Spatial Correlated Data Collection in Wireless Sensor Networks With Multiple Sinks

Bin Cheng, Zhezhuang Xu, Cailian Chen, Xinping Guan
Department of Automation, School of Electronic, Information and Electrical Engineering
Shanghai Jiao Tong University, Shanghai, P.R.China
Email: {cb3974, jmmouse, cailianchen, xpguan}@sjtu.edu.cn

***Abstract*—Due to the high density of node deployment in wireless sensor network, the sensing data of nodes in spatially proximate locations are highly correlated. By effectively exploiting this spatial correlation in the data collection process, unnecessary energy costs for redundant data transmission can be largely reduced. In this paper, we focus on collecting spatial correlated data in multi-sink scenario. The main challenge in this scenario is that data collection process should consider how to exploit the spatial correlation and decide which sink the data are transmitted to at the same time. To address this challenge, we propose an algorithm to select a subset of sensor nodes to represent the whole multi-sink sensor network based on the spatial correlated sensing readings. In this algorithm, only these representatives named *sources* need to upload their data to the chosen sinks. The problem is firstly formulated as a Binary Integer Linear Programming (BILP). Since the problem is proved to be NP-Complete, two heuristic algorithms are designed for approximation. The simulation results show that the proposed algorithms can largely reduce the number of the sources and then significantly improve energy efficiency.**

I. Introduction

Wireless sensor networks (WSNs) are composed of a large amount of sensor nodes. These sensor nodes are wildly deployed to construct a sensor network and accomplish the preassigned tasks. One of the fundamental functions of wireless sensor networks is data collection. Each sensor node periodically collects local data of interest, such as temperature and humidity, and reports the samples to the sink nodes. Via the reported sensing data, the sink nodes can estimate or reconstruct the interest phenomenon in the sensing region. In such application, all the sensor nodes are required to continuously sample and report local environmental data. Therefore energy efficiency becomes a key challenge for data collection. However, due to the high density of the deployment, the observations of spatially proximal sensors are highly correlated. It means collecting data from all sensor nodes may cause information redundancy and consume a large amount of unnecessary energy. Therefore this inherent spatial correlation can be exploited to develop efficient approaches for decreasing traffic redundancy and reducing energy consumption.

To describe this correlation and bring significant potential advantages for energy efficiency, several algorithms and protocols have been proposed. In [1], a theoretical framework is developed to model the spatial correlations. Based on this framework several approaches are discussed to exploit spatial correlation for efficient medium access. In recent research, there are two major research directions to capture the spatial correlation and design energy efficient data collection strategies. The first one is to compress the volume of reported information of every node. This category includes Slepian-Wolf coding and explicit communication coding (conditional coding). Slepian and Wolf [2] prove that it is theoretically possible that distributed sources can encode the correlated information at the rate of their joint entropy even if there are no information exchanges among the sources. Unfortunately the method requires a perfect prior knowledge of the whole network which is always not easy to provide in the real application. On the other hand, Cristescu et al. [3], Pattern et al. [4] exploit the correlation through explicit communication among sensor nodes, which a node encodes its data depends on those data relayed by itself. The other direction is to select a small subset of sensor nodes as the representatives of the whole network to transmit samples to the sink. The samples from these representative nodes are sufficient for the reconstruction of phenomenon in the sensing region. Gupta et al. [5] develop a set of energy efficient distributed algorithms and competitive centralized heuristics for constructing the correlation-dominating set in small size. Liu et al. [6] design a cluster algorithm to group sensor nodes into several clusters based on correlations and randomly choose cluster members as cluster head to represent the whole cluster. Xu et al. [7] consider the correlated data collection with mobile sinks, therefore the representative nodes selection problem turns to be a sink route schedule problem.

To facilitate the efficient data collection and expand deployment in large scale, it can be envisioned that sensor networks will consist of multiple sink nodes. For these reasons, we concentrate our attention on spatially correlated data collection in multi-sink scenario. As shown before, most of related work in this field focus on the data collection with one sink, and only few work is multi-sink supported. In [8], authors exploit the correlation by localized Slepian-Wolf Coding and the multi-sink supported collected data transmission structure is constructed by solving an optimization problem. In [9], Cristescu et al. jointly optimize the transmission structure and Slepian-Wolf Coding rate allocation across the source node in several transmission scenarios including multi-sink scenario. Different from prior work, we exploit spatial correlation in multi-sink scenario by selecting a subset of sensor nodes as *sources*. These sources represent the whole multi-sink wireless

978-1-4244-9920-5/11/$26.00 ©2011 IEEE

sensor network to sample the environmental surroundings and report data to the sink nodes. The readings of non-source nodes can be estimated by using the reported data from the sources. The challenge of this algorithm is that the sensor node should decide whether it is a source and which sink its data is reported at the same time. In this paper, the problem is firstly formulated as a Binary Integer Linear Programming (BILP). The objective of this BILP is to minimize energy consumption of entire data collection process. In BILP, by choosing feasible sources and their optimal reported paths to the sinks, both the source and the sink selection are achieved. Due to the high complexity of BILP, two greedy heuristic algorithms are developed for approximation. To assess the spatial correlation, the conditional entropy is introduced in this paper. It has been regarded as an effective technique to quantize the degree of similarity. We assume that a node can be represented by other nodes if their the conditional entropy is less than a threshold.

Main contributions of this paper is listed as follows:

- Multi-sink supported: Different from prior works, our work exploits the spatial correlation in multi-sink scenario by selecting several nodes as representatives to complete the data collection task. The proposed algorithm inherits the strong points of multiple sinks and takes advantage of the spatial correlation. Both characteristic can reduce the number of sensor nodes taking part in the data collection process.
- Joint optimization: We formulate a optimization problem jointly considering the source node selection and the reported sink selection. A data collection process with minimized energy consumption is established by solving this problem.

The rest of this paper is organized as follows. In the next section, several important problem definitions are introduced. In section III, the correlation set selection problem is modeled as a BILP, and it is proved to be NP-Complete. In section IV, two heuristic algorithms are proposed for approximation of BILP. Results of simulation are shown in the section V. Finally, some conclusions are discussed in section VI.

II. Preliminaries

A. Network model and Assumptions

The network of sensors and sinks is represented as a graph *G = (V, E)*, where *V* is the set of nodes, and *E* is the set of wireless links. Let S_N denote the set of sensor nodes and S_K denote the set of sink nodes, $S_N \cap S_K = V$. We define $\Omega = \{(s,k) : s \in S_N,\ k \in S_K\}$ as a set of source-sink pairs. All the sensor nodes have a fixed transmission range r_{xt}. Let d_{ij} denote the distance between node *i* and node *j*. A link (*i, j*) $\in E$ exists only if $d_{ij} < r_{xt}$. A sensor node is called *source* if it is assigned to report its sensing readings. We assume that a source generates reports at a fixed rate and the reports can be transmitted to sinks through single or multi-hop transmissions. A sensor node is considered as a *relay* if it is on a route from a source to the sink. The nodes which are neither sources nor relays will be at sleep state to reduce the energy consumption.

It is assumed that the transmission power is automatically managed by the sensor nodes. During the transmission, the sensor nodes are capable to adjust their transmission power depending on the transmission distance. Consequently, the energy consumption for sending a bit data is a function of transmission distance. Therefore, we formulate energy model as $c = e_{trans} + \beta d^{\alpha} + e_{rec}$, where e_{trans} and e_{rec} are distance independent and can be fixed as a constant. βd^{α} is distance dependent. It indicates the radiated power necessary to transmit one bit over a distance *d*, where α is the exponent of the path loss ($2 \leq \alpha \leq 5$), β is a constant [J/(bit·m$^{\alpha}$)]. Additionally, for a source, it is assumed that the amount of energy consumed for sensing a bit data is constant, denoted by c_s.

B. Correlation Model

In a given sensor network, we assume that the spatial correlation degree of two nodes is proportional to the distance between them. In this case, a model frequently encountered in practice is the Gaussian Random Process, which spatial data *X* measured at *N* sensor nodes follow an *N*-dimensional multivariate normal distribution [10]:

$$f(X) = \frac{1}{(\sqrt{2\pi})^N |\Sigma|^{1/2}} e^{-\frac{1}{2}(X-\mu)^T \Sigma^{-1}(X-\mu)} \quad (1)$$

where Σ is the covariance matrix and μ is mean vector. The diagonal elements of Σ are the variances $\Sigma_{ii} = \sigma_i^2$. The rest of Σ_{ij} are the covariances of readings from node *i* and node *j*. Under our assumption, Σ_{ij} is a function of the distance of two nodes. In this paper, we use $\Sigma_i j = \sigma_i^2\ e^{(d_{ij}/\theta_1)^{\theta_2}}$ ($\theta_1 > 0$, $\theta_2 \in (0, 2]$) [11] for analysis and simulations.

Suppose the sensor set $\mathcal{N}_u^m$ is the *m*-hop open neighbor set of node *u*, i.e. the elements in $\mathcal{N}_u^m$ are all within m hops from node u. The power set of $\mathcal{N}_u^m$ is denoted by $P(\mathcal{N}_u^m)$. To evaluate the correlation between node *u* and set $W \in P(\mathcal{N}_u^m)$, we utilize the conditional entropy evaluation technique. Since data at all nodes are assumed to quantized with the same quantization step and actually the differential entropy differs from discrete one only by a constant [10], the differential entropy is introduced instead of discrete entropy in this paper. The differential entropy of a *N*-dimensional multivariate normal distribution $G_N(\mu, \Sigma)$:

$$h(G_N(\mu, \Sigma)) = \frac{1}{2} \log(2\pi e)^N |\Sigma| \quad (2)$$

Given *W*, the conditional entropy of node *u* can be calculated by:

$$\begin{aligned} h(u|W) &= h(u, W) - h(W) \\ &= \frac{1}{2} \log(2\pi e)^{N+1} |\Sigma_{W^*}| - \frac{1}{2} \log(2\pi e)^N |\Sigma_W| \\ &= \frac{1}{2} \log 2\pi e \frac{|\Sigma_{W^*}|}{|\Sigma_W|} \end{aligned} \quad (3)$$

where $W^* = W \cup \{u\}$, and *N* is the number of elements in set *W*.

Definition 1.(Correlation Set) Given Node *u* and a subset of its m-hop neighbors *W*, i.e. $W \in P(\mathcal{N}_u^m)$, if $h(u\,|W\,)$ is smaller

than a correlation threshold ε, then W is called a correlation set of node u. It means that the sensing readings of node u can be estimated by the readings from nodes in W with high confidence. All the correlation set of node u consist of a collection D_u.

Remark 1.(Correlation Set Selection Problem) By effectively exploiting the spatial correlation among nodes, correlation sets are selected to represent the whole network and report the data to the chosen sinks. The correlation set selection problem jointly considers selecting correlation sets and their reported sinks to minimize energy consumption.

III. Problem Formulation

In this section, we formulate the correlation set selection problem as a Binary Integer Linear Programming (BILP). This BILP is a twofold problem: 1) selecting correlation sets as sources. The process decides whether a node can be represented by its correlation set. 2) minimize the transmission cost from these sources to the sinks. This process selects the optimal sink to upload data. The network is assumed to be connected, that is, at least one path exists between sensors and sinks. We also assume that data from a sensor can be collected by any sink.

A. BILP formulation

First, we need to introduce the following notations:

- c_{ij} is the energy cost of the link (*i,j*), calculated by the energy model $c_{ij} = e_{trans} + \beta d_{ij}{}^{\alpha} + e_{rec}$, where d_{ij} is the distance between node *i* and node *j*.
- c_s is the energy consumption for sensing, which is a constant.
- l^{sk} is a binary variable equal to 1 if node *s* is selected as a source, and it transmits data to the sink *k*.
- f_{ij}^{sk} is a binary variable. It equals 1 only when node *s* selected as a source sends data to the sink *k* and also the link (*i,j*) is on the path from *s* to *k*.
- W_s^v is a correlation set of node *s* and the number of elements in W_s^v is $|W_s^v|$. D_s denotes the collection of set W_s^v.
- χ_v^{sk} is a binary variable equal to 1 if correlation set W_s^v is selected to represent node *s*.

The correlation set selection problem can be formulated as follows:

$$\begin{aligned} Given: &\quad c_{ij},\ c_s,\ D_s \\ Find: &\quad \chi_v^{sk},\ l^{sk},\ f_{ij}^{sk} \\ Minimize: &\quad C = \sum_{(s,k)\in\Omega}\Big(\sum_{(i,j)\in E} f_{ij}^{sk} c_{ij} + l^{sk} c_s\Big) \qquad (4) \\ Subject\ to: & \end{aligned}$$

$$\sum_{j\in S_N}(f_{sj}^{sk} - f_{js}^{sk}) = l^{sk}, \quad \forall (s,k)\in\Omega \quad \forall s\in S_N \qquad (4.a)$$

$$\sum_{j\in S_N}(f_{kj}^{sk} - f_{jk}^{sk}) = l^{sk}, \quad \forall (s,k)\in\Omega \quad \forall k\in S_K \qquad (4.b)$$

$$\sum_{j\in S_N}(f_{ij}^{sk} - f_{ji}^{sk}) = 0, \quad \forall (s,k)\in\Omega, \quad \forall i\in V, \quad s.t \quad i\neq s, i\neq k \qquad (4.c)$$

$$\chi_v^{sk} \le \frac{1}{|W_s^v|}\sum_{j\in W_s^v} l^{jk} \qquad \forall W_s^v \in D_s \qquad (4.d)$$

$$\sum_{k\in S_K}\left(\sum_{v:\, W_s^v\in D_s}\chi_v^{sk} + l^{sk}\right) = 1 \qquad \forall s\in S_N \qquad (4.e)$$

$$\sum_{k\in S_A} l^{sk} \le 1 \qquad \forall s\in S_N \qquad (4.f)$$

$$f_{ij}^{sk} \le l^{sk}, \qquad \forall (s,k)\in\Omega, \forall (i,j)\in E \qquad (4.g)$$

The objective function in (4) minimizes the overall energy consumption which includes two parts. One part is the sensing cost consumed by all sources. The other part is the transmission cost consumed by nodes on the paths from the source to the sinks. Hence, in order to minimize the energy consumption, we need to reduce the number of sources and limit the number of non-source nodes as relays taking part in the transmission. Constraints (4.a), (4.b) and (4.c) express the conservation of traffic flows. Each source generates a flow, which is collected by a sink. Constraint (4.d) ensures that if correlation set W_s^v is selected by node *s* as the representatives to send data to sink *k*, all the nodes in set W_s^v will be selected as sources. In other words, until all the nodes in W_s^v are sources, W_s^v can be used to estimate node *s*. Constraint (4.e) imposes that node *s* is either selected as a source or represented by a correlation set. Constraint (4.f) ensures that data from node *s* can be collected only by one sink. Constraint (4.g) expresses that all the flow variables from node *s* to sink *k* are 0 unless node *s* is selected as a source and reports data to sink *k*.

B. Problem Complexity Analysis

It is straightforward that the correlation set selection problem in this paper is NP, that is, we can examine whether a given set of sources can be used to represent the whole network and whether the paths from them to the sinks satisfy energy efficiency constraints in polynomial time. To prove the problem is NP-hard, firstly we make some assumptions. Suppose that there exists a virtual link between any two sinks, and the energy cost of these virtual links are extremely high. In order to conserve the energy, no sources will choose a path including a virtual link. Therefore, the addition of the virtual links will not affect the report path selection from sources to sinks. Considering a special case, all the sink nodes are also sources and the energy cost of these actual links are equal. Then in this case, the correlation set selection problem turns to be a connected correlation dominating set problem. All the sources from the correlation sets and a subset of non-source nodes who are on data report paths consist of a dominating set and the other non-source nodes can be estimated by at least one subset of dominating set. The connected correlation dominating set problem is NP-hard as

the minimum dominating set problem which is well known to be NP-hard [12]. Therefore, the correlation set selection problem is NP-Complete.

IV. Heuristic Algorithm

Due to the high computation complexity of BILP, we develop two heuristic algorithms for correlation set selection problem. Each algorithm tries to select a set of sources and choose the optimal sink to collect the data. Firstly some assumptions are made for the algorithms: 1) all the sensor nodes have a prior knowledge about the locations of the sinks. 2) each sensor node has three status: source, represented and undecided. 3) each node has a weight associated with a sink node, denoted by $\omega(s_j, a_n) = \frac{E_j}{Dist(s_j,a_n)}$, where E_j is the residual energy of node s_j, $Dist(s_j, a_n)$ is the distance between node s_j and sink nodes a_n. Before introducing two algorithms, a weight-based correlation set construction algorithm is shown as follow:

Algorithm 1 Weight-Based Correlation Set Construction

1: With *m*-round message exchanging, sensor node u acquires its *m*-hop neighbors' information, including node ID, status, residual energy, location, etc. The information is stored in NeighborList $\mathcal{N}_u^m$.
2: Node *u* selects a subset of $\mathcal{N}_u^m$ as $\mathcal{N}_u^{*m} = \{s_j | s_j \in \mathcal{N}_u^m$, the status of s_j is source or undecided$\}$. Then node u calculates the weight of the nodes in the set $\mathcal{N}_u^{*m}$.
3: For each sink node, node *u* selects the nodes with higher weight as sets $\mathcal{N}_u^{*m}(a_n) = \{s_j | \omega(s_j, a_n) > \omega(u, a_n), s_j \in \mathcal{N}_u^{*m}\}$. $\bigcup_{a_n \in S_K} \mathcal{N}_u^{*m}(a_n) = \mathcal{N}_u^{*m}$.
4: Let $P(\mathcal{N}_u^{*m}(a_n))$ denote the power set of $\mathcal{N}_u^{*m}(a_n)$. $W_u^{*v}(a_n)$ is vth element of $P(\mathcal{N}_u^{*m}(a_n))$. If $h(u|W_u^{*v})$ is less than a correlation threshold ε, $W_u^{*v}(a_n)$ is selected into CSList D_u^*.
5: After calculating all weight-based correlation set, node *u* broadcasts a CSNotify message, including its ID and CSList D_u^*.
6: When receiving a CSNotify message, a neighbor node s_j records the received information if its ID in the CSList.

A weight-based correlation set of a sensor node is composed of the nodes who have energy advantages compared to itself. The energy advantage indicates more residual energy or less distance to a sink. Both parameters are related to the improvement of energy efficiency. Therefore, it is more feasible to represent a node by one of its weight-based correlation set for saving energy and balancing the energy cost. Based on this idea, we have designed two heuristic algorithms to select correlation sets on the behalf of the whole network.

A. Correlation First Algorithm

In Correlation First algorithm, each node selects the weight-based correlation set with smallest conditional entropy to represent itself. As discussed in previous section, the conditional entropy is referred as to an index of similarity of two elements. Therefore, the node can be estimated with highest confidence by the set with smallest conditional entropy. Algorithm 2 shows the details of Correlation First Algorithm.

Algorithm 2 Correlation First Algorithm

1: Node u whose status is undecided runs Algorithm 1, and sorts all the weight-based correlation sets in D_u^* according to the conditional entropy calculated by (3).
2: Node u selects $W_u^{*v}(a_n)$, the set with smallest entropy, to represent itself, then marks its status represented.
3: Node u broadcasts a selection result, including its ID, status, and $W_u^{*v}(a_n)$. The nodes in $W_u^{*v}(a_n)$ change their status to source.
4: Source s_j in the selected set $W_u^{*v}(a_n)$ chooses a_n as its destination sink node. Among the one-hop neighbors with same destination a_n, nodes s_j selects node s_c whose weight $\omega(s_c, a_n)$ is largest as the next-hop node. If there are several candidates for the next-hop, nodes s_j randomly selects a source node among them to break even.

Since each node chooses most correlated set as its representative, the estimated distortion is theoretically minimized. It suggests that the algorithm can be designed to reduce energy consumption with estimation accuracy guaranteed. However, the analysis of estimate error between the source and the represented nodes is out of the scope of this paper.

B. Distance First Algorithm

In Distance First algorithm, each node chooses the nearest sink as its optimal destination, and selects weight-based correlation sets with the same destination as the representative candidates. Among the candidates, the one whose average weight is largest will eventually be chosen. The average weight is defined as

$$\omega_{average}(W_u^v(a_n)) = \frac{\sum_{s_j \in W_u^v(a_n)} \omega(s_j, a_n)}{|W_u^v(a_n)|}$$

where $|W_u^{*v}(a_n)|$ is the number of nodes in $W_u^{*v}(a_n)$. Algorithm 3 demonstrates the details of Distance First Algorithm.

Algorithm 3 Distance First Algorithm

1: If the status is undecided, node u runs Algorithm 1. Suppose sink a_n is the nearest sink from u, then node u marks a_n as the optimal destination.
2: Node u sorts $W_u^{*v}(a_n)$ based on the average weight, and the most weighted one is chosen as the best set to represent node u.
3: Node u marks itself represented and broadcasts a message including its status and the selected set $W_u^{*v}(a_n)$. Then node s_j in $W_u^{*v}(a_n)$ assigns its status as source.
4: Source node s_j greedily chooses a node nearer to sink a_n as the next-hop node. If there are several candidates in the neighborhood, a source node is chosen with a priority.

The Distance First algorithm is designed to minimize the report hops from the sources to the sinks. By represented by nodes nearer to a sink, the algorithm can reduce the total energy consumption of entire data collection process. The performance of the two algorithm is evaluated by simulations in the next section.

V. Performance Evaluation

In this section, we evaluate the performance of the proposed two algorithms. The results are derived from three sets of simulations that elicit various interesting properties of our algorithms. In all three simulations, the multi-sink sensor network is randomly generated. Each node collects information from 2-hop neighbors. The energy model parameters are set as e_{trans}=e_{rec}=50nJ/bit, β=100pJ/bit/m$^{\alpha}$, and α=4 [11]. The transmission range is set to 8m. The correlation threshold ε is set at the neighborhood of 0.1. It is adjustable in different scenarios. In these simulations, the power model ${\sigma_i}^2\ e^{(d_{ij}/\theta_1)^{\theta_2}}$ is used to model the correlation of the sensor readings. ${\sigma_i}^2$ is set to 5, θ_2 is fixed as 2. Since θ_1 is directly decides the relationship between correlation and distance, it varies in different scenarios.

A. The impact of the number of sink nodes

In this simulation, 200 sensor nodes are randomly deployed in a 50m×50m square area. θ_1 is set as 30. The number of sink nodes varies from 1 to 8. Fig.1 shows that energy consumption is decreased with a increase in the number of sinks. When the number of sinks is less than 4, the energy consumption drops more than 50%. However, this trend of the decrease is limited. The drop of the energy consumption becomes slight after the number of sinks is larger than 6. The increasing amount of sinks can shorten the distance between the sensor nodes and the sinks. While the number of sinks exceeds a threshold, its impact on energy consumption is weaken. It is because that most of the sources have been within a short range of sinks. Therefore the improvement brought by increasing the number of sink nodes is limited. Comparing the two algorithms, Correlation First Algorithm consumes more energy than Distance First Algorithm. The reason is that in order to maintain a theoretically estimated accuracy, more nodes are required to report its data.

B. The impact of sensor node density

In this scenario, we study the impact of the number of deployed sensor nodes on the number of sources and the total network energy consumption. The number of sensor nodes varies from 150 to 400 in a 100m×100m area. The other parameters are set as: θ_1 is 30, the number of sink nodes is 3.

Fig.2(a) shows that both algorithms can significantly reduce the number of sources. Only 25%-40% of nodes are selected as the sources. In Fig.2(b), we notice that the increment of sensor nodes slightly influence the energy consumption when the number of sensor nodes is larger than 250. The reason is that with the increased amount of sensor nodes, the number of sources approaches to saturate. Newly added sensor nodes can be estimated by an existed correlation set. Therefore these newly added sensor nodes only result in the redundancy with little contribution to energy consumption in our algorithms.

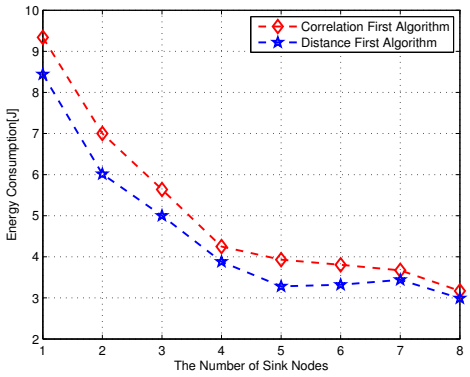


Fig. 1. Energy consumption under different number of sink nodes

C. The impact of correlation parameter θ_1

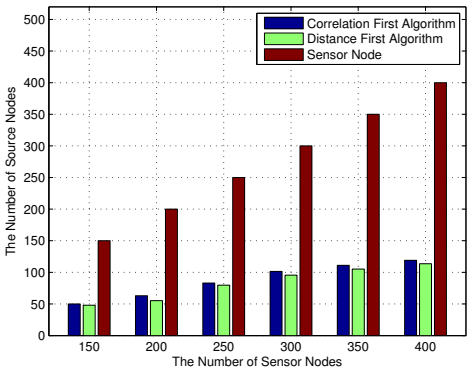
In this scenario, 200 nodes are randomly deployed in a 50m×50m area and the number of sink nodes is fixed as 3. We study the impact of θ_1 on the number of sources and total energy consumption by varying value of θ_1 from 10 to 60. The large value of θ_1 indicates a high correlation between two nodes.

Fig.3(a) demonstrates the number of sources over different value of θ_1. When θ_1 is smaller than 40, the number of sources decreases dramatically by increasing θ_1. However, the trend of this decline is limited. After θ_1 is greater than a threshold, the number of sources does not vary with the increase of θ_1. It is because the larger θ_1 is, the more nodes within 2-hop range (since we only consider the correlation relationship of nodes within 2 hops) can be represented by one correlation set and less new sources are needed to be generated. Therefore, the number of the sources intends to saturate. The change of energy consumption follows the similar changing discipline. Therefore increasing correlation degree can only improve the network performance in a certain extent.

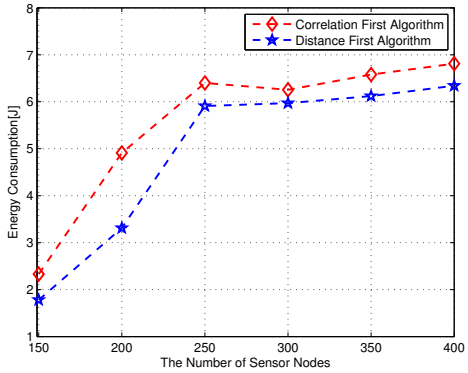
Based on simulation results given above, we can conclude that both two algorithms can significantly reduce the number of the sources and improve the energy efficiency of entire network. The energy performance of Correlation First Algorithm is a little bit worse than Distance First Algorithm. The gap between two algorithms maintains at 7%-8% on average.

VI. Conclusion

In this paper, we studied the energy efficient data collection problem in the multi-sink wireless sensor network. By exploiting the spatial correlation, only a small subset of sensor nodes are selected to upload their data to the optimal sinks. We defined this problem as the correlation set selection problem and formulate it as a BILP. The BILP is proved to be NP-Complete, and two greedy heuristic algorithms based on

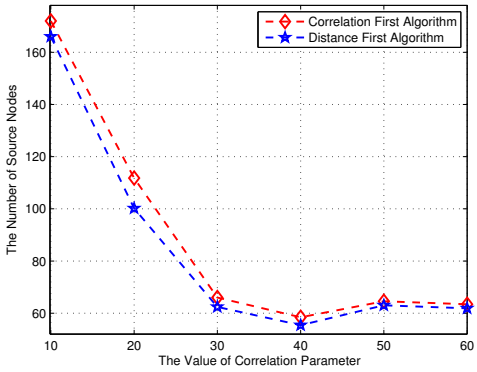


(a) The Number of sources

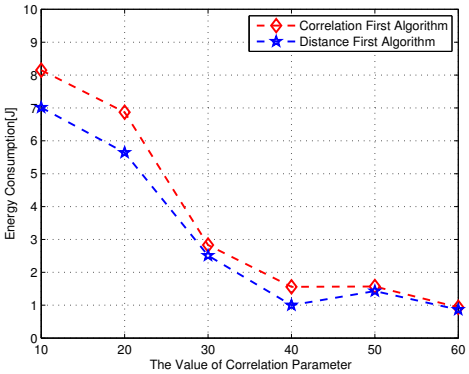


(b) The Energy Consumption

Fig. 2. The Impact of Sensor Node Density



(a) The Number of sources



(b) The Energy Consumption

Fig. 3. The Impact of Correlation Parameter θ_1

different selection criterion are proposed for approximation. Finally, the simulation results show that both two algorithms can significantly improve the energy efficiency of the data collection process.

ACKNOWLEDGMENT

The work was partially supported by National Basic Research Program of China under the grant No.2010CB731803, by NSF of China under 60804010, 60934003 and 60904123, by Science and Technology Communication of Shanghai Municipality (STCSM), China under 08511501600, by Shanghai "Pujiang" Program under 09PJ1406100, and "Chenguang" Program under 09CG06.

REFERENCES

[1] M. Vuran, O. Akan, and I. Akyildiz, "Spatio-temporal correlation: theory and applications for wireless sensor networks," *Computer Networks*, vol. 45, pp. 245–259, 2004.

[2] D. Slepian and J. Wolf, "Noiseless coding of correlated information sources," *IEEE Trans. Inf. Theory*, vol. 19, no. 4, pp. 471–480, 1973.

[3] R. Cristescu, B. Beferull-Lozano, M. Vetterli, and R. Wattenhofer, "Network correlated data gathering with explicit communication: Np-completeness and algorithms," *IEEE/ACM Trans. Netw.*, vol. 14, no. 1, pp. 41–54, 2006.

[4] S. Pattem, B. Krishnamachari, and R. Govindan, "The impact of spatial correlation on routing with compression in wireless sensor networks," *ACM Trans. Sensor Networks(TOSN)*, vol. 4, no. 8, pp. 24–33, 2008.

[5] H. Gupta, V. Navda, S. Das, and V.Chowdhary, "Efficient gathering of correlated data in sensor networks," in *MobiHoc*, Urbana-Champaign, Illinois, USA, May 2005.

[6] C. Liu, K. Wu, and J. Pei, "An energy-efficient data collection framework for wireless sensor networks by exploiting spatiotemporal correlation," *IEEE Trans. Parallel Distrib. Syst.*, vol. 18, no. 7, pp. 1011–1023, 2007.

[7] X. Xu, J. Luo, and Q. Zhang, "Delay tolerant event collection in sensor networks with mobile sink," in *INFOCOM*, San Diego, CA, USA, 2010.

[8] K. Yuan, B. Li, and B. Liang, "A distributed framework for correlated data gathering in sensor networks," *IEEE Trans. Veh. Technol.*, vol. 57, no. 1, pp. 578–593, 2008.

[9] R. Cristescu, B. Beferull-Lozano, and M. Vetterli, "Networked slepian-wolf: Theory, algorithm, and scaling laws," *IEEE Trans. Inf. Theory*, vol. 51, no. 12, pp. 4057–4073, 2005.

[10] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. John Wiley, 1991.

[11] J. Berger, V. de Oliviera, and B. Sanso, "Objective bayesian analysis of spatially correlated data," *J. Am. Statist. Assoc.*, vol. 96, pp. 1361–1374, 2001.

[12] S. Guha and S. Khuller, "Approximation algorithms for connected dominating sets," *Algorithmica*, vol. 20, no. 4, 1998.