Data Quality Maximization in Sensor Networks With a Mobile Sink

Xu Xu, Weifa Liang
School of Computer Science
The Australian National University
Canberra, Australia
{grace.xu, wliang}@cs.anu.edu.au

Tim Wark
Autonomous Systems Laboratory
CSIRO ICT Centre
Brisbane, Australia
Tim.Wark@csiro.au

***Abstract*—In this paper we consider a wireless sensor network with a mobile sink moving along a fixed trajectory without stop to collect data. We assume that a few powerful and high-capacity sensors deployed nearby the trajectory are the gateways, which are able to communicate with the sink directly when the sink is within their transmission ranges. Gateways play the role of relay nodes for the other sensors in the network. We also assume that the time of data uploading from the gateways to the mobile sink is limited. Thus, data from only a subset of sensors, called packet nodes, can be collected and used to estimate those of the others. The data quality is measured by the estimation accuracy. The upper bound on the number of packet nodes for a gateway is defined as the gateway quota. We formulate the optimization problem with the objective to identify the set of packet nodes, allocate them to gateways subject to gateway quotas, and devise an energy-efficient routing protocol, such that the sink can efficiently collect data generated from packet nodes with the maximum quality. Due to the NP-hardness of this problem, we propose a heuristic with low computation complexity. We also conduct extensive experiments by simulation to evaluate the performance of the proposed heuristic.**

I. Introduction

Wireless sensor networks (WSNs) and their applications have incurred intensive research interests over the past few years. WSNs are expected to operate for an extended period of time but the fundamental constraint is the limited energy supplies on sensors. In traditional sensor networks, the major part of energy is consumed on multi-hop data transmission and this causes *single sink neighborhood problem*, where sensors near to the sink bear more data relay workload and thus deplete their energy much faster than the others [3]. Such unbalanced energy consumption among sensors will compromise the network lifetime and data delivery reliability. Recent studies have explored the use multiple sinks or mobile element(s) to address this problem and proven its effectiveness [7], [1], [5], [3]. Mobile element(s) shifts the data relay workload from individual sensors to itself(themselves) by collecting data directly from each sensor when it is moving within the sensor's transmission range.

In this paper, we consider a densely deployed sensor network with a mobile sink travelling along a pre-defined fixed trajectory to collect data. The mobile sink moves along the trajectory at a constant speed without stop. This scenario falls into some realistic applications, e.g., the sink installed on a shuttle bus following fixed routes, or carried by a truck moving along a paved track in the forest. These vehicles may be just on patrol and their speeds are considered constant during the normal moving. The sensors that are within the transmission range of the mobile sink are referred to as *gateways*. During each tour, the sink visits the gateways one by one. The other sensors beyond the transmission range of the mobile sink are referred to as *common sensors*, as illustrated by Fig. 1. There are intersections between the trajectory and transmission ranges of gateways. While the sink moving through an intersection of a gateway, the gateway is able to upload the data stored at it to the sink.

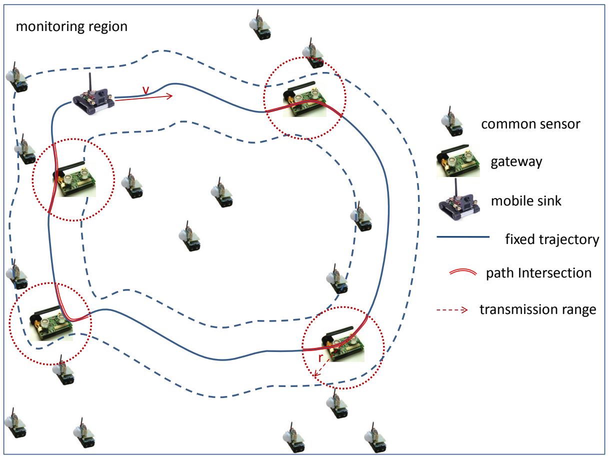


Fig. 1. A wireless sensor network with gateways, common sensors, and a mobile sink moving along a fixed trajectory

The network thus can be treated as three tiers: common sensors at the bottom tier transmit their data to the middle tier; gateways at the middle tier play the role of gateway between the bottom tier and the top tier; the mobile sink at the top tier is the data collector and collects data by directly communicating with gateways. The two extreme cases of this hierarchical network are as follows. One is that all sensors are the gateways uploading their data to the sink in one-hop data transmission mode. This is the most energy-efficient way to eliminate energy consumption on multi-hop routing, at the

978-1-4577-0513-7/11/$26.00 ©2011 Crown

cost of long data delivery latency. The other is that there is only one gateway in the entire network and it has to relay data for all common sensors. This is the traditional static sink paradigm since the sink needs to gather data from the gateway and thus does not have to move. There will be less data delivery delay but much more severe energy unbalance among the sensors. The hierarchical structure in our paper achieves a desirable trade-off between energy conservation and data collection latency.

Although several previous works have exploited the hybrid architecture for mobile collectors [1], [2], [5], [6], very few of them has taken the time of uploading data from gateways to the mobile sink into account. Indeed, such uploading time has been considered to be significant [8], [9]. We incorporate this constraint into consideration in this paper. Since the length of each intersection is limited, the amount of data the sink collects from a gateway is also limited when it travels through the corresponding intersection. Thus, the sink may not be able to collect data sensed by all common sensors. That is, data generated from some sensors, referred to as *packet nodes*, could be collected by the sink via gateways, while those from others, referred to as *non-packet nodes*, have to be discarded. Since sensors are densely deployed in the network, we assume that the sensing data from sensors are highly spatially and temporally correlated, especially among neighboring sensors [11]. The sensing data of packet nodes can be used to estimate those of non-packet nodes and the estimation accuracy is used to measure the data quality. The upper bound on the number of packet nodes for each gateway is related to the length of its intersection. We define such an upper bound as the gateway *quota* and aim to allocate packet nodes to gateways subject to the quotas. Also, packet nodes should be able to route their data to the corresponding gateways, supported by an energy-efficient routing protocol. Therefore, the objective of this paper is to *identify packet nodes, allocate the packet nodes to gateways subject to gateway quotas, and devise an energy-efficient routing protocol, such that the mobile sink moving along the fixed trajectory is able to efficiently collect data generated from packet nodes with the maximum quality*.

The main contributions of this paper are as follows:

- We propose a joint optimization frame for WSNs with a mobile sink moving along a fixed trajectory to collect data by visiting gateways such that the data quality is maximized.
- We propose a heuristic to identify the set of packet nodes, considering the constraint of the sum of quotas on gateways, such that the data of packet nodes can be used to estimate that of non-packet nodes accurately.
- We develop a strategy to partition packet nodes into disjoint sets and allocate them to gateways subject to gateway quotas, with the objective to minimize the total distance between packet nodes and corresponding gateways.
- We devise an energy-efficient routing protocol to ensure the data collection from packet nodes to the mobile sink.
- We conduct extensive experiments to demonstrate the effectiveness of the proposed approaches.

The rest of the paper is organized as follows. Section II discusses the related works. Section III proposes the problem formulation. Section IV presents a novel heuristic algorithm for the problem, and Section V evaluates the performance of the proposed algorithm through simulations. Section VI concludes the paper.

II. Related Work

There are three categories of exploiting the mobile sink to prolong the network lifetime. (i) The mobile sink visits individual sensors and collects data by one-hop data transmission. Sensors buffer data until the sink moves into their transmission ranges. This mode of data collection is the most energy efficient since no data relay is required. However, the data delivery latency is long, which is the round trip time of the mobile sink. (ii) The mobile sink moves along a trajectory and collects data via multiple hops from all sensors each time. This mode of data collection would minimize the data latency but cause high energy consumption. Moreover, the overhead of maintaining routing structure is also non-ignorable. (iii) In the hybrid mode, the sink visits a subset of nodes, which relay data for the other nodes, and collects data from them. In this way, the trade-off between data delay and energy consumption can be achieved. Our paper falls into the third category.

Most existing studies aim to find the set of nodes serving as gateways and plan the route for the mobile sink to visit the gateways. For example, in [1], a subset of nodes called rendezvous points (RPs) buffer and aggregate data from the other nodes and then transfer them to the mobile sink. The sink visits these RPs and picks up data without travelling a long distance. The optimal locations of RPs are identified subject to constraints on data delivery latency, such that the energy consumption on data collection is minimized. In our paper, we assume there is no data aggregation on any sensor in the network. In [5], the k multi-hop data routing is considered, where k is a tunable parameter. By adjusting k, the desirable balance between energy efficiency and data delivery delay can be achieved. Navigation algorithms and data transmission schemes are designed for the mobile sink to collect data from all sensors within k hops, rather than visiting each of them. Another example is [6], where Ma *et al.* propose a heuristic to design the moving path of the mobile sink called *SenCar*. SenCar moves along the designed path and some sensors close to the path transmit their data to SenCar within one hop while the others through multiple hops. The network lifetime is prolonged significantly by devising such a moving path and balancing the traffic load from sensors to SenCar. And this algorithm can be used in either connected or disconnected networks.

In some practical applications, however, the trajectory for the mobile sink is fixed and the locations for the sink to pick up data are pre-arranged. Yun *et al.* in [4] assume that there are a set of possible locations for the mobile sink to stop. The sink is supposed to collect data from all sensors by visiting a partial, or all, of these possible locations. In our paper, we consider the mobile sink moving at a constant speed without stop. We discuss the problem of maximizing the data quality when it is not possible for the sink to collect data from all sensors due to the limited time of data uploading from gateways to the sink. To the best of our knowledge, this is the first time to address the problem of data quality maximization for sensor networks with a mobile sink moving along a fixed trajectory.

III. Preliminaries

In this paper, we consider a heterogeneous network $G(GS \cup V, E)$ with a pre-defined fixed trajectory for the mobile sink, where GS is the set of gateways with $|GS| = m$, V is the set of common sensors with $|V| = n$, and E is the set of links. We assume that gateways have enough buffer sizes to store the data relayed to them until the sink collects the data, and they can be recharged by the mobile sink directly or through the infrared ray within their transmission ranges. Common sensors are randomly and densely deployed in the monitoring region and they transmit sensing data to the mobile sink via gateways. All common sensors have identical initial energy capacities which are assumed not to be rechargeable during the network lifetime. Gateways play the role of relay nodes for common sensors that will store their data temporarily until the sink passes by to collect them. Each sensor in GS and V equipped with an omni-directional antenna has a fixed, identical transmission range r. Assume the locations of GS and V are fixed and known a *priori*. There is a link between two sensors if they are within the transmission range of each other. Assume that each sensor in V has identical data generation rate r_g.

The mobile sink moves along the fixed trajectory at a constant speed v and collects sensing data of common sensors through gateways. Assuming that the length of the trajectory is L, it takes the mobile sink L/v time units to finish each tour, which corresponds the delay of data delivery. Thus, the tunable parameter v depends on the tolerant data delivery latency. Besides, since the mobile sink is powered by battery or petrol and needs to be recharged after a while, v is also related to the maximum working time of the mobile sink.

Let $GS = \{g_1, g_2, \ldots, g_m\}$ be the set of gateways. The mobile sink collects data from g_l when it moves within the intersection of the fixed trajectory and the transmission range of g_l. We define the length of such intersection I_l, $1 \leq l \leq m$. Assuming there is no data aggregation on gateways or common sensors, the maximum amount of data that can be collected by the sink from g_l per tour is

$$data(g_l) = \frac{I_l}{v} \cdot r_l, \ 1 \leq l \leq m, \tag{1}$$

where the data transmission rate on gateway g_l is r_l. Thus, the upper bound on the number of packet nodes for g_l is the number of common sensors that can transmit their data to the sink via g_l during this period,

$$c(g_l) = \lfloor \frac{data(g_l)}{\frac{L}{v} \cdot r_g} \rfloor = \lfloor \frac{I_l \cdot r_l}{v \cdot \frac{L}{v} \cdot r_g} \rfloor = \lfloor \frac{I_l \cdot r_l}{L \cdot r_g} \rfloor, \ 1 \leq l \leq m \tag{2}$$

$c(g_l)$ is referred to as the quota of g_l.

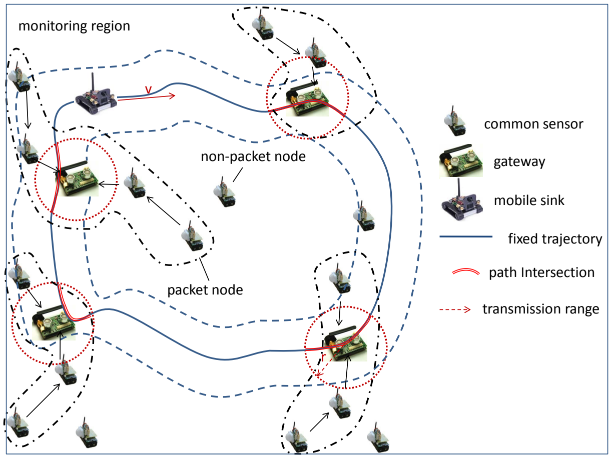


Fig. 2. $n = 16$, $m = 4$, and $c(g_1) + c(g_2) + c(g_3) + c(g_4) = 12$, only 12 common sensors can be chosen as the packet nodes. And the rest 4 nodes are non-packet nodes whose sensing data are discarded.

If $\sum_{l=1}^{m} c(g_l) < n$, the mobile sink is unable to collect data generated from all sensors per tour and the data generated from some of them have to be discarded. The set of *packet nodes*, denoted by V', transmit their data to the sink via gateways while data generated from *non-packet nodes* are ignored. Such example is shown in Fig. 2. We consider applications of collecting data such as temperature, humidity, etc., where we expect there are highly spatially and temporally correlated measurements among sensors, especially among neighboring sensors. Thus, sensing data of non-packet nodes can be estimated by those of packet nodes. Assuming time is slotted and x_{tj} represents the sensing data generated from sensor v_j at time slot t, $1 \leq t \leq T$, $1 \leq j \leq n$, the original readings generated from n sensors within a period T will form the following $T \times n$ matrix.

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1n} \\ x_{21} & x_{22} & \ldots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{T1} & x_{T2} & \ldots & x_{Tn} \end{pmatrix}$$

The squared estimated error $(x_{tj} - \widehat{x}_{tj})^2$ is used to measure the estimation error of v_j at time t. Note that for each packet node $v_j \in V'$, $(x_{tj} - \widehat{x}_{tj})^2 = 0$ at any time slot. We use the mean squared estimation error of all nodes per time unit over T to measure the *data quality*.

$$mse(T) = \frac{\sum_{v_j \in V} \sum_{t=1}^{T} (x_{tj} - \widehat{x}_{tj})^2}{n \cdot T} \tag{3}$$

The set of packet nodes V' should be divided into m disjoint subsets D_l, and then allocated to gateways under the constraint $|D_l| \leq c(g_l)$, $1 \leq l \leq m$. Besides, as shown in Fig. 2, nodes in D_l should be located to g_l as closely as possible since it is more energy-efficient to transmit data.

Having m sets of packet nodes associated with m gateways, a feasible routing protocol is required to ensure that the mobile sink can collect data from packet nodes via corresponding gateways efficiently.

The *quota-constrained data quality maximization problem* in a sensor network $G(GS \cup V, E)$ with a mobile sink moving along a fixed trajectory at a constant speed is to jointly identify packet nodes, partition them into m disjoint sets for gateways subject to gateway quotas, and design an energy-efficient routing protocol for data collection, such that the mobile sink is able to efficiently gather data generated from all these packets nodes per tour through gateways and the estimation error is minimized.

Consider a special case of the problem assuming all gateways have the same quota Q. This problem can be reduced from the *Capacitated Minimum Spanning Tree Problem* (CMSTP) [14], by assigning each edge with weight 1 and assuming Q to be the cardinality constraint on the number of nodes in any subtree. The CMSTP is known to be NP-hard when $3 \leq Q \leq \lfloor \frac{n}{2} \rfloor$ [12]. Thus, the problem of concern in this paper is NP-hard, too.

IV. Heuristic

Due to NP-hardness of the problem, in this section we propose a heuristic by decomposing the problem into three sub-problems: finding the set of packet nodes to maximize the data quality; partitioning the packet nodes into m disjoint sets to meet the quota of each gateway; and devising an energy-efficient routing protocol to ensure the data collection.

A. *Finding the set of packet nodes*

To find a subset of V, denoted by V', $|V'| = \sum_{l=1}^{m} c(g_l)$, such that the estimation error is minimized. Note that $V' = V$ and the error is zero if $\sum_{l=1}^{m} c(g_l) = n$.

To identify the set of packet nodes, an *initial training phase* $P1$ is needed to explore the correlation of sensing data from all common sensors. During this phase, sensors transmit sensing data to their neighboring nodes. According to [10], data generated from v_i at time slot t can be used to estimate those of its neighboring node v_j at t with the minimum error by the following equation:

$$\widehat{x_{tj}} = a_{i,j} \cdot x_{ti} + b_{i,j} \tag{4}$$

where $a_{i,j}$ together with $b_{i,j}$ is referred to as the *estimation model* and can be calculated as follows:

$$b_{i,j} = \frac{\sum_{t=1}^{P1} x_{tj} - a_{i,j} \cdot \sum_{t=1}^{P1} x_{ti}}{P1} \tag{5}$$

and

$$a_{i,j} = \frac{n \sum_{t=1}^{P1} x_{ti} x_{tj} - \sum_{t=1}^{P1} x_{ti} \sum_{t=1}^{P1} x_{tj}}{n \sum_{t=1}^{P1} (x_{ti})^2 - (\sum_{t=1}^{P1} x_{ti})^2} \tag{6}$$

We define the tolerant error threshold ε. A common sensor v_i can *represent* its neighboring sensor v_j (including itself) if and only if $\frac{\sum_{t=1}^{P1} (x_{tj} - \widehat{x_{tj}})^2}{P1} \leq \varepsilon$.

For a given ε, each common sensor builds a set $Cand\_v_i$ containing nodes that it can represent, $1 \leq i \leq n$. That is, for each node in $Cand\_v_i$, its generated data within $P1$ can be estimated by those of v_i with the mean error no greater than ε. Note that $v_i \in Cand\_v_i$ always holds. Let $\mathcal{C} = \{Cand\_v_i | v_i \in V\}$ be the collection of sets derived by the set of common sensors. We aim to find $V' \subseteq V$ with $|V'| = \sum_{l=1}^{m} c(g_l)$ such that $\bigcup_{v_i \in V'} Cand\_v_i = V$.

The algorithm proceeds iteratively. Initially, ε is assigned with a small positive number and its value is increased by a constant Δ_ε each iteration. Each time, a V' will be selected to represent all common sensors and the size of V' will be checked. For example, at the very beginning with a small ε, every $Cand\_v_i$ only contains v_i itself, and thus all common sensors will be chosen as packet nodes. With the increase of ε, each $Cand\_v_i$ will contain more sensors and the size of the resultant V' will decrease. The iteration continues until $|V'| \leq \sum_{l=1}^{m} c(g_l)$. If $|V'| < \sum_{l=1}^{m} c(g_l)$, $\sum_{l=1}^{m} c(g_l) - |V'|$ nodes are randomly chosen from $V - V'$ to be packet nodes. The current value of ε is an upper bound of data estimation error of non-packet nodes per time unit over the initial training phase. The detailed description of the proposed algorithm to identify packet nodes is as follows.

```
Find_Packet_Nodes($V, \Delta_\varepsilon$)
begin
        $V' \leftarrow V$;
        $\varepsilon \leftarrow 0$;
        while $|V'| > \sum_{l=1}^{m} c(g_l)$ do
                $\varepsilon \leftarrow \varepsilon + \Delta_\varepsilon$;
                build $Cand\_v_i$ for each $v_i \in V$
                $\mathcal{C} \leftarrow \{Cand\_v_i | v_i \in V\}$;
                $V' \leftarrow$Set_Cover($V, \mathcal{C}$);
        endwhile
        if $|V'| < \sum_{l=1}^{m} c(g_l)$
                randomly choose $\sum_{l=1}^{m} c(g_l) - |V'|$
                sensors from $V - V'$ as packet nodes
                and add them into $V'$;
        endif
        return $V'$;
end
```

```
Set_Cover($V, F$)
begin
        $U \leftarrow V$;
        $\mathcal{S} \leftarrow F$;
        $S' \leftarrow \emptyset$; /* the set of packet nodes */
        while $U \neq \emptyset$ do
                select a set $Cand\_v_i \in \mathcal{S}$ such that
                $|U \cap Cand\_v_i|$ is maximized;
                $S' \leftarrow S' \cup \{v_i\}$;
                $U \leftarrow U - Cand\_v_i$;
                $\mathcal{S} \leftarrow \mathcal{S} - \{Cand\_v_i\}$;
        endwhile
        return $S'$;
end
```

Algorithm Set_Cover can be implemented in time $O(n^3)$. Since the number of iterations in Find_Packet_Nodes is bounded by n, the time complexity of `Find_Packet_Nodes` is $O(n^4 m)$. The algorithm `Find_Packet_Nodes` delivers a set of packet nodes V' with $|V'| = \sum_{l=1}^{m} c(g_l)$, and the estimation model used to estimate data of non-packet nodes for the following *operation phase* $P2$. Note that due to the dynamic nature of sensing data, such estimation model may be outdated and no longer accurate enough to reflect the correlation of sensing data after a while. Thus, the model is supposed to be updated during the network lifetime to ensure the data quality. The impact of the ratio of $P1$ on the network performance will be analyzed in Section V.

B. Partitioning packet nodes into m disjoint subsets

We then partition V' into m disjoint subsets $D_1, D_2, \dots, D_m$ for gateways subject to $c(g_l)$, $1 \le l \le m$. Since the energy consumption on data transmission from one source to a destination is related to the distance between them, it will be more energy-efficient if the packet nodes in D_l and g_l are close to each other. Thus, we aim to find m disjoint subsets of V', that is, $\bigcup_{l=1}^{m} D_i = V', D_l \cap D_k = \emptyset$, $1 \le l, k \le m, l \ne k$, such that: (i)$|D_l| \le c(g_l)$, $1 \le l \le m$; (ii)$\sum_{l=1}^{m} \sum_{u \in D_l, g_l \in GS} dist(u, g_l)$ is minimized. The following three-step strategy will solve this problem.

1) Constructing a bipartite graph: First, we construct a weighted bipartite graph $G' = (V', GS', E', \omega)$, where $GS' = \bigcup_{l=1}^{m} g_{aux_l}$ is the collection of m sets of auxiliary gateways, $E' = \{\langle u, g \rangle | u \in V', g \in GS'\}$, and $\omega(u, g) = dist(u, g)$ is the Euclidean distance between u and g. Let $g_{aux_l} = \{g_{l1}, g_{l2}, \dots, g_{lc(g_l)}\}$ be a set containing $c(g_l)$ copies of g_l and $dist(u, g_{li}) = dist(u, g_l)$, for $u \in V'$ and $1 \le i \le c(g_l)$.

2) Finding a perfect matching with the minimum total weight: We then find a perfect matching E_M of G' with the minimum $\sum_{e \in E_M} w(e)$, using the Hungarian algorithm [15], where E_M is the set of matched edges with $|E_M| = m$. The computation complexity of the Hungarian algorithm is $O(|V'|^4) \le O(n^4)$. In the resultant matching, all $g \in GS'$ and $u \in V'$ are involved in m matched edges. For each matched $u \in V'$, there is a matched edge with $g \in GS'$ as the other endpoint, representing that each packet node is assigned with an auxiliary gateway.

3) Forming m disjoint subsets: We finally merge auxiliary gateways $g_{l1}, g_{l2}, \dots, g_{lc(g_l)}$ to their original node g_l and put their matched nodes into D_l. Since each auxiliary gateway $g \in GS'$ is associated with one packet node $u \in V'$, the number of matched nodes added into D_l is exactly $c(g_l)$. That is, $|D_l| = c(g_l)$, meeting the quota constraint.

C. Designing the routing protocol for data collection

Having allocated m sets of packet nodes for gateways, we then devise an energy-efficient routing protocol to enable successful data transmission from packet nodes to corresponding gateways. If the transmission range of common sensors can be adjusted by changing their transmission power, direct single-hop data transmission from packet nodes to gateways is feasible. Several papers [19], [20] investigated the problem of finding the optimal number of hops between a source node and the sink, given their distance, to minimize the energy consumption on data transmission.

In this paper, however, we assume the common sensors are simply functioned and their transmission power is fixed. Thus, multi-hop data transmission is needed if g_l is not within the transmission range of packet nodes in D_l.

Consider the original network graph $G(GS \cup V, E)$ with weight $d(u, v)$ for each edge $(u, v) \in E$. For every node $u \in D_l$, find a shortest path from u to g_l in G. The shortest path may include nodes not in D_l. Since the network is densely deployed, such shortest path always exists and the Floyd-Warshall algorithm [18] can be adopted.

Having shortest paths between each packet node $u \in D_l$ and its corresponding gateway g_l, for all $1 \le l \le m$, data transmission to the mobile sink can be guaranteed. We denote by dt_i and dr_i the amount of data transmitted from and received by v_i per time unit, $v_i \in V$. For a common sensor $v_i \in V$, its dt_i and dr_i are related to the number of nodes $c(v_i)$ which relay their data to the gateway via v_i.

$$dt_i = \begin{cases} r_g \cdot (c(v_i) + 1), & \text{if } v_i \in V' \\ r_g \cdot c(v_i), & \text{if } v_i \notin V' \end{cases} \tag{7}$$

$$dr_t = r_g \cdot c(v_i) \tag{8}$$

Since gateways can be recharged, we only concern the total energy consumption of common sensors on data collection per tour:

$$e\_tour = \sum_{v_i \in V} (e_t \cdot dt_i + e_r \cdot dr_i) \cdot \frac{L}{v}, \tag{9}$$

where e_t and e_r are energy consumed on transmitting and receiving one bit of data. The computation complexity of the Floyd-Warshall algorithm is $O(n^3)$. Together with the complexity analysis of the previous two steps, the complexity of the algorithm is $O(n^4)$. For convenience, the proposed heuristic for the data quality maximization problem is referred to as algorithm `Data_Quality_Maximization`, `DQM` for short.

V. Performance Evaluation

In this section, we evaluate the performance of algorithm `DQM` through extensive experiments by simulations. We vary network topologies with different numbers of gateways and common sensors, and evaluate the data quality in terms of data estimation error, and the energy consumption on data collection.

A. Simulation environment

Considering a $100m \times 100m$ square region, we assume that there is a fixed trajectory with $L = 200m$ for the mobile sink moving at speed $v = 2m/s$. It takes the mobile sink $100s$ to finish each tour. All gateways have identical data transmission rates $r_l = 100bit/s$, $1 \le l \le m$. Common sensors are randomly deployed and their transmission range r is fixed to be 10 meters and the initial energy capacity

IE is $100 Jules$. We adopt the energy consumption parameters of real sensors - MICA2 motes [21], where the transmission energy $e_t = 14.4 \times 10^{-6} J/bit$ and the reception energy $e_r = 5.76 \times 10^{-6} J/bit$. We also assume that the data generation rate of each common sensor is $r_g = 1 bit/s$. All experiment results in the following subsections are mean values of 50 different network topologies with identical parameter settings.

B. Sensing data generation

The synthetic data used for simulations is generated, following a random walking pattern [10], [16], [17]. The initial value of each common sensor is chosen uniformly in the range $[0, 100]$. All common sensors are partitioned into K classes and the values of sensors in the same class are generated by making random step with the same probability uniformly chosen in the range $[0.2, 1]$. That is, sensors in the same class have the same behaviors in terms of sensing data generation. When $K = 1$, all common sensors have the same behaviors and all generated data are highly correlated. With the growth of K, the correlation among data generated from all sensors decreases. For all experiments in this paper, K is fixed to be 5.

Having a set of synthetic data generated within P time units, a partial of them, which is generated within the initial training phase $P1$, is used to train the estimation model, while the remaining data generated within the operation phase $P2$ is used to evaluate the effectiveness of the model. In our default simulation setting, $P = 800$ and the initial training phase ratio $R = P1/P$ is set to be 10%. That is, data generated from all common sensors during the first 10% time units is used to build the estimation model and determine the set of packet nodes. Using the model and data generated from packet nodes within the following 90% time units, we can estimate data for non-packet nodes and compare the estimated data with the actual ones using Eq.(3) to evaluate the estimation error.

C. Impact of the initial training phase ratio R on the data quality

We evaluate the impact of the initial training phase ratio on the mean estimation error over each tour $mse(L/V)$. While fixing $n = 100$, $m = 10$, and the total gateway quota to be 90, we vary R from 2.5% to 15%. Fig. 3 shows the estimation error per tour. It is observed that with a fixed R, the estimation error goes up as the number of tour increases. As mentioned in Section IV, the estimation model is outdated and not able to accurately reflect the correlation among sensing data after a while, resulting in larger estimation error.

Also, the estimation model delivered with a larger R causes a smaller error in the same tour. That is because the longer the initial training phase, the more accurately the estimation model is built, and the longer the model can be used. For instance, with $R = 7.5\%$, the error of the 4^{th} tour is about 70, which is similar to the error of the 3^{rd} tour with $R = 2.5\%$. Thus, the estimation model with a smaller R requires to be updated more frequently.

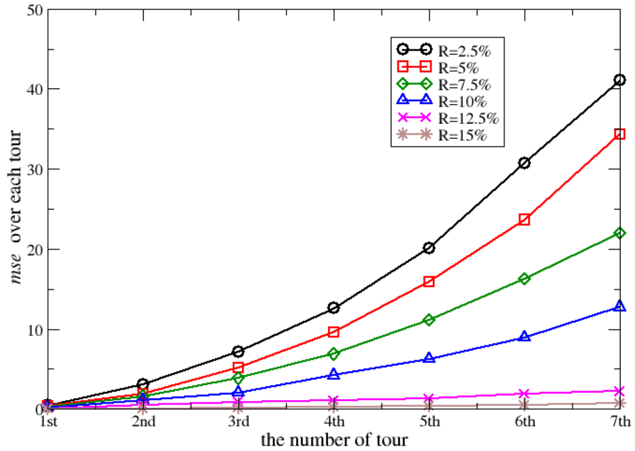


Fig. 3. Estimation error over each tour with different initial training phase ratios

D. Impact of the number of gateways m and network size n on the network performance

In this section we evaluate the mean estimation error over the operation period $mse(P2)$ as well as the total energy consumption on data collection per tour, by varying the number of gateways m and network size n.

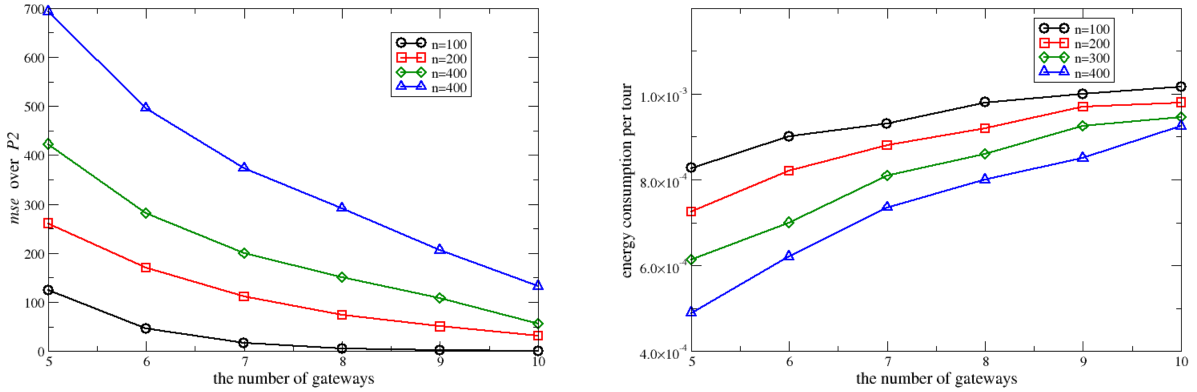
1) Impact of m: We first fix $n = 100$ and vary m from 5 to 10. Assume all gateway quotas are equal to 10. With the growth of m, the mean estimation error decreases, whereas the energy consumption increases, shown in Fig. 4(a) and (b).

For the given network size, a larger number of gateways are able to relay data for more common sensors (packet nodes) to the mobile sink, thus these packet nodes can estimate data of non-packet nodes more accurately, resulting in smaller estimation error. Such data quality improvement is at the cost of higher energy consumption on data collection per tour (Fig. 4(b), $n = 100$). This is because more energy is required to transmit data from more packet nodes to the mobile sink.

2) Impact of n: We then conduct the same experiments by varying n from 200 to 400. The delivered mean estimation error over $P2$ and the energy consumption are shown in Fig. 4(a) and (b).

Fig. 4(a) demonstrates that for a fixed m, the estimation error rises as the network size increases. The reason behind is that with the same number of packet nodes constrained by gateway quotas, the data estimation error is greater when these packet nodes have to estimate data for more non-packet nodes. For example, fixing $m = 5$ (50 packet nodes are chosen), when $n = 100$, 50% sensors estimate data for the rest 50% ones while when $n = 200$, 25% sensors estimates data for the remaining 75% of sensors, causing greater estimation error.

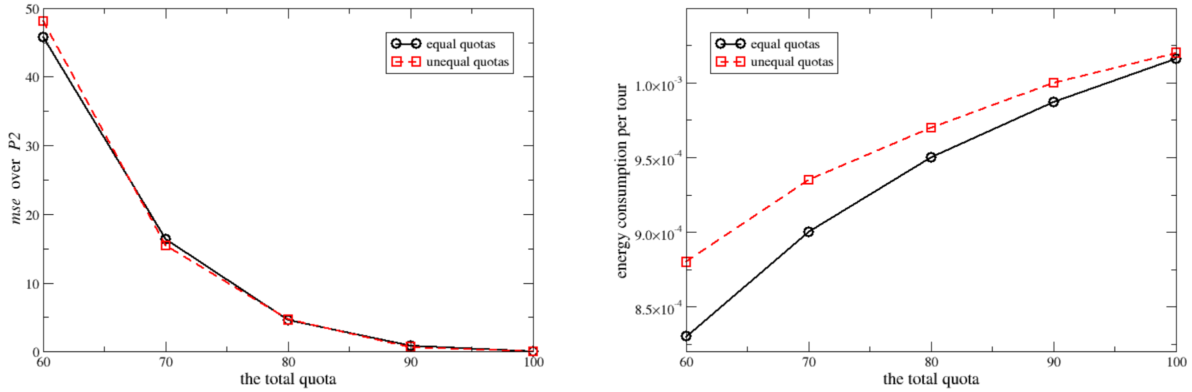
It is observed from Fig. 4(b) that the energy consumed on data collection per tour decreases as the value of n increases while m is constant. This is because the energy consumption



(a) Impact of m and n on the estimation error

(b) Impact of m and n on energy consumption

Fig. 4. Impact of m and n on the network performance



(a) Impact of the gateway quota on the estimation error

(b) Impact of the gateway quota on the energy consumption

Fig. 5. Impact of the gateway quota on the network performance

is more evenly distributed to sensors when the network scale is large.

E. Impact of the gateway quota on the network performance

We vary the total gateway quota from 60 to 100 with the increment of 10 and evaluate the network performance. Fix $n = 100$ and $m = 10$. For a given value of total quota, the 10 gateways are first assigned with equal quotas and then with unequal ones randomly generated with the fixed total value. For example, with total value 60, the 10 gateway quotas are equally to be 6 in the *equal quota* scenario. Whereas in the *unequal quota* scenario, the 10 gateway quotas vary from each other but the sum is fixed as 60.

Fig. 5(a) shows the mean estimation error over $P2$ with different total quotas. It is observed that the larger the total quota, the more packet nodes are chosen, and the more accurate data estimation is delivered. Also, with the same total quota, the estimation error with equal quotas is almost identical to that with unequal ones. It shows that the estimation error does not depend on the individual quota assignment but the total quota value.

Fig. 5(b) illustrates the energy consumption on data collection per tour with different total quotas. The consumed energy goes up with the increase of the total quota since data from more common sensors needs to be transmitted. We further notice that unequal quota distribution results in higher energy consumption because data from packet nodes has to be relayed to distant sensors to reach the corresponding gateways. The two curves intersect at total quota=100, since all common sensors are packet nodes and the routing structures are the same. However, the gap becomes larger as the total quota decreases.

VI. Conclusion

In this paper we have considered a wireless sensor network with a mobile sink moving along a fixed trajectory. The sink has been supposed to collect data by visiting gateways. We formulated the problem as a joint optimization problem, consisting of identifying the set of packet nodes, allocating them to gateways subject to gateway quotas, and devising an energy-efficient routing protocol, such that the mobile sink efficiently collects data from packet nodes with the maximized quality. We proposed a heuristic to solve it with low computation complexity. We finally evaluated the data quality and energy consumption delivered by the proposed heuristic through extensive experiments.

References

[1] G. Xing, T. Wang, W. Jia, and M. Li. Rendezvous design algorithms for wireless sensor networks with a mobile base station. *Proceedings of ACM MobiHoc*, 2008.

[2] G. Xing, T. Wang, Z. Xie, and W. Jia. Rendezvous planning in mobility-assisted wireless sensor networks. *Proceedings of IEEE Real-Time Systems Symposium (RTSS)*, 2007.

[3] W. Liang, J. Luo, X. Xu. Prolonging network lifetime via a controlled mobile sink in wireless sensor networks. *Proceedings of IEEE GLOBECOM*, 2010.

[4] Y. Yun, Y. Xia. Maximizing the lifetime of wireless sensor networks with mobile sink in delay-tolerant applications. *IEEE Transactions on Mobile Computing*, vol. 9, pp. 1308-1318, 2010.

[5] J. Rao, S. Biswas. Joint routing and navigation protocols for data harvesting in sensor networks. *Proceedings of IEEE International Conference on Mobile Ad Hoc and Sensor Systems (MASS)*, 2008.

[6] M. Ma, and Y. Yang. SenCar: an energy-efficient data gathering mechanism for large-scale multihop sensor networks. *IEEE Transactions on Parallel and Distributed Systems (TPDS)*, vol. 18, pp. 1476-1488, 2007.

[7] X. Xu and W. Liang. Placing optimal number of sinks in sensor networks for network lifetime maximization. To appear in *Proceedings of IEEE ICC*, 2011.

[8] M. Zhao, M. Ma, and Y. Yang. Efficient data gathering with mobile collectors and space-division multiple access technique in wireless sensor networks *IEEE Transactions on Computers (TC)*, vol. 60, pp.400-415, 2010.

[9] R. Sugihara, R. K. Gupta. Optimizing energy-latency trade-off in sensor networks with controlled mobility. *Proceedings of IEEE INFOCOM*, 2009.

[10] Y. Kotidis. Snapshot queries: towards data-centric sensor networks. *Proceedings of IEEE ICDE*, 2005.

[11] X. Xu, Y. Hu, J. Bi, W. Liu. Adaptive nodes scheduling approach for clustered sensor networks. *Proceedings of IEEE Symposium on Computers and Communications (ISCC)*, 2009.

[12] C. Papadimitriou. The complexity of the capacitated tree problem. *Networks*, vol. 8, pp. 217-230, 1978.

[13] L. Kou, G. Markowsky, and L. Berman. A fast algorithm for Steiner trees. *Acta Informatica*, vol. 15, pp. 141-145, 1981.

[14] R. Jothi and B. Raghavachari. Approximation algorithms for the capacitated minimum spanning tree problem and its variants in network design. *ACM Transactions on Algorithms (TALG)*, vol. 1, pp. 265-282, 2005.

[15] Harold W. Kuhn. The Hungarian Method for the assignment problem. *Naval Research Logistics Quarterly*, vol. 2, pp. 83-97, 1955.

[16] A. Deligiannakis, Y. Kotidis, and N. Roussopoulos. Hierarchical In-Network Data Aggregation with Quality Guarantees. *Proceedings of International Conference on Extending Database Technology (EDBT)*, 2004.

[17] C. Olston, J. Jiang, and J. Widom. Adaptive filters for continuous queries over distributed data streams. *Proceedings of ACM SIGMOD*, 2003.

[18] T. H. Cormen, C. E. Leiserson, R. L. Rivest. *Introduction to Algorithms (1st ed.).* MIT Press and McGraw-Hill. 1990.

[19] C. Chen, J. Ma, K. Yu. Designing energy efficient wireless sensor networks with mobile sinks. *Proceedings of ACM Sensys'06 workshop WSW*, 2006.

[20] V. Mhatre and C. Rosenberg. Design guidelines for wireless sensor networks communication: clustering and aggregation. *Ad Hoc Networks Journal*, vol. 2, pp. 45-63, 2004.

[21] Crossbow Inc. MPR-Mote Processor Radio Board User's Manual.