Collect Spatiotemporally Correlated Data in IoT Networks With an Energy-Constrained UAV

Wenzheng Xu, *Member, IEEE*, Heng Shao, Qunli Shen, Jian Peng, Wen Huang, Weifa Liang, *Senior Member, IEEE*, Tang Liu, *Member, IEEE*, Xin-Wei Yao, *Senior Member, IEEE*, Tao Lin, and Sajal K. Das, *Fellow, IEEE*

***Abstract*—Unmanned aerial vehicles (UAVs) are promising tools for efficient data collections of sensors in Internet of Things networks. Existing studies exploited both spatial and temporal data correlations to reduce the amount of collected redundant data, in which sensors are first partitioned into different clusters, a master sensor in each cluster then collects raw data from other sensors and compresses the received data. An energy-constrained UAV finally collects the maximum amount of compressed data from different master sensors. We however notice that the compressed data from only a portion of clusters are collected by the UAV in the existing studies, while the data from other clusters are not collected at all. In this article, we study a problem of finding a data collection trajectory for an energy-constrained UAV, so that the accumulative utility of collected data is maximized, where the accumulative utility measures the quality of spatiotemporally correlated data collected from different clusters. We propose a novel $[1/(6+\epsilon)]$-approximation algorithm for the problem, where ϵ is a given constant with $\epsilon > 0$. Experimental results with real data sets show that the accumulative utility by the proposed algorithm is at least 23% larger than those by the existing studies, and the number of clusters collected by the proposed algorithm is from 45% to 105% larger than those by the existing studies.**

Manuscript received 14 December 2023; revised 1 February 2024 and 11 February 2024; accepted 23 February 2024. Date of publication 22 March 2024; date of current version 23 May 2024. The work of Wenzheng Xu was supported in part by the National Natural Science Foundation of China under Grant 62272328, and in part by the Sichuan Science and Technology Program under Grant 24NSFJQ0152. The work of Jian Peng was supported in part by the Cooperative Program of Sichuan University and Yibin under Grant 2020CDYB-30; in part by the Cooperative Program of Sichuan University and Zigong under Grant 2022CDZG-6; in part by the Key Research and Development Program of Sichuan Province of China under Grant 22ZDYF3599; and in part by the Sichuan Science and Technology Program under Grant 2022ZDZX0011. *(Corresponding authors: Jian Peng; Wen Huang.)*

Wenzheng Xu, Heng Shao, Qunli Shen, Jian Peng, Wen Huang, and Tao Lin are with the College of Computer Science, Sichuan University, Chengdu 610065, China (e-mail: wenzheng.xu3@gmail.com; 2022223040029@stu.scu.edu.cn; shenqunli@stu.scu.edu.cn; jianpeng@scu.edu.cn; wen@scu.edu.cn; lintao@scu.edu.cn).

Weifa Liang is with the Department of Computer Science, City University of Hong Kong, Hong Kong, China (e-mail: weifa.liang@cityu.edu.hk).

Tang Liu is with the College of Computer Science, Sichuan Normal University, Chengdu 610101, China (e-mail: liutang@sicnu.edu.cn).

Xin-Wei Yao is with the School of Computer Science and Technology and the Institute for Frontier and Interdisciplinary Sciences, Zhejiang University of Technology, Hangzhou 310023, China (e-mail: xwyao@zjut.edu.cn).

Sajal K. Das is with the Department of Computer Science, Missouri University of Science and Technology, Rolla, MO 65409 USA (e-mail: sdas@mst.edu).

This article has supplementary downloadable material available at https://doi.org/10.1109/JIOT.2024.3370295, provided by the authors.

Digital Object Identifier 10.1109/JIOT.2024.3370295

***Index Terms*—Approximation algorithms, mobile data collections, spatial data correlations, unmanned aerial vehicles (UAVs).**

I. Introduction

The last decade has witnessed the unprecedented explosion of various Internet of Things (IoT) applications, such as smart city monitoring, disaster monitoring, intelligent transportation monitoring, battlefield monitoring, environment monitoring, etc. [3], [16], [26]. Since the battery energy of sensors in IoT networks is limited, it is promising to enable sensors to harvest energy from their surrounding environment, e.g., solar energy or wind energy [4], [19], [21].

In this article, we study the efficient data collections of sensors in an IoT network, where sensors are sparsely placed, e.g., in an area for several square kilometers. Due to the limited transmission range between sensors (e.g., tens of meters), a large number of relay sensors need to be deployed to ensure the network connectivity, which incurs a high deployment cost. In addition, the amounts of harvested energy by sensors usually are limited and thus may not be enough for a large amount of data transmissions [7], [12]. On the other hand, unmanned aerial vehicles (UAVs) are promising tools for efficient data collections of sensors, due to their high flexibility, low cost, and ease of deployment [10], [13], [14], [15], [27].

Since a UAV is energy constrained and there may be many sensors in an IoT network, existing studies exploited both spatial and temporal data correlations to reduce the amount of collected data by the UAV [10], [14], [15], [27], where the spatial data correlations mean that sensing data from nearby sensors are highly correlated [24], and the temporal data correlations indicate that sensing data from the same sensor in a short period are also highly correlated [6], [7].

The existing studies [10], [14], [15], [27] exploited a coarse-grained data collection model, in which sensors in an IoT network usually are first partitioned into different clusters, a master sensor in each cluster then collects raw data from other sensors and compresses the received data. An energy-constrained UAV finally collects the maximum amount of compressed data from different master sensors. For example, Fig. 1(a) shows an IoT network consisting of 12 sensors and there are 15 MB to-be-collected raw data in each of the 12 sensors. The sensors are partitioned into three clusters

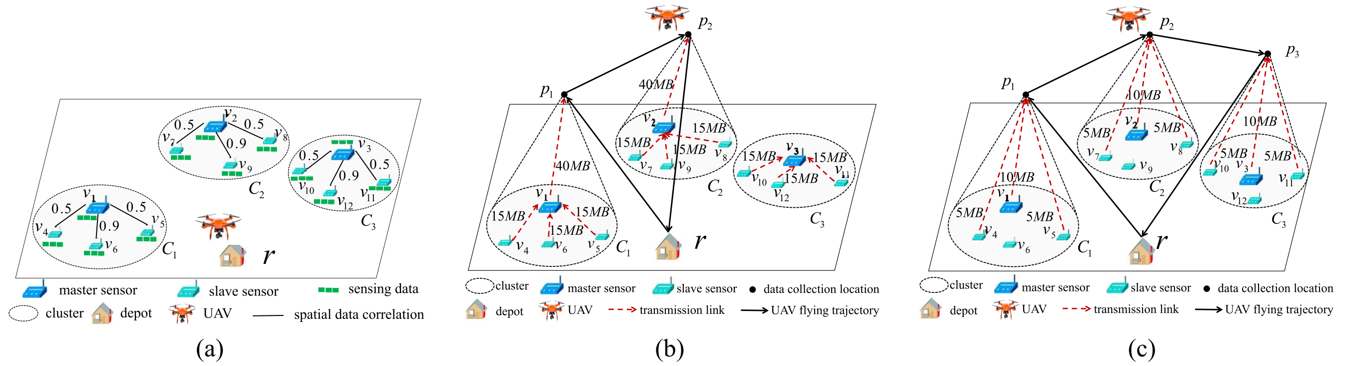


Fig. 1. Illustration of the difference between the data collection trajectory found by existing studies and the one found in this article. (a) 12 sensors in an IoT network are partitioned into 3 clusters, where there are 15 MB to-be-collected raw data in each of the 12 sensors, and the edge weight between the master sensor v_i of cluster C_i ($1 \leq i \leq 3$) and a slave sensor v_j is their spatial data correlation. (b) Data collection trajectory found by existing studies, where only the compressed data from master sensors v_1 and v_2 are collected, and there are 40-MB compressed data in each of the three master sensors v_1, v_2, and v_3. (c) Data collection trajectory found in this article, where 20-MB nonredundant data are collected from each of the three clusters, and the 20-MB data is able to recover 80% raw data of the four sensors in each cluster.

C_1, C_2, and C_3, and there is a master sensor in each cluster, see masters v_1, v_2, and v_3. In Fig. 1(a), the edge weight between the master sensor v_i of cluster C_i ($1 \leq i \leq 3$) and a slave sensor v_j is the spatial data correlation between v_i and v_j. For example, the spatial data correlation between sensors v_1 and v_4 is 0.5, which means that the probability of the difference between the sensing data from v_1 and v_4 being smaller than a small threshold at any time is 0.5 [7], [14], [15], [27]. Slave sensors in each cluster send their raw data to their master sensor, and the master sensor compresses raw data, e.g., by applying the compressive sensing theory and/or the sparsity-optimization method K-SVD [10]. For example, Fig. 1(b) shows that slave sensors v_4, v_5, and v_6 send their raw data to their master sensor v_1 in cluster C_1. The master sensor v_1 then compresses the 60-MB raw data from the four senors in the cluster into, e.g., 40-MB data, as there are 15-MB raw data from each of the four sensors in C_1. Notice that all the compressed 40-MB data should be collected if the UAV collects data from cluster C_1. Otherwise (only a portion of the compressed data is collected), the raw data from the four sensors in cluster C_1 cannot be recovered from the partially collected compressed data [7], [14], [15], [27]. Since a UAV is energy-constrained and it may not be able to collect compressed data from all the clusters. Fig. 1(b) shows the data collection trajectory found by the existing studies, where only the compressed data from masters sensors v_1 and v_2 are collected, while the data from master sensor v_3 are not collected at all.

In spite of the pioneering studies on the efficient UAV data collection by exploiting spatial and/or temporal data correlations [10], [14], [15], [27], we notice that there is still a major problem to be addressed. That is, the compressed data from only a portion of clusters are collected by the UAV in the existing studies, while the data from other clusters are not collected at all, e.g., the compressed data in master sensor v_3 are not collected in Fig. 1(b).

In this article, we adopt a light-weight fine-grained data collection model as follows. In each cluster, the master sensor first sends some of its raw data to the UAV, while the slave sensors in the cluster can overhear the data transmitted by the master sensor at the same time. Each slave sensor then removes its redundant data with the overheard data, and finally sends a portion of their rest nonredundant data to the UAV. Fig. 1(c) shows that the master sensor v_1 first sends 10 MB of its 15-MB data to the UAV, slave sensors v_4, v_5, and v_6 then removes their redundant data with the transmitted 10-MB data by v_1. Both sensors v_4 and v_5 further send 5-MB nonredundant data to the UAV, while sensor v_6 does not send any data, due to its high spatial data correlation 0.9 with the master sensor v_1. Therefore, 20 MB (=10 MB + 5 MB + 5 MB) nonredundant data in cluster C_1 are collected by the UAV and the 20-MB data is able to recover, e.g., 80% the raw data from the four sensors in cluster C_1. Fig. 1(c) demonstrates the data collection trajectory found in this article, where 20-MB nonredundant data are collected from each of the three clusters.

The *novelty* of this article is that, unlike existing studies that exploited a coarse-grained data collection model, in which all compressed data in each cluster must be collected to recover the original raw data, if the UAV collects data from the cluster, in this article we adopt a light-weight fine-grained data collection model in which compressed data can be partially collected. In addition, we propose a novel approximation algorithm for finding the data collection trajectory of a UAV, and determining the amount of collected data from each sensor by the UAV, such that the amount of nonredundant data collected from different clusters is maximized, thereby increasing the diversity of collected data.

The *contributions* of this article are summarized as follows. We first study a problem of finding a data collection trajectory for an energy-constrained UAV and determining the amount of collected data from each sensor of every cluster, so that the accumulative utility of collected data is maximized, where the accumulative utility measures the quality of spatiotemporally correlated data collected from different clusters. We then propose a novel $[1/(6+\epsilon)]$-approximation algorithm for the problem, where ϵ is a given constant with $\epsilon > 0$. We finally evaluate the proposed algorithm with real data sets that consist of temperature sensing data and energy-harvesting data. Experimental results show that the accumulative utility by the proposed algorithm is at least 23% larger than those by the existing studies, and the number of clusters collected by the proposed algorithm is from 45% to 105% larger than those by the existing studies.

The remainder of this article is organized as follows. We review related studies in Section II. We introduce preliminaries and the problem definition in Section III. We propose a novel approximation algorithm for the problem in Section IV. We evaluate the performance of the proposed algorithm in Section V, and conclude this article in Section VI.

II. Related Work

The study on UAV data collections in IoT network has drawn many attentions. Most studies considered neither the temporal correlations nor spatial correlations of sensing data [8], [13], [23], [29], [30]. For example, Li et al. [13] studied a UAV-assisted data collection problem, in which a UAV is able to collect sensing data from multiple IoT devices within its communication range simultaneously, e.g., MIMO. They proposed an algorithm to maximize the amount of collected data, under the constraint on the UAV energy capacity. Hu et al. [8] founded a UAV-assisted data collection tour to minimize the Age of Information (AoI) of collected data, where the UAV can not only charge sensors but also collect their data. Tsai et al. [23] devised an algorithm to find the flying tour for a UAV to monitor multiple restricted regions, such that the consumed time in the tour is minimized. Zhan et al. [29] designed a reinforcement learning algorithm for data collection in a UAV-assisted multicell cellular network, which balances the operation time of a UAV and the AoI of collected data. Zhang et al. [30] founded a flying trajectory for a UAV, so that the UAV energy consumption for collecting data from marine buoys sensors is minimized. Notice that these studies ignored both the spatial and temporal data correlations, and UAVs may collect redundant data from different sensors.

Some studies exploited spatial data correlation to reduce the amount of collected redundant data [14], [15], [27]. Liu et al. [15] adopted a matrix completion theory for UAV data collections, which first chooses some representative sensors, then dispatches a UAV to collect their data, and finally recovers some uncollected data from the collected data. Yu et al. [27] proposed a spatial data collection strategy, where the network is first partitioned into clusters, some slave sensors are chosen to send their data to their master sensor in each cluster, the master sensor forwards the received data to a UAV, and the UAV finally recovers some uncollected data by a denoising autoencoder, based on neural networks. Liu et al. [14] designed a data collection scheme for a mobile sink, where the mobile sink randomly visits a portion of static sensors, collects their data, and recovers sensing data based on the compressive sensing theory. It can be seen that, since these studies ignored temporal data correlations and the collected data by UAVs are still highly redundant, they are not suitable for data collections in large-scale networks.

Only a very few studies focused on both spatial and temporal data correlations, and studied efficient data collections with a UAV [10]. For example, Li et al. [10] proposed a spatially and temporally correlated data aggregation model to reduce data redundancy in UAV-assisted WSNs. Specifically, they first partition sensors into clusters, and slave sensors in each cluster send their data to its master sensor in the cluster. Since the data received from slave sensors and from itself are not only spatially correlated but also temporally correlated, the master sensor then compresses data by applying the compressive sensing theory and a sparsity-optimization method K-SVD. The UAV finally collects compressed data from masters. Although these studies exploited both spatial and temporal data correlations, they only used a coarse-grained data collection model, in which all compressed data in each cluster must be collected to recover the original raw data, if the UAV collects data from the cluster, in this article, we adopt a light-weight fine-grained data collection model in which compressed data can be partially collected, thereby increasing the diversity of collected data.

There are also some studies that considered both spatial and temporal data correlations, but UAVs or mobile sinks are not employed to collect data. Instead, the data of each sensor is sent to a base station directly or via the relay of other sensors. For example, Guo et al. [7] devised a fine-grained clustering method based on spatial data correlations, and designed a routing protocol to maximize the utility of collected spatiotemporally correlated data. Fattoum et al. [6] proposed an adaptive sampling approach, which minimizes sampling rates of sensor nodes while ensuring a high quality of collected data. Xie et al. [24] chose some important sampling points to collect data and estimated the data of not-sampled points from the collected points, based on spatial data correlations. However, it is unknown how to extend the proposed algorithms for UAV data collections.

III. Preliminaries

In this section, we introduce the network model, spatial data correlation model, channel model, UAV data collection framework, UAV energy consumption model, and define the problem precisely.

A. Network Model

We consider an IoT network deployed in a critical area, which is used for, e.g., smart city monitoring or environmental monitoring. Assume that there are n sensors in the network with $n \geq 1$. Let V be the set of the n sensors, i.e., $V = \{v_1, v_2, \ldots, v_n\}$. Denote by $(x_i, y_i, 0)$ the coordinate of a sensor v_i with $1 \leq i \leq n$. Assume that the locations of the n sensors are known in the phase of network deployment. Each sensor v_i in V is powered by a rechargeable battery with a capacity of B_i, and it can harvest energy from its surrounding environment, e.g., solar energy or wind energy.

In this article, for the sake of simplicity, we study the employment of a single UAV to collect data from the sparsely located sensors, which is applicable to a small-scale network, e.g., one square kilometer. In case that the area of the network may be very large (e.g., tens of square kilometers), we need to employ multiple UAVs. The proposed algorithm in this article can be easily extended to the case with multiple UAVs, e.g., finding the data collection tours of multiple UAVs one by one in a greedy manner, by applying the strategy in work [25].

We consider the data collection of the IoT network, e.g., for one month. For every fixed period T, e.g., one day, the UAV is dispatched to collect sensor data. The period T is divided into equal-length time slots, where the duration of each time slot is short, e.g., 5 min. Each sensor v_i generates data with a rate $r_i(t)$ at time slot t with $1 \leq t \leq T$.

Denote by D_i^R the amount of residual data in sensor v_i at the beginning of period T. The amount of to-be-collected data in sensor v_i at the end of period T is $\min\{D_i^R + \sum_{t=1}^{T} r_i(t),\ D_i^S\}$, where $r_i(t)$ is the data rate of sensor v_i at time slot t and D_i^S is the storage capacity of sensor v_i. Denote by $\hat{r}_i$ the predicted average data rate of sensor v_i in period T, for example, its past average data rate before the period T. The predicted amount of to-be-collected data in sensor v_i at the end of period T then is $D_i^C = \min\{D_i^R + \hat{r}_i T,\ D_i^S\}$.

Each sensor can harvest energy from its surrounding environment, such as solar energy or wind energy. Although environmental conditions, e.g., weather condition, affect the harvesting rate, it is predictable by adopting a machine learning-based prediction algorithm, e.g., the NARNET model in [20]. Denoted by $e_H(v_i, t)$ the amount of harvested energy by sensor v_i at time slot t. Also, denote by $e_R(v_i, t)$ the amount of residual energy of sensor v_i at the beginning of time slot t. It can be seen that the amount $P_C(v_i, t)$ of consumed energy by sensor v_i at time slot t should be no greater than the sum of its amount $e_R(v_i, t)$ of residual energy at the beginning of time slot t and its harvested energy $e_H(v_i, t)$ at time slot t, i.e., $P_C(v_i, t) \leq e_R(v_i, t) + \hat{e}_H(v_i, t)$. The amount of residual energy of sensor v_i at the end of period T then is $e_i^{\text{Budget}} = \min\{e_R(v_i, 0) + \sum_{t=1}^{T} \hat{e}_H(v_i, t) - \alpha \hat{r}_i T,\ B_i\}$, where $e_R(v_i, 0)$ is the amount of residual energy of sensor v_i at the beginning of period T, $\sum_{t=1}^{T} \hat{e}_H(v_i, t)$ is its amount of harvested energy in period T, $\alpha \hat{r}_i T$ is its amount of consumed energy for data sensing in period T, α is the energy consumption rate for sensing data [11], and B_i is its battery capacity.

On the other hand, each sensor consumes its energy on data sensing, data transmission, and data reception. Following the exiting work in [11], we model the energy consumptions $P_S(v_i, t)$, $P_{Tx}(v_i, t)$, $P_{Rx}(v_i, t)$ of sensor v_i on data sensing, data transmission, and data reception at time slot t as follows. $P_S(v_i, t) = \alpha \cdot r_i(t)$, $P_{Tx}(v_i, t) = (\beta_1 + \beta_2 d^2(v_i, u_t)) \cdot f_{it}$, and $P_{Rx}(v_i, t) = \gamma \cdot \sum_{v_j \in N(v_i)} \cdot f_{jit}$, where α, β_1, β_2 and γ are constants [11], $r_i(t)$ is the data rate of sensor v_i, $d(v_i, u_t)$ is the Euclidean distance between sensor v_i and the UAV at time slot t and f_{it} is the data transmission rate from sensor v_i to the UAV at time slot t, f_{jit} is the reception rate of sensor v_i from its neighbor v_j at time slot t and $N(v_i)$ is the set of neighbors of sensor v_i.

B. Spatial Data Correlation Model

It is recognized that nearby sensors may generate similar sensing data. Following the study in [7], we can measure the spatial data correlation between sensors v_i and v_j as $c(v_i, v_j) = (N_{ij}/N)$, where N_{ij} is the number of time slots so that the difference between the data generated by sensors v_i and v_j is no more than a small threshold in the past N time slots, e.g., $N = 100$. Notice that the value of N_{ij} can be calculated with historical data [7].

To collect nonredundant information from the n sensors in the network, we partition the n sensors into several clusters, and each cluster consists of a master sensor and some slave sensors. Then, the data collected from slave sensors should be nonredundant with the data collected from the master sensor. By doing so, less redundant data will be collected from each cluster and the energy-constrained UAV is able to collect nonredundant data from more clusters.

We apply the method in [7] to partition the n sensors into, e.g., Q clusters $C_1, C_2, \ldots, C_Q$, and find a master sensor v_i for each cluster, where the value of Q is determined by the method. Specifically, let v_i be the master sensor of cluster C_i. Denote by $D_i^{\max}$ the maximum amount of potentially collected data from sensor v_i with its residual energy and harvested energy in the period T [7], where $D_i^{\max} = \min\{R_i^{\max}(e_i^{\text{Budget}}/P_t^i),\ D_i^C\}$, $R_i^{\max}$ is the maximum data transmission rate from sensor v_i to the UAV, e_i^{Budget} is its energy budget, P_t^i is its transmission power, and D_i^C is its amount of to-be-collected data.

The amount of suppressed redundant data transmission by a slave sensor v_j in cluster C_i is $D_i^{\max} \cdot c(v_i, v_j)$, and the total amount of suppressed redundant data transmission in the network is $\sum_{i=1}^{Q} \sum_{v_j \in C_i \setminus \{v_i\}} D_i^{\max} \cdot c(v_i, v_j)$. The redundant data suppression maximization problem is to partition the n sensors into Q clusters, e.g., $C_1, C_2, \ldots, C_Q$, and find a master sensor v_i in each cluster C_i, such that the amount of suppressed redundant data transmission is maximized, where the value of Q is to be determined. Note that there is a $(0.5 - \varepsilon)$-approximation algorithm in [7] for the redundant data suppression maximization problem, where ε is a given constant with $0 < \varepsilon < 0.5$.

C. Channel Model

Consider a ground sensor v_i and the UAV that hovers at a location p_j in the air. Following the study in [18], the average signal-to-noise ratio SNR_{ij} is $\text{SNR}_{ij} = P_{ij}^{\text{LoS}} \text{SNR}_{ij}^{\text{LoS}} + P_{ij}^{\text{NLoS}} \text{SNR}_{ij}^{\text{NLoS}}$, where P_{ij}^{LoS} is the probability of Line-of-Sight (LoS) link between the sensor and the UAV, and its value can be calculated by the method in [1], $P_{ij}^{\text{NLoS}} (= 1 - P_{ij}^{\text{LoS}})$ is the probability of Non-Line-Sight (NLoS) link, $\text{SNR}_{ij}^{\text{LoS}}$ and $\text{SNR}_{ij}^{\text{NLoS}}$ are the signal-to-noise ratios of LoS and NLoS links, respectively. Specifically, $\text{SNR}_{ij}^{\text{LoS}} = [(P_t^i g_t^i)/P_N][c/(4\pi f_C d(v_i, p_j))]^2 (1/\eta_{\text{LoS}})$ and $\text{SNR}_{ij}^{\text{NLoS}} = [(P_t^i g_t^i)/P_N][c/(4\pi f_C d(v_i, p_j))]^2 (1/\eta_{\text{NLoS}})$, where P_t^i is the transmission power of the sensor v_i, g_t^i is its antenna gain, P_N is the power of Gaussian white noise, c is the speed of light, f_C is the frequency of the radio carrier (e.g., $f_C = 2.4$ GHz), $d(v_i, p_j)$ is the Euclidean distance between sensor v_i and location p_j, η_{LoS} and η_{NLoS} are the average shadow fadings for the LoS and NLoS links, respectively, e.g., $\eta_{\text{LoS}} = 1$ dB and $\eta_{\text{NLoS}} =$

20 dB in an urban environment [1]. Then, the data collection rate R_{ij} from sensor v_i to the UAV at location p_j is $R_{ij} = W \cdot \log_2(1 + \mathrm{SNR}_{ij})$, where W is the bandwidth allocated to the sensor v_j by the UAV, e.g., 20 MHz.

D. UAV Data Collection Framework

At the end of every fixed period T, e.g., one day, a UAV is dispatched to quickly collect data from sensors in the network. The UAV is located at a depot r initially. The UAV departs from the depot, flies along a planned trajectory for the data collection task, and returns to the depot to replenish its energy after completing the task. The UAV hovers at an optimal height h to collect sensor data, e.g., $h = 300$ m, as the data transmission rate from a sensor to the UAV at a higher or a lower height is smaller than the optimal height h [1].

Due to the limited energy capacity of the UAV, it may not be able to collect data from all sensors. Instead, it collects only nonredundant information from the sensors. Assume that the UAV collects data from Q' of Q clusters with $Q' \le Q$, and Q' is to be determined later. For the sake of convenience, we assume that the UAV collects data from clusters $C_1, C_2, \ldots, C_{Q'}$ one by one. For example, Fig. 1(c) in Section I shows that the network is divided into $Q = 3$ clusters, and UAV collects data from clusters C_1, C_2, and C_3 one by one, where $Q' = Q = 3$.

Denote by $P_r = r \to p_1 \to p_2 \to \cdots \to p_{Q'} \to r$ the UAV data collection trajectory, where p_i is the hovering location above the master sensor v_i in cluster C_i at height h. In other words, assume that the coordinate of the master sensor v_i is $(x_i, y_i, 0)$, the UAV hovering location p_i for collecting data from sensors in cluster C_i then is (x_i, y_i, h). The rationale behind the assumption that the UAV hovers above the master sensor v_i for data collections in cluster C_i is that, the data from the master sensor is most representative and more data should be collected from the master sensor than another slave sensor in the cluster, and the data transmission rate from the master sensor to the UAV is maximized when the UAV hovers above the sensor, e.g., an LoS link [1].

For any cluster C_i, the UAV may collect data from both the master sensor v_i and slave sensors in the cluster. When the UAV hovers above cluster C_i, the master sensor v_i first sends its data to the UAV, and every slave sensor v_j in C_i overhears the transmitted data by v_i at the same time. Slave sensor v_j thus knows its similar data with the transmitted data by the master sensor v_i, and then sends its nonredundant data to the UAV. Denote by t_i and t_{ij} the data collection durations allocated by the UAV to the master sensor v_i and a slave sensor v_j in cluster C_i, respectively. Then, the amounts of collected data from the master sensor v_i and the slave sensor v_j are $D_i = R_i \cdot t_i$ and $D_j = R_{ji} \cdot t_{ij}$, respectively, where R_i and R_{ji} are the data transmission rates from the master sensor v_i and the slave sensor v_j to the UAV, respectively.

Due to some spatial data similarity $c(v_i, v_j)$ between the master sensor v_i and the slave sensor v_j, the amount D_{ij} of obtained data by the slave sensor v_j is the sum of the amount of collected data from v_j itself and the similar data collected from v_i, i.e., $D_{ij} = D_j + D_i \cdot c(v_i, v_j)$.

It can be seen that, in the proposed data collection model, each sensor is required to have very limited computing ability. On one hand, the sensing data generated by each sensor usually is stored in its memory in order of their sensing time. Then, the transmitted data by the master sensor of each cluster also is in order of their sensing time. On the other hand, when a slave sensor overhears the transmitted data from its master sensor, it can remove its redundant data with the overheard data, by simply performing numerical comparisons.

Notice that many existing studies required that a sensor has some processing ability to remove redundant data, e.g., [10] and [31], and the cost of such a sensor usually is small, e.g., no more than ten dollars. For example, consider a sensor consists of a DHT22 sensing unit, an ESP8266 WiFi SoC, and an STM32F103C8T6 micro CPU, where the DHT22 sensing unit is able to sense temperature and humidity data, the ESP8266 WiFi SoC can transmit and overhear data, and the STM32F103C8T6 micro CPU can read from and write to the memory, compare and remove similar data by performing only about ten lines of codes. Of course, it is desirable to have less requirements on sensors to remove redundant data, and this is one of our future work.

E. UAV Energy Consumption Model

The UAV consumes its energy on flying and data collections. Recall that the UAV data collection trajectory is $P_r = r \to p_1 \to p_2 \to \cdots \to p_{Q'} \to r$, where p_i is the hovering location above the master sensor v_i in cluster C_i at height h. For each cluster C_i, its total data collection time is $t(C_i) = t_i + \sum_{v_j \in C_i \setminus \{v_i\}} t_{ij}$, where t_i and t_{ij} are data collection durations of master sensor v_i and a slave sensor v_j in cluster C_i, respectively. The amount of energy consumption for data collection in cluster C_i then is $t(C_i) \cdot \xi$, where ξ is the UAV energy consumption rate for hovering and data collection.

On the other hand, during the flying from locations p_i to p_{i+1} in the data collection trajectory P_r with $0 \le i \le Q'$, the UAV first accelerates from zero to a speed $v_{\max}$ (e.g., 10 m/s) with an acceleration speed, then flies with the speed $v_{\max}$ toward location p_{i+1}, and finally decelerates from $v_{\max}$ to zero when the UAV arrives at p_{i+1}. Following existing studies in [22] and [28], the UAV flying energy consumption $e^{fly}_{i,i+1}$ from locations p_i to p_{i+1} then is $e^{fly}_{i,i+1} = \eta[(d_{i,i+1} - d_{\mathrm{acc}} - d_{\mathrm{dec}})/v_{\max}] + e^{\mathrm{acc}} + e^{\mathrm{dec}}$, where η is the energy consumption rate at speed $v_{\max}$, $d_{i,i+1}$ is the Euclidean distance from locations p_i to p_{i+1}, d_{acc} and d_{dec} are the flying distances for acceleration and deceleration, respectively, e^{acc} and e^{dec} are its energy consumptions for acceleration and deceleration, respectively. The energy consumption of the UAV in its data collection trajectory P_r then is $w(P_r) = \sum_{i=1}^{Q'} t(C_i) \cdot \xi + \sum_{i=0}^{Q'} e^{fly}_{i,i+1}$. Notice that the UAV energy consumption $w(P_r)$ should be no greater than its battery capacity e^{CAP}.

F. Problem Definition

Since sensing data is not only spatially correlated but also temporally correlated, we introduce a utility function $f(\cdot)$ to measure the quality of spatiotemporally correlated data [7], where the utility function $f(\cdot)$ has a diminishing return property. Assume that function $f(\cdot)$ is increasing, twice-differentiable, and strictly concave, e.g., $f(x) = \log_2(1 + x)$. For a cluster C_i, the accumulative utility of collected data from its master sensor v_i and slave sensors is $f(D_i) + \sum_{v_j \in C_i \setminus \{v_i\}} f(D_{ij}) = f(D_i) + \sum_{v_j \in C_i \setminus \{v_i\}} f(D_j + c(v_i, v_j) \cdot D_i)$. The accumulative utility of collected data from the Q clusters of the network then is

$$\sum_{i=1}^{Q} \left(f(D_i) + \sum_{v_j \in C_i \setminus \{v_i\}} f(D_{ij}) \right). \quad (1)$$

In this article, we consider a *data collection utility maximization problem* as follows. Given the amounts of to-be-collected data D_i^C and residual energy e_i^{Budget} of each sensor v_i with $1 \le i \le n$, assume that the n sensors have been partitioned into Q clusters $C_1, C_2, \ldots, C_Q$ (see Section III-B). The problem is to find a data collection flying trajectory P_r for a UAV, and calculate the data collection durations t_i and t_{ij} for the master sensor v_i and each slave sensor v_j in each cluster C_i, respectively, so that the accumulative utility $\sum_{i=1}^{Q} (f(D_i) + \sum_{v_j \in C_i \setminus \{v_i\}} f(D_{ij}))$ of collected data from sensors in the Q clusters is maximized, where the energy consumption of the UAV in the trajectory P_r is no greater than its energy capacity e^{CAP}.

Notice that if the data spatial correlation $c(v_i, v_j)$ between any two sensors is zero, each cluster C_i consists of only the master sensor v_i in the solution delivered by the algorithm [7]. The number Q of clusters then is equal to the number n of sensors. The accumulative utility of collected data from sensors in the network thus is $\sum_{i=1}^{n} f(D_i)$, where D_i is the amount of collected data from sensor v_i. By utilizing the utility function $f(.)$, we are able to collect data from more sensors, since the data from the same sensor are temporally correlated. For example, consider two data collection strategies. The first strategy is to collect 10-MB data from a sensor v_1 and none data from another sensor v_2, while the second strategy is to collect 5-MB data from both sensors v_1 and v_2. Although the amount of collected data in the two strategies is equal, i.e., 10 MB, the accumulative utility of collected data in the first strategy is less than that in the second strategy, i.e., $f(10) + f(0) = \log_2(1 + 10) + \log_2(1 + 0) = 3.46 < f(5) + f(5) = 2\log_2(1 + 5) = 5.17$.

G. Related Problem

We introduce a related problem—the orienteering problem. Given an undirected and weighted graph $G(V \cup \{r\}, E)$ and a length budget $L_{\max}$, there is a length $w(u, v)$ associated with each edge (u, v) in E, where r is the root and edge weights in G satisfy the triangle inequality. The problem is to find an r-rooted closed tour with its length no greater than $L_{\max}$, so that the number of nodes visited in the tour is maximized. Notice that there is a (1/3)-approximation algorithm for the problem [2], which will serve as a subroutine of the proposed algorithm in this article.

IV. Approximation Algorithm

In this section, we propose a novel $[1/(6 + \epsilon)]$-approximation algorithm for the data collection utility maximization problem, where ϵ is a given constant with $\epsilon > 0$.

A. Algorithm Basic Idea

The basic idea behind the proposed algorithm is to transform the data collection utility maximization problem to the orienteering problem. There are however two major differences between the two problems. The first difference is that we do not know the data collection utility of each cluster in the former problem since the data collection duration of each sensor in the cluster is unknown, while the utilities of visiting different nodes in the latter problem are given and uniform, i.e., equal to one. The second difference is that both edge weights (i.e., the UAV flying energy consumption) and node weights (i.e., data collection energy consumption) are taken into consideration in the UAV tour of the former problem, whereas only edges weights are counted in the found tour of the latter problem.

We devise a series of smart graph transformations to address the differences. Specifically, in the first graph transformation, we transform each cluster into several virtual nodes so that each virtual node is associated with a node weight (i.e., UAV data collection energy consumption) and a profit (i.e., the utility of visiting the virtual node). In the second graph transformation, we transform the both node-weighted and edge-weighted graph obtained in the first graph transformation into an only edge-weighted graph. In the last transformation, we transform each virtual node into several nodes with uniform profits, through a novel scaling technique.

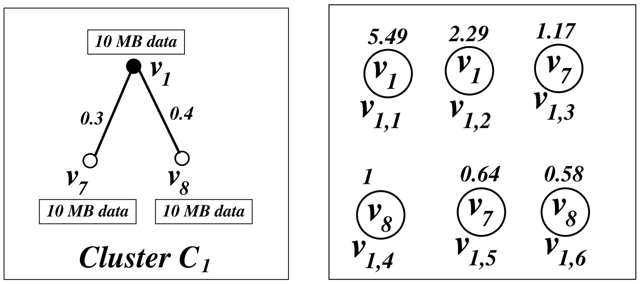
B. First Graph Transformation

In the first graph transformation, we transform each cluster C_i into several virtual nodes so that each virtual node is associated with a node weight (i.e., UAV data collection energy consumption) and a profit (i.e., the utility of visiting the node). For each sensor v_j in a cluster C_i, its maximum data collection time is $t_j^{\max} = \min\{(D_j^C/R_{ji}),\ (e_j^{\text{Budget}}/P_t^j)\}$, where D_j^C is its amount of to-be-collected data, R_{ji} is its data transmission rate, e_j^{Budget} is its amount of residual energy, and P_t^j is its transmission power.

Denote by δ the data collection time unit, e.g., $\delta = 5$ s. This implies that if the UAV collects data from a sensor v_j in cluster C_i, its data collection time is one of the values in set $\{\delta, 2\delta, \ldots, k_j\delta\}$, where k_j is the maximum number of time units for collecting all data from sensor v_j. That is

$$k_j = \left\lceil \frac{t_j^{\max}}{\delta} \right\rceil. \quad (2)$$

Then, the number of time units for collecting data from sensors in cluster C_i is no greater than $K_i = k_i + \sum_{v_j \in C_i \setminus \{v_i\}} k_j$, where k_i and k_j are the maximum numbers of time units for collecting all data from the master sensor v_i and a slave sensor v_j in cluster C_i, respectively. For example, assume that there are three sensors v_1, v_7, and v_8 in cluster C_1, and v_1 is the master sensor in C_1, see Fig. 2(a). The amount of to-be-collected raw data in each of the three sensors is 10 MB. Assume that the data transmission rate from each sensor to the UAV is 8 Mb/s (i.e., one MB per second). Then, the maximum number of time slots for collecting all data from each sensor in C_1 is $\lceil(10 \text{ MB}/8 \text{ Mb/s})/5 \text{ s}\rceil = 2$. Then, the number of



(a) (b)

Fig. 2. Illustration of the first graph transformation. (a) Cluster C_1 that consists of master sensor v_1 and two slave sensors v_7 and v_8. (b) Transform cluster node C_1 into six virtual nodes $v_{1,1}, v_{1,2}, \ldots, v_{1,6}$.

time units for collecting data from the three sensors in cluster C_1 is no more than $K_1 = 2 \times 3 = 6$.

For each cluster C_i, we transform it into K_i virtual nodes $v_{i,1}, v_{i,2}, \ldots, v_{i,K_i}$. The node weight $h(v_{i,l})$ of each virtual node $v_{i,l}$ is the UAV data collection energy consumption for one-time slot, i.e., $h(v_{i,l}) = \xi \cdot \delta$, where ξ is the UAV energy consumption rate for hovering and data collection, and δ is the duration of the time unit.

On the other hand, a profit $p_{i,l}$ is associated with each virtual node $v_{i,l}$ with $1 \leq l \leq K_i$, which is the increased accumulative utility if the UAV collects data from a sensor at the lth time slot. To see from which sensor the UAV should collect data at the lth time slot, the increased accumulative utility of each sensor in C_i is first calculated. Then, the data from the sensor with the maximum increased accumulative utility is collected at the lth time slot. Specifically, recall that D_i and D_j are the amounts of collected data from the master sensor v_i and a slave sensor v_j in C_i, respectively. Initially, $D_i = 0$ and $D_j = 0$. Following (1), the increased accumulative utility of the master sensor v_i in C_i at the lth time slot is $f(D_i + R_i\delta) - f(D_i) + \sum_{v_j \in C_i \setminus \{v_i\}}(f(D_j + c(v_i, v_j) \cdot (D_i + R_i\delta)) - f(D_j + c(v_i, v_j) \cdot D_i))$, where $R_i\delta$ is the amount of data that sensor v_i can send to the UAV at the lth time slot. Similarly, the increased accumulative utility of a slave sensor v_j in C_i at the lth time slot is $f(D_j + R_{ji}\delta + c(v_i, v_j) \cdot D_i) - f(D_j + c(v_i, v_j) \cdot D_i)$. For example, Fig. 2(b) shows that six virtual nodes are transformed from cluster C_1. To see from which sensor the UAV should collect data at the first time slot, the increased accumulative utility of each sensor in C_1 is calculated as follows. Assume that the utility function is $f(x) = \log_2(1+x)$. In addition, the spatial data correlation $c(v_1, v_7)$ between v_1 and v_7 is 0.3, while the correlation $c(v_1, v_8)$ between v_1 and v_8 is 0.4. The increased accumulative utility of the master sensor v_1 is $f(0 + 5 \text{ MB}) - f(0) + \sum_{v_j \in C_1 \setminus \{v_1\}}(f(0 + c(v_1, v_j) \cdot (0 + 5 \text{ MB})) - f(0 + c(v_1, v_j) \cdot 0)\ ) = \log_2(1 + 5 \text{ MB}) + \log_2(1 + 0.3 \cdot 5 \text{ MB}) + \log_2(1 + 0.4 \cdot 5 \text{ MB}) = 5.49$. The increased accumulative utility of slave sensor v_7 is $f(0 + 5 \text{ MB} + c(v_1, v_7) \cdot 0) - f(0 + c(v_1, v_7) \cdot 0) = 2.58$, while the increased accumulative utility of slave sensor v_8 also is 2.58. Therefore, at the first time slot, the data from the master sensor v_1 should be collected by the UAV and the increased accumulative utility is $5.49 (= \max\{5.49, 2.58, 2.58\})$, see the profit of virtual node $v_{1,1}$ in Fig. 2(b). Similarly, the profits of rest five virtual nodes $v_{1,2}, v_{1,3}, \ldots, v_{1,6}$ can be calculated.

There is an important property for the profits of the virtual nodes, that is, the profits of the K_i virtual nodes $v_{i,1}, v_{i,2}, \ldots, v_{i,K_i}$ constructed from each cluster C_i are in a decreasing order, see Fig. 2(b).

A graph $G_1(V_1 \cup \{r\}, E_1)$ is constructed from the original IoT network, where V_1 is the set of virtual nodes constructed from the Q clusters, i.e., $V_1 = \cup_{i=1}^{Q}\{v_{1,1}, v_{1,2}, \ldots, v_{1,K_i}\}$. There is an edge in E_1 between any two nodes in V_1. The node weight $h(v_{i,l})$ of each virtual node $v_{i,l}$ is the UAV data collection energy consumption for one time slot. The edge weight $w_1(v_{i,l_1}, v_{i,l_2})$ between two virtual nodes v_{i_1,l_1} and v_{i_2,l_1} constructed from two clusters C_{i_1} and C_{i_2} is the UAV flying energy consumption between the two hovering locations in the two clusters. In addition, the profit $p_{i,l}$ of visiting virtual node $v_{i,l}$ is the increased accumulative utility if the UAV collects data at the lth time slot. It can be seen that the data collection utility maximization in the original network can be transformed to a problem $\mathcal{P}_1$ of finding an r-rooted tour in G_1, such that the profit sum of nodes in the tour is maximized, while ensuring that the sum of node weights and edge weights in the tour is no more than the UAV energy capacity.

C. Second Graph Transformation

In the second graph transformation, we transform the both node-weighted and edge-weighted graph $G_1(V_1 \cup \{r\}, E_1)$ into an only edge-weighted graph $G_2(V_2 \cup \{r\}, E_2)$, where $V_2 = V_1$ and $E_2 = E_1$. The profit $p_{i,l}$ of each virtual node $v_{i,l}$ in G_2 is equal to its profit in G_1.

The edge weight $w_2(v_{i_1,l_1}, v_{i_2,l_2})$ between any two virtual nodes v_{i_1,l_1} and v_{i_2,l_2} is

$$w_2\left(v_{i_1,l_1}, v_{i_2,l_2}\right) = w_1\left(v_{i_1,l_1}, v_{i_2,l_2}\right) + \frac{h\left(v_{i_1,l_1}\right) + h\left(v_{i_2,l_2}\right)}{2} \quad (3)$$

where $w_1(v_{i_1,l_1}, v_{i_2,l_2})$ is the UAV flying energy consumption between the hovering locations of two clusters C_{i_1} and $C_{i,2}$, $w_1(v_{i_1,l_1}, v_{i_2,l_2}) = 0$ if $i_1 = i_2$, $h(v_{i_1,l_1})$ and $h(v_{i_2,l_2})$ are the node weights of virtual nodes v_{i_1,l_1} and v_{i_2,l_2}, respectively. For example, Fig. 3(a) shows a graph G_1 that consists of root r and only three virtual nodes. Fig. 3(b) shows the edge weight $w_2(v_{1,1}, v_{2,1})$ between nodes $v_{1,1}$ and $v_{2,1}$ in graph G_2 is $w_1(v_{1,1}, v_{2,1}) + [(h(v_{1,1}) + h(v_{2,1}))/2] = 10 + [(5+5)/2] = 15$.

We later will show that the problem $\mathcal{P}_1$ in graph G_1 is equivalent to a problem $\mathcal{P}_2$ of finding an r-rooted tour in G_2 such that the profit sum of nodes in G_2 is maximized, under the constraint that the weighted sum of edges in the tour in no greater than the UAV energy capacity.

D. Third Graph Transformation

Notice that the problem $\mathcal{P}_2$ in G_2 is still different from the orienteering problem, where the profits of different nodes in G_2 are different, while the orienteering problem requires that the profits of different nodes are uniform. Then, we still cannot apply the $(1/3)$-approximation algorithm for the orienteering problem [2] to solve problem $\mathcal{P}_2$ in G_2.

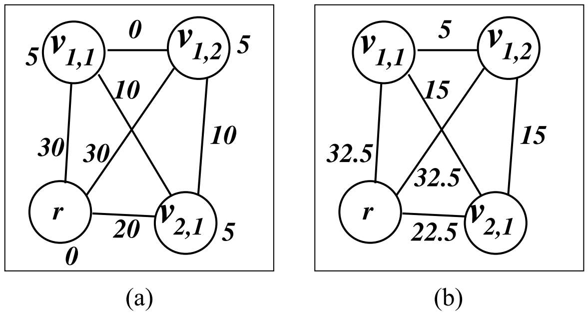


Fig. 3. Illustration of the second graph transformation. (a) Graph G_1 that consists of root r and three virtual nodes. (b) Constructed graph G_2.

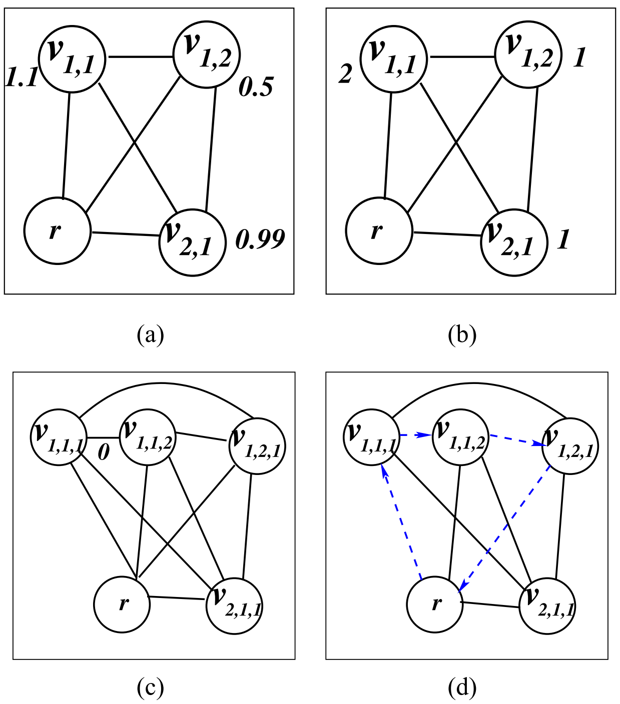


Fig. 4. Illustration of the third graph transformation. (a) Graph G_2, where the profits of three nodes $v_{1,1}$, $v_{1,2}$, and $v_{2,1}$ are 1.1, 0.5, and 0.99, respectively. (b) Constructed graph G'_2, where the scaling factor $\lambda = 0.5$. (c) Constructed graph G_3. (d) Found r-rooted tour P_r^3 by the algorithm in [2].

In the third graph transformation, we transform G_2 to another auxiliary graph G_3 such that the profits of different nodes are uniform in G_3, by losing a performance of only θ, where $\theta = (\epsilon/2)$, and ϵ is a given constant with $\epsilon > 0$.

Given graph $G_2(V_2 \cup \{r\}, E_2)$, let n_2 be the number of virtual nodes in G_2, i.e., $n_2 = |V_2|$. Denote by $p_{\max}$ the maximum profit of nodes in V_2, i.e., $p_{\max} = \max_{a_i \in V_2}\{p_i\}$, where p_i is the node profit of a virtual node a_i in V_2. We first transform G_2 to a graph $G'_2(V_2 \cup \{r\}, E_2)$, where the edge weight $w'_2(a, b)$ between any two nodes a and b in G'_2 is equal to its weight $w_2(a, b)$ in G_2, i.e., $w'_2(a, b) = w_2(a, b)$. However, a node a_i in G'_2 is associated with a profit p'_i different from its profit p_i in G_2. Consider a scaling factor λ with $\lambda > 0$, the profit of each node a_i in G'_2 is $p'_i = \lfloor (p_i/\lambda) \rfloor$, and the value of λ is at least $(p_{\max}/\lceil n_2/\theta \rceil)$, i.e., $\lambda \geq (p_{\max}/\lceil n_2/\theta \rceil)$. For example, Fig. 4(a) shows that the profits of the three nodes $v_{1,1}$, $v_{1,2}$, and $v_{2,1}$ in G_2 are 1.1, 0.5, and 0.99, respectively, where the edge weights in G_2 were shown in Fig. 3(b). Assume that the scaling factor λ is 0.5, Fig. 4(b) demonstrates that the profits of the three nodes $v_{1,1}$, $v_{1,2}$, and $v_{2,1}$ in G'_2 are $2(= \lfloor (1.1/0.5) \rfloor)$, $1(= \lfloor (0.5/0.5) \rfloor)$, and $1(= \lfloor (0.99/0.5) \rfloor)$, respectively.

We construct a graph $G_3(V_3 \cup \{r\}, E_3)$ from G'_2 as follows. For each node a_i in the node set V'_2 of G'_2, p'_i nodes $a_{i,1}, a_{i,2}, \ldots, a_{i,p'_i}$ are added to V_3. For example, Fig. 4(c) shows that two nodes $v_{1,1,1}$ and $v_{1,1,2}$ are added to V_3 for node $v_{1,1}$ in V'_2, since the profit of node $v_{1,1}$ is two. On the other hand, only one node $v_{1,2,1}$ (or $v_{2,1,1}$) is added to V_3 for node $v_{1,2}$ (or $v_{2,1}$) in V'_2, since the profit of node $v_{1,2}$ (or $v_{2,1}$) is one. Then, there are $\sum_{i=1}^{n_2} p'_i$ nodes in V_3. The profit of each node in V_3 is one. There is an edge in E_3 between any two nodes in $V_3 \cup \{r\}$. The edge weight $w_3(a_{i,j}, a_{i,l})$ between any two nodes $a_{i,j}$ and $a_{i,l}$ constructed from the same node a_i in G'_2 is zero, i.e., $w_3(a_{i,j}, a_{i,l}) = 0$, while the edge weight $w_3(a_{i_1,j}, a_{i_2,l})$ between any two nodes $a_{i_1,j}$ and $a_{i_2,l}$ constructed from two different nodes a_{i_1} and a_{i_2} in G'_2 is the edge weight $w'_2(a_{i_1}, a_{i_2})$ in G'_2, i.e., $w_3(a_{i_1,j}, a_{i_2,l}) = w'_2(a_{i_1}, a_{i_2})$ with $1 \leq j \leq p'_{i_1}$ and $1 \leq l \leq p'_{i_1}$.

Having obtained graph G_3, we can apply the (1/3)-approximation algorithm in [2] for the orienteering problem in G_3, which is to find an r-rooted tour P_r^3 to visit as many nodes in G_3 as possible, while ensuring that the weighted sum of edges in the tour is no greater than the UAV energy capacity. For example, Fig. 4(d) shows that the found tour by the algorithm is $r \rightarrow v_{1,1,1} \rightarrow v_{1,1,2} \rightarrow v_{1,2,1} \rightarrow r$. The found tour indicates that the UAV will collect data only from cluster C_1 for two time units, since virtual nodes $v_{1,1}$ and $v_{1,2}$ are constructed from cluster C_1, while the UAV do not collect any data from cluster C_2.

We refer to the approximation algorithm as Algorithm 1.

E. Algorithm Analysis

Theorem 1: Given the amounts of to-be-collected data D_i^C and residual energy e_i^{Budget} of each sensor v_i with $1 \leq i \leq n$, an energy-constrained UAV located at depot r, and a data collection time unit δ, assume that the n sensors are partitioned into Q clusters $C_1, C_2, \ldots, C_Q$. There is a $[1/(6+\epsilon)]$-approximation algorithm, i.e., Algorithm 1, for the data collection utility maximization problem, where ϵ is a given constant with $\epsilon > 0$.

Proof: The proof is contained in the supplementary file. ■

V. Performance Evaluation

A. Experimental Environment

We consider an IoT network deployed in a 1 km × 1 km area. The number of sensors in the network is from 100 to 500, which are randomly placed in the area. We adopt real temperature sensing data of 54 sensors for 31 days (from 1 March to 31 March in 2004) from the Intel Berkeley Research Lab [9]. Then, the sensing data of a sensor is randomly chosen from the 54 sensors. The battery energy capacity of each sensor is 10.8 kJ [7]. Each sensor is equipped with a solar panel and its size is 37 mm × 37 mm. We use real 89 energy harvesting profiles from the National Renewable Energy Laboratory [17] (from 1 March to 31 March in 2023).

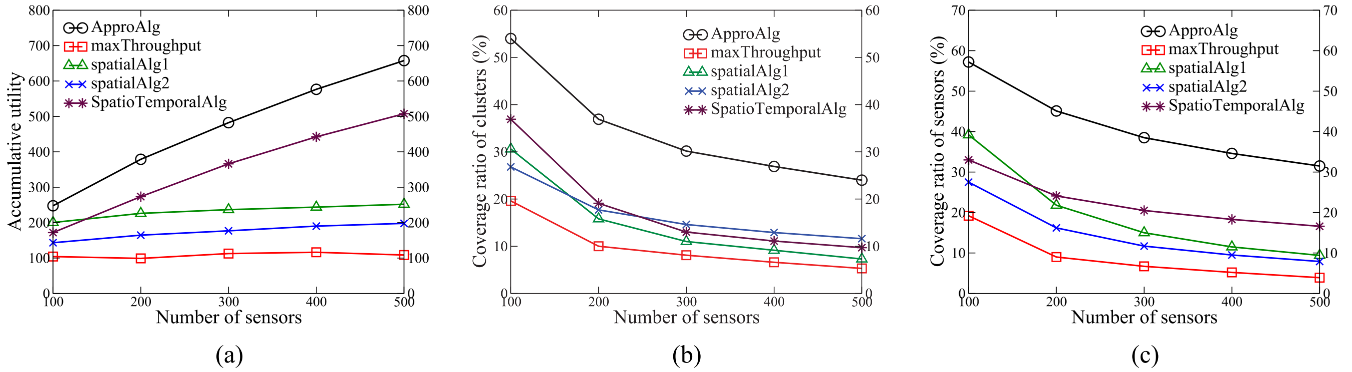


Fig. 5. Performance of different algorithms by varying the number of sensors from 100 to 500. (a) Accumulative utility of collected data. (b) Coverage ratio of clusters. (c) Coverage ratio of sensors.

Algorithm 1 Algorithm `ApproAlg` for the Data Collection Utility Maximization Problem

Input: The amounts of to-be-collected data D_i^C and residual energy e_i^{Budget} of each sensor v_i with $1 \le i \le n$, Q clusters $C_1, C_2, \ldots, C_Q$ of the n sensors, a data collection time unit δ, and a performance loss value ϵ with $\epsilon > 0$

Output: A data collection flying tour of a UAV, and the data collection duration of each sensor in each cluster

/* First graph transformation */

For each cluster C_i, transform it into K_i virtual nodes $v_{i,1}, v_{i,2}, \ldots, v_{i,K_i}$, where the node weight $h(v_{i,l})$ of each virtual node $v_{i,l}$ is the UAV data collection energy consumption for one time slot, and its profit $p_{i,l}$ is the increased accumulative utility when the UAV collects data from a sensor at the lth time slot. Construct graph $G_1(V_1 \cup \{r\}, E_1)$, where V_1 is the set of the virtual nodes constructed from the Q clusters;

/* Second graph transformation */

Construct a graph $G_2(V_2 \cup \{r\}, E_2)$ from G_1, where $V_2 = V_1$, $E_2 = E_1$, the edge weight $w_2(v_{i_1,l_1}, v_{i_2,l_2})$ between any two virtual nodes v_{i_1,l_1} and v_{i_2,l_2} is $w_2(v_{i_1,l_1}, v_{i_2,l_2}) = w_1(v_{i_1,l_1}, v_{i_2,l_2}) + \frac{h(v_{i_1,l_1})+h(v_{i_2,l_2})}{2}$;

/* Third graph transformation */

Let $\theta \leftarrow \frac{\epsilon}{2}$, $n_2 \leftarrow |V_2|$;

Let $\lambda \leftarrow \frac{p_{\max}}{\lceil n_2/\theta \rceil}$, where $p_{\max}$ is the maximum profit of nodes in G_2;

Construct a graph $G_2'(V_2 \cup \{r\}, E_2)$ from G_2, where the edge weight $w_2'(a, b)$ between any two nodes a and b in G_2' is its weight $w_2(a, b)$ in G_2, and the profit of a node $a_i \in V_2$ in graph G_2' is $p_i' = \lfloor \frac{p_i}{\lambda} \rfloor$;

Construct a graph $G_3(V_3 \cup \{r\}, E_3)$ from G_2', where p_i' nodes $a_{i,1}, a_{i,2}, \ldots, a_{i,p_i'}$ are added to V_3 for each node $a_i \in V_2$ in G_2';

Find an r-rooted tour P_r^3 in G_3 by applying the $\frac{1}{3}$-approximation algorithm in [2] for the orienteering problem;

Construct a tour P_r^2 in graph G_2 from P_r^3;

Construct a tour P_r^1 in G_1 and calculate the data collection duration of each sensor in every cluster from P_r^2.

The energy harvesting profile of a sensor is randomly chosen from the 89 energy profiles. We consider the employment of a DJI Phantom 4 RTK UAV to collect data from sensors [5], its battery capacity is 89.2 Wh and its maximum flying time is 30 min. The UAV hovers at a height $h = 300$ m to collect data from sensors in each cluster.

We compare the proposed algorithm `ApproAlg` with following four benchmark algorithms.

1) Algorithm `maxThroughput` [13] considers neither spatial nor temporal data correlations, which schedules an energy-constrained UAV to maximize the amount of collected data from sensors.
2) Algorithm `spatialAlg1` [27] takes spatial data correlations into consideration.
3) Algorithm `spatialAlg2` [15] also considers spatial data correlations, but adopts a different approach from the one in [27].
4) Algorithm `SpatiotemporalAlg` [10] incorporates both spatial and temporal data correlations.

We consider the following three performance metrics.

1) *Accumulative utility* measures the quality of collected data with spatiotemporal correlations and was defined in Section III-F.
2) *Coverage ratio of clusters*. A cluster is collected by the UAV if the sensing data of a sensor in the cluster is collected by the UAV. The coverage ratio of clusters is defined as the ratio of the number of clusters collected by the UAV to the number of clusters Q in the network.
3) *Coverage ratio of sensors* is the ratio of the number of sensors collected by the UAV to the total number n of sensors.

B. Algorithm Performance

We first study the performance of different algorithms by varying the number of sensors from 100 to 500. Fig. 5(a) shows that the accumulative utility by each of the five algorithms increases with the number n of sensors in the network, as the UAV flying energy consumption is smaller in a denser network, and the UAV thus has more energy to collect data from more sensors. Fig. 5(a) also shows that the accumulative utility by the proposed algorithm `ApproAlg` is from 23% to 38% larger than those by the other four algorithms, as the UAV in the solution found by algorithm `ApproAlg` collects data from more clusters in the fine-grained data collection model.

Fig. 5(b) demonstrates that the coverage ratio of clusters collected by algorithm `ApproAlg` is from 45% to 105% larger than those by the other four algorithms. For example, Fig. 5(b) plots that the coverage ratios of clusters by

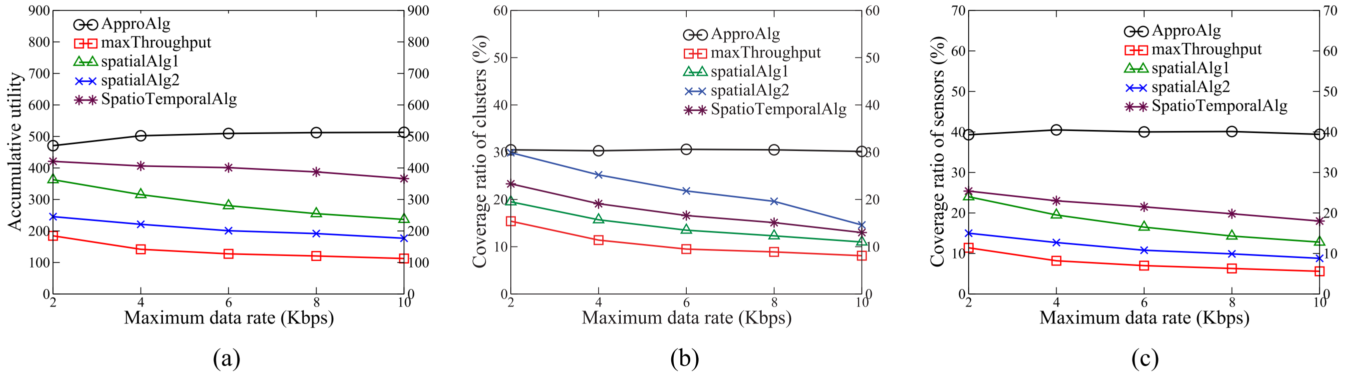


Fig. 6. Performance of different algorithms by increasing the maximum data rate $R_{\max}$ from 2 to 10 kb/s, when there are 300 sensors. (a) Accumulative utility of collected data. (b) Coverage ratio of clusters. (c) Coverage ratio of sensors.

algorithms `ApproAlg`, `maxThroughput`, `spatialAlg1`, `spatialAlg2`, and `SpatiotemporalAlg` are 54%, 37%, 30%, 27%, and 24%, respectively, when there are 100 sensors in the network. Fig. 5(b) also shows that the coverage ratio of clusters by each of the five algorithms decreases when there are more sensors, as the UAV is energy constrained and the number of clusters collected by the UAV will not increase too much with the growth of the number of sensors. Fig. 5(c) plots that the coverage ratio of sensors by `ApproAlg` is from 46% to 90% larger than those by the other four algorithms. In addition, it can be seen from Fig. 5(b) and (c) that the coverage ratio of clusters by algorithm `ApproAlg` is smaller than its coverage ratio of sensors. For example, its coverage ratio of clusters is 24% while its coverage ratio of sensors is 31%, when there are 500 sensors. The rationale behind is that the UAV in the solution delivered by algorithm `ApproAlg` is more likely to collect data from clusters with many sensors, and barely collect data from clusters with only a few sensors, when the energy in the UAV is constrained.

We then evaluate the algorithm performance by varying the maximum data rate $R_{\max}$ from 2 to 10 kb/s, where the data rate of a sensor is randomly chosen in an interval $[0, R_{\max}]$. Fig. 6(a) shows that the accumulative utility by algorithm `ApproAlg` is from 12% to 40% larger than those by the other four algorithms. In addition, the accumulative utility by algorithm `ApproAlg` slightly increases with the growth of the maximum data rate $R_{\max}$, whereas the accumulative utilities by the other four algorithms decrease with the growth of $R_{\max}$. The rationale behind the phenomenon is that, although there are more to-be-collected data in each sensor with the growth of $R_{\max}$, the UAV in the solution by the algorithm `ApproAlg` collects data from more clusters than the other four algorithms, and both the coverage ratios of clusters and sensors do not change too much with the growth of $R_{\max}$, see Fig. 6(b) and (c). In contrast, the UAV by algorithm `maxThroughput` collects more data from each sensor, while the UAV by each of the three algorithms `spatialAlg1`, `spatialAlg2`, and `SpatiotemporalAlg` collects more compressed data from each sensor or each cluster. Then, the UAV by each of the four algorithms `maxThroughput`, `spatialAlg1`, `spatialAlg2`, and `SpatiotemporalAlg` will collect data from less clusters and less sensors, see Fig. 6(b) and (c).

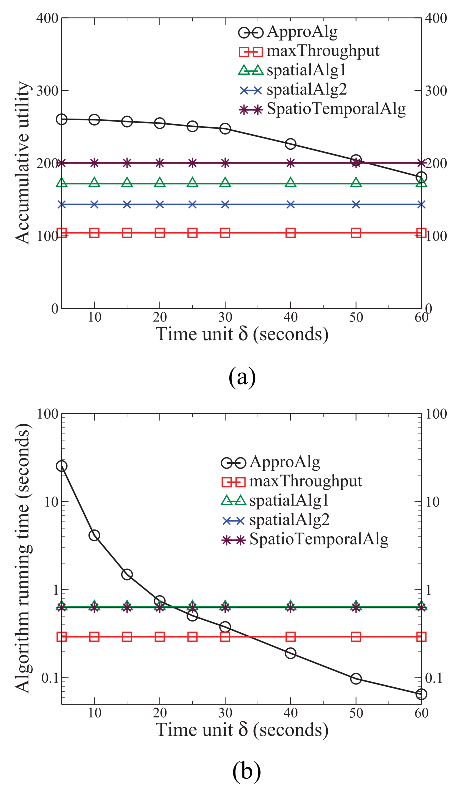


Fig. 7. Performance of different algorithms by varying the data collection time unit δ from 5 to 60 s, when there are 100 sensors. (a) Accumulative utility of collected data. (b) Algorithm running time.

We finally investigate the algorithm performance by increasing the data collection time unit δ from 5 to 60 s, when there are 100 sensors. Fig. 7(a) plots that the accumulative utility by each of the four existing algorithms `maxThroughput`, `spatialAlg1`, `spatialAlg2`, and `SpatiotemporalAlg` does not change with the growth of δ, as the raw data or compressed data in each cluster in the solutions by the four algorithms will be fully collected, if the UAV collects data from the cluster. In contrast, Fig. 7(a) demonstrates that the accumulative utility by algorithm `ApproAlg` slightly decreases when the time unit δ decreases from 5 to 20 s, and the utility becomes even smaller than that by algorithm SpatiotemporalAlg when δ is 60 s. The rationale behind the phenomenon is as follows. Recall that, in the proposed algorithm ApproAlg, if the UAV collects data from a sensor v_j, its data collection time is one of the values in set $\{\delta, 2\delta, \ldots, k_j\delta\}$, where $k_j = \lceil (t_j^{\max}/\delta) \rceil$, and $t_j^{\max}$ is the maximum time for collecting all data from sensor v_j. Then, if the value of the data collection time unit δ is larger, the UAV collects more data in each time unit, and thus has less options on the data collection time. In addition, the difference between the two values $k_j\delta$ and $t_j^{\max}$ is larger, since $0 \leq k_j\delta - t_j^{\max} < \delta$, where $k_j\delta = \lceil t_j^{\max}/\delta \rceil \delta$. For example, the maximum data collection time $t_j^{\max}$ of a sensor v_j is 50 s, while the data collection time unit δ is 60 s. The UAV will spend one data collection time unit for 60 s to collect data from sensor v_j, and waste $10 (= 60 - 50)$ s on the data collection. This indicates that, if the data collection time unit δ is too large, the performance of the algorithm ApproAlg may be worse than the existing algorithms that collect all raw data or compressed data from each cluster.

On the other hand, Fig. 7(b) shows that the running time of algorithm ApproAlg becomes smaller when the time unit δ is larger, since less number k_j of virtual nodes are constructed from each sensor in the first graph transformation of the algorithm ApproAlg, and the running time of the algorithm is proportional to the number of constructed virtual nodes from all sensors, where $k_j = \lceil t_j^{\max}/\delta \rceil$. In contrast, the running time of each of the other four algorithms does not change with the increase of δ, as no virtual nodes are constructed in the four algorithms.

VI. Conclusion

Unlike existing studies that adopted the coarse-grained data collection model, in which all compressed data in each cluster must be collected to recover the original raw data, if the UAV collects data from the cluster, in this article we used a light-weight fine-grained data collection model in which compressed data can be partially collected.

Under the fine-grained data collection model, we investigated a problem of finding a data collection trajectory for an energy-constrained UAV, so that the accumulative utility of collected data is maximized. We proposed a novel $[1/(6 + \epsilon)]$-approximation algorithm for the problem, by transforming the problem to the orienteering problem via a series of three smart graph transformations, where ϵ is a given constant with $\epsilon > 0$. In the first graph transformation, we exploited spatial-temporal similarities and transformed each cluster into several virtual nodes so that each virtual node is associated with a node weight (i.e., UAV data collection energy consumption) and a profit (i.e., the utility of visiting the virtual node). In the second graph transformation, we transformed the both node-weighted and edge-weighted graph obtained in the first graph transformation into an only edge-weighted graph. In the last transformation, we transformed each virtual node into several nodes with uniform profits, through a scaling technique, so that the approximation algorithm for the orienteering problem can be applied in the graph obtained in the third graph transformation.

Experimental results with real data sets showed that the accumulative utility by the proposed algorithm is at least 23% larger than those by the existing studies, and the number of clusters collected by the proposed algorithm is from 45% to 105% larger than those by the existing studies.

References

[1] A. Al-Hourani, S. Kandeepan, and S. Lardner, “Optimal LAP altitude for maximum coverage,” *IEEE Wireless Commun. Lett.*, vol. 3, no. 6, pp. 569–572, Dec. 2014.
[2] N. Bansal, A. Blum, S. Chawla, and A. Meyerson, “Approximation algorithms for deadline-TSP and vehicle routing with time-windows,” in *Proc. 36th Annu. ACM Symp. Theory Comput. (STOC)*, 2004, pp. 166–174.
[3] A. Boubrima, W. Bechkit, and H. Rivano, “Optimal WSN deployment models for air pollution monitoring,” *IEEE Trans. Wireless Commun.*, vol. 16, no. 5, pp. 2723–2735, May 2017.
[4] S. Chen, P. Sinha, N. B. Shro, and C. Joo, “A simple asymptotically optimal joint energy allocation and routing scheme in rechargeable sensor networks,” *IEEE/ACM Trans. Netw.*, vol. 22, no. 4, pp. 1325–1336, Aug. 2014.
[5] “DJI Phantom 4 RTK.” Accessed: Apr. 15, 2023. [Online]. Available: https://www.dji.com/cn/phantom-4-rtk
[6] M. Fattoum, Z. Jellali, and L. N. Atallah, “Adaptive sampling approach exploiting spatio-temporal correlation and residual energy in periodic wireless sensor networks,” *IEEE Access*, vol. 11, pp. 7670–7681, 2023.
[7] Z. Guo et al., “Minimizing redundant sensing data transmissions in energy-harvesting sensor networks via exploring spatial data correlations,” *IEEE Internet Things J.*, vol. 8, no. 1, pp. 512–527, Jan. 2021.
[8] H. Hu, K. Xiong, G. Qu, Q. Ni, P. Fan, and K. Letaief, “AoI-minimal trajectory planning and data collection in UAV-assisted wireless powered IoT networks,” *IEEE Internet Things J.*, vol. 8, no. 2, pp. 1211–1223, Jan. 2021.
[9] “Intel Berkeley Research Lab.” Accessed: Apr. 15, 2023. [Online]. Available: http://db.csail.mit.edu/labdata/labdata.html
[10] G. Li, B. He, Z. Wang, X. Cheng, and J. Chen, “Blockchain-enhanced spatiotemporal data aggregation for UAV-assisted wireless sensor networks,” *IEEE Trans. Ind. Informat.*, vol. 18, no. 7, pp. 4520–4530, Jul. 2022.
[11] J. Li and P. Mohapatra, “Analytical modeling and mitigation techniques for the energy hole problem in sensor networks,” *Pervasive Mobile Comput.*, vol. 3, no. 3, pp. 233–254, 2007.
[12] S. Lin et al., “ATPC: Adaptive transmission power control for wireless sensor networks,” *ACM Trans. Sensor Netw.*, vol. 12, no. 1, pp. 1–31, 2016.
[13] Y. Li et al., “Data collection maximization in IoT-sensor networks via an energy-constrained UAV,” *IEEE Trans. Mobile Comput.*, vol. 22, no. 1, pp. 159–174, Jan. 2023.
[14] F. Liu, M. Lin, Y. Hu, C. Luo, and F. Wu, “Design and analysis of compressive data persistence in large-scale wireless sensor networks,” *IEEE Trans. Parallel Distrib. Syst.*, vol. 26, no. 10, pp. 2685–2698, Oct. 2015.
[15] X. Liu, Y. Liu, N. Zhang, W. Wu, and A. Liu, “Optimizing trajectory of unmanned aerial vehicles for efficient data acquisition: A matrix completion approach,” *IEEE Internet Things J.*, vol. 6, no. 2, pp. 1829–1840, Apr. 2019.
[16] L. Ma, X. Wang, X. Wang, L. Wang, Y. Shi, and M. Huang, “TCDA: Truthful combinatorial double auctions for mobile edge computing in Industrial Internet of Things,” *IEEE Trans. Mobile Comput.*, vol. 21, no. 11, pp. 4125–4138, Nov. 2022.
[17] “Measurement and instrumentation data center.” Accessed: Apr. 15, 2023. [Online]. Available: http://midcdmz.nrel.gov/
[18] M. Mozaffari, W. Saad, M. Bennis, and M. Debbah, “Unmanned aerial vehicle with underlaid device-to-device communications: Performance and tradeoffs,” *IEEE Trans. Wireless Commun.*, vol. 15, no. 6, pp. 3949–3963, Jun. 2016.
[19] D. K. Sah et al., “Renewable energy harvesting schemes in wireless sensor networks: A survey,” *Inf. Fusion*, vol. 63, pp. 223–247, Nov. 2020.
[20] S. Sarang, G. M. Stojanović, M. Drieberg, S. Stankovski, K. Bingi, and V. Jeoti, “Machine learning prediction based adaptive duty cycle MAC protocol for solar energy harvesting wireless sensor networks,” *IEEE Access*, vol. 11, pp. 17536–17554, 2023.
[21] J. Singh, R. Kaur, and D. Singh, “Energy harvesting in wireless sensor networks: A taxonomic survey,” *Int. J. Energy Res.*, vol. 45, no. 1, pp. 118–140, 2021.

[22] F. Shan, J. Luo, R. Xiong, W. Wu, and J. Li, “Looking before crossing: An optimal algorithm to minimize UAV energy by speed scheduling with a practical flight energy model,” in *Proc. IEEE Int. Conf. Comput. Commun. (INFOCOM)*, 2020, pp. 1758–1767.

[23] H.-C. Tsai, Y.-W. P. Hong, and J.-P. Sheu, “Completion time minimization for UAV-enabled surveillance over multiple restricted regions,” *IEEE Trans. Mobile Comput.*, vol. 22, no. 12, pp. 6907–6920, Dec. 2023.

[24] X. Xie, X. Liu, H. Qi, B. Xiao, K. Li, and J. Wu, “Geographical correlation-based data collection for sensor-augmented RFID systems,” *IEEE Trans. Mobile Comput.*, vol. 19, no. 10, pp. 2344–2357, Oct. 2020.

[25] W. Xu et al., “Approximation algorithms for the generalized team orienteering problem and its applications,” *IEEE/ACM Trans. Netw.*, vol. 29, no. 1, pp. 176–189, Feb. 2021.

[26] G. Xu, E. C. H. Ngai, and J. Liu, “Ubiquitous transmission of multimedia sensor data in Internet of Things,” *IEEE Internet Things J.*, vol. 5, no. 1, pp. 403–414, Feb. 2018.

[27] T. Yu, X. Wang, and A. Shami, “UAV-enabled spatial data sampling in large-scale IoT systems using denoising autoencoder neural network,” *IEEE Internet Things J.*, vol. 6, no. 2, pp. 1856–1865, Apr. 2019.

[28] Y. Zeng, J. Xu, and R. Zhang, “Energy minimization for wireless communication with rotary-wing UAV,” *IEEE Trans. Wireless Commun.*, vol. 18, no. 4, pp. 2329–2345, Apr. 2019.

[29] C. Zhan, H. Hu, J. Wang, Z. Liu, and S. Mao, “Tradeoff between age of information and operation time for UAV sensing over multi-cell cellular networks,” *IEEE Trans. Mobile Comput.*, early access, Apr. 17, 2023, doi: 10.1109/TMC.2023.3267656.

[30] Y. Zhang, J. Lyu, and L. Fu, “Energy-efficient trajectory design for UAV-aided maritime data collection in wind,” *IEEE Trans. Wireless Commun.*, vol. 21, no. 12, pp. 10871–10886, Dec. 2022.

[31] T. Zhu, J. Li, H. Gao, and Y. Li, “Data aggregation scheduling in battery-free wireless sensor networks,” *IEEE Trans. Mobile Comput.*, vol. 21, no. 6, pp. 1972–1984, Jun. 2022.

Wenzheng Xu (Member, IEEE) received the B.Sc., M.E., and Ph.D. degrees in computer science from Sun Yat-Sen University, Guangzhou, China, in 2008, 2010, and 2015, respectively.

He currently is an Associate Professor with Sichuan University, Chengdu, China. Also, he was a Visitor with both The Australian National University, Canberra, ACT, Australia, and the Chinese University of Hong Kong, Hong Kong. He has published more than 100 papers on prestigious journals and conferences, such as IEEE/ACM TRANSACTIONS ON NETWORKING, IEEE TRANSACTIONS ON COMPUTERS, IEEE TRANSACTIONS ON MOBILE COMPUTING, IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, INFOCOM, and ICDCS. His research interests include Internet of Things, UAV networks, mobile computing, approximation algorithms, combinatorial optimization, online social networks, and graph theory.

Heng Shao received the B.E. degree in computer science and technology from Sichuan University, Chengdu, China, in 2022, where he is currently pursuing the master’s degree with the College of Computer Science.

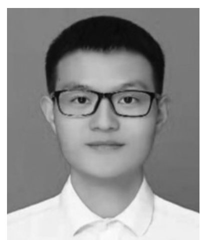
His current research interests include IoT networks and social networks.

Qunli Shen received the B.E. degree in network engineering from Zhejiang Normal University, Jinhua, China, in 2021. He is currently pursuing the master’s degree with the College of Computer Science, Sichuan University, Chengdu, China.

His current research interests include UAV networking and IoT networks.

Jian Peng received the B.A. and Ph.D. degrees from the University of Electronic Science and Technology of China, Chengdu, China, in 1992 and 2004, respectively.

He is a Professor with the College of Computer Science, Sichuan University, Chengdu. His recent research interests include Internet of Things, UAV networks, big data, and cloud computing.



Wen Huang received the Ph.D. degree from the School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu, China, in 2022.

He is working a Postdoctoral Fellow with the College of Computer Science, Sichuan University, Chengdu. His main research interests include cryptography, differential privacy, and IoT networks.

Weifa Liang (Senior Member, IEEE) received the B.Sc. degree in computer science from Wuhan University, Wuhan, China, in 1984, the M.E. degree in computer science from the University of Science and Technology of China, Hefei, China, in 1989, and the Ph.D. degree in computer science from The Australian National University, Canberra, ACT, Australia, in 1998.

He currently is a Professor with the Department of Computer Science, City University of Hong Kong, Hong Kong. Prior to the current position, he was a Professor with The Australian National University. His research interests include design and analysis of energy-efficient routing protocols for wireless ad hoc and sensor networks, Internet of Things, edge and cloud computing, network function virtualization and software-defined networking, design and analysis of parallel and distributed algorithms, approximation algorithms, combinatorial optimization, and graph theory.

Prof. Liang serves as an Associate Editor for the IEEE TRANSACTIONS ON COMMUNICATIONS.

Tang Liu (Member, IEEE) received the B.S. degree in computer science from the University of Electronic and Science of China, Chengdu, China, in 2003, and the M.S. and Ph.D. degrees in computer science from Sichuan University, Chengdu, in 2009 and 2015, respectively.

He is a Professor with the College of Computer Science, Sichuan Normal University, Chengdu. From 2015 to 2016, he was a Visiting Scholar with the University of Louisiana at Lafayette, Lafayette, LA, USA. His research interests include wireless charging and wireless sensor networks.

Xin-Wei Yao (Senior Member, IEEE) received the Ph.D. degree in information engineering from the Zhejiang University of Technology, Hangzhou, China, in 2013.

He is currently an Associate Professor with the College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou. From March 2012 to February 2013, he was a Visiting Scholar with Loughborough University, Leicestershire, U.K. From August 2015 to July 2016, he was a Visiting Professor with the University of Buffalo, Amherst, NY, USA; and The State University of New York, Albany, NY, USA. His current research interests include the area of Terahertz-band communication networks, electromagnetic nanonetworks, wireless ad hoc and sensor networks, wireless power transfer, and the Internet of Things.

Dr. Yao was the recipient of the Distinguished Associate Professor Award and the Outstanding Doctoral Thesis Award with the Zhejiang University of Technology. He has served on technical program committees of many IEEE/ACM conferences. He is a member of ACM.

Tao Lin received the M.Sc. degree in computer science from Sichuan University, Chengdu, China, in 2003, and the Ph.D. degree in information science from the University of Yamanashi, Kofu, Japan, in 2007.

He is a Full Professor with the College of Computer Science, Sichuan University. His recent research interests include intelligent software engineering and optimization.

Sajal K. Das (Fellow, IEEE) received the B.S. degree from the University of Calcutta, Kolkata, India, in 1983, the M.S. degree from the Indian Institute of Science, Bengaluru, India, in 1984, and the Ph.D. degree from the University of Central Florida, Orlando, FL, USA, in 1988.

He directed numerous funded projects in these areas totaling over $15M and published extensively with more than 600 research articles in high quality journals and refereed conference proceedings. His current research interests include theory and practice of wireless sensor networks, big data, cyber–physical systems, smart health-care, distributed and cloud computing, security and privacy, biological and social networks, applied graph theory and game theory.

Dr. Das is the Chair of Computer Science Department and the Daniel St. Clair Endowed Chair of the Missouri University of Science and Technology in the States. He serves as the Founding Editor-in-Chief of the *Pervasive and Mobile Computing*, and as an Associate Editor of IEEE TRANSACTIONS ON MOBILE COMPUTING and *ACM Transactions on Sensor Networks*. He is a Co-Founder of the IEEE PerCom, IEEE WoWMoM, and ICDCN conferences, and served on numerous conference committees as a general chair, a program chair, or a program committee member.