bmatrix} \quad i\in[1,m],\quad j\in[1,T] \tag{4}$$

in which, $d_i^j=0\ or 1$.

$d_i^j=1$ indicates that the data reporter U_k can collect data e_i^j at location a_i and time a_i; $d_i^j=0$ indicates that the data reporter U_k cannot collect data e_i^j at location a_i and time a_i.

At each sampling time, a data reporter can only submit a data at one location, which means:

$$\sum_{i=1}^{m} d_i^j\leq 1 \tag{5}$$

B. PROBLEM STATEMENTS

When collecting data, the platform hopes to reduce the cost as much as possible. Data reporter will have expenses of time, energy, electric power etc. To obtain stable data with high quality, a certain reward is paid to motivate data reporter.

Assure the candidate data reporters set volunteered to participate in the data collection is:

$$U=\{U_1,U_2,\ldots,U_k,\ldots,U_q\}$$

The reward set proposed by the data reporter is:

$$R=\{r_1,r_2,\ldots,r_k,\ldots,r_q\}$$

The proposed reward of different data reporter is different. Because even at the same sampling time and location, the cost of different data reporters may still be different. In some remote area which data reporters are difficult to reach, due to the weak signal strength or other environment limitations, high data costs may be generated when collecting data. In the same sampling grid, the cost may differ due to the data collection time. For example, even in areas where data reporters are densely distributed, if data collection is done at late night, data reporters will expect higher rewards. In addition, different data reporters have different expectations for reward. Therefore, the expected rewards of reporters with the same cost may be different. The platform hopes to select the lowest-cost, best-collected collection of data collectors that meet the minimum requirements for building applications. The platform hopes to select the reporter group with low cost and good cooperative effect. If the platform selects p data reporters to participate in the data collection, the selected data reporter set is:

$$W=\left\{W_{(1)},W_{(2)},\ldots,W_{(h)},\ldots,W_{(p)}\right\},$$

The corresponding reward of the selected data reporters is:

$$R=\{r_{(1)},r_{(2)},\ldots,r_{(h)},\ldots,r_{(p)}\}$$

The total payment the system need to pay can be calculated as

$$P=\sum_{h=1}^{p} r_{(h)} \tag{6}$$

The first goal of the platform is to minimize the data collection cost of the system:

$$\min(P)=\min(\sum_{h=1}^{p} r_{(h)}) \tag{7}$$

However, while reducing the cost of data collection, the platform must ensure the cooperative effect of the data reporter group, that is, to ensure the coverage of the sampling data matrix. In the most ideal case, all the sampling points in

the matrix is valid. That is, for each $x_i^j, i \in [1, m], j \in [1, T]$, there is:

$$x_i^j = 1$$

And thus, the ideal sampling data matrix is:

$$X_{ideal}^{m\times T} = \begin{bmatrix} 1 & 1 & \cdots & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 & \cdots & 1 \end{bmatrix} \tag{8}$$

But in fact, it is difficult to collect data for all sampling points. For example, in some remote areas, data reporters may only arrive and collect data at specific time. Even in the center of the city, there may be no data reporters at late nigh. Therefore, in the sampling data matrix, only part of the sampling points are valid sampling points. From the valid sampling point matrix of the platform, the number of valid sampling points can be obtained as follows:

$$A = \sum_{i=1}^{m} \sum_{j=1}^{T} x_i^j \tag{9}$$

Then the coverage of the sampling matrix is:

$$C = \frac{A}{m \times T} = \frac{\sum_{i=1}^{m} \sum_{j=1}^{T} x_i^j}{m \times T} \tag{10}$$

The higher the coverage of the sampling matrix, the higher the quality of the application. Therefore, to improve the service quality of the application, another goal of the platform is to increase the coverage of the sampling matrix as much as possible, which can be denoted as:

$$\max(C) = \max(\frac{\sum_{i=1}^{m} \sum_{j=1}^{T} x_i^j}{m \times T})$$

Therefore, how to select data reporters, to reduce the cost of the platform as much as possible, and to improve the coverage to ensure the service quality of the application, is the main research goal of this paper:

$$\begin{cases} \min(P) = \min\left(\sum_{h=1}^{p} r_{(h)}\right) \\ \max(C) = \max(\frac{\sum_{i=1}^{m} \sum_{j=1}^{T} x_i^j}{m \times T}) \end{cases}$$

IV. SCHEME DESIGN

To state the parameter of this paper clearly, the main notions introduced in this paper can be found in table 1.

A. MOTIVATION

(1) In the data collection, due to environment limitations, there are some sampling points where data cannot be collected. Even if the platform continues to increase the reward, there are still some sampling points where data cannot be collected. These sampling points may not be able to have data due to the remoteness of the location or the particularity of the sampling time. For example, in the fine-grained observation of long-term migratory patterns of migratory birds, due to the limitations of the traditional technology and costs, large-scale, long-term observations of wide areas (spatial areas spanning 3,000 kilometers) are difficult to obtain accurately first-hand information. Such application is difficult to raise the coverage of the sampling matrix by the incentive.

TABLE 1. Parameter description.

Parameter	State
$\mathbb{W}$	The selected set of data collector.
U_k	The k-th data collector.
$\mathbb{N}$	The collected data amount of U_k.
A	The least data amount to meet the requirement of matrix completion technology
F_k	The efficiency of data collector U_k
EC_k	The Efficiency Cost of data collector U_k.
r_k	The repeated data set of data collector U_p.
$\mathcal{r}$	The rank of the sampling matrix
$\boldsymbol{U}_{(z)}$	The z-th data collector in the sorted set.
R_P	The repeated data set of data collector U_p
U_s	The sorted data collector set

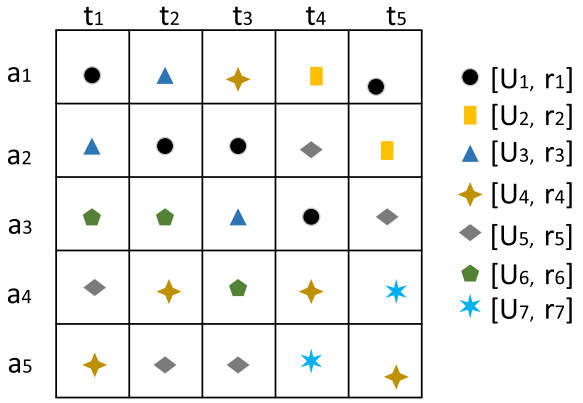


FIGURE 2. The ideal status of traditional data collection scheme.

In the traditional data collection scheme, to obtain the ideal sampling matrix, the application has to collect at least m × T data as shown in Figure 2. Due to the limitation of the environment, the actual data collection status of the traditional scheme is shown in Figure 3 in which the shadow grids are the unideal sampling points. In these sampling points, part of the sampling points doesn't have any data, while some sampling points collect multiple data. The redundant data will bring additional cost to the platform. And the points without data will lead to the failure of application construction. Therefore, how to break through the limitations of the traditional data collection scheme is the key to further reduce costs.

The key to breaking through the limitation of traditional data collection scheme is to reduce the data amount. The application of matrix completion technology to data

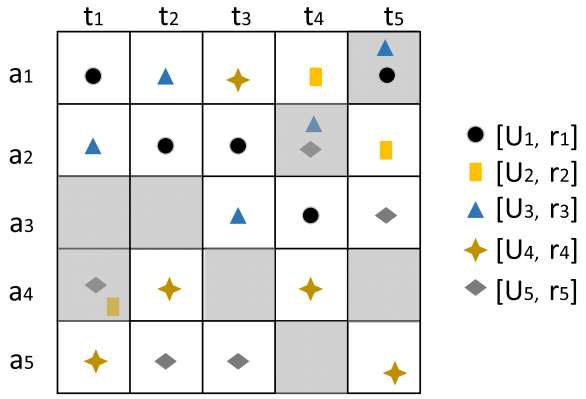


FIGURE 3. The actual data collection situation of traditional scheme.

collection can compensate the problems of traditional data collection scheme in this respect. Matrix completion technology is a relatively mature data recovery technology, which can recover the missing data through partial data in the matrix. In matrix with low rank, a certain correlation exists in the data. And many data-based applications have similar properties. For example, when performing meteorological monitoring, the change of data often has a certain regularity in the same period. Therefore, the matrix completion technology can be used to restore a complete sampling matrix, so that the lack of partial data in the sampling matrix can be tolerated. The data collection scheme after the application of matrix technology is shown in Figure 4. To construct a complete sampling data matrix, only part of data reporters need to be selected. However, the application of matrix completion technology has certain conditions. The sampling data matrix shown in Figure 5 does not satisfy the application conditions of the matrix completion.

Matrix completion technology has two conditions for sampling data matrix:

a. Each row and each column of the sampling matrix must have data, which is:

$$\begin{cases} \sum_{i=1}^{n} x_i^j \neq 0, & \forall j \in [1, T] \\ \sum_{t=1}^{T} x_i^j \neq 0, & \forall i \in [1, m] \end{cases} \tag{11}$$

b. The number of valid sampling points in the sampling data matrix must bigger than the lowest value, which is:

$$A = \sum_{i=1}^{m} \sum_{j=1}^{T} x_i^j > Ch^{\frac{6}{5}} r \log h, \quad h = \max\{m, T\} \tag{12}$$

where r is the rank of the matrix and C is a constant.

The data collection situation shown in Figure 5 does not meet the first condition that each row and each column of sampling matrix must have data. Therefore, how to select suitable data reporters to collect data efficiently while satisfying the conditions is a key issue to be solved.

(2) How to select the appropriate data reporter set to reduce the overhead and maximize the collaboration effect of data

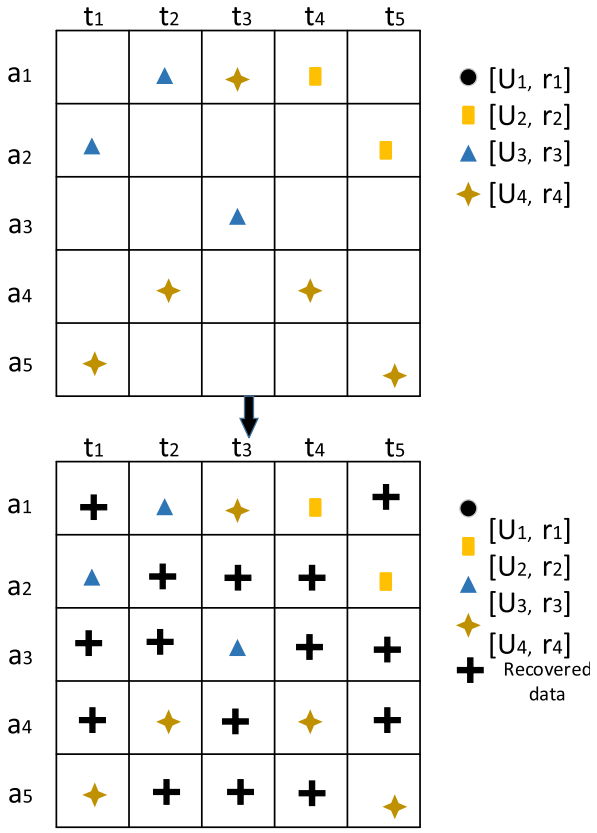


FIGURE 4. The data collection scheme after the application of matrix completion technology.

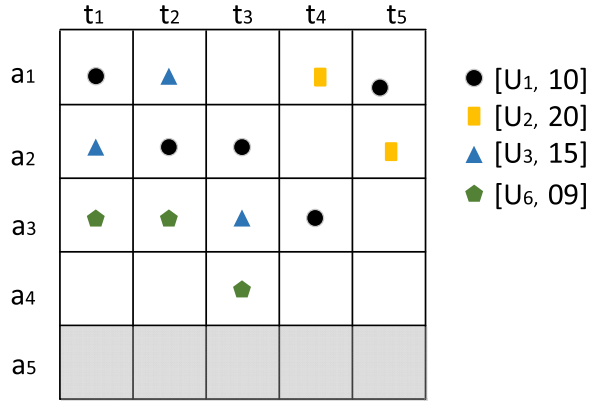


FIGURE 5. The data collection situation which doesn't meet the conditions of matrix completion.

reporter set is also a key issue that the platform needs to solve when collecting data. Different data reporters have different behavior patterns. For example, when taking a smart phone to collect data, the device holder acts as a data reporter and its behavioral pattern has a critical impact on the results of data collection. Some data reporters are very active and can report data to the system frequently during the sampling time. During the sampling time $\mathcal{T}$, the data reporter may be unable to submit data to the system at each unit sampling time. Therefore, for different data reporters, the amount of data

collected during the sampling time is different. In addition to the difference in the amount of data, different data reporters have different expectations for the reward.

The overhead generated by different data reporters when collecting a single data is different. Due to the type of the device, the strength of the signal, the difference in configuration, etc., the cost of data reporters at the same sampling point may be different. And different data reporters have different expectations for the reward. Some data reporters have lower expectations for the reward so low reward can meet their expectation. Some data reporters have higher expectations for the reward so high reward can meet their expectation. From the perspective of the platform, the cost incurred during data collection is expected to be as lower as possible. The platform expects to select data reporters with more data at low cost.

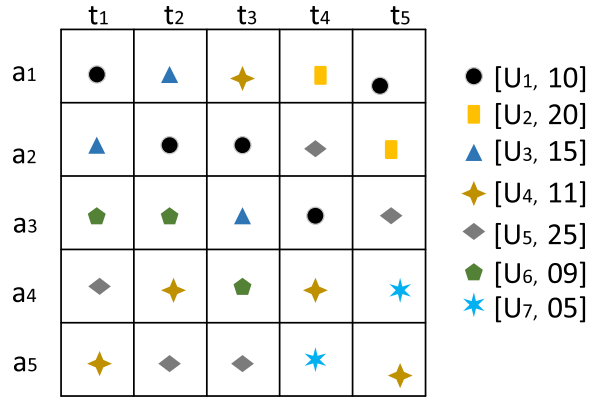


FIGURE 6. The ideal data collection pattern of traditional scheme.

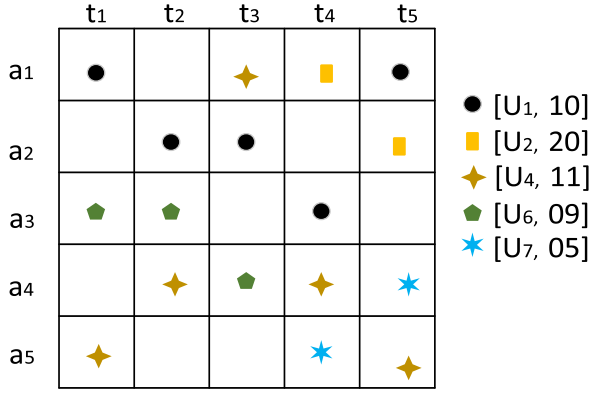


FIGURE 7. Data collection situation 1 utilizing matrix completion.

Figure 6 shows the ideal data collection situation of traditional scheme. The number in the bracket represent the rewards of the selected data reporter. Figure 7, and Figure 8 are possible data collection situations applying matrix completion technology. Compared to the traditional data collection scheme, the application of matrix completion reduced the cost of the system but the cost of different data reporter groups is quite different. The total payment of the

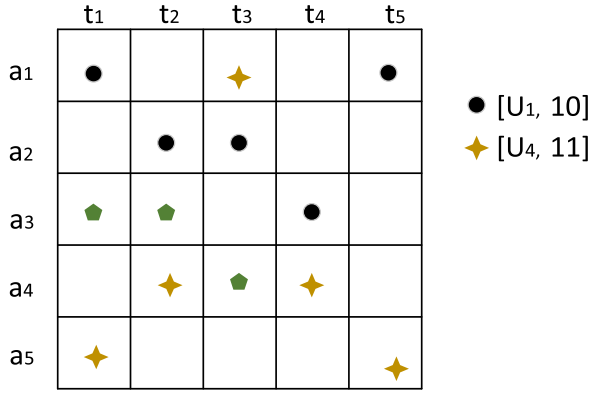


FIGURE 8. Data collection situation 3 utilizing matrix completion.

traditional scheme shown in Figure 6 is 95, while the total payment shown in Figure 7, and Figure 8 which utilize the matrix completion technology is 55, 21. Therefore, how to establish an effective data reporter selection method is very important for sensing tasks.

B. ECDCS

1) EFFICIENCY COST

Different smart devices have different reporting frequencies and locations. And data reporters have different expectations for the reward. The platform hopes to reduce the cost as much as possible while meeting the basic requirements of constructing applications.

In the IoT network, the selection of data reporter is very important. The reward proposed by the data reporter may not truly reflect the actual workload and the contribution to the application. Therefore, if only considering the reward which is the cost of the system, it is not always possible to achieve good data collection effects. For example, as shown in Figure 1, assume that the platform selects data reporters according to the reward proposed by the data reporter. The platform will select the reporter with the lowest expected reward according to the sorted set. Sort all the data reporters in Figure 1 according to the reward:

$$\{U_7, U_6, U_1, U_4, U_3, U_2, U_5\}$$

Then U_7, U_6 and U_1 will be selected. After the selection of U_1, the conditions of matrix completion will be satisfied. The platform need to pay 24 to the data reporters selected in this way. However, the payment of the situation shown in Figure 8 is only 21. There may also be data collection situation with lower cost. Therefore, the reward can be the selection standard.

For data reporter U_k, the data amount it can collect can be denoted as:

$$\mathbb{N} = \sum_{j=1}^{T} \sum_{i=1}^{m} d_i^j \tag{13}$$

At each sampling time, a data reporter can only submit a data at one location, which means:

$$\sum_{i=1}^{m} d_i^j \leq 1$$

Therefore, for a single data reporter, the maximum amount of data collected during the sampling period is T. And thus, for a single data reporter U_k, the data collection efficiency can be defined as:

$$F_k = \frac{\mathbb{N}}{T} = \frac{\sum_{j=1}^{T}\sum_{i=1}^{m} d_i^j}{T} \in [0, 1] \tag{14}$$

Data collection efficiency reflects the activeness of data reporters. If the data reporter U_k submits data at each unit sampling time, the efficiency of U_k is 1 which means U_k is an active data reporter. The cost of different data reporters is different, and it may produce higher costs when take the efficiency as the only standard. The system needs a selection standard to evaluate data reporters in terms of efficiency and overhead. And thus, Efficiency Cost (EC) is proposed to evaluate the value of each data reporter. The EC value of a data reporter is the ratio of the reward it receives to its efficiency. For a single data reporter U_k, the value of EC can be denoted as:

$$EC_k = \frac{r_k}{F_k} \tag{15}$$

When defining data collection efficiency, F_k is used as denominator rather than r_k because $F_k \in [0, 1]$ which can magnify the difference of different data reporters.

Therefore, the application can select suitable data reporters according the EC. The smaller the EC value of the data reporter, the less overhead it will generate when collecting single data. Sort all the data reporters according to the EC. The sorted data reporter set is:

$$U_s = \{U_{(1)}, U_{(2)}, \ldots, U_{(k)}, \ldots, U_{(q)}\}$$

For example, the EC value of data reporters in Figure 6 is calculated in Table 2.

TABLE 2. The EC value of data reporters in Figure 6.

	U_1	U_2	U_3	U_4	U_5	U_6	U_7
r_k	10	20	15	11	25	09	05
F_k	1	0.4	0.6	1	1	0.6	0.4
EC	10	50	25	11	25	15	12.5

Sort all the data reporters in Figure 6 according to the value of EC. The sorted set is:

$$\{U_1, U_4, U_7, U_6, U_3, U_5, U_2\}$$

According to value of EC, the data reporter with smaller EC value is preferentially selected to perform data collection. Data reporter U_1 and U_4 will be selected. After the selection of reporter U_4, the condition of matrix completion will be satisfied. And the selection will stop.

2) EC UPDATING

The situation shown in Figure 6 is ideal where data collection range of each data reporter does not coincide with each other. However, in the practice, multiple data reporters often collect redundant data in the same sampling point while some data sampling points have no data as shown in Figure 9.

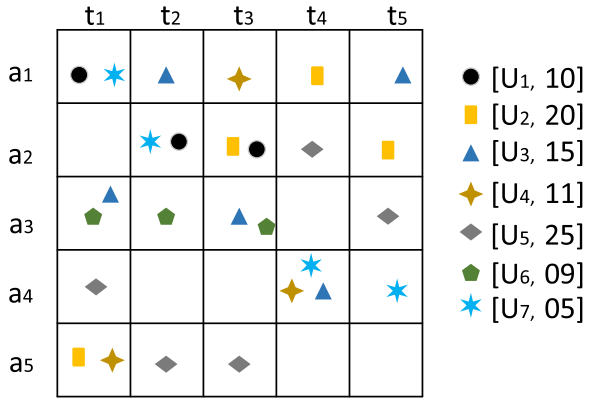


FIGURE 9. The unideal data collection situation.

TABLE 3. The EC value of data reporters in Figure 10.

	U_1	U_2	U_3	U_4	U_5	U_6	U_7
r_k	10	20	15	11	25	09	05
F_k	0.6	0.8	1	0.6	1	0.6	0.8
EC	16.7	25	15	18.3	25	15	6.25

The EC value in Figure 9 is calculated in Table 3. According to the value of EC, U_7, U_3, U_6 and U_4 will be selected. From Figure 9, it can be observed that U_7, U_3, U_6 have higher repeatability between data which means the selected group has higher redundancy. Therefore, in order to reduce the redundancy, each time a data reporter is selected, the EC values of all remaining data reporters are recalculated. The key to recalculating the EC value is to recalculate the data reporter's efficiency. For the data reporter U_p in the remaining data reporter set, if data it collects have duplicated part R_P with the selected data reporters, the efficiency of data reporter U_p will be recalculated as:

$$F_k = \frac{(\sum_{j=1}^{T}\sum_{i=1}^{m} d_i^j) - |R_P|}{T} \in [0, 1] \tag{16}$$

Therefore, the data redundancy will be reduced. In Figure 9, U_7 is first selected. The updated EC value of remaining data reporter is shown in Table 4. After the value of is updated, U_6 will be selected. And after the selection of U_6, the updated EC value of remaining data reporter is shown in Table 5. And from Table 5, U_2 will be selected. And data reporters finally selected are U_7, U_6 and U_2.

The main steps of ECDCS algorithm based on EC value can be expressed as:

TABLE 4. The EC value of all the data reporters after the first updating.

	U_1	U_2	U_3	U_4	U_5	U_6	U_7
r_k	10	20	15	11	25	09	05
F_k	0.2	0.8	0.8	0.4	1	0.6	0.8
EC	50	25	18.75	27.5	25	15	6.25

TABLE 5. The EC value of all the data reporters after the second updating.

	U_1	U_2	U_3	U_4	U_5	U_6	U_7
r_k	10	20	15	11	25	09	05
F_k	0.2	0.8	0.4	0.4	1	0.6	0.8
EC	50	25	37.5	27.5	25	15	6.25

Step 1: First, the data amount $\mathbb{N}$ that data reporter U_k collects during sampling time T is calculated. And the efficiency F_k of data reporter U_k is also calculated.

Step 2: Calculate the value of EC according to the payment and efficiency of data reporter U_k.

Step 3: Sort all the data reporter according to the value of EC.

Step 4: Select data reporter with smallest EC value and add it to the set W.

Step 5: ECDCS judges whether the selected data reporters could satisfy the basic conditions of matrix completion:

(1) Each row and each column must have data:

$$\begin{cases} \sum_{i=1}^{n} x_i^j \neq 0, & \forall j \in [1, T] \\ \sum_{t=1}^{T} x_i^j \neq 0, & \forall i \in [1, m] \end{cases}$$

(2) The amount of data sampling points must satisfy the condition:

$$A = \sum_{i=1}^{m} \sum_{j=1}^{T} x_i^j > Ch^{\frac{6}{5}} r \log h, \quad h = \max\{m, T\}$$

If the selected data reporter group can satisfy the two conditions above, then the algorithm comes to an end. And the selected data reporter set W will be output. If the conditions are not satisfied, Step 4 will be executed.

Step 6: Recalculate the EC value of the remaining data reporters and go back to the Step3.

The details of the ECDCS is shown in Algorithm 1.

After the selection of data reporter, the complete data sampling matrix can be recovered by the mature matrix completion technology such as OptSpace [51], SVT [52], SET [50].

Algorithm 1: Efficiency Cost Based Data Collection Scheme (ECDCS)

Input: Reported trajectory data by the data collector
Output: Selected data collector set $\mathbb{W}$
$\mathbb{W} = \emptyset$ //to initialize the selected data collector set
For each data collector $U_k, k \in [1, m]$
  For i=1 to m
    For j=1 to T
  // if data collector collect data in the location i at time t
    If $d_i^j = 1$
  // To compute the data amount collected by the data collector
      $\mathbb{N} = \mathbb{N} + 1$
    End
  End for
 End for
End for
For each U_k in $U, k \in [1, m]$
  // to compute the data collection rate of each data collector
  $\boldsymbol{F_k} = \frac{\mathbb{N}}{\boldsymbol{T}}$
  // to compute the EC value for each data collector
  $EC_k = \frac{r_k}{F_k}$
End for
  // to get the ordered data collector set
$U_s =$ Sort U according to EC in ascending order
While $(\sum_{i=1}^{m} x_i^t = 0, \forall t \in [1, T]$
    $\| \sum_{j=1}^{T} x_i^t = 0, \forall i \in [1, n]$
    $\|A < Ch^{\frac{6}{5}} r \log h, h = \max\{m, T\}$
// the collected data amount is less than the requested
    &&$|\mathbb{W}| < |U|$)
    z=1;
  // jump the selected participant
  While $U_{(z)}$ is in $\mathbb{W}$ && z $< |U|$
    z=z+1
  End While
  $\mathbb{W} = \mathbb{W} \bigcup \{\boldsymbol{U}_{(z)}\}$
  // to update the contribution degree of each participant
  For each U_k in $U, k \in [1, m]$
  $\boldsymbol{F_k} = \frac{(\sum_{j=1}^{T} \sum_{i=1}^{m} \boldsymbol{d_i^j}) - |\boldsymbol{R_P}|}{\boldsymbol{T}} \in [\mathbf{0}, \mathbf{1}]$
  $\boldsymbol{EC_k} = \frac{\boldsymbol{r_k}}{\boldsymbol{F_k}}$
  End For
  // to get the ordered data collector set
$U_s =$ Sort U according to EC in ascending order
End While
Return $\mathbb{W}$

V. EXPERIMENTAL STUDY

To verify the effectiveness of ECDCS, we use different data reporter group and compare it with three other schemes: Price First Data Collection Scheme (PFDCS), Random Selection Data Collection Scheme (RSDCS) and No Matrix Completion Data Collection Scheme (NMCDCS).

The PFDCS is based on the matrix completion technology but only considers cost of data collection and data reporter with low expected reward is selected firstly. PFDCS doesn't take the coverage into account. RSDCS also apply the matrix completion technology and use traditional way to select data reporter randomly. NMCDC does not apply the matrix

completion and selects data reporters randomly which is the traditional data collection way.

A. THE EVALUATION OF DATA COLLECTOR

In the experiment, assume the amount of data sampling is100, and the number of sampling time slots is also 100. So, the sampling data matrix is: $E^{100\times100}$. Figures 10, 12, and 14 are data statistics of all data reporters at each sampling point when the number of data reporters is 200, 500, and 1000, respectively. The application needs one data in the sampling point to construct the application. Therefore, it can be seen from the three figures that the total data amount of each sampling point of the candidate reporter group is sufficient. And the distribution of candidate data reporters is evenly distributed. As the number of data reporters increases, the amount of data in each sampling point increases which means the application have bigger space for the choice. But it also means that if there is no reasonable selection strategy, redundancy will be generated.

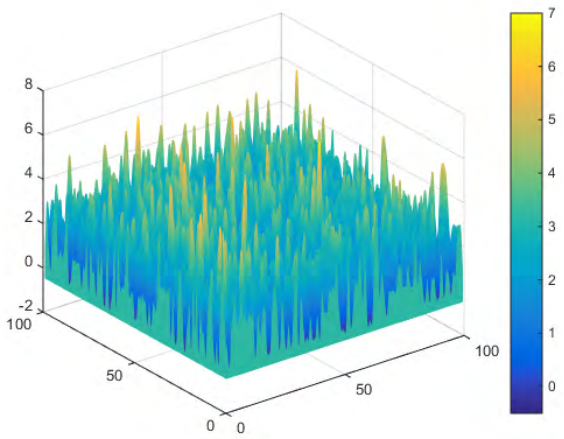


FIGURE 10. The data amount of each sampling point when amount of data reporter is 200.

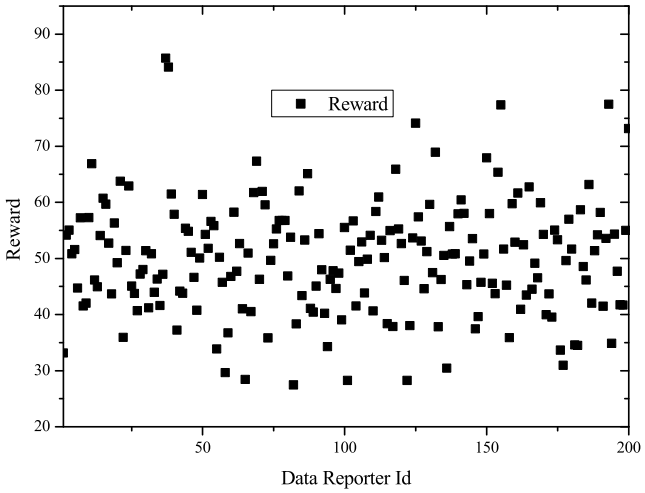


FIGURE 11. The reward distribution when reporter amount is 200.

Figures 11, 13, and 15 are the reward distributions of three data reporter groups respectively. The reward proposed by

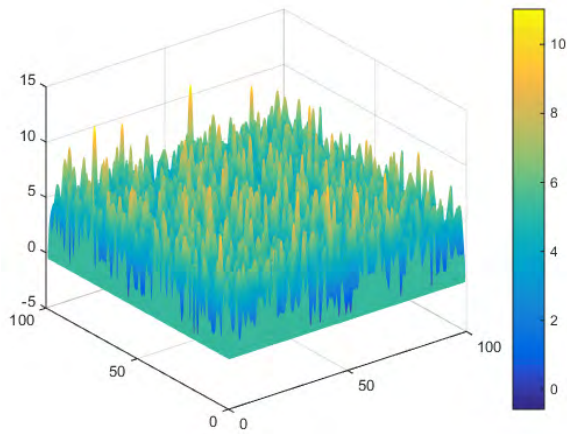


FIGURE 12. The data amount of each sampling point when amount of data reporter is 500.

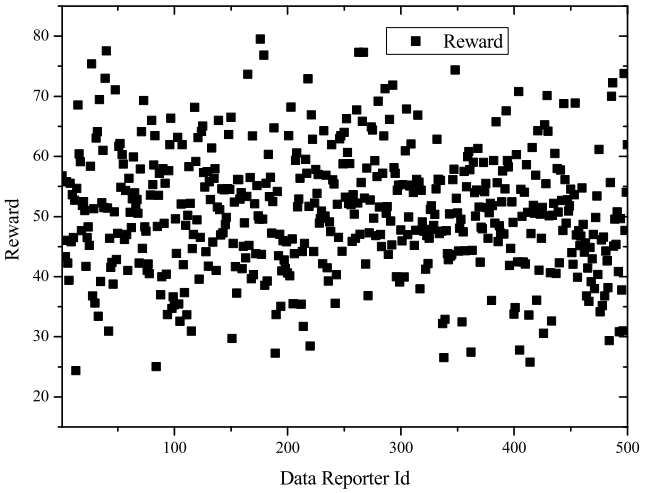


FIGURE 13. The reward distribution when reporter amount is 500.

data reporters is concentrated in the interval [30], [80] and follows a normal distribution. High cost data reporters and low cost data reporters are both a minority in the three data reporter sets.

Figure 16 is part of the trajectory data of five reporters which are selected from 200 data reporters randomly. The figure shows the trajectory data of the first ten time slots. From the figure, it can be seen that different data reporter collects different data due to the difference in the trajectory. Moreover, it can be seen from the data reporter labeled 5 that the data reporter does not necessarily submit data at all unit sampling time. Therefore, different data reporters have different efficiency.

Different reporters may have different data efficiency due to the different data amount they collect. Figure 17 shows the efficiency of different data reporters when the number of data reporters is 200. The efficiency of most data reporters is concentrated in the interval [0.3, 0.9].

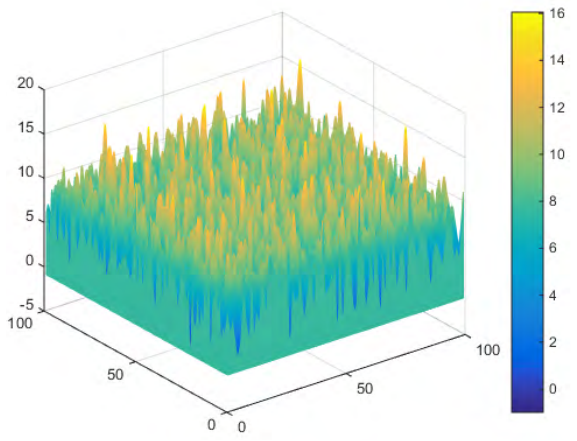


FIGURE 14. The data amount of each sampling point when amount of data reporter is 1000.

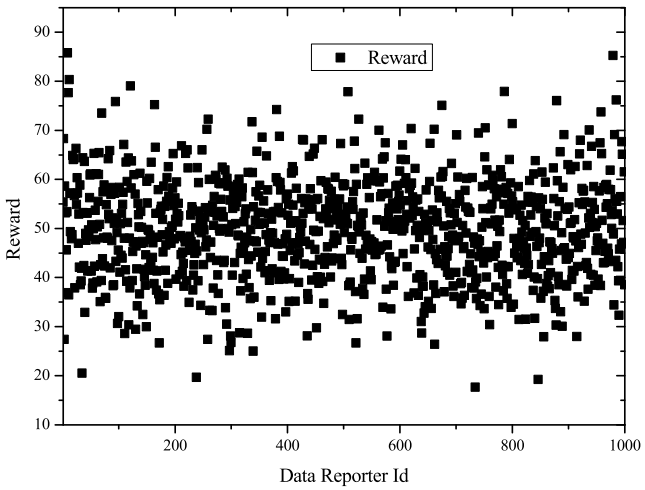


FIGURE 15. The reward distribution when reporter amount is 1000.

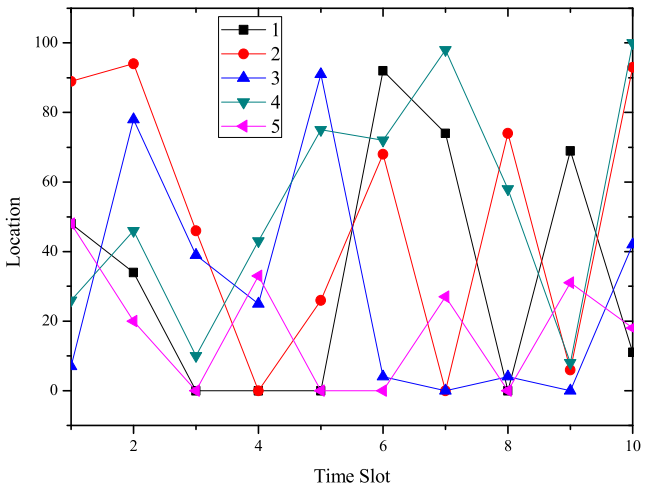


FIGURE 16. Part trajectory data of five data reporters.

B. THE EVALUATION OF ECDCS

The data distribution of data reporters selected by ECDCS is shown in Figure 18. From Figure 18 it can be seen that not all the sampling points have data. Because ECDCS can use

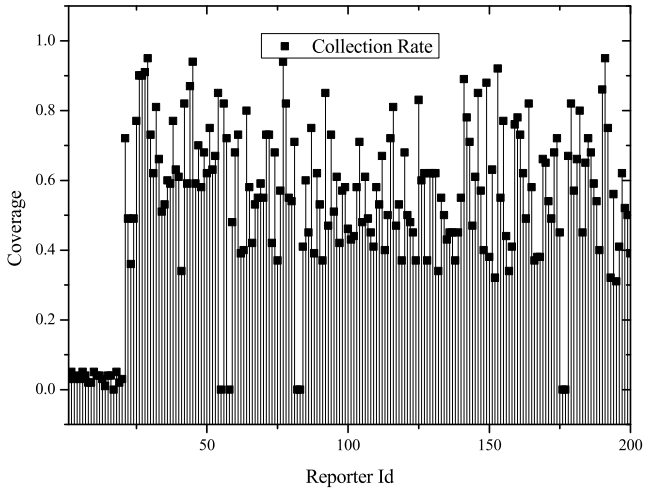


FIGURE 17. The efficiency of data reporter when data reporter amount is 200.

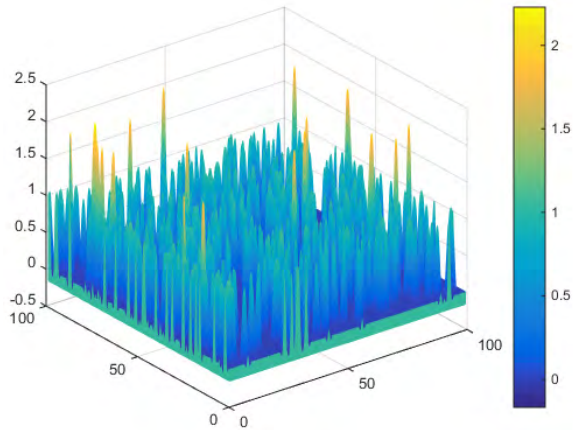


FIGURE 18. Data distribution of data reporters selected by ECDCS when the amount of data reporters is 200.

matrix completion technology to recover the missing data, it can reduce the overall data amount by collecting part of the data, thereby reducing costs. As can be seen from the Figure 18, the most data amount of the data sampling point is 1. Figure 19 is the distribution of data reporters selected by the NMCDC. As can be seen from the figure, most of the sampling points have 2-3 data which means there is much data redundancy. In addition to a single sampling point, the amount of data collected by NMCDC is greater than that of ECDCS, and NMCDC needs to gather data at almost all the sampling points to meet the basic requirements of application. ECDCS only needs to collect data at some sampling points to meet the construction needs of the application.

In order to apply the matrix completion technology to reduce the cost, the collected data amount must satisfy the condition:

$$A = \sum_{i=1}^{m} \sum_{j=1}^{T} x_i^j > Ch^{\frac{6}{5}} r \log h, \quad h = \max\{m, T\}$$

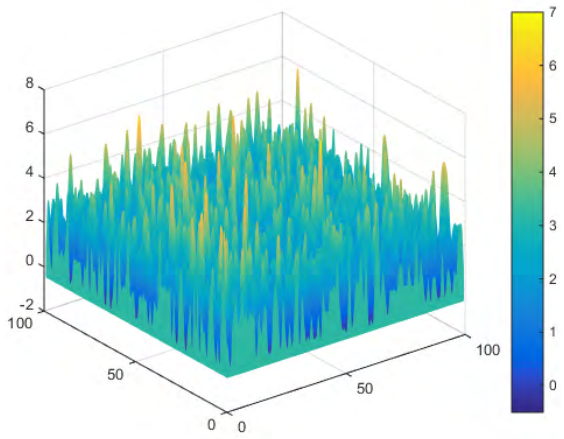


FIGURE 19. Data distribution of data reporters selected by NMCDC when the amount of data reporters is 200.

In the experiment, the scale of sampling matrix used is 100×100 and the constant $C = 1$. And thus the collected data amount of the sampling matrix must satisfy:

$$A = \sum_{i=1}^{m} \sum_{j=1}^{T} x_i^j > 503$$

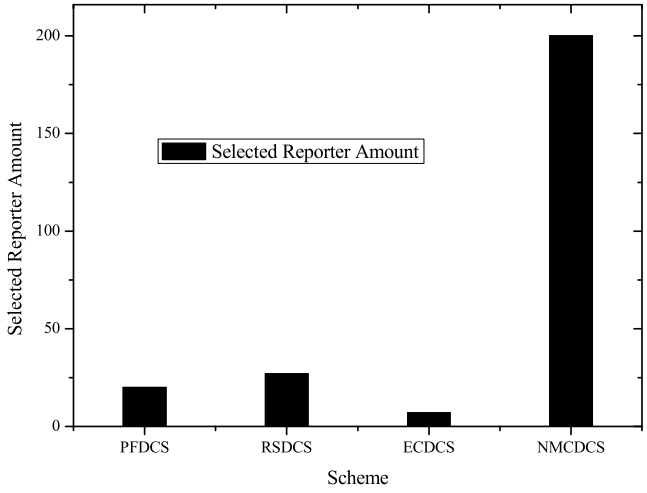


FIGURE 20. The amount of selected data reporters, when data reporter amount = 200.

Figure 20 is the comparison of the amount of selected data reporters selected by the four schemes when the number of candidate data reporters is 200. Among them, PFDCS, RSDCS and ECDCS are data collection schemes based on matrix completion technology. NMCDCS is a traditional data collection scheme without matrix completion technology. As can be seen from the figure, NMCDCS without matrix completion technology needs to select all data reporters to satisfy the requirement of constructing applications. ECDCS based on the EC value needs the least data reporters.

Figure 21 is the comparison of the four schemes. The overhead of the four schemes is similar to that of the amount of

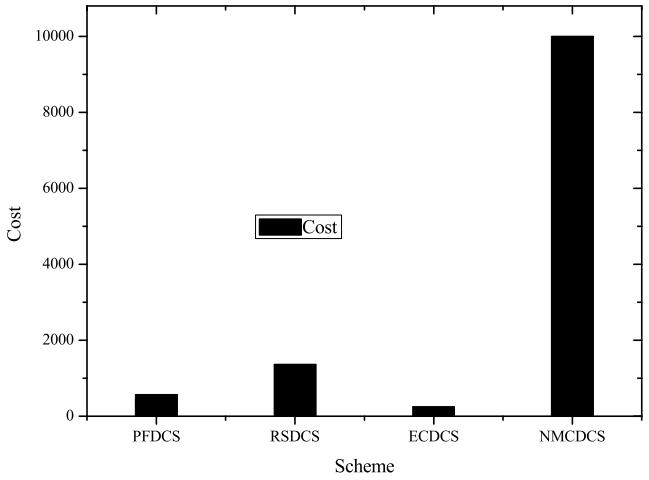


FIGURE 21. The cost of the fours schemes, when data reporter amount = 200.

data reporters. The cost of NMCDCS strategy without matrix completion technology is the highest and the cost of ECDCS is the lowest. Compared with the traditional NMCDCS, ECDCS reduces the overhead by 97.53%. Compared with PFDCS and RSDCS, which also use matrix completion technology, ECDCS reduces the overhead of 56.76% and 81.94% respectively.

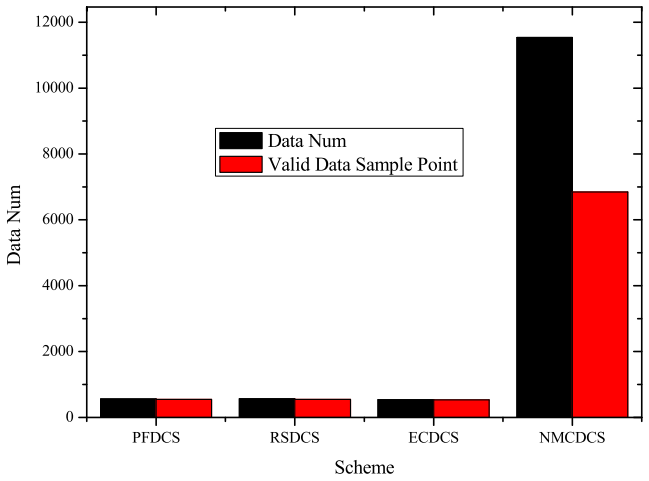


FIGURE 22. The comparison between the collected data amount and the number of valid sampling points of the four strategies when data reporter amount = 200.

Figure 22 is a comparison between the collected data amount and the number of valid sampling points of the four strategies. It can be seen from the figure that NMCDCS has the maximum redundancy both in data amount and valid sampling points. The collected data amount of ECDCS is the least.

Figures 23 and 24 show the comparison of the amount of data collected and the total cost of the PFDCS, RSDCS and ECDCS schemes when the number of data reporters is 200 as the lower limit of collected data amount changes.

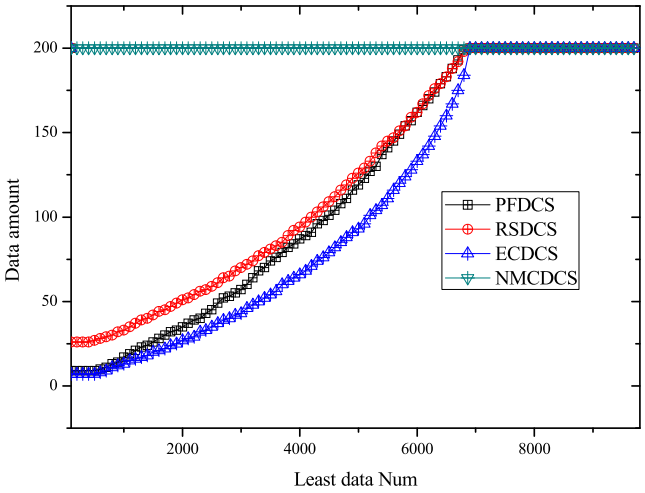


FIGURE 23. Comparison of the amount of data collected as the lower limit of data amount changes when data reporter amount is 200.

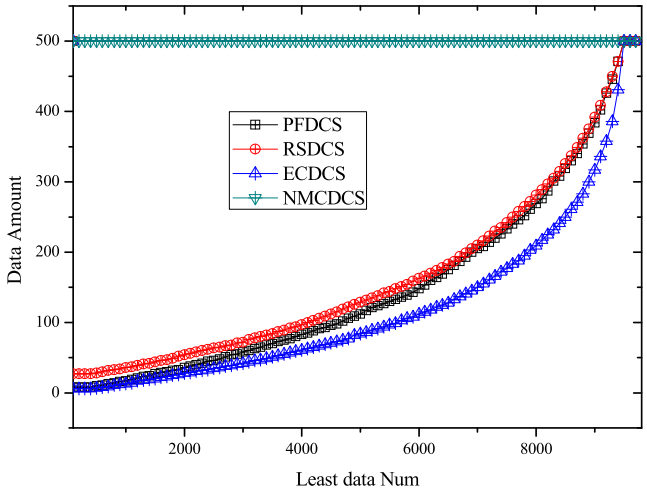


FIGURE 25. Comparison of the amount of data collected as the lower limit of data amount changes when data reporter amount = 500.

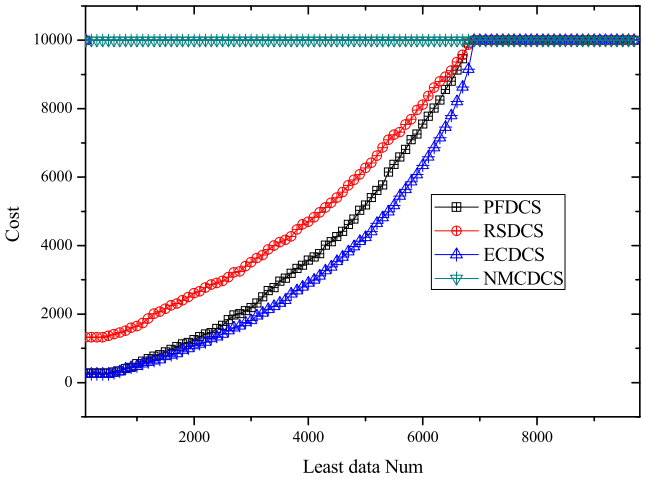


FIGURE 24. Comparison of the cost as the lower limit of data amount changes when data reporter amount is 200.

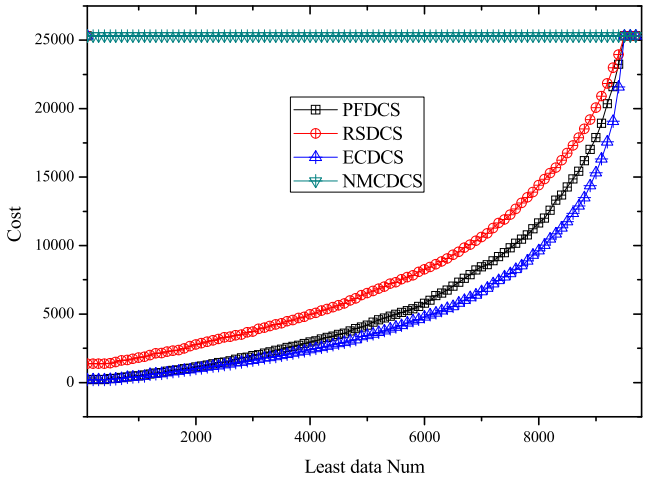


FIGURE 26. Comparison of the cost as the lower limit of data amount changes when data reporter amount is 500.

Since NMCDCS does not use matrix completion technology, its target data amount is 100×100, and no change is required. As can be seen from Figure 23, the lower limit of the number of data is changed, and the amount of data that ECDCS collected is the least before the amount of data exceeds the boundary of the system. Figure 24 illustrates that the total cost of ECDCS is also the lowest. When the amount of data required reaches a certain level, all the four schemes need to select all the data reporters to meet the need of construction requirements.

Figures 23 and 24 show the comparison of the amount of data collected and the total cost of the PFDCS, RSDCS and ECDCS schemes when the number of data reporters is 500 as the lower limit of collected data amount changes. As can be seen from Figure 25, as the lower limit of the number of data changed, the amount of data that ECDCS collected is the least. Figure 26 illustrates that the total cost of ECDCS is also the lowest. When the amount of data required reaches a certain level, all the four schemes need to select all the data reporters to meet the need of construction requirements. But compared to the number of data reporters is 200, when the number of data reporters is 500, the application has more space to choose. When the number of data reporters is 200, and the needed minimum amount of data is about 7000, the boundary of the system has been reached. Four strategies require all data reporters to participate in order to meet application requirements.

Figures 27 and 28 show the comparison of the amount of data collected and the total cost of the PFDCS, RSDCS and ECDCS schemes when the number of data reporters is 1000 as the lower limit of collected data amount changes. As can be seen from Figure 27 and Figure 28, as the lower limit of the number of data changed, the amount of data that ECDCS collected and the cost of the ECDCS is the least. As the number of data reporters increases, the gap between ECDCS and the other three schemes will be reduced.

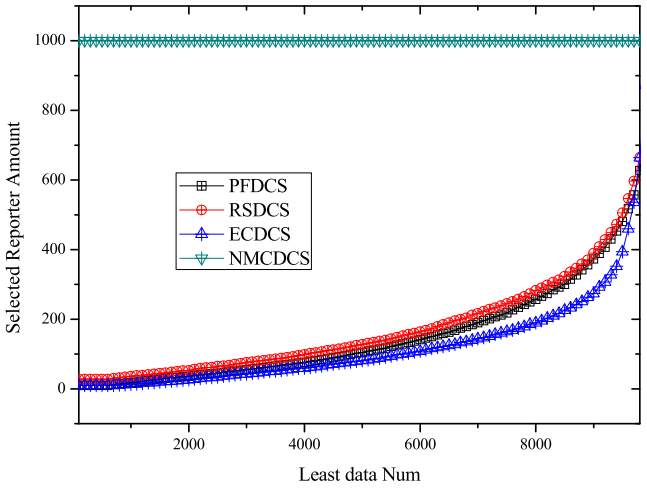


FIGURE 27. Comparison of the amount of data collected as the lower limit of data amount changes when data reporter amount = 1000.

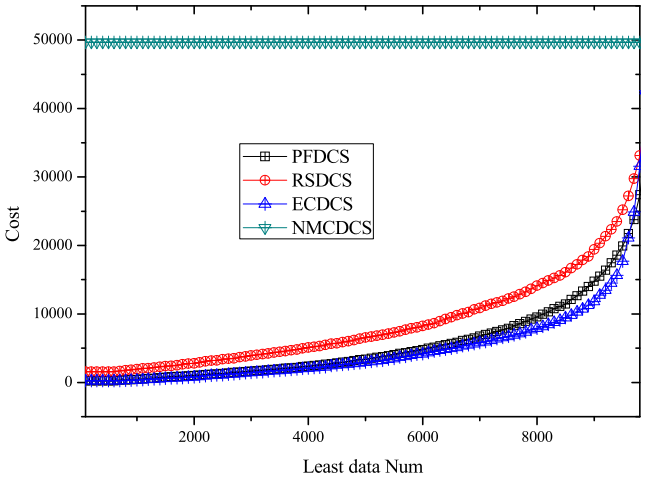


FIGURE 28. Comparison of the cost as the lower limit of data amount changes when data reporter amount is 1000.

When the number of data reporters is small, the advantage of ECDCS strategy is more obvious.

VI. CONCLUSION

The development of smart devices has enabled more and more data-based applications to be developed. The smart devices change the traditional data collection mode. Complex applications based on big data usually have a huge number of tasks. It is difficult to build complex applications with a single or a small number of smart devices. Fortunately, the collaboration of a large number of smart devices can greatly reduce the difficulty of application construction. Many researchers have conducted relevant research on this data collection scheme but there are still some problems.

First, there may be a large amount of redundant data generated during the actual data collection process. These redundant data will bring additional overhead to the system. Second, there may be no data in some remote areas. How to recover missing data while reducing data redundancy is a key issue that needs to be addressed.

Utilizing the matrix completion technology, this paper takes reporter as the basic data collection unit. And EC value is proposed to select suitable reporters to participate in the data collection. Matrix completion technology can recover the missing data with partial data in the sampling matrix. The proposed EDCDS can select data reporter group which can maximize the cooperative effect while satisfy the requirements of the matrix completion.

REFERENCES

[1] X. Hu *et al.*, "Emotion-aware cognitive system in multi-channel cognitive radio ad hoc networks," *IEEE Commun. Mag.*, vol. 56, no. 4, pp. 180–187, Apr. 2018.
[2] L. Guo, Z. Ning, W. Hou, B. Hu, and P. Guo, "Quick answer for big data in sharing economy: Innovative computer architecture design facilitating optimal service-demand matching," *IEEE Trans. Autom. Sci. Eng.*, to be published, doi: 10.1109/TASE.2018.2838340.
[3] Z. Ning, J. Huang, and X. Wang, "Vehicular fog computing: Enabling real-time traffic management for smart cities," *IEEE Wireless Commun.*, to be published, doi: 10.1109/MWC.2018.1700441.
[4] M. Wu *et al.*, "An effective delay reduction approach through a portion of nodes with a larger duty cycle for industrial WSNs," *Sensors*, vol. 18, no. 5, p. 1535, 2018, doi: 10.3390/s18051535.
[5] Z. Ning, X. Kong, F. Xia, W. Hou, and X. Wang, "Green and sustainable cloud of things: Enabling collaborative edge computing," *IEEE Commun. Mag.*, to be published, doi: 10.1109/MCOM.2018.1700895.
[6] X. Liu, M. Dong, Y. Liu, A. Liu, and N. Xiong, "Construction low complexity and low delay CDS for big data code dissemination," *Complexity*, vol. 2018, Jun. 2018, Art no. 5429546, doi: 10.1155/2018/5429546.2018.
[7] X. Liu *et al.*, "Construction of large-scale low-cost delivery infrastructure using vehicular networks," *IEEE Access*, vol. 6, no. 1, pp. 21482–21497, 2018.
[8] M. Z. A. Bhuiyan, G. Wang, J. Wu, J. Cao, X. Liu, and T. Wang, "Dependable structural health monitoring using wireless sensor networks," *IEEE Trans. Depend. Sec. Comput.*, vol. 14, no. 4, pp. 363–376, Jul./Aug. 2017.
[9] X. Wang *et al.*, "A city-wide real-time traffic management system: Enabling crowdsensing in social Internet of vehicles," *IEEE Commun. Mag.*, to be published, doi: 10.1109/MCOM.2018.1701065.2018.
[10] S. Sarkar, S. Chatterjee, and S. Misra, "Assessment of the suitability of fog computing in the context of Internet of Things," *IEEE Trans. Cloud Comput.*, vol. 6, no. 1, pp. 46–59, Jan./Mar. 2018.
[11] Cisco. *Internet of Things Market Forecast*. Accessed: May 8, 2018. [Online]. Available: http://postscapes.com/internet-of-things-market-size
[12] X. Liu, G. Li, S. Zhang, and A. Liu, "Big program code dissemination scheme for emergency software-define wireless sensor networks," *Peer-to-Peer Netw. Appl.*, vol. 11, no. 5, pp. 1038–1059, 2018.
[13] X. Xu *et al.*, "A cross-layer optimized opportunistic routing scheme for loss-and-delay sensitive WSNs," *Sensors*, vol. 18, no. 5, p. 1422, 2018, doi: 10.3390/s18051422.
[14] X. Liu, M. Dong, K. Ota, L. T. Yang, and A. Liu, "Trace malicious source to guarantee cyber security for mass monitor critical infrastructure," *J. Comput. Syst. Sci.*, vol. 98, pp. 1–26, Dec. 2018, doi: 10.1016/j.jcss.2016.09.008.2016.
[15] Z. Li, Y. Liu, M. Ma, A. Liu, X. Zhang, and G. Luo, "MSDG: A novel green data gathering scheme for wireless sensor networks," *Comput. Netw.*, vol. 142, no. 4, pp. 223–239, 2018.
[16] Z. Ding *et al.*, "Orchestrating data as services based computing and communication model for information-centric Internet of Things," *IEEE Access*, vol. 6, pp. 38900–38920, 2018, doi: 10.1109/ACCESS.2018.2853134.2018.
[17] H. Teng *et al.*, "Adaptive transmission range based topology control scheme for fast and reliable data collection," *Wireless Commun. Mobile Comput.*, vol. 2018, Jul. 2018, Art. no. 4172049, doi: 10.1155/2018/4172049.2018.
[18] Y. Ren, Y. Liu, N. Zhang, A. Liu, N. N. Xiong, and Z. Cai, "Minimum-cost mobile crowdsourcing with QoS guarantee using matrix completion technique," *Pervasive Mobile Comput.*, vol. 49, pp. 23–44, Sep. 2018, doi: 10.1016/j.pmcj.2018.06.012.2018.

[19] T. Li, Y. Liu, N. N. Xiong, A. Liu, Z. Cai, and H. Song, "Privacy-preserving protocol for sink node location in telemedicine networks," *IEEE Access*, vol. 6, pp. 42886–42903, 2018, doi: 10.1109/ACCESS.2018.2858274.2018.

[20] X. Li *et al.*, "Differentiated data aggregation routing scheme for energy conserving and delay sensitive wireless sensor networks," *Sensors*, vol. 18, no. 7, p. 2349, 2018, doi: 10.3390/s18072349

[21] Q. Liu and A. Liu, "On the hybrid using of unicast-broadcast in wireless sensor networks," *Comput. Elect. Eng.*, to be published, doi: 10.1016/j.compeleceng. 2017.03.004.

[22] M. Huang *et al.*, "A services routing based caching scheme for cloud assisted CRNs," *IEEE Access*, vol. 6, no. 1, pp. 15787–15805, 2018.

[23] S. Yu, X. Liu, A. Liu, N. Xiong, Z. Cai, and T. Wang, "An adaption broadcast radius-based code dissemination scheme for low energy wireless sensor networks," *Sensors*, vol. 18, no. 5, p. 1509, 2018, doi: 10.3390/s18051509.

[24] J. Li, Y. Zhang, X. Chen, and Y. Xiang, "Secure attribute-based data sharing for resource-limited users in cloud computing," *Comput. Secur.*, vol. 72, pp. 1–12, Jan. 2018, doi: 10.1016/j.cose.2017.08.007.

[25] A. Liu and S. Zhao, "High performance target tracking scheme with low prediction precision requirement in WSNs," *Int. J. Ad Hoc Ubiquitous Comput.*, to be published.

[26] J. Li, J. Li, X. Chen, C. Jia, and W. Lou, "Identity-based encryption with outsourced revocation in cloud computing," *IEEE Trans. Comput.*, vol. 64, no. 2, pp. 425–437, Feb. 2015.

[27] B. Huang, A. Liu, C. Zhang, N. Xiong, Z. Zeng, and Z. Cai, "Caching joint shortcut routing to improve quality of service for information-centric networking," *Sensors*, vol. 18, no. 6, p. 1750, 2018, doi: 10.3390/s18061750.

[28] Q. Liu, Y. Guo, J. Wu, and G. Wang, "Effective query grouping strategy in clouds," *J. Comput. Sci. Technol.*, vol. 32, no. 6, pp. 1231–1249, Nov. 2017.

[29] X. Chen, J. Li, J. Weng, J. Ma, and W. Lou, "Verifiable computation over large database with incremental updates," *IEEE Trans. Comput.*, vol. 65, no. 10, pp. 3184–3195, Oct. 2016.

[30] X. Wang *et al.*, "A privacy-preserving message forwarding framework for opportunistic cloud of things," *IEEE Internet Things J.*, to be published, doi: 10.1109/JIOT.2018.2864782.2018.

[31] J. Li, Y. K. Li, X. Chen, P. P. C. Lee, and W. Lou, "A hybrid cloud approach for secure authorized deduplication," *IEEE Trans. Parallel Distrib. Syst.*, vol. 26, no. 5, pp. 1206–1216, May 2015.

[32] W. Jiang, G. Wang, M. Z. A. Bhuiyan, and J. Wu, "Understanding graph-based trust evaluation in online social networks: Methodologies and challenges," *ACM Comput. Surv.*, vol. 49, no. 1, 2016, Art. no. 10.

[33] X. Wang, Z. Ning, and L. Wang, "Offloading in Internet of vehicles: A fog-enabled real-time traffic management system," *IEEE Trans. Ind. Informat.*, to be published, doi: 10.1109/TII.2018.2816590.2018.

[34] M. Z. A. Bhuiyan, J. Wu, G. Wang, T. Wang, and M. M. Hassan, "e-sampling: Event-sensitive autonomous adaptive sensing and low-cost monitoring in networked sensing systems," *ACM Trans. Auton. Adapt. Syst.*, vol. 12, no. 1, 2017, Art. no. 1.

[35] S. Fang *et al.*, "Feature selection method based on class discriminative degree for intelligent medical diagnosis," *Comput., Mater. Continua*, vol. 55, no. 3, pp. 419–433, 2018.

[36] Q. Liu, G. Wang, X. Liu, T. Peng, and J. Wu, "Achieving reliable and secure services in cloud computing environments," *Comput. Elect. Eng.*, vol. 59, pp. 153–164, Apr. 2017.

[37] X. Liu and P. Zhang, "Data drainage: A novel load balancing strategy for wireless sensor networks," *IEEE Commun. Lett.*, vol. 22, no. 1, pp. 125–128, Jan. 2018.

[38] X. Liu, "A novel transmission range adjustment strategy for energy hole avoiding in wireless sensor networks," *J. Netw. Comput. Appl.*, vol. 67, pp. 43–52, May 2016.

[39] 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.

[40] L. Duan, T. Kubo, K. Sugiyama, J. Huang, T. Hasegawa, and J. Walrand, "Incentive mechanisms for smartphone collaboration in data acquisition and distributed computing," in *Proc. INFOCOM*, Mar. 2012, pp. 1701–1709.

[41] R. Azzam, R. Mizouni, H. Otrok, S. Singh, and A. Ouali, "A stability-based group recruitment system for continuous mobile crowd sensing," *Comput. Commun.*, vol. 119, pp. 1–14, Apr. 2018.

[42] Y. Wang, Z. Cai, G. Yin, Y. Gao, X. Tong, and G. Wu, "An incentive mechanism with privacy protection in mobile crowdsourcing systems," *Comput. Netw.*, vol. 102, pp. 157–171, Jun. 2016.

[43] W. Gong, B. Zhang, and C. Li, "Task assignment in mobile crowdsensing: Present and future directions," *IEEE Netw.*, vol. 32, no. 4, pp. 100–107, Jul./Aug. 2018.

[44] S. He, D.-H. Shin, J. Zhang, and J. Chen, "Toward optimal allocation of location dependent tasks in crowdsensing," in *Proc. INFOCOM*, Apr./Mar. 2014, pp. 745–753.

[45] Y. Zheng, J. Wang, G. Li, R. Cheng, and J. Feng, "QASCA: A quality-aware task assignment system for crowdsourcing applications," in *Proc. ACM SIGMOD Int. Conf. Manage. Data.*, 2015, pp. 1031–1046.

[46] W. Wang, H. Gao, C Liu, and K. K. Leung, "Credible and energy-aware participant selection with limited task budget for mobile crowd sensing," *Ad Hoc Netw.*, vol. 43, pp. 56–70, Jun. 2016.

[47] H. Zhang, Z. Xu, X. Du, Z. Zhou, and J. Shi, "CAPR: Context-aware participant recruitment mechanism in mobile crowdsourcing," *Wireless Commun. Mobile Comput.*, vol. 16, no. 15, pp. 2179–2193, 2016.

[48] L. Pu, X. Chen, J. Xu, and X. Fu, "Crowdlet: Optimal worker recruitment for self-organized mobile crowdsourcing," in *Proc. 35th Annu. IEEE Int. Conf. Comput. Commun. (INFOCOM)*, Apr. 2016, pp. 1–9.

[49] X. Xu, N. Zhang, H. Song, A. Liu, M. Zhao, and Z. Zeng, "Adaptive beaconing based MAC protocol for sensor based wearable system," *IEEE Access*, vol. 6, pp. 29700–29714, 2018.

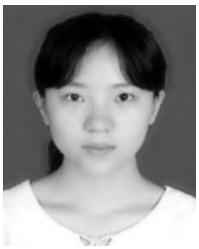
[50] A. Antonić, M. Marjanović, K. Pripužić, and I. P. Žarko, "A mobile crowd sensing ecosystem enabled by CUPUS: Cloud-based publish/subscribe middleware for the Internet of Things," *Future Generat. Comput. Syst.*, vol. 56, pp. 607–622, Mar. 2016.

[51] W. Dai, E. Milenkovic, and E. Kerman, "Subspace evolution and transfer (SET) for low-rank matrix completion," *IEEE Trans. Signal Process.*, vol. 59, no. 7, pp. 3120–3132, Jul. 2011.

[52] J.-F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM J. Optim.*, vol. 20, no. 4, pp. 1956–1982, 2010.

YINGYING REN is currently pursuing the master's degree with the School of Information Science and Engineering, Central South University, China. Her research interests include services-based network, crowd sensing networks, and wireless sensor networks.

WEI LIU received the Ph.D. degree in computer application technology from Central South University, China, in 2014. He is currently an Associate Professor and a Senior Engineer with the School of Informatics, Hunan University of Chinese Medicine, China. He has published over 20 papers in the related fields. His research interests include software engineering, data mining and medical informatics.



YUXIN LIU is currently pursuing the degree with the School of Information Science and Engineering, Central South University, China. Her research interests are services-based networks, crowd sensing networks, and wireless sensor networks.

NEAL N. XIONG received the Ph.D. degrees with Wuhan University (about sensor system engineering), and the Japan Advanced Institute of Science and Technology, respectively. Before he attends Northeastern State University, OK, USA, he was with Georgia State University, Wentworth Technology Institution, and Colorado Technical University about 10 years. He is currently an Associate Professor with the Department of Mathematics and Computer Science, Northeastern State University. His research interests include cloud computing, security and dependability, parallel and distributed computing, networks, and optimization theory.

He published over 200 international journal papers and over 100 international conference papers. Some of his works were published in IEEE JSAC, IEEE or ACM Transactions, ACM Sigcomm Workshop, IEEE INFOCOM, ICDCS, and IPDPS. He is a Senior Member of the IEEE Computer Society. He received the Best Paper Award in the 10th IEEE International Conference on High Performance Computing and Communications in 2008 and the Best student Paper Award in the 28th North American Fuzzy Information Processing Society Annual Conference in 2009. He has been a general chair, program chair, publicity chair, PC member, and OC member over 100 international conferences, and as a reviewer of about 100 international journals, including the IEEE JSAC, the IEEE SMC (Park: A/B/C), the IEEE TRANSACTIONS ON COMMUNICATIONS, the IEEE TRANSACTIONS ON MOBILE COMPUTING, the IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS. He is serving as an editor-in-chief, an associate editor, or an editor member for over 10 international journals (including an Associate Editor for the IEEE TRANSCTIONS ON SYSTEMS, MAN AND CYBERNETICS: SYSTEMS, an Associate Editor for *Information Science*, an Editor-in-Chief for the *Journal of Internet Technology*, and an Editor-in-Chief for the *Journal of Parallel & Cloud Computing*), and a guest editor for over 10 international journals, including *Sensor* Journal, WINET, and MONET.

ANFENG LIU received the M.Sc. and Ph.D. degrees in computer science from Central South University, China, in 2002 and 2005, respectively. He is currently a Professor of the School of Information Science and Engineering, Central South University. His major research interest is wireless sensor networks. He is also a member (E200012141M) of the China Computer Federation.

XUXUN LIU (M'14) received the Ph.D. degree in communication and information system from Wuhan University, Wuhan, China, in 2007. He is currently an Associate Professor with the School of Electronic and Information Engineering, South China University of Technology, Guangzhou, China. He has authored or co-authored over 30 scientific papers in international journals and conference proceedings. His current research interests include wireless sensor networks, wireless communications, computational intelligence, and mobile computing. His research has been supported by the National Natural Science Foundation of China for three times. He serves as an Associate Editor of the IEEE ACCESS, and as a workshop chair, a publication chair, or a TPC member of a number of conferences. He has served as a reviewer of over 30 journals, including 10 IEEE journals and five Elsevier journals.