Minimizing Redundant Sensing Data Transmissions in Energy-Harvesting Sensor Networks via Exploring Spatial Data Correlations

Zhenjie Guo, Jian Peng, Wenzheng Xu, *Member, IEEE*, Weifa Liang, *Senior Member, IEEE*, Weigang Wu, *Member, IEEE*, Zichuan Xu, *Member, IEEE*, Bing Guo, and Yue Ivan Wu, *Member, IEEE*

***Abstract*—Energy harvesting rates of sensors in renewable (e.g., solar energy) wireless sensor networks are not only lower than their energy consumption rates but also temporally varying. Existing studies exploited spatial data correlations among sensors to reduce their energy consumptions, where the data correlations mean that the sensing data of nearby sensors have high similarities. They assumed that the sensing data of nearby sensors are very likely to highly correlated. They adopted a coarse-grained spatial-correlation model, in which sensors are partitioned into different clusters such that the sensors in the same cluster have high data similarities with each other. Then, only the sensor with the maximum residual energy in each cluster sends its sensing data, while the other sensors do not. We, however, notice that the data similarities among nearby sensors in real sensor networks may vary significantly, i.e., ranging from very similar to not similar at all. Since the existing algorithms require that the sensors in the same cluster have high data similarities with each other, the sensors in a network may be partitioned into many clusters and each cluster consists of only a few sensors, where two nearby sensors belong to two different clusters if the sensing data of the two sensors are not highly correlated. Therefore, in the existing studies, many sensors have to send all their data as there are many clusters. Unlike the existing studies, in this article, we first propose a fine-grained spatial correlation model, in which sensors are partitioned into only a few clusters and each cluster consists of many sensors. Then, each cluster master sensor sends all its data to the sink, while the majority of other sensors in the cluster transmit only their nonredundant data, thereby significantly saving sensor energy consumptions. We formulate a novel sensor clustering problem under the proposed model, which is to partition sensors into different clusters and choose a representative sensor for each cluster such that the amount of suppressed redundant data transmissions is maximized. We propose a randomized $(0.5-\epsilon)$-approximation algorithm for the clustering problem, where ϵ is a given constant with $0 < \epsilon \leq 0.5$. To further reduce sensor energy consumption, we consider temporal data correlations, where the sensing data by a sensor in a short period are likely to be highly correlated. We investigate a data utility maximization problem that allocates sensor data rates and routing so that the accumulative utility of both spatially and temporally correlated data received by the sink is maximized. We devise a near-optimal algorithm for the problem. We finally evaluate the performance of the proposed algorithms through experiments. the experimental results show that the proposed algorithms are very promising.**

***Index Terms*—Approximation algorithm, energy-harvesting sensor networks, redundant sensing data, spatial data correlations.**

I. Introduction

Wireless sensor networks (WSNs) have wide applications, including industrial automation systems monitoring, environmental monitoring, smart grid network monitoring, Internet of Things (IoT), etc. [9], [11]. In conventional WSNs, sensors are powered by energy-limited batteries, the network lifetime thus is restricted by sensor batteries, as sensors will not function when they run out of energy. To address the network lifetime issue, some pioneering researchers proposed to enable sensors to harvest energy from their surrounding environment, such as solar energy, wind energy, thermal energy, etc. [4], [28].

It is well known that data transmission and reception between sensors are very energy-consuming [19], while the energy harvesting rate of a sensor usually is lower than its energy consumption rate, since its attached energy harvesting component cannot be too large, due to practical restrictions on

Manuscript received March 9, 2020; revised May 15, 2020; accepted June 20, 2020. Date of publication June 24, 2020; date of current version December 21, 2020. The work of Jian Peng was supported in part by the National Key Research and Development Program of China under Grant 2017YFB0202403; in part by the Sichuan Key Research and Development Program under Grant 2018GZ0098; in part by the Sichuan Science and Technology Program under Grant 2017GZDZX0003; and in part by the Sichuan University–Zigong City Program under Grant 2018CDZG-15. The work of Wenzheng Xu was supported by the National Natural Science Foundation of China (NSFC) under Grant 61602330. The work of Weifa Liang was supported by the Australian Research Council Discovery Project under Grant DP200101985. The work of Weigang Wu was supported by the Guangdong Natural Science Foundation of China under Grant 2018A030313492. The work of Bing Guo was supported in part by NSFC under Grant 61772352, and in part by the Science and Technology Planning Project of Sichuan Province under Grant 2019YFG0400, Grant 2018GZDZX0031, Grant 2018GZDZX0004, Grant 2017GZDZX0003, Grant 2018JY0182, and Grant 19ZDYF1286. *(Zhenjie Guo and Jian Peng contributed equally to this work.) (Corresponding author: Wenzheng Xu.)*

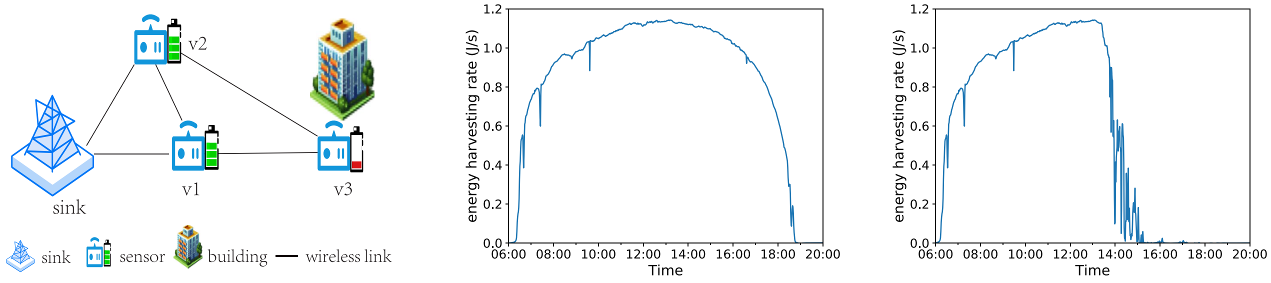
Zhenjie Guo, Jian Peng, Wenzheng Xu, Bing Guo, and Yue Ivan Wu are with the College of Computer Science, Sichuan University, Chengdu 610065, China (e-mail: zhenjieguo1@163.com; jianpeng@scu.edu.cn; wenzheng.xu3@gmail.com; guobing@scu.edu.cn; y.i.wu@ieee.org).

Weifa Liang is with the Research School of Computer Science, Australian National University, Canberra, ACT 2601, Australia (e-mail: wliang@cs.anu.edu.au).

Weigang Wu is with the School of Data and Computer Science, Guangdong Province Key Laboratory of Big Data Analysis and Processing, Key Laboratory of Machine Intelligence and Advanced Computing, and Ministry of Education, Sun Yat-sen University, Guangzhou 510006, China (e-mail: wuweig@mail.sysu.edu.cn).

Zichuan Xu is with the School of Software, Dalian University of Technology, Dalian 116024, China (e-mail: z.xu@dlut.edu.cn).

Digital Object Identifier 10.1109/JIOT.2020.3004554



(a) (b) (c)

Fig. 1. Energy harvesting profiles in a renewable sensor network. (a) Sensor network. (b) Energy harvesting rates of sensors v_1 and v_2 at different time from [37]. (c) Energy harvesting rates of sensor v_3 at different time points from [37].

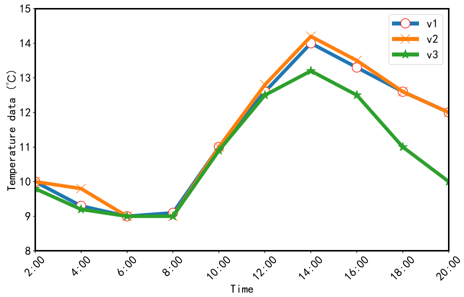


Fig. 2. Temperature sensing data by three sensors v_1, v_2, and v_3 from 2:00 to 20:00 in a day.

TABLE I
TEMPERATURE DATA MONITORED BY SENSORS v_1, v_2, AND v_3 FROM 2:00 TO 20:00 IN A DAY

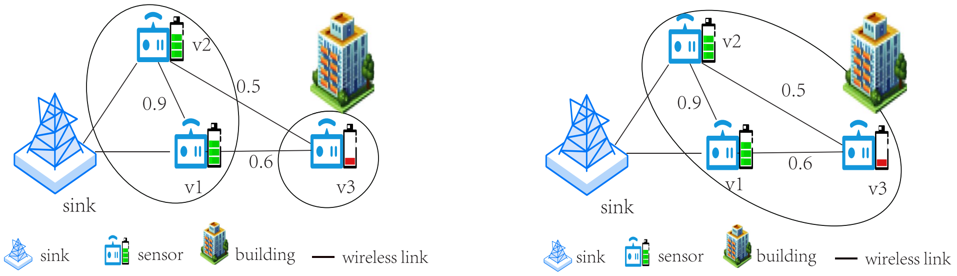
Time	data of v_1	data of v_2	data of v_3
2:00	10	10	9.8
4:00	**9.3**	**9.8**	9.2
6:00	9	9	9
8:00	9.1	9	9
10:00	11	11	10.9
12:00	12.6	12.8	12.5
14:00	**14**	14.2	**13.2**
16:00	**13.3**	13.5	**12.5**
18:00	**12.6**	12.6	**11**
20:00	**12**	12	**10**

sensor size or its cost [9], [11]. In addition, the energy harvesting rate of a sensor is temporally and spatially varying [37]. For example, Fig. 1(a) shows a sensor network, which consists of three sensors v_1, v_2, and v_3 and a sink for temperature monitoring. Fig. 1(b) plots the energy harvesting rates of sensors v_1 and v_2 at different time points in a day [6], [37]. In contrast, Fig. 1(c) plots that the energy harvesting rate of sensor v_3 in the afternoon is very low since it is within the shadow area of a nearby building. Due to the variability of the harvested energy in renewable sensor networks, one fundamental question is how to make full use of the harvested energy such that the amount of data collected from sensors is maximized.

It is recognized that sensing data generated by nearby sensors exhibit high similarities, which is referred to as *spatial data correlations* [8], [29], [35]. For example, Fig. 2 shows the temperature sensing data by three sensors v_1, v_2, and v_3 from 2:00 to 20:00 in a day, and Table I keeps their detailed records [16]. It can be seen from Fig. 2 that the temperature data by sensors v_1 and v_2 are very similar with each other, while they have some deviations only at time 4:00. In contrast, the sensing data of v_3 after 14:00 have large differences from that of both v_1 and v_2, where the difference of the temperature data from two sensors at a time slot is referred to be *large* if the difference is larger than a given threshold γ, e.g., $\gamma = 0.5$. Following the existing studies [16], [22], the data similarities among the three sensors can be calculated as follows. The data similarity $c(1, 2)$ between sensors v_1 and v_2 is $[(10 - 1)/10] = 0.9$, since their data have a large difference only at time point 4:00 among the 10 time points in Table I. The data similarities $c(1, 3)$ between sensors v_1 and v_3, and $c(2, 3)$ between v_2 and v_3 can be calculated similarly, which are $[(10 - 4)/10] = 0.6$ and $[(10 - 5)/10] = 0.5$, respectively.

Existing studies proposed to explore the high spatial data correlations among sensors to reduce the transmission of redundant sensing data, thereby saving the energy consumption of energy-limited sensors [22], [32], [35]. They partition sensors into different clusters such that the sensors in each cluster have high spatial data correlations with each other. The sensor with the maximum residual energy in each cluster transmits its sensing data to the sink (acting as a representative node), while the others do not transmit their sensing data. For example, the three sensors v_1, v_2, and v_3 in Fig. 3(a) will be partitioned into two clusters $C_1 = \{v_1, v_2\}$ and $C_2 = \{v_3\}$, due to the high data similarity between v_1 and v_2. Then, sensor v_1 sends its sensing data to the sink, while sensor v_2 in C_1 does not. Also, sensor v_3 transmits its data to the sink, too. Assume that the data generating rate of each of the three sensors is 1 kb/s. Then, the throughput received by the sink is 1 kb/s + 1 kb/s = 2 kb/s, which is less than the accumulative data generating rate of the three sensors, i.e., 1 kb/s $\times 3 = 3$ kb/s, thereby saving the energy consumption of sensor v_2. On the other hand, the data collected by the sink still maintain high quality.

Although existing studies have conducted excellent work on the exploitation of spatial data correlations, such exploitation is in the *coarse-grained level*, i.e., each sensor either transmits its sensing data to the sink or not. In this article, we observe



(a) (b)

Fig. 3. Coarse-grained spatial-correlation model in the existing studies versus the fine-grained spatial-correlation model in this article. (a) Example of the coarse-grained spatialcorrelation model, where $v1$ and $v2$ form a cluster, $v3$ itself forms another cluster, and only sensors $v1$ and $v3$ transmit their data to the sink. and (b) Example of the fine-grained spatialcorrelation model in this paper, where sensor $v1$ sends all its data back to the sink, while both $v2$ and $v3$ transmit their nonredundant data to the sink, and data sent by the three sensors are shown in Table. 2.

TABLE II
TEMPERATURE DATA TRANSMITTED BY v_1, v_2, AND v_3 WHEN CONSIDERING THE SPATIAL DATA CORRELATIONS AMONG THE THREE SENSORS

Time	data of v_1	data of v_2	data of v_3
2:00	10		
4:00	9.3	9.8	
6:00	9		
8:00	9.1		
10:00	11		
12:00	12.6		
14:00	14		13.2
16:00	13.3		12.5
18:00	12.6		11
20:00	12		10

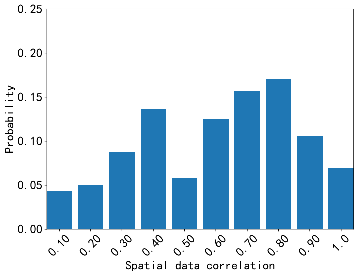


Fig. 4. Distribution of real spatial data correlations between nearby sensors in a sensor network from [12].

that although some sensor v may not have a very high spatial correlation with its nearby sensor u, their sensing data indeed have some similarity to a certain degree. We make statistics of real spatial temperature data correlations between nearby sensors in a sensor network deployed by the Intel Berkeley Research laboratory [12], where 54 sensors are deployed in the network. Fig. 4 shows the distribution of the spatial data correlations, from which it can be seen that the spatial data correlations vary from very small to very large.

In addition, Fig. 3(a) shows that sensor v_3 has a data similarity $c(1, 3) = 0.6$ with sensor v_1. Following the existing studies [22], [32], [35], sensor v_3 should transmit its sensing data to the sink, since the data similarity $c(1, 3)$ between v_1 and v_3 is not high enough, e.g., $c(1, 3)$ should be no less than 0.9, while $c(1, 3) = 0.6$.

In contrast, in this work, we exploit a *fine-grained spatial correlation model*, in which some sensors send all their sensing data, while the others transmit only nonredundant data. For example, in Fig. 3(b), sensor v_1 sends all its data back to the sink, while both v_2 and v_3 send their nonredundant data with respect to the data by v_1. Table II shows the data sent back to the sink by sensors v_1, v_2, and v_3, respectively. It can be seen that the total throughput received by the sink is 1.5 kb/s (= 1 kb/s +0.1 kb/s +0.4 kb/s), which is less than that by the existing studies, i.e., 1.5 kb/s< 2 kb/s. On the other hand, the data sent by v_1, v_2, and v_3 under this fine-grained spatial-correlation model still maintain a similar high quality against the data by the existing studies under the coarse-grained spatial-correlation model.

We consider an important clustering problem in the fine-grained model, which is to partition sensors into different clusters, and one sensor in each cluster is chosen to serve as the master sensor, while the others serve as slave sensors. The master sensor in each cluster sends its data to the sink, while the other slave sensors in the cluster transmit only their nonredundant data. The objective is to maximize the amount of suppressed transmissions of redundant data for the entire network.

The clustering problem is very challenging since each chosen master sensor in a cluster should not only have high spatial data correlations with its slave sensors but also have a large amount of available energy. Otherwise, even if the chosen master sensor with a little energy has high spatial data correlations with the data by its slave sensors, it can only generate representative data for a short period. On the other hand, if the chosen master sensor with a large amount of available energy has low spatial data correlations with the data by its slave sensors, then the slave sensors have to send more nonredundant data by themselves, thereby consuming more energy. In this article, we address the challenge by proposing a novel $(0.5 - \epsilon)$-approximation algorithm for the clustering problem, where ϵ is a given constant with $0 < \epsilon \leq 0.5$.

In addition to the spatial data correlation, some researchers also exploited *temporal data correlations*, where sensing data by a sensor in a short period are also highly correlated [6], [37]. For example, it can be seen that the temperature data by v_1 at 6:00 and 8:00 in Table I are similar, i.e., 9 °C and 9.1 °C, respectively.

To further enhance energy savings of sensors, we also explore the temporal data correlations. We refer to this problem as the *data utility maximization problem*, which is to allocate a data rate for each sensor and the routing among sensors for sending the data to the sink such that the utility of the spatially temporally correlated sensing data received by the sink is maximized, subject to the time-varying energy constraints on sensors.

The main contributions of this article are summarized as follows.

1) Unlike the existing studies that adopted the coarse-grained spatial-correlation model, in which each sensor either sends its data to the sink or not, in this article, we introduce a novel fine-grained spatial-correlation model, in which only a small portion of sensors send their representative data and the other majority sensors transmit their nonredundant data, thereby reducing redundant data transmissions within the network.
2) We first consider a clustering problem under the proposed fine-grained spatial-correlation model, which is to partition sensors into different clusters and choose a master sensor for each cluster such that the amount of suppressed redundant data transmissions is maximized. We propose a novel approximation algorithm for the problem, which delivers a $(0.5 - \epsilon)$-approximate solution with high probability, where ϵ is a given constant with $0 < \epsilon \leq 0.5$.
3) We then devise an efficient algorithm for the data utility maximization problem, by allocating a near-optimal data rate for each sensor and routing among sensors such that the utility of the spatially temporally correlated data received by the sink is maximized.
4) We finally evaluate the performance of the proposed algorithms, using real data sets. The experimental results show that the amount of suppressed redundant data transmissions and the accumulative utility by the proposed algorithms are approximately 50% and 17% larger than those by the existing algorithms, respectively.

The remainder of the article is organized as follows. Section II introduces the network model and defines the problems. Section III proposes an approximation algorithm for the clustering problem. Section IV devises a near-optimal algorithm for the data utility maximization problem. Section V evaluates the performance of the proposed algorithms. Section VI reviews related work and Section VII concludes the article.

II. Preliminaries

In this section, we introduce the network model, energy model, spatial data correlation model, and define the problems precisely.

A. Network Model

Consider an energy harvesting sensor network $G = (V \cup \{s\}, E)$, which consists of n sensors $v_1, v_2, \ldots, v_n$ and a sink s, where $n = |V|$ and $v_i \in V$ with $1 \leq i \leq n$. The sensors are randomly deployed in a 2-D space and the sink s is located at the center of the area for sensing data collection. We assume that each sensor v_i in V is powered by a rechargeable battery with capacity B_i, which can harvest energy from its surrounding environment, such as solar energy, wind energy, etc. There is an edge in E between two sensors or between one sensor and the sink if they are within the transmission range of each other. Let $m = |E|$. Denote by $N(v_i)$ the set of neighbors of sensor v_i in G, i.e., $N(v_i) = \{v_j \mid v_j \in V \cup \{s\},\ (v_i, v_j) \in E\}$.

We consider the data collection in the network G for a monitoring period of T (e.g., $T = 24$ h) that is divided into equal time slots $1, 2, \ldots, T$. Each time slot lasts for a duration of τ (e.g., $\tau = 1$ h). The sensing rate of each sensor v_i is R_i, but it sends a portion of its sensing data at a rate of $r_i(t)$ at time slot t since there are some redundant data in the sensing data and these data are not sent to the sink. Then, $0 \leq r_i(t) \leq R_i$. Denote by $f_{ij}(t)$ and $f_{ji}(t)$ the data transmission rates from sensors v_i to v_j, and from v_j to v_i, at time slot t, respectively, where $(v_i, v_j) \in E$. Notice that the data transmitted from sensors v_i to v_j include the data from v_i itself and the data received from other sensors.

B. Energy Model

It is well known that each sensor v_i consumes energy on data sensing, data transmission, and data reception. Denote by $P(v_i, t)$ the total energy consumption of sensor v_i at time slot t, then

$$P(v_i, t) = P_S(v_i, t) + P_T(v_i, t) + P_R(v_i, t) \tag{1}$$

where $P_S(v_i, t)$, $P_T(v_i, t)$, and $P_R(v_i, t)$ are the amounts of energy consumption of v_i for data sensing, data transmission, and reception at time slot t, respectively. The values of $P_S(v_i, t)$, $P_T(v_i, t)$, and $P_R(v_i, t)$ can be precisely calculated as follows. By adopting a real energy consumption model from [14], we have

$$P_S(v_i, t) = \lambda \cdot R_i \cdot \tau \tag{2}$$

$$P_T(v_i, t) = \sum_{v_j \in N(v_i)} f_{ij}(t) \cdot \left(\beta_1 + \beta_2 d_{ij}^{\alpha}\right) \cdot \tau \tag{3}$$

$$P_R(v_i, t) = \sum_{v_j \in N(v_i)} f_{ji}(t) \cdot e_R \cdot \tau \tag{4}$$

where λ, R_i, and τ in (2) are the energy consumption for sensing per unit data, the sensing data rate of sensor v_i, and the duration of a time slot, respectively, $f_{ij}(t)$ and $(\beta_1 + \beta_2 d_{ij}^{\alpha})$ in (3) are the data transmission rate from v_i to v_j, and the energy consumption from v_i to v_j, d_{ij} is the Euclidean distance between sensors v_i and v_j, and $f_{ji}(t)$ and e_{Rx} in (4) are the data transmission rate from v_j to v_i and the energy consumption for receiving per unit data, respectively. Notice that the values of $\lambda, \tau, \beta_1, \beta_2, \alpha$, and e_R are constants [14].

Sensors can harvest energy from their surrounding environments, such as solar energy. Denote by $H(v_i, t)$ the amount of harvested energy in each sensor v_i at each time slot t. Also, denote by $RE(v_i, t-1)$ and $RE(v_i, t)$ the amounts of residual energy of sensor v_i after time slots $t-1$ and t, respectively. Then

$$RE(v_i, t) = \min\{RE(v_i, t-1) + H(v_i, t) - P(v_i, t),\ B_i\} \tag{5}$$

TABLE III
GENERATED DATA BY BOTH SENSORS v_1 AND v_2 IN A PERIOD P

Time slot	Sensing data of v_1	Sensing data of v_2
1	15	14.9
2	13.1	–
3	–	12
4	–	–
5	–	–
6	8	7.9
7	–	7.7
8	8	8.1

TABLE IV
ESTIMATED DATA BY BOTH SENSORS v_1 AND v_2

Time slot	Sensing data of v_1	Sensing data of v_2
1	15	14.9
2	13.1	**13.43**
3	**11.45**	12
4	–	–
5	–	–
6	8	7.9
7	**7.35**	7.7
8	8	8.1

where $H(v_i, d, t)$ and $P(v_i, t)$ are the amounts of generated and consumed energy by sensor v_i at time slot t, and B_i is the battery capacity of sensor v_i.

C. Spatial Data Correlation Model

It is recognized that a group of sensors close with each other may have similar sensing data at the same time. For example, in a WSN for temperature monitoring, the temperature data monitored by two nearby sensors usually are highly correlated. In this case, we may just collect temperature data from one of them, thereby saving the energy of the other for future data collection.

We measure the spatial data correlation between two sensors v_i and v_j by their data similarity for a given long period P, e.g., one month. Let $r_{i,1}, r_{i,2}, \ldots, r_{i,k_i}$ be the sensing data of sensor v_i at time slots $t_{i,1}, t_{i,2}, \ldots, t_{i,k_i}$ with $t_{i,1} \leq t_{i,2} \leq t_{i,3} \leq \cdots \leq t_{i,k_i}$, and let $r_{j,1}, r_{j,2}, \ldots, r_{j,k_j}$ be the sensing data of sensor v_j at time slots $t_{j,1}, t_{j,2}, \ldots, t_{j,k_j}$ with $t_{j,1} \leq t_{j,2} \leq t_{j,3} \leq \cdots \leq t_{j,k_j}$. For example, Table III shows that sensor v_1 generates data at time slots $1, 2, 6,$ and 8, but no data at time slots $3, 4, 5,$ and 7, while sensor v_2 generates data at time slots $1, 3, 6,$ and 8, but no data at time slots $2, 4,$ and 5.

It can be seen from Table III that there are some time slots that one sensor generates data, whereas the other does not, e.g., sensor v_2 has data at time slot $t = 3$, while v_1 does not. In order to measure the spatial data correlation of v_i and v_j, we need to estimate the missed data such that for each time slot t, either both v_i and v_j have data or none of them have data.

Let $r'_{i,t}$ and $r'_{j,t}$ be the predicted data of sensors v_i and v_j at time slot t, respectively. There are three cases to be considered.

Case 1: Both sensors v_i and v_j have data $r_{i,t}$ and $r_{j,t}$, then $r'_{i,t} = r_{i,t}$ and $r'_{j,t} = r_{j,t}$. *Case 2:* None of v_i and v_j have data at time slot t. Then, we do not estimate their data. *Case 3:* Sensor v_i has data at time slot t, but sensor v_j does not. Then, $r'_{i,t} = r_{i,t}$ and we estimate the value of $r'_{j,t}$ as follows. We can adopt a quadratic function or a cubic function to curve fit the missed data. For example, consider v_2 has data $r'_{2,3} = 12$ at time slot $t = 3$, whereas v_1 does not. We estimate $r'_{1,3}$, by noticing that v_1 has data at time slots $t = 1,\ t = 2$, and $t = 6$. Then, we can obtain a quadratic function $y = ax^2 + bx + c$ that passes the three points $(1,\ 15)$, $(2,\ 13.1)$, and $(6,\ 8)$, where $a = 0.125,\ b = -2.275$, and $c = 17.15$, i.e., $y = 0.125x^2 - 1.275x + 17.15$. The missed data of v_1 at $t = 3$ are estimated as $r'_{1,3} = 0.125 \cdot 3^2 - 1.275 \cdot 3 + 17.15 = 11.45$. Table IV shows the estimated data of both sensors v_1 and v_2.

We measure the spatial data correlation between two sensors v_i and v_j by their data similarity for a given long period P, e.g., one month. Let $r_{i,1}, r_{i,2}, \ldots, r_{i,P}$ be the sensing data of sensor v_i at time slots $1, 2, \ldots, P$, respectively. Also, let $r_{j,1}, r_{j,2}, \ldots, r_{j,P}$ be the sensing data of sensor v_j at these P time slots. Denote by $K_{i,j}$ the number of time slots that the difference of the data between v_i and v_j is no greater than a given small threshold γ, i.e., $K_{i,j} = \sum_{l=1}^{P} I(|r_{i,l} - r_{j,l}|)$, where $I(|r_{i,l} - r_{j,l}|) = 1$ if $|r_{i,l} - r_{j,l}| \leq \gamma$; otherwise, $I(|r_{i,l} - r_{j,l}|) = 0$. Table IV indicates that $K_{i,j} = 5$ when $\gamma = 0.5$.

Denote by P' the number of time slots in period P such that both sensors v_i and v_j have data. For example, Table IV shows that $P' = 6$. The spatial data correlation $c(i, j)$ between sensors v_i and v_j is defined as the ratio of $K_{i,j}$ to P', i.e.,

$$c(i, j) = \frac{K_{i,j}}{P'}. \tag{6}$$

For example, the spatial data correlation $c(1, 2)$ between v_1 and v_2 in Table IV then is $c(1, 2) = (5/6)$.

D. Problem Definitions

In an energy-harvesting WSN, it is well known that the energy harvesting rate of each sensor is lower than its energy consumption rate [20]. In order to make full use of limited harvested energy, we can reduce the energy consumption of sensors, by enabling only energy-sufficient sensors to generate representative data for their nearby sensors with less energy due to their data similarities, while the nearby sensors send only a small amount of their nonredundant data to the sink. By doing so, the quality of the collected data may be compromised. To maximize the quality of collected data in energy harvesting sensor networks, we here consider two novel problems. The first one is to partition sensors into different clusters and a representative sensor is chosen for each cluster. The second one is to allocate sensor data rates and routing, such that the accumulative utility of both spatially correlated and temporally correlated data received by the sink is maximized. We formally define the two problems as follows.

1) Redundant Data Suppression Maximization Problem: Given an energy-harvesting WSN $G = (V, E)$, and spatial data correlations $c(v_i, v_j)$ between sensors, we consider the problem of partitioning the sensors in V into k clusters $C_1, C_2, \ldots, C_k$,

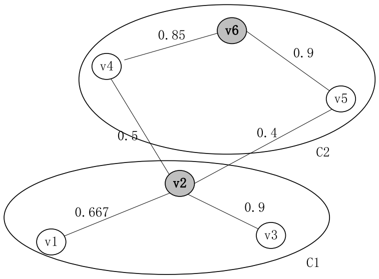


Fig. 5. Six sensors are partitioned into $k = 2$ clusters C_1 and C_2, where $C_1 = \{v_1, v_2, v_3\}$, $C_2 = \{v_4, v_5, v_6\}$, v_2, and v_6 are the master sensors of C_1 and C_2, respectively.

where k is not given and is determined by the data correlation relationships. One sensor u_i in each C_i is chosen as the master sensor, and the rest sensors in $C_i \backslash \{u_i\}$ are the slave sensors of u_i, where each slave sensor v_j in C_i is a neighbor of the master sensor u_i (i.e., $v_j \in N(u_i)$). Then, $C_i \cap C_j = \emptyset$ if $i \neq j$, $\cup_{i=1}^{k} C_i = V$. For example, Fig. 5 shows that six sensors in a sensor network are partitioned into $k = 2$ clusters C_1 and C_2, where $C_1 = \{v_1, v_2, v_3\}$, $C_2 = \{v_4, v_5, v_6\}$, and v_2 and v_6 are the master sensors of C_1 and C_2, respectively.

In the process of data sensing, for each cluster C_i with its master sensor u_i, assume that u_i generates data at a rate of $r_i(t)$ at time slot t. For each slave sensor $v_j \in C_i \backslash \{u_i\}$, v_j can generate at a lower data rate $r_j(t)$ due to the data similarity $c(i, j)$ between v_j and its master sensor u_i.

For each cluster C_i with its master sensor u_i, the total energy budget of u_i in period T is

$$E_i^b = RE(u_i, 0) + \sum_{t=1}^{T} H(u_i, t) \tag{7}$$

where $RE(u_i, 0)$ is the residual energy at the beginning of the monitoring period and $H(u_i, t)$ is the amount of harvested energy at time slot t with $1 \leq t \leq T$. On the other hand, the per sensor energy-consuming rate ρ_i of sensor u_i can be calculated by following the realistic energy model from [14, eq. (7), p. 8], which incorporates the energy consumptions for sensing, data transmission, and data reception. It can be seen that the residual lifetime of sensor u_i is $l_i = (E_i^b/\rho_i)$. The maximum amount of data $D_i^{\max}$ that can be generated by u_i in period T then is

$$D_i^{\max} = R_i \cdot l_i \tag{8}$$

where R_i is the sensing data rate of u_i.

Due to the data similarity $c(i, j)$ between the master sensor u_i of C_i and each slave sensor v_j in $C_i \backslash \{u_i\}$, the amount of *suppressed redundant data transmissions* by v_j due to the generated data $D_i^{\max}$ by u_i is $D_i^{\max} \cdot c(i, j)$, where $0 \leq c(i, j) \leq 1$. The total amount of suppressed redundant data transmissions by sensors in $C_i \backslash \{u_i\}$ then is $\sum_{v_j \in C_i \backslash \{u_i\}} D_i^{\max} \cdot c(i, j)$. Given clusters $C_1, C_2, \ldots, C_k$ with their master sensors $u_i, u_2, \ldots, u_k$, it can be seen that the total amount of suppressed redundant data transmissions in the network is

$$\sum_{i=1}^{k} \sum_{v_j \in C_i \backslash \{u_i\}} D_i^{\max} \cdot c(i, j). \tag{9}$$

In this article, we consider a *redundant data suppression maximization problem*, which is to partition n sensors into k disjoint clusters $C_1, C_2, \ldots, C_k$ and find a master sensor u_i for each cluster C_i, such that the total amount of suppressed redundant data transmissions is maximized, where k is not given and it is dynamically determined by an algorithm for this problem. Then, the problem is

$$\max \sum_{i=1}^{k} \sum_{v_j \in C_i \backslash \{u_i\}} D_i^{\max} \cdot c(i, j). \tag{10}$$

2) Utility Maximization Problem: To further enhance the energy saving of sensors, we also explore the temporal data correlations. We term this problem as the data utility maximization problem, which is to allocate a data rate for each sensor and the routing among sensors for sending generated data to the sink such that the utility of the spatially temporally correlated sensing data received by the sink is maximized, subject to the temporally varying energy constraints on sensors.

Recall that $r_i(t)$ is the data rate of each sensor v_i in V at each time slot t. Also, $f_{ij}(t)$ is the data transmission rate from sensor v_i to v_j at time slot t. For each cluster C_i with its master sensor u_i, it can be seen that the amounts of data sent by u_i and each slave sensor v_j in C_i for the monitoring period are $D_i = \sum_{t=1}^{T} r_i(t) \cdot \tau$ and $D_j = \sum_{t=1}^{T} r_j(t) \cdot \tau$, respectively, where τ is the duration of each time slot. On the other hand, due to the spatial correlation $c(i, j)$ between the master sensor u_i and slave sensor v_j, the amount of obtained data D_{ij} by v_j is the sum of the amount of data sent by v_j itself and the amount of similar data between v_j and its master sensor u_i, i.e., $D_{ij} = D_j + D_i \cdot c(i, j)$.

We introduce a utility function $U(\cdot)$ to measure the quality of collected data from each sensor, and we aim to maximize the accumulative utility of the spatially temporally correlated data collected from sensors. Let $U(\cdot)$ be a network utility function, which is defined as an increasing, twice-differentiable, and strictly concave function. For example, $U(x) = \log_2(x+1)$, where x is a nonnegative number. The utility function can characterize the diminishing return property, i.e., collecting an additional amount of data from a sensor results in a smaller increase in the nonredundant data. This encourages to collect data from many sensors, rather than from a few sensors. For each cluster C_i, the utilities of the data from the master sensor u_i and each slave sensor $v_j \in C_i \backslash \{u_i\}$ are $U(D_i)$ and $U(D_{ij})$, respectively. Therefore, the *accumulative utility* of the data received by the sink is

$$\sum_{i=1}^{k} \left(U(D_i) + \sum_{v_j \in C_i \backslash \{u_i\}} U\big(D_{ij}\big) \right). \tag{11}$$

We consider a novel *utility maximization problem*, which is to allocate a data rate $r_i(t)$ for each sensor and routing $f_{ij}(t)$ among sensors such that the accumulative utility of the spatially temporally correlated data received by the sink is maximized, subject to the energy constraints on sensors.

Notice that when there are no spatial correlations among sensors, i.e., $c(i, j) = 0$, each sensor then forms a cluster by itself, and the objective function in (11) degenerates to $\sum_{i=1}^{n} U(D_i)$, which is identical to the existing studies that took only temporal correlations into account [6].

III. Approximation Algorithm for the Redundant Data Suppression Maximization Problem

In this section, we deal with the redundant data suppression maximization problem, by proposing a novel approximation algorithm.

A. Algorithm

Recall that the redundant data suppression maximization problem is to partition n sensors $v_1, v_2, \ldots, v_n$ into, say k, clusters $C_1, C_2, \ldots, C_k$, and find a master sensor for each cluster C_i such that the total amount of suppressed redundant data transmissions by slave sensors is maximized.

The basic idea behind the proposed approximation algorithm is that we reduce the problem to another problem. We then show that the objective function in the latter is submodular, and the latter problem is an unconstrained submodular maximization problem. An approximate solution to the latter problem then is obtained, which, in turn, returns an approximate solution to the former.

We first reduce the redundant data suppression maximization problem to another equivalent problem in a novel way. Given any two sensors v_i and v_j, denote by p_{ij} the amount of suppressed redundant data transmissions if v_i is the master node of v_j, i.e., $p_{ij} = D_i^{\max} \cdot c(i, j)$, where $D_i^{\max}$ is the maximum amount of generated data by v_i and $c(i, j)$ is the data similarity between v_i and v_j. Especially, $p_{ij} = 0$ if v_j is not a neighbor of v_i or $i = j$.

Denote by M the set of master sensors in the network. Then, $M \subseteq V$, where $V = \{v_1, v_2, \ldots, v_n\}$. For each slave sensor $v_j \in V \backslash M$, it can be seen that v_j may be a slave of multiple master nodes. In order to maximize the amount of suppressed redundant data transmissions by v_j, its master sensor should be a sensor u in M such that $p_{u,j} = \max_{v_i \in M}\{p_{ij}\}$. Then, the amount of suppressed redundant data transmissions by slave sensors in $V \backslash M$ can be calculated as $f(M) = \sum_{v_j \in V \backslash M} \max_{v_i \in M}\{p_{ij}\}$. Therefore, the redundant data suppression maximization problem is equivalent to find a set $M \subseteq V$ of master sensors such that

$$\max_{M \subseteq V}\{f(M)\} = \max_{M \subseteq V}\left\{\sum_{v_j \in V \backslash M} \max_{v_i \in M}\{p_{ij}\}\right\}. \quad (12)$$

We later will show that function $f(M)$ is submodular [3]. That is, given any two subsets $A \subseteq B \subseteq V$ and any sensor $v_k \in V \backslash B$, we have $f(A \cup \{v_k\}) - f(A) \geq f(B \cup \{v_k\}) - f(B)$. It is worthy to mention that although function $f(M)$ is submodular, $f(M)$ is not monotonically increasing. In other words, for any two subsets $A \subseteq B \subseteq V$, it is possible that $f(A) > f(B)$.

We finally solve the redundant data suppression maximization problem, by reducing it to the unconstrained submodular maximization problem, and solve the problem with a $(0.5 - \epsilon)$-approximate solution with high probability, by invoking an algorithm in [3], where ϵ is a given constant with $0 < \epsilon \leq 0.5$.

B. Finding a Randomized Solution

In the following, we briefly describe the algorithm in [3], which finds a randomized approximate solution such that the expected value of the solution is at least half of the optimal value.

Let $V = \{v_1, v_2, \ldots, v_n\}$. The algorithm in [3] starts with two trivial solutions M and N to the problem: one solution M is that there are no master nodes, i.e., $M = \emptyset$; and the other solution N is that all sensors are master nodes, i.e., $N = V$. It can be seen that, in each of the two solutions, the spatial data correlations are not considered at all. The algorithm considers one sensor in V at a time in an arbitrary order, assuming that it considers sensors $v_1, v_2, \ldots, v_n$ one by one. At the ith iteration, the algorithm determines whether sensor v_i is a master node or a slave node with $1 \leq i \leq n$, by constructing two better solutions as follows.

Denote by M_i and N_i the constructed solutions after considering sensor v_i, where $0 \leq i \leq n$. Especially, $M_0 = \emptyset$ and $N_0 = V$. The algorithm first calculates the marginal gain a_i by adding sensor v_i to M_{i-1}, i.e., $a_i = f(M_{i-1} \cup \{v_i\}) - f(M_{i-1})$. It can be seen that a_i means the increased amount of suppressed redundant data transmissions if sensor v_i changes its role from a slave node to a *master node*.

The algorithm also calculates the marginal gain $b_i = f(N_{i-1} \backslash \{v_i\}) - f(N_{i-1})$ by removing v_i from N_{i-1}, where the value of b_i indicates the increased amount of suppressed redundant data transmissions if sensor v_i changes its role from a master node in N_{i-1} to a *slave node*.

The algorithm then determines whether sensor v_i is a master node or a slave node by considering four cases for the values of a_i and b_i: 1) $a_i \geq 0$ and $b_i \geq 0$; 2) $a_i \geq 0$ and $b_i < 0$; 3) $a_i < 0$ and $b_i \geq 0$; and 4) $a_i < 0$ and $b_i < 0$.

For case 1 that $a_i \geq 0$ and $b_i \geq 0$, this means that both changing the role of sensor v_i from a slave node in solution M_{i-1} to a master node and changing the role of v_i from a master node in N_{i-1} to a slave mode will increase the amount of suppressed redundant data transmissions. The algorithm enables sensor v_i to be a master node with a probability $\frac{a_i}{a_i + b_i}$ by adding v_i to M_{i-1}, i.e., $M_i \leftarrow M_{i-1} \cup \{v_i\}$ and $N_i \leftarrow N_{i-1}$; or enables v_i to be a slave node with a probability $1 - [a_i/(a_i + b_i)] = [b_i/(a_i + b_i)]$ by removing v_i from N_{i-1}, i.e., $M_i \leftarrow M_{i-1}$ and $N_i \leftarrow N_{i-1} \backslash \{v_i\}$.

We define that $[a_i/(a_i + b_i)] = 1$ when $a_i = 0$ and $b_i = 0$. Notice that the algorithm either adds v_i to M_{i-1} or removes v_i from N_{i-1} exclusively, but not both of them at the same time.

For case 2 that $a_i \geq 0$ and $b_i < 0$, it can be seen that changing the role of sensor of v_i from a slave node in solution M_{i-1} to a master node will *increase* the amount of suppressed redundant data transmissions $f(M_{i-1})$, but changing the role of v_i from a master node in N_{i-1} to a slave node will *decrease* the amount of suppressed redundant data transmissions $f(N_{i-1})$. Then, the algorithm adds v_i to M_{i-1}, i.e., $M_i \leftarrow M_{i-1} \cup \{v_i\}$ and $N_i \leftarrow N_{i-1}$.

For case 3 that $a_i < 0$ and $b_i \geq 0$, it removes v_i from N_{i-1}, i.e., $M_i \leftarrow M_{i-1}$ and $N_i \leftarrow N_{i-1} \backslash \{v_i\}$.

For case 4 that $a_i < 0$ and $b_i < 0$, Buchbinder *et al.* [3] showed that this case is impossible.

Algorithm 1 Randomized Algorithm for the redundant Data Suppression Maximization Problem

```
Input: $n$ sensors $v_1, v_2, \ldots, v_n$, the amount $p_{ij}$ of suppressed
    redundant data transmissions by sensor $v_j$ if $v_i$ is its master
Output: A subset $M \subseteq V$ of master nodes
Let $M_0 = \emptyset$ and $N_0 = V$;
for $i \leftarrow 1$ to $n$ do
    Let $a_i = f(M_{i-1} \cup \{v_i\}) - f(M_{i-1})$; /* $a_i$ is the marginal gain by adding $v_i$ to $M_{i-1}$ */
    Let $a'_i = \max\{a_i, 0\}$;
    Let $b_i = f(N_{i-1} \backslash \{v_i\}) - f(N_{i-1})$; /* $b_i$ is the marginal gain by removing $v_i$ from $N_{i-1}$ */
    Let $b'_i = \max\{b_i, 0\}$;
    /* Add $v_i$ to $M_{i-1}$ or remove $v_i$ from $N_{i-1}$, but do not perform both actions at the same time*/
    Add $v_i$ to $M_{i-1}$ with a probability $\frac{a'_i}{a'_i+b'_i}$, i.e., $M_i \leftarrow M_{i-1} \cup \{v_i\}$ and $N_i \leftarrow N_{i-1}$; or complementarily remove $v_i$ from $N_{i-1}$ with a probability $1 - \frac{a'_i}{a'_i+b'_i}$, i.e., $M_i \leftarrow M_{i-1}$ and $N_i \leftarrow N_{i-1} \backslash \{v_i\}$, where $\frac{a'_i}{a'_i+b'_i} = 1$ if $a'_i = b'_i = 0$;
end for
Let $M = M_n$; /* in the end, $M_n = N_n$*/
Each $v_j \in V \backslash M$ selects a master node $u \in M$ such that $u = \arg\max_{v_i \in M}\{p_{ij}\}$;
return $M$.
```

After n iterations, Buchbinder *et al.* [3] showed that $M_n = N_n$. Let $M = M_n$ be the set of master sensors in the final solution. Buchbinder *et al.* [3] showed that the expected amount of suppressed redundant data transmissions is at least half of the optimal value, i.e., $\mathbf{E}[f(M_n)] \geq (OPT/2)$, where OPT is the optimal value of the problem. The detailed algorithm for finding a randomized solution is presented in Algorithm 1.

We illustrate the execution of Algorithm 1 with a simple example. We assume that there are three sensors v_1, v_2, and v_3 in a sensor network. Initially, $M_0 = \emptyset$ and $N_0 = \{v_1, v_2, v_3\}$. First, consider sensor v_1, assume that a_1 is much larger than b_1 and $b_1 \geq 0$, this implies that $[a_1/(a_1 + b_1)] \approx 1$. Therefore, v_1 should be added to M_0, i.e., $M_1 = \{v_1\}$ and $N_1 = \{v_1, v_2, v_3\}$. Then, consider sensor v_2, assume that $a_2 < 0$ and $b_2 > 0$, which implies that v_2 should be removed from N_1, i.e., $M_2 = M_1 = \{v_1\}$ and $N_2 = N_1 \backslash \{v_2\} = \{v_1, v_3\}$. Finally, consider sensor v_3, assume that $a_3 < 0$ and $b_3 > 0$, which indicates that v_3 should be removed from N_2, i.e., $M_3 = M_2 = \{v_1\}$ and $N_3 = N_2 \backslash \{v_3\} = \{v_1\}$. It can be seen that $M_3 = N_3 = \{v_1\}$, sensor v_1 is the only master node while v_2 and v_3 are its slave nodes.

C. Find an Approximate Solution With High Probability

Notice that although Algorithm 1 is able to find a randomized solution by lower bounding its expected value [3] [i.e., $\mathbf{E}[f(M_n)] \geq (OPT/2)$], it does not specify how much chance such a solution can be achieved. To deliver an approximate solution with high probability, we invoke Algorithm 1 multiple times, and choose the best solution among the multiple randomized solutions as the final solution to the problem. Assume that Algorithm 1 is invoked L times, and it delivers a randomized solution $M^{(l)}$ at the lth time with $1 \leq l \leq L$, where L is a given time. The final solution M to the problem then is the one with the maximum profit, i.e., $f(M) = \max_{l=1}^{L} f(M^{(l)})$. The detailed algorithm for the redundant data suppression maximization problem is presented in Algorithm 2. We later show that Algorithm 2 is able to deliver a $(0.5 - \epsilon)$-approximate solution with a probability of $1 - (1/n^{\alpha})$, by invoking $L = \lceil \alpha \cdot \log_{1+2\epsilon} n \rceil$ times of Algorithm 1, where ϵ and α are given two constants with $0 < \epsilon \leq 0.5$ and $\alpha > 0$.

Algorithm 2 Approximation Algorithm for the Redundant Data Suppression Maximization Problem (`maxSuppression`)

```
Input: $n$ sensors $v_1, v_2, \ldots, v_n$, the amount $p_{ij}$ of suppressed
    redundant data transmissions by sensor $v_j$ if $v_i$ is its master
    node, an error ratio $\epsilon$ with $0 < \epsilon \leq 0.5$, and a success
    probability $1 - \frac{1}{n^{\alpha}}$
Output: A $(0.5 - \epsilon)$-approximate solution $M \subseteq V$ of master
    nodes with a probability of $1 - \frac{1}{n^{\alpha}}$
Let $L = \lceil \alpha \cdot \log_{1+2\epsilon} n \rceil$; /* the number of invokings */
Let $M \leftarrow \emptyset$;
for $l \leftarrow 1$ to $L$ do
    Find a randomized solution $M^{(l)}$ by invoking Algorithm 1;
    if $f(M^{(l)}) > f(M)$ then
        Let $M \leftarrow M^{(l)}$;
    end if
end for
return the set $M$ of master sensors.
```

For example, to find a $0.45(= 0.5 - 0.05)$ approximate solution with a probability of $1 - (1/n^1)$ in a network with $n = 500$ sensors, Algorithm 1 will be invoked $L = \lceil \alpha \cdot \log_{1+2\epsilon} n \rceil = \lceil 1 \cdot \log_{1+2 \cdot 0.05} 500 \rceil = 66$ times.

D. Algorithm Analysis

Lemma 1: Given any subset M of V, define function $f(M) = \sum_{v_j \in V \backslash M} \max_{v_i \in M}\{p_{ij}\}$. Then, $f(M)$ is submodular.

Proof: The proof is contained in the Appendix. ■

Theorem 1: Given a network $G = (V, E)$ with n sensors $v_1, v_2, \ldots, v_n$ in V, and the amount p_{ij} of suppressed redundant data transmissions by sensor v_j if v_i is its master node with $1 \leq i, j \leq n$, there is an approximation algorithm, Algorithm 2, for the redundant data suppression maximization problem, which delivers a $(0.5 - \epsilon)$-approximate solution with a probability $1 - (1/n^{\alpha})$ in time $O(\alpha m (\log \Delta) \log_{1+2\epsilon} n) = O(m(\log n)^2)$, where $n = |V|$ and $m = |E|$, ϵ and α are two given constants with $0 < \epsilon \leq 0.5$ and $\alpha > 0$, and Δ is the maximum number of neighbors among sensors in G, i.e., $\Delta = \max_{v_i \in V}\{|N(v_i)|\}$.

Proof: The proof is contained in the Appendix. ■

IV. Near-Optimal Algorithm for the Data Utility Maximization Problem

In the previous section, we partitioned sensors into different clusters and selected a master sensor for each cluster, where the master sensor sends representative data to the sink, while the other slave sensors in the cluster transmit only their nonredundant data. Given the clusters and chosen master sensors, in this section, we address the data utility maximization problem, by proposing a near-optimal algorithm for it.

A. Algorithm

Recall that the data utility maximization problem is to allocate the data rate $r_i(t)$ to each sensor $v_i \in V$ and the data transmission rate $f_{ij}(t)$ to each link $(v_i, v_j) \in E$ such that the accumulative utility of spatially temporally correlated data is maximized. We formally formulate the problem as

$$\mathbf{P1:}\ \max_{r_i(t),f_{ij}(t)} \left\{ \sum_{i=1}^{k} \left(U(D_i) + \sum_{v_j \in C_i \setminus \{u_i\}} U\big(D_{i,j}\big) \right) \right\} \tag{13}$$

subject to

$$\sum_{v_j \in N(v_i)} f_{ji}(t) + r_i(t) = \sum_{v_j \in N(v_i)} f_{ij}(t) \quad \forall v_i \in V,\ 1 \le t \le T \tag{14}$$

$$\begin{aligned} RE(v_i, t) = \min\{&RE(v_i, t-1) + H(v_i, t) \\ &- P(v_i, t),\ B_i\} \\ &\forall v_i \in V,\ 1 \le t \le T \end{aligned} \tag{15}$$

where constraint (14) indicates that sensor v_i will forward its received data and its sensing data to the sink via the relay of its neighbors; constraint (15) describes the relationship between the residual energy $RE(v_i, t)$ and $RE(v_i, t-1)$ of each sensor v_i at time slot t and its previous time slot $t-1$;B_i is the battery capacity of v_i; $0 \le RE(v_i, t) \le B_i$; $f_{ij}(t) \ge 0$; and $0 \le r_i(t) \le R_i$.

Notice that problem **P1** is not a convex optimization problem, as both its objective function and constraint (15) are not convex. Fortunately, we can transform it into a convex optimization problem **P2**

$$\mathbf{P2:}\ \min_{r_i(t),f_{ij}(t),RE(v_i,t)} -\left(\sum_{i=1}^{k} \left(U(D_i) + \sum_{v_j \in C_i \setminus \{u_i\}} U\big(D_{i,j}\big) \right) \right) \tag{16}$$

subject to constraint (14) and

$$RE(v_i, t) \le RE(v_i, t-1) + H(v_i, t) - P(v_i, t). \tag{17}$$

Following a recent study [37], the optimal values of problem **P1** and **P2** are equal. Then, we can find a near-optimal solution to the convex optimization problem **P2**, by applying any existing algorithm, such as interior-point methods, subgradient methods, ellipsoid methods, cutting-planes methods, etc. [1]. The solution to **P2** in turn returns a near-optimal solution to **P1**. The detailed algorithm for problem **P1** is termed as algorithm `maxUtility`.

Theorem 2: Given a sensor network $G = (V \cup \{s\},\ E)$, a data collection period T, the residual energy $RE(v_i, 0)$ of each sensor v_i at the beginning of period T, the harvested energy $H(v_i, t)$ of each sensor v_i at time slot t with $1 \le t \le T$, there is a near-optimal algorithm for the data utility maximization problem in G.

Proof: Following a recent study in [37], the optimal values of problem **P1** and **P2** are equal. Following the work [1], the barrier method in the existing algorithm can find a solution with its total utility only ϵ smaller than the maximum utility, where ϵ is a given additive error with $0 < \epsilon < 1$. The theorem then follows. ■

V. Performance Evaluation

In this section, we evaluate the performance of the proposed algorithms, by comparing them with the state-of-the-art algorithms for the clustering problem and the data utility maximization problem. We also investigate the impact of different parameters on the performance.

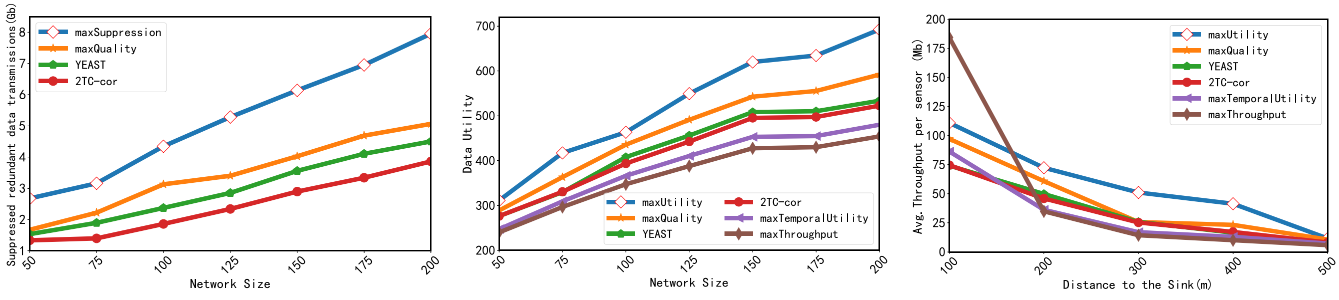
A. Simulation Settings

We consider a sensor network, which consists of from 50 to 200 sensors. The sensors are deployed in a 1000 × 1000 square area randomly. The sink s is located at the center of the area. Assume that the battery capacity of each sensor is $B = 10.8$ kJ and the maximum sensing rate of the sensor is $R = 10$ kb/s [37]. We consider a monitoring period $T = 24$ h that is divided into equal time slots with an equal duration of τ (e.g., $\tau = 1$ h). Each sensor is powered by a 37 mm × 33 mm solar panel. We adopt real 89 solar power harvesting profiles from the baseline measurement system at the National Renewable Energy Laboratory [25]. (from February 1 to April 30 in 2019). The harvested energy profile of each sensor in a day is randomly chosen from the 89 energy profiles. Specifically, the amount of harvested energy $H(v_i, d, t)$ of sensor v_i on day d at time t is $H(v_i, d, t) = \lambda \cdot H_{\text{sample}}(t)$, where λ is a constant with $\lambda > 0$ and its default value is 1, and $H_{\text{sample}}(t)$ is the real energy generation rate at time slot t in a randomly chosen energy profile. Recall that we used the energy harvesting profiles in spring. To consider the energy profiles of sensors in other seasons, we can adopt different values of λ. For instance, we can obtain the energy profiles of sensors in summer by setting $\lambda = 2$. λ is referred to as the *energy scaling coefficient* [37].

We adopt both real and simulation sensing data in our experiments. On the one hand, we use real temperature sensing data collected from 54 sensors for 38 days (from February 28 to April 5 in 2004) [12]. In our experiments, the sensing data sequence of a sensor v_i on day d is randomly chosen from the 54 sensor data sequence on a random day of the 38 days. Then, the spatial data correlations of nearby sensors follow the distribution in Fig. 4. On the other hand, we use simulation data as follows. For any two neighboring sensors v_i and v_j, the value of their spatial data correlation $c(i, j)$ is randomly chosen from an interval $[0,\ C_{\max}]$ with $C_{\max} = 1$.

To evaluate the performance of the proposed algorithms `maxSuppression` and `maxUtility`, we adopt five existing algorithms.

1) Algorithm `maxThroughput` [24] does not take temporal and spatial data correlations, and it aims to maximize the total amount of data collected in a period.
2) Algorithm `maxTemporalUtility` [37] considered only temporal data correlation and focused the problem



(a) (b) (c)

Fig. 6. Performance of the different algorithms by increasing the network size n from 50 to 200. (a) Amount of suppressed redundant data transmissions. (b) Accumulative utility of collected data. (c) Average throughput per sensor with 200 sensors.

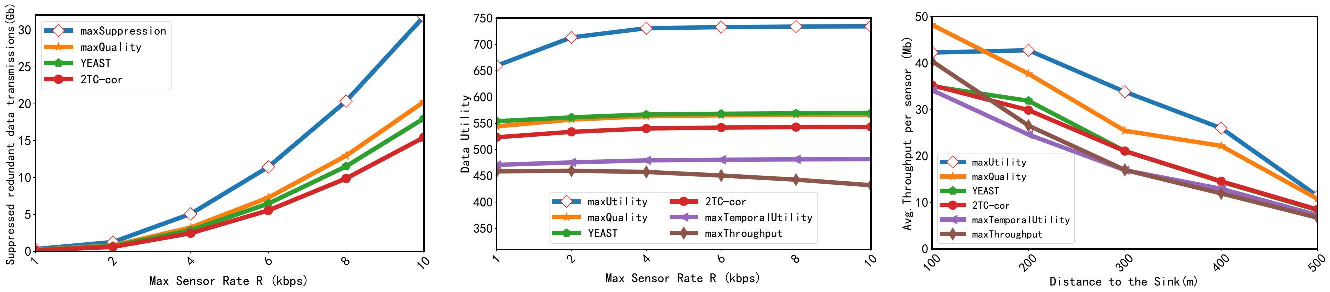
of sensor data rates and routing allocations, such that the accumulative utility of temporally correlated data is maximized.

3) Algorithm YEAST [32] exploited only spatial data correlation, by partitioning sensors into different clusters and selecting representative nodes to construct a dynamic scalable routing tree to minimize sensor energy consumption.
4) Algorithm maxQuality [16], [23] incorporated both temporal and spatial data correlations, through fair rate allocations and flow routing to maximize the quality of collected data in an energy harvesting sensor network.
5) Algorithm 2TC-cor [31] also incorporated both temporal and spatial data correlations and proposed a subclustering algorithm based on spatial data correlation, in which each sensor first compares its generated data with its historical data, so as to suppress transmissions of similar temporal data, cluster heads then select some representative sensors for further energy saving by considering spatial correlation.

B. Algorithm Performance

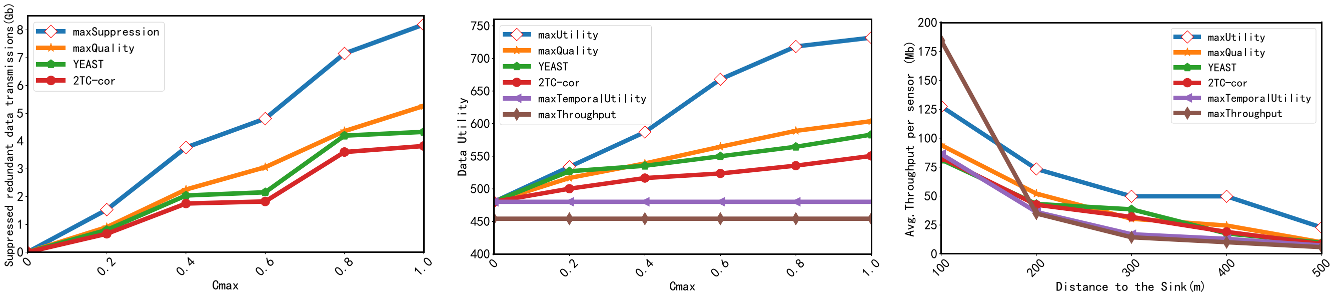
We first study the performance of the different algorithms by increasing the network size n from 50 to 200. Fig. 6(a) shows that the amount of suppressed redundant data transmissions by the proposed algorithm maxSuppression is from 50% to 55% larger than those by algorithms maxQuality, YEAST, and 2TC-cor, when n increases from 50 to 200. Notice that we here did not plot the performance of algorithms maxThroughput and maxTemporalUtility, as they did not take the spatial data correlations among sensors into consideration, and thus the amounts of suppressed redundant data transmissions by them are zeros. Fig. 6(b) demonstrates that the accumulative utility of the collected data by the proposed algorithm maxUtility is upto 17% larger than those by the other algorithms. For example, the accumulative utilities by maxUtility, maxQuality, YEAST, 2TC-cor, maxTemporalUtility, and maxThroughput are 690, 590, 533, 522, 480, and 454, respectively, when there are 200 sensors. Fig. 6(c) further plots the average throughput per sensor by the different algorithms, from which it can be seen that the average throughput per sensor by algorithm maxThroughput decreases dramatically with the distance to the sink. In contrast, the average throughput per sensor by the proposed algorithm maxUtility is much larger than those by the other five algorithms when the distance to the sink is longer than 200 m. For example, the average throughputs by maxUtility, maxQuality, YEAST, 2TC-cor, maxTemporalUtility, and maxThroughput are 41, 23, 16, 17, 12, and 10 Mb, respectively, when the distance to the sink is 400 m. The rationale behind is that the proposed algorithm maxUtility can exploit much more spatial data correlations, thereby collecting fewer amounts of data from nearby sensors and saving their energy for relaying the nonredundant data from remote sensors.

We then investigate the algorithm performance by increasing the maximum sensing rate R from 1 to 10 kb/s. Fig. 7(a) shows that the difference between the amount of suppressed redundant data transmissions by the proposed algorithm maxSuppression and the amounts by other algorithms becomes larger with the increase of the maximum sensing rate R, as there are more redundant data if each sensor generates data faster. Fig. 7(a) also demonstrates that the amount of suppressed redundant data transmissions by algorithm maxSuppression is about 50% larger than those by other existing algorithms. Fig. 7(b) shows that the data utility by the algorithm maxUtility is from 19% to 29% larger than those by other algorithms, and the utility by each algorithm increases with the growth of R since more nonredundant information will be sent. Fig. 7(c) demonstrates that the throughputs of different sensors by algorithm maxUtility distributes more fairly than those by other algorithms. Notice that the average throughput per sensor by the proposed algorithm maxUtility is smaller than those by the existing algorithms when the distance to the sink is small. The rationale behind is as follows. This article addresses the problem of collecting nonredundant data from sensors, by incorporating both spatial data correlations and temporal data correlations. The solutions delivered by the existing algorithms will collect a large amount of data from the sensors near to the sink, but less amount of data from remote sensors. However, the collected data from the nearby sensors are highly redundant. In contrast, the solution delivered by the proposed algorithm maxUtility collects more data from remote sensors, though the amount of data collected from nearby sensors is less than those by the existing studies. The utility of the data collected,



(a) (b) (c)

Fig. 7. Performance of the different algorithms by increasing the maximum sensing rate R from 1 to 10 kb/s when $n = 200$. (a) Amount of suppressed redundant data transmissions. (b) Accumulative utility of collected data. (c) Average throughput per sensor with $R = 1$ kb/s.



(a) (b) (c)

Fig. 8. Performance of the different algorithms by increasing the maximum spatial data correlation $C_{\max}$ from 0 to 1 when $n = 200$. (a) Amount of suppressed redundant data transmissions. (b) Accumulative utility of collected data. (c) Average throughput per sensor with $C_{\max} = 0.8$.

i.e., the amount of nonredundant data, by the proposed algorithm maxUtility is larger than those by the existing algorithm.

We finally study the algorithm performance by increasing the maximum spatial data correlation $C_{\max}$ from 0 to 1. Fig. 8(a) shows that the amount of suppressed redundant data transmissions by the proposed algorithm maxSuppression is about from 50% to 60% larger than those by the other algorithms when $C_{\max}$ increases from 0.2 to 1. However, the amounts of suppressed redundant data transmissions by each algorithm is 0 when $C_{\max} = 0$ since $C_{\max} = 0$ indicates that there are no spatial data correlations. Fig. 8(b) shows that the accumulative data utility by each of the four algorithms maxUtility, maxQuality, YEAST, and 2TC-cor increase with the growth of $C_{\max}$, while the utilities by both maxTemporalUtility and maxThroughput do not change with the increase of $C_{\max}$, as both they do not consider spatial data correlations among sensors. In addition, Fig. 8(b) plots that the difference between the data utility by algorithm maxUtility and those by the other algorithms increase with the growth of $C_{\max}$. Especially, the utility by algorithm maxUtility is 20% larger than those by the other algorithm when $C_{\max} = 1$. Fig. 8(c) finally demonstrates that the average throughput per sensor by the proposed algorithm maxUtility is much larger than those by the other five algorithms when the distances to the sink are longer than 200 m, but less amount of data from the sensors with their distance to the sink being less than 100 m. This implies that the solution delivered by the proposed algorithm maxUtility collects data from different sensors more fairly, rather than from only a few nearby sensors. Therefore, the utility by the proposed algorithm is larger than those by the existing algorithms.

VI. Related Work

There are extensive studies on efficient data collection in WSNs in the past. Some existing studies mainly focused on optimizing sensor energy saving when performing sensing data collection, without the consideration of temporal and spatial data correlations [5], [13], [15], [24], [26], [34]. For example, Liew *et al.* [18] proposed a fast, adaptive, and energy-efficient data collection protocol for only battery-powered sensor networks, by utilizing multichannel and multipath data transmission, thereby reducing delay and packet loss rates. He *et al.* [10] proposed a data gathering scheme with a mobile sink to achieve efficient data collection in sensor networks. Liu *et al.* [21] designed a distributed routing algorithm for data collection in low duty cycle WSNs, which achieves a better tradeoff between data transmission latency and sensor energy conservation than existing schemes. Also, there are some studies focusing on replenishing sensor energy through harvesting energy from their surrounding environments [24], [26]. Gallego *et al.* [7] presented a medium access control protocol that combines distributed queuing and energy harvesting. Mao *et al.* [24] studied the problem of allocating energy for data sensing and data transmitting, and used the infinite horizon Markov decision process to maximize the amount of collected data. Mehrabi and Kim [26] and Wang *et al.* [33] studied the use of mobile sinks to collect data from energy-harvested sensors such that the amount of data collected is maximized. Xu *et al.* [34] investigated the employment of multiple energy-constrained sinks to collect sensor data to maximize the amount of data collected, subject to the constraint on the energy capacity of each sink.

Although extensive studies have been conducted for data collections and energy conservation in sensor networks [24], [26], they did not take into account sensing data correlations. Deng *et al.* [6] and Zhang *et al.* [37] recognized that sensing data are quite similar during a short period of time, that is, sensing readings are *temporally correlated*. Deng *et al.* [6] studied the network utility maximization problem in static routing rechargeable sensor networks and allocated sensing rates for sensors. They devised a distributed algorithm to obtain globally optimal data rates. Zhang *et al.* [37] devised a near-optimal algorithm for the utility maximization problem, by considering time-varying energy harvesting rates of sensors.

We also noticed that some researchers found that sensing data are spatially correlated and they aimed to maximize data quality by utilizing spatial correlation while reducing energy consumption [22], [32], [35]. For example, Yoon and Shahabi [35] designed an algorithm to partition sensors into multiple clusters such that sensors in each cluster are close with each other and their sensing data are highly similar. Then, only the cluster heads transmit their data to the sink, whereas the other sensors do not generate data, thereby saving their energy. Villas *et al.* [32] focused on the cluster head choice by considering the residual energy of sensors and designed an efficient routing algorithm for data collection. Liu *et al.* [22] strived for the fine balance between the quality of collected data and sensor energy consumption, by choosing a sensor in each cluster to work with a given probability p. That is, a smaller value of p is chosen if the residual energy of a sensor is low, while a larger p is adopted when the monitoring quality needs to be improved. Brahmi *et al.* [2] devised a hybrid strategy for energy-efficient data gathering: local aggregation and global aggregation, where the local aggregation means that data from member sensors in the same cluster are aggregated by the cluster head, while the global aggregation indicates a cluster head also aggregates the data received from cluster heads. In order to further improve data collection and energy efficiency, some studies considered both energy harvesting and spatial correlations. Qamaji and Atakan [27] studied the tradeoff between energy conservation and data distortion level.

There are quite a few studies that considered both the temporal and spatial data correlations [16], [30], [31], [36]. For example, Liang *et al.* [16] first clustered sensors by their spatial data correlation such that each cluster has no more than two sensors, one sensor serves as the master sensor while the other serves the slave sensor. They then investigated a fair data rate allocation problem. Tayeh *et al.* [30] assumed that a sufficient number of powerful cluster heads have been deployed so that each sensor can communicate with at least one cluster head, and then each cluster head adjusts the data rates of the sensors associated with it. However, it is very challenging to select cluster heads in a sensor network, which must consider both the spatial data correlations among sensors and their residual energy. Tsai and Huang [31] proposed a subclustering algorithm based on spatial data correlation, in which each member sensor first compares its generated data with its historical data, so as to suppress the transmissions of similar data, cluster heads then select some representative sensors for further energy saving by considering spatial correlation. Yoon and Shahabi [36] extended their work in [35], respectively, by incorporating temporal data correlation, whereas the latter study in [35] considered only spatial correlation. Liang *et al.* [17] studied a problem of scheduling a UAV to collect spatially correlated data from PoIs such that the amount of nonredundant information collected is maximized, subject to the energy capacity constraint on the UAV.

It can be seen that although the existing studies considered both spatial and temporal data correlation, they adopted a coarse-grained spatial correlation model. That is, each sensor either transmits its data to the sink or not. Unlike the aforementioned studies, we here proposed a fine-grained spatial correlation model, in which some representative sensors send all their sensing data, while the others transmit only nonredundant data from that sent by the representative sensors. Therefore, our work is promising in reducing the transmissions of redundant data, while maintaining a high quality of collected data.

VII. Conclusion

Unlike the existing studies that adopted the coarse-grained spatial-correlation model, in this article, we introduced a novel fine-grained spatial-correlation model, in which only a small portion of sensors send their representative data and the other majority sensors transmit their nonredundant data, thereby reducing redundant data transmissions. We first studied the redundant data suppression maximization problem under the proposed model, by proposing a novel approximation algorithm for it, which delivers a $(0.5-\epsilon)$-approximate solution with high probability, where ϵ is a given constant with $0 < \epsilon \leq 0.5$. We then devised a near-optimal algorithm for the data utility maximization problem such that accumulative utility of the spatially temporally correlated data received by the sink is maximized. We finally evaluated the performance of the proposed algorithms, using real data sets. The experimental results show that the proposed algorithms outperform the existing algorithms.

Appendix

A. Algorithm Analysis

We show that function $f(M)$ is submodular by the following lemma.

Lemma 2: Given any subset M of V, define function $f(M) = \sum_{v_j \in V\setminus M} \max_{v_i \in M}\{p_{ij}\}$. Then, $f(M)$ is submodular.

Proof: Given any two subsets $A \subseteq B \subseteq V$ and any sensor $v_k \in V\setminus B$, we show that $f(A\cup\{v_k\}) - f(A) \geq f(B\cup\{v_k\}) - f(B)$ as follows.

For any $v_j \in V\setminus(B\cup\{v_k\})$, it can be seen that

$$\max_{v_i\in A\cup\{v_k\}}\{p_{ij}\} - \max_{v_i\in A}\{p_{ij}\}$$
$$= \max\left\{p_{kj} - \max_{v_i\in A}\{p_{ij}\},\ 0\right\}$$

$$
\begin{aligned}
&\geq \max\left\{p_{kj} - \max_{v_i \in B}\{p_{ij}\},\ 0\right\}\\
&\text{as } \max_{v_i \in A}\{p_{ij}\} \leq \max_{v_i \in B}\{p_{ij}\}, \text{ due to } A \subseteq B\\
&= \max_{v_i \in B \cup \{v_k\}}\{p_{ij}\} - \max_{v_i \in B}\{p_{ij}\}.
\end{aligned}
\tag{18}
$$

We then have

$$
\begin{aligned}
&f(A \cup \{v_k\}) - f(A)\\
&= \sum_{v_j \in V \setminus (A \cup \{v_k\})} \max_{v_i \in A \cup \{v_k\}}\{p_{ij}\} - \sum_{v_j \in V \setminus A} \max_{v_i \in A}\{p_{ij}\}\\
&= \sum_{v_j \in V \setminus (A \cup \{v_k\})} \max_{v_i \in A \cup \{v_k\}}\{p_{ij}\} - \sum_{v_j \in V \setminus (A \cup \{v_k\})} \max_{v_i \in A}\{p_{ij}\}\\
&\quad - \max_{v_i \in A}\{p_{ik}\}\\
&= \sum_{v_j \in V \setminus (A \cup \{v_k\})} \left(\max_{v_i \in A \cup \{v_k\}}\{p_{ij}\} - \max_{v_i \in A}\{p_{ij}\}\right) - \max_{v_i \in A}\{p_{ik}\}\\
&\geq \sum_{v_j \in V \setminus (B \cup \{v_k\})} \left(\max_{v_i \in A \cup \{v_k\}}\{p_{ij}\} - \max_{v_i \in A}\{p_{ij}\}\right) - \max_{v_i \in A}\{p_{ik}\}\\
&\text{as } V \setminus (A \cup \{v_k\}) \supseteq V \setminus (B \cup \{v_k\}), \text{ due to } A \subseteq B\\
&\geq \sum_{v_j \in V \setminus (B \cup \{v_k\})} \left(\max_{v_i \in B \cup \{v_k\}}\{p_{ij}\} - \max_{v_i \in B}\{p_{ij}\}\right) - \max_{v_i \in A}\{p_{ik}\}\\
&\text{due to (18)}\\
&\geq \sum_{v_j \in V \setminus (B \cup \{v_k\})} \left(\max_{v_i \in B \cup \{v_k\}}\{p_{ij}\} - \max_{v_i \in B}\{p_{ij}\}\right) - \max_{v_i \in B}\{p_{ik}\}\\
&\text{since } \max_{v_i \in A}\{p_{ik}\} \leq \max_{v_i \in B}\{p_{ik}\}, \text{ due to } A \subseteq B\\
&= \sum_{v_j \in V \setminus (B \cup \{v_k\})} \max_{v_i \in B \cup \{v_k\}}\{p_{ij}\} - \sum_{v_j \in V \setminus (B \cup \{v_k\})} \max_{v_i \in B}\{p_{ij}\}\\
&\quad - \max_{v_i \in B}\{p_{ik}\}\\
&= \sum_{v_j \in V \setminus (B \cup \{v_k\})} \max_{v_i \in B \cup \{v_k\}}\{p_{ij}\} - \sum_{v_j \in V \setminus B} \max_{v_i \in B}\{p_{ij}\}\\
&= f(B \cup \{v_k\}) - f(B).
\end{aligned}
\tag{19}
$$

Then, function $f(.)$ is submodular, the lemma then follows. ■

Theorem 3: Given a network $G = (V, E)$ with n sensors $v_1, v_2, \ldots, v_n$ in V, and the amount p_{ij} of suppressed redundant data transmissions by sensor v_j if v_i is its master node with $1 \leq i, j \leq n$, there is an approximation algorithm, Algorithm 2, for the redundant data suppression maximization problem, which delivers a $(0.5 - \epsilon)$-approximate solution with a probability $1 - \frac{1}{n^\alpha}$ in time $O(\alpha m (\log \Delta) \log_{1+2\epsilon} n) = O(m(\log n)^2)$, where $n = |V|$, $m = |E|$, ϵ and α are two given constants with $0 < \epsilon \leq 0.5$ and $\alpha > 0$, and Δ is the maximum number of neighbors among sensors in G, i.e., $\Delta = \max_{v_i \in V}\{|N(v_i)|\}$.

Proof: Following Lemma 2, the redundant data suppression maximization problem can be cast as an unconstrained submodular maximization problem. Following Buchbinder *et al.* [3], Algorithm 1 can find a randomized solution such that its expected value is at least half of the optimal value OPT.

Recall that Algorithm 2 finds $L = \lceil \alpha \cdot \log_{1+2\epsilon} n \rceil$ randomized solutions $M^{(1)}, M^{(2)}, \ldots, M^{(L)}$, where each solution $M^{(l)}$ is obtained by invoking Algorithm 1. The final solution in Algorithm 2 is $M = \arg\max_{l=1}^{L} f(M^{(l)})$. In the following, we first analyze the probability that $f(M) \geq (0.5 - \epsilon) \cdot OPT$.

Let π_l be the probability that the value of the $M^{(l)}$ is no more than $(0.5 - \epsilon) \cdot OPT$, i.e.,

$$
\pi_l = Pr\left[f(M^{(l)}) \ \leq\ (0.5 - \epsilon) \cdot OPT\right]
\tag{20}
$$

where $1 \leq l \leq L$. The expected value of $M^{(l)}$ then can be upper bounded by

$$
\mathbf{E}\left[f(M^{(l)})\right] \leq \pi_l \cdot (0.5 - \epsilon) \cdot OPT + (1 - \pi_l) \cdot OPT.
\tag{21}
$$

Following Buchbinder *et al.* [3], we have $\mathbf{E}[f(M^{(l)})] \geq (OPT/2)$. Then, the value of probability π_l thus is upper bounded by

$$
\pi_l \leq \frac{1}{1 + 2\epsilon}.
\tag{22}
$$

The probability that the value of each of the L randomized solutions is no more than $(0.5 - \epsilon) \cdot OPT$ then is no more than

$$
\begin{aligned}
\prod_{l=1}^{L} \pi_l &\leq \left(\frac{1}{1 + 2\epsilon}\right)^{L}\\
&= \left(\frac{1}{1 + 2\epsilon}\right)^{\lceil \alpha \cdot \log_{1+2\epsilon} n \rceil}\\
&\leq \left(\frac{1}{1 + 2\epsilon}\right)^{\alpha \cdot \log_{1+2\epsilon} n}\\
&= \frac{1}{n^\alpha}.
\end{aligned}
\tag{23}
$$

This indicates that the probability that Algorithm 2 delivers a $(0.5 - \epsilon)$-approximate solution is no less than $1 - (1/n^\alpha)$.

The rest is to analyze the time complexity of Algorithm 2. In the following, we show that Algorithm 1 can be efficiently implemented in time $O(m \log \Delta)$. Then, the time complexity of Algorithm 2 is $O(Lm \log \Delta) = O(\alpha m (\log \Delta) \log_{1+2\epsilon} n)$.

Notice that the running time of Algorithm 1 is dominated by calculating a_i and b_i with $1 \leq i \leq n$, which is analyzed as follows.

We first analyze the time complexity of the calculation of a_i for $1 \leq i \leq n$. Before each iteration i, assume that we already knew the value $\max_{v_k \in M_{i-1}}\{p_{kj}\}$ for each $v_j \in V \setminus M_{i-1}$. Notice that the addition of v_i to M_{i-1} will change the values $\max_{v_k \in M_{i-1}}\{p_{kj}\}$ for only neighbors v_j of v_i, but not change the value $\max_{v_k \in M_{i-1}}\{p_{kj'}\}$ for a sensor $v_{j'}$ such that $v_{j'} \notin N(v_i)$, since $\max_{v_k \in M_{i-1} \cup \{v_i\}}\{p_{kj'}\} = \max_{v_k \in N(v_{j'}) \cap (M_{i-1} \cup \{v_i\})}\{p_{kj'}\} = \max_{v_k \in N(v_{j'}) \cap M_{i-1}}\{p_{kj'}\} = \max_{v_k \in M_{i-1}}\{p_{kj'}\}$, due to that v_i is not a neighbor of $v_{j'}$. Then

$$
\begin{aligned}
a_i &= f(M_{i-1} \cup \{v_i\}) - f(M_{i-1})\\
&= \sum_{v_j \in N(v_i) \setminus M_{i-1}} \left(\max_{v_k \in M_{i-1} \cup \{v_i\}}\{p_{kj}\} - \max_{v_k \in M_{i-1}}\{p_{kj}\}\right)\\
&\quad - \max_{v_k \in M_{i-1}}\{p_{ki}\}\\
&= \sum_{v_j \in N(v_i) \setminus M_{i-1}} \max\left\{p_{ij} - \max_{v_k \in M_{i-1}}\{p_{kj}\},\ 0\right\}\\
&\quad - \max_{v_k \in M_{i-1}}\{p_{ki}\}.
\end{aligned}
\tag{24}
$$

Since the value $\max_{v_k\in M_{i-1}}\{p_{kj}\}$ for each $v_j \in V\backslash M_{i-1}$ is already known, the calculation of a_i takes time $O(|N(v_i)|) = O(\delta_i)$, where $\delta_i = |N(v_i)|$. Then, the calculation for all a_is takes time $\sum_{i=1}^{n} O(\delta_i) = O(m)$, where m is the number of edges in G, i.e., $m = |E|$.

We then analyze the time complexity of the calculations of all b_is. Similarly, we assume that we already know the value $\max_{v_k\in N_{i-1}}\{p_{kj}\}$ for each $v_j \in V\backslash N_{i-1}$ before each iteration i. Also, if a sensor v_i is removed from N_{i-1} at the ith iteration in Algorithm 1, we construct a list L_i of neighbors of v_i, where a sensor v_k is in L_i if and only if $v_k \in N(v_i)$ and $v_k \in N_i = N_{i-1}\backslash\{v_i\}$. We sort the neighbors in list L_i in decreasing order of their values p_{ki}s. Then, the construction of the lists in the n iterations of Algorithm 1 takes time $\sum_{i-1}^{n} O(\delta_i \log \delta_i) = \sum_{i-1}^{n} O(\delta_i \log \Delta) = O(m \log \Delta)$, where $\delta_i = |N(v_i)|$ and $\Delta = \max_{i=1}^{n}\{\delta_i\}$.

Notice that the removal of v_i from N_{i-1} will change the value $\max_{v_k\in N_{i-1}}\{p_{kj}\}$ for only neighbors v_j of v_i, but do not change the value $\max_{v_k\in N_{i-1}}\{p_{kj'}\}$ for a sensor $v_{j'}$ such that $v_{j'} \notin N(v_i)$, since $\max_{v_k\in N_{i-1}\backslash\{v_i\}}\{p_{kj'}\} = \max_{v_k\in N(v_{j'})\cap(N_{i-1}\backslash\{v_i\})}\{p_{kj'}\} = \max_{v_k\in N(v_{j'})\cap N_{i-1}}\{p_{kj'}\} = \max_{v_k\in N_{i-1}}\{p_{kj'}\}$, due to the fact that v_i is not a neighbor of $v_{j'}$. Then

$$\begin{aligned}
b_i &= f(N_{i-1}\backslash\{v_i\}) - f(N_{i-1})\\
&= \sum_{v_j\in N(v_i)\backslash N_{i-1}} \left(\max_{v_k\in N_{i-1}\backslash\{v_i\}}\{p_{kj}\} - \max_{v_k\in N_{i-1}}\{p_{kj}\}\right)\\
&\quad + \max_{v_k\in N_{i-1}}\{p_{ki}\}\\
&= \sum_{v_j\in N(v_i)\backslash N_{i-1}} \left(\max_{v_k\in N(v_j)\cap(N_{i-1}\backslash\{v_i\})}\{p_{kj}\}\right.\\
&\qquad\qquad \left. - \max_{v_k\in N(v_j)\cap N_{i-1}}\{p_{kj}\}\right) + \max_{v_k\in N_{i-1}}\{p_{ki}\}\\
&= \sum_{v_j\in N(v_i)\backslash N_{i-1}} \theta_j + \max_{v_k\in N(v_i)\cap N_{i-1}}\{p_{ki}\} \qquad (25)
\end{aligned}$$

where $\theta_j = \max_{v_k\in N(v_j)\cap(N_{i-1}\backslash\{v_i\})}\{p_{kj}\} - \max_{v_k\in N(v_j)\cap N_{i-1}}\{p_{kj}\}$. To calculate the value of θ_j, we search the neighbor list L_j of v_j from the beginning to the end since the neighbors of v_j in L_j are sorted in decreasing order of p_{kj}s. Denote by $L_j = v_{j,1}, v_{j,2}, \ldots, v_{j,n_j}$, where n_j is the number of nodes in L_j. If the first node $v_{j,1}$ is in set $N_{i-1}\backslash\{v_i\}$, then $\max_{v_k\in N(v_j)\cap(N_{i-1}\backslash\{v_i\})}\{p_{kj}\} = \max_{v_k\in N(v_j)\cap N_{i-1}}\{p_{kj}\} = p_{v_{j,1},j}$, and $\theta_j = 0$. Otherwise ($v_{j,1}$ is not in $N_{i-1}\backslash\{v_i\}$), we remove $v_{j,1}$ from L_j and search the next node, until we find a node $v_{j,k}$ such that $v_{j,k}$ is in $N_{i-1}\backslash\{v_i\}$, where $1 \le k \le n_j$. Then, $\max_{v_k\in N(v_j)\cap(N_{i-1}\backslash\{v_i\})}\{p_{kj}\} = p_{v_{j,k},j}$. Denote by n_{ij} the number of nodes removed from L_j before we find the node $v_{j,k}$. Then, the calculation of b_i takes time

$$\begin{aligned}
&\sum_{v_j\in N(v_i)\backslash N_{i-1}} \Big(O\big(n_{ij}\big) + O(1)\Big) + O(\delta_i)\\
&\qquad \le \sum_{v_j\in N(v_i)} \Big(O\big(n_{ij}\big) + O(1)\Big) + O(\delta_i)\\
&\qquad = \sum_{v_j\in N(v_i)} O\big(n_{ij}\big) + O(\delta_i) \qquad (26)
\end{aligned}$$

where $\delta_i = |N(v_i)|$ and let $n_{ij} = 0$ for $v_j \notin N(v_i)\backslash N_{i-1}$.

The calculation of all b_is takes time

$$\begin{aligned}
&O(m\log\Delta) + \sum_{i=1}^{n}\left(\sum_{v_j\in N(v_i)} O\big(n_{ij}\big) + O(\delta_i)\right)\\
&= O(m\log\Delta) + \sum_{i=1}^{n}\sum_{v_j\in N(v_i)} O\big(n_{ij}\big) + \sum_{i=1}^{n} O(\delta_i)\\
&= O(m\log\Delta) + \sum_{j=1}^{n}\sum_{v_i\in N(v_j)} O\big(n_{ij}\big) + O(m)\\
&\le O(m\log\Delta) + \sum_{j=1}^{n} O\big(\delta_j\big) + O(m) \qquad (27)\\
&= O(m\log\Delta) + O(m) + O(m)\\
&= O(m\log\Delta) \qquad (28)
\end{aligned}$$

where (27) holds since each node in list L_j will be removed no more than once, which indicates that the number of removed nodes from L_j is no more than the number of nodes in L_j when L_j is constructed, i.e., $\sum_{v_i\in N(v_j)} O(n_{ij}) \le O(\delta_j)$.

Therefore, the time complexity of Algorithm 1 is $O(m) + O(m\log\Delta) = O(m\log\Delta)$. Then, the time complexity of Algorithm 2 is $O(\alpha m(\log\Delta)\log_{1+2\epsilon} n) = O(m(\log n)^2)$, as $\Delta \le n$ and α and ϵ are given two constants. The theorem then follows. ■

Acknowledgement

The authors would like to thank the anonymous referees and the associate editor for their constructive comments and valuable suggestions, which helped to improve the quality and presentation of the article greatly.

References

[1] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[2] I. H. Brahmi, S. Djahel, D. Magoni and J. Murphy, "A spatial correlation aware scheme for efficient data aggregation in wireless sensor networks," in *Proc. IEEE 40th Local Comput. Netw. Conf. Workshops*, 2015, pp. 847–854.

[3] N. Buchbinder, M. Feldman, J. Seffi, and R. Schwartz, "A tight linear time (1/2)-approximation for unconstrained submodular maximization," *SIAM J. Comput.*, vol. 44, no. 5, pp. 1384–1402, 2015.

[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] L. Deng *et al.*, "Approximation algorithms for min-max cycle cover problems with neighborhoods," *IEEE/ACM Trans. Netw.*, early access, Jun. 16, 2020, doi: 10.1109/TNET.2020.2999630.

[6] R. Deng, Y. Zhang, S. He, J. Chen, and X. Shen, "Maximizing network utility of rechargeable sensor networks with spatiotemporally coupled constraints," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 5, pp. 1307–1319, May 2016.

[7] F. V. Gallego, P. T. Peir, L. Alonso, and J. A. Zarate, "Combining distributed queuing with energy harvesting to enable perpetual distributed data collection applications," *Trans. Emerg. Telecommun. Technol.*, vol. 29, no. 7, pp. 1–19, 2018.

[8] B. Gedik, L. Liu, and P. S. Yu, "ASAP: An adaptive sampling approach to data collection in sensor networks," *IEEE Trans. Parallel Distrib. Syst.*, vol. 18, no. 12, pp. 1766–1783, Dec. 2007.

[9] V. C. Gungor and G. P. Hancke, "Industrial wireless sensor networks: challenges, design principles, and technical approaches," *IEEE Trans. Ind. Electron.*, vol. 56, no. 10, pp. 4258–4265, Oct. 2009.

[10] X. He, S. Liu, G. Yang, and N. Xiong, "Achieving efficient data collection in heterogeneous sensing WSNs," *IEEE Access*, vol. 6, pp. 63187–63199, 2018.

[11] IEC White Paper. *Internet of Things: Wireless Sensor Networks.* Accessed: Mar. 1, 2020. [Online]. Available: http://www.iec.ch/whitepaper/internetofthings

[12] (2013). *Intel Berkeley Research Lab*. [Online]. Available: http://db.csail.mit.edu/labdata/labdata.html

[13] Y. Li, W. Liang, W. Xu, and X. Jia, "Data collection of IoT devices using an energy-constrained UAV," in *Proc. 34th IEEE Int. Parallel Distrib. Process. Symp. (IPDPS)*, 2020. [Online]. Available: http://www.ipdps.org/ipdps2020/2020-advance-program.html

[14] 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.

[15] W. Liang, P. Schweitzer, and Z. Xu, "Approximation algorithms for capacitated minimum spanning forest problems in wireless sensor networks with a mobile sink," *IEEE Trans. Comput.*, vol. 62, no. 10, pp. 1932–1944, Oct. 2013.

[16] W. Liang, X. Ren, X. Jia, and X. Xu, "Monitoring quality maximization through fair rate allocation in harvesting sensor networks," *IEEE Trans. Parallel Distrib. Syst.*, vol. 24, no. 9, pp. 1827–1840, Sep. 2013.

[17] Y. Liang *et al.*, "Nonredundant information collection in rescue applications via an energy-constrained UAV," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 2945–2958, Apr. 2019.

[18] S.-Y. Liew, C.-K. Tan, M.-L. Gan, and H. G. Goh, "A fast, adaptive, and energy-efficient data collection protocol in multi-channel-multi-path wireless sensor networks," *IEEE Comput. Intell. Mag.*, vol. 13, no. 1, pp. 30–40, Feb. 2018.

[19] 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.

[20] R. Liu, K. Fan, Z. Zheng, and P. Sinha, "Perpetual and fair data collection for environmental energy harvesting sensor networks," *IEEE/ACM Trans. Netw.*, vol. 19, no. 4, pp. 947–960, Aug. 2011.

[21] F. Liu, Y. Wang, M. Lin, K. Liu, and D. Wu, "A distributed routing algorithm for data collection in low-duty-cycle wireless sensor networks," *IEEE Internet Things J.*, vol. 4, no. 5, pp. 1420–1433, Oct. 2017.

[22] Z. Liu, W. Xing, B. Zeng, Y. Wang, and D. Lu, "Distributed spatial correlation-based clustering for approximate data collection in WSNs," in *Proc. IEEE 27th Int. Conf. Adv. Inf. Netw. Appl. (AINA)*, 2013, pp. 56–63.

[23] T. Lu, G. Liu, W. Li, S. Chang, and W. Guo, "Distributed sampling rate allocation for data quality maximization in rechargeable sensor networks," *J. Netw. Comput. Appl.*, vol. 80, pp. 1–9, Feb. 2017.

[24] S. Mao, M. H. Cheung, and V. W. Wong, "Joint energy allocation for sensing and transmission in rechargeable wireless sensor networks," *IEEE Trans. Veh. Technol.*, vol. 63, no. 6, pp. 2862–2875, Jul. 2014.

[25] *Measurement and Instrumentation Data Center*. Accessed: Mar. 30, 2019. [Online]. Available: http://midcdmz.nrel.gov/

[26] A. Mehrabi and K. Kim, "Maximizing data collection throughput on a path in energy harvesting sensor networks using a mobile sink," *IEEE Trans. Mobile Comput.*, vol. 15, no. 3, pp. 690–704, Mar. 2016.

[27] A. A. Qamaji and B. Atakan, "On exploiting spatial correlation for energy harvesting wireless sensor networks," in *Proc. 25th Signal Commun. Appl. Conf. (SIU)*, 2017, pp. 1–4.

[28] X. Ren, W. Liang, and W. Xu, "Quality-aware target coverage in energy harvesting sensor networks," *IEEE Trans. Emerg. Topics Comput.*, vol. 3, no. 1, pp. 8–21, Mar. 2015.

[29] G. A. Shah and M. Bozyigit, "Exploiting energy-aware spatial correlation in wireless sensor networks," in *Proc. Int. Conf. Commun. Syst. Softw. Middleware*, 2007, pp. 1–6.

[30] G. B. Tayeh, A. Makhoul, C. Perera, and J. Demerjian, "A spatial-temporal correlation approach for data reduction in cluster-based sensor networks," *IEEE Access*, vol. 7, pp. 50669–50680, 2019.

[31] M. H. Tsai and Y. M. Huang, "A sub-clustering algorithm based on spatial data correlation for energy conservation in wireless sensor networks," *Sensors*, vol. 14, no. 11, pp. 21858–21871, 2014.

[32] L. Villas, A. Boukerche, H. Oliveira, R. Araujo, and A. Loureiro, "A spatial correlation aware algorithm to perform efficient data collection in wireless sensor networks," *Ad Hoc Netw.*, vol. 12, pp. 69–85, Jan. 2014.

[33] C. Wang, S. Guo, and Y. Yang, "An optimization framework for mobile data collection in energy-harvesting wireless sensor networks," *IEEE Trans. Mobile Comput.*, vol. 15, no. 2, pp. 2969–2986, Dec. 2016.

[34] W. Xu *et al.*, "Approximation algorithms for the team orienteering problem," in *Proc. 39th IEEE Int. Conf. Comput. Commun. (INFOCOM)*, 2020.

[35] S. Yoon and C. Shahabi, "Exploiting spatial correlation towards an energy efficient clustered aggregation technique (CAG)," in *Proc. IEEE Int. Conf. Commun. (ICC)* , 2005, pp. 3307–3313. [Online]. Available: https://infocom2020.ieee-infocom.org/accepted-paper-list-main-conference

[36] S. Yoon and C. Shahabi, "The clustered aggregation (CAG) technique leveraging spatial and temporal correlations in wireless sensor networks," *ACM Trans. Sensor Netw.*, vol. 3, no. 1, pp. 1–39, 2007.

[37] R. Zhang, J. Peng, W. Xu, W. Liang, Z. Li, and T. Wang, "Utility maximization of temporally-correlated sensing data in energy harvesting sensor networks," *IEEE Internet Things J.*, vol. 6, no. 3, pp. 5411–5422, Jun. 2019.

Zhenjie Guo received the B.Sc. degree in computer science from China West Normal University, Nanchong, China, in 2016. He is currently pursuing the master's degree in computer science with Sichuan University, Chengdu, China.

His research interests include wireless sensor networks and mobile computing.

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 wireless sensor networks, big data, and cloud computing.

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 is currently an Associate Professor with Sichuan University, Chengdu, China. His research interests include wireless *ad hoc* and sensor networks, mobile computing, approximation algorithms, combinatorial optimization, online social networks, and graph theory.

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 Australian National University, Canberra, ACT, Australia, in 1998.

He is currently a Full Professor with the Research School of Computer Science, Australian National University. His research interests include wireless *ad hoc* and sensor networks, mobile-edge computing and cloud computing, approximation algorithms, and graph theory.

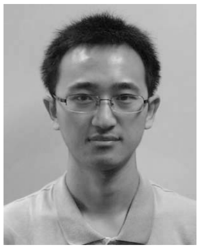
Weigang Wu (Member, IEEE) received the B.Sc. and M.Sc. degrees from Xi'an Jiaotong University, Xi'an, China, in 1998 and 2003, respectively, and the Ph.D. degree in computer science from Hong Kong Polytechnic University, Hong Kong, in 2007.

He is currently a Full Professor with the School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China. He has published more than 60 papers in major conferences and journals. His research interests include distributed systems and wireless networks, especially, cloud computing platforms, and *ad hoc* networks.

Prof. Wu has served as a member of the editorial board of two international journals, *Frontiers of Computer Science* and *Ad Hoc and Sensor Wireless Networks*. He is also an Organizing/Program Committee Member for many international conferences. He is a member of ACM.

Bing Guo received the B.S. degree in computer science from Beijing Institute of Technology, Beijing, China, in 1991, and the M.S. and Ph.D. degrees in computer science from the University of Electronic Science and Technology of China, Chengdu, China, in 1999 and 2002, respectively.

He is currently a Professor and Vice Dean with the School of Computer Science, Sichuan University, Chengdu. His current research interests include embedded real-time system and green computing.



Zichuan Xu (Member, IEEE) received the B.Sc. and M.E. degrees in computer science from Dalian University of Technology, Dalian, China, in 2011 and 2008, respectively, and the Ph.D. degree from Australian National University, Canberra, ACT, Australia, in 2016.

He was a Research Associate with University College London, London, U.K. He is currently an Associate Professor with the School of Software, Dalian University of Technology. His research interests include cloud computing, software-defined networking, wireless sensor networks, algorithmic game theory, and optimization problems.

Yue Ivan Wu (Member, IEEE) received the B.Eng. degree in communication engineering and the M.Eng. degree in communication and information system from the University of Electronic Science and Technology of China, Chengdu, China, in 2004 and 2007, respectively, and the Ph.D. degree in electronic and information engineering from Hong Kong Polytechnic University, in 2010.

He was a Postdoctoral Fellow or a Research Fellow in 2011 and from 2015 to 2016 with Hong Kong Polytechnic University. He was a Lecturer with the University of Electronic Science and Technology of China from May 2010 to June 2013. He joined VALENS with Nanyang Technological University, Singapore, as a Research Fellow in June 2011, and then the University of Florida, Gainesville, FL, USA, as a Postdoctoral Associate in July 2012. Since July 2013, he has been an Associate Professor with the College of Computer Science, Sichuan University, Chengdu. He currently works as a Project Director with the Department of Information Science, National Natural Science Foundation of China. His research interests include space-time signal processing, wireless sensor networks, and wireless channel modeling.

Dr. Wu is on the editorial boards of *IET Signal Processing*, IEEE ACCESS, and *Telecommunication Systems*.