Online Data Gathering for Maximizing Network Lifetime in Sensor Networks

Weifa Liang, *Senior Member*, *IEEE*, and Yuzhen Liu

Abstract—Energy-constrained sensor networks have been deployed widely for monitoring and surveillance purposes. Data gathering in such networks is often a prevalent operation. Since sensors have significant power constraints (battery life), energy efficient methods must be employed for data gathering to prolong network lifetime. We consider an online data gathering problem in sensor networks, which is stated as follows: Assume that there is a sequence of data gathering queries, which arrive one by one. To respond to each query as it arrives, the system builds a routing tree for it. Within the tree, the volume of the data transmitted by each internal node depends on not only the volume of sensed data by the node itself, but also the volume of data received from its children. The objective is to maximize the network lifetime without any knowledge of future query arrivals and generation rates. In other words, the objective is to maximize the number of data gathering queries answered until the first node in the network fails. For the problem of concern, in this paper, we first present a generic cost model of energy consumption for data gathering queries if a routing tree is used for the query evaluation. We then show the problem to be NP-complete and propose several heuristic algorithms for it. We finally conduct experiments by simulation to evaluate the performance of the proposed algorithms in terms of network lifetime delivered. The experimental results show that, among the proposed algorithms, one algorithm that takes into account both the residual energy and the volume of data at each sensor node significantly outperforms the others.

Index Terms—Sensor network, data gathering, energy consumption optimization, network lifetime, sensor database, sensornet query optimization.

1 Introduction

Recent advances in microelectronic technology have made it possible to construct compact and inexpensive wireless sensors. Networks formed by such sensors, termed wireless sensor networks, have been receiving significant attention due to their potential applications in environmental surveillance, military operations, and other domains [22]. In such networks, each sensor not only serves as a host to generate sensed data and to process the collected data, but also as a router to transmit messages to and receive messages from other sensors within its transmission range. The main constraint of sensor nodes, however, is their low finite battery energies, which limit the network lifetime and impact on the quality of the network. Therefore, energy efficiency in the design of routing protocols for sensor networks is of paramount importance.

To prolong the network lifetime, many different energy optimization metrics have been proposed [24]. One typical optimization objective in the design of routing protocols is to minimize the total energy consumption, while in many practical applications, the performance measure of actual interest is not only to optimize the overall energy consumption but also to maximize the lifetime of each node in the network because a node failure in a network can cause the network partitioned and any further service will be interrupted. To avoid the extinction of nodes due to the exhaustion of their batteries, energy efficient routing algorithms should evenly distribute transmission energy load among the nodes, thereby prolonging the network lifetime. Thus, the *network lifetime* of a wireless sensor network is defined as the time of the first node failure in the network [3].

A sensor network can usually be treated as a *database* [8]. In such a database, each sensor produces one or more tuples. The node that generates tuples is termed *the source*. A collection of similarly-typed tuples from a group of sensors forms a ''snapshot.'' This snapshot constitutes a relational table which is horizontally partitioned across the sensors in the group. For example, the tuples generated by a collection of temperature sensors form a temperature table. A sensor network database allows any user to issue a data gathering query to the network and obtain a response to the query as if it is a database system. There are two typical forms of data gathering queries. One is the periodic collection of information from the sensor nodes, where the query result is updated periodically for a specified interval. Another is event-driven [19], [26], in which the occurrence of an event (a specific data gathering query arrives) triggers a data gathering query. In this paper, we will focus on event-driven data gathering queries.

When a user poses a query at a base station (the sink node) to the sensor network, the query is disseminated across the network. In response to the query, the system builds a routing tree rooted at the sink for it, each node generates tuples that match the query, and the matched tuples are transmitted toward the origin (the sink) of the query. As the tuples are routed using the routing tree consisting of all nodes, relay nodes in the tree might apply one or more database operators (e.g., aggregation operators). We refer to this kind of query processing in sensor

• *The authors are with the Department of Computer Science, Australian National University, Canberra, ACT 0200, Australia. E-mail: {wliang, yliu}@cs.anu.edu.au.*

Manuscript received 10 Feb. 2005; revised 22 Nov. 2005; accepted 14 Mar. 2006; published online 15 Nov. 2006.
For information on obtaining reprints of this article, please send e-mail to: tmc@computer.org, and reference IEEECS Log Number TMC-0028-0205.

networks as the *in-network processing*. It has shown that in-network processing is fundamental to achieve energy efficiency in energy-constrained sensor networks [18], [20], [29].

1.1 Motivations

In the following, we use two examples to illustrate our motivations in this paper. The first one is a simple data gathering query. Consider the following SQL query on a temperature table consisting of temperature tuples stored in each sensor during a certain time period:

SELECT `node_id`
FROM `sensors`
WHERE $18 \leq$ `temperature` ≤ 25
DURATION `30s`

Assume that there are many nodes in the sensor network whose temperatures are within 18C to 25C. If a routing tree is built to evaluate the above data gathering query, then the volume of the data transmitted by a relay node near to the sink is proportional to the number of nodes in the network whose temperatures are between 18C and 25C. In other words, given two nodes u and v and assuming that u is an ancestor of v in the tree, then the length of the message transmitted by u is no less than that transmitted by v.

The second example is an aggregation data gathering query. Assume that a sensor network is deployed to monitor a specific phenomenon in a given region and each sensor monitors its vicinity by taking photos of its covered area. Now, there is a data gathering query to request the up-to-date phenomenon landscape of this region. To respond to this query by the system, a routing tree is built. Within the tree, each relay node may or may not perform an aggregation operation (for an aggregation, it can remove some redundant data, e.g., overlapping areas sensed by the node and its children); thus, the volume of the data transmitted by the relay node will be no less than that received from any of its children. In the worst case, the data transmitted by the relay node is the union of its own sensed data and the data received from its children. The up-to-date phenomenon landscape of that region can then be constructed at the tree root.

In both of the above cases, the length of the message transmitted by a relay node in the routing tree depends on not only the length of the message sensed by the node itself, but also on the lengths of messages received from its children.

In this paper, we will focus on devising algorithms to find energy efficient routing trees for this type of query evaluation. Specifically, we consider the following online data gathering problem:

Given a sequence of data gathering queries which arrive one by one, the response by the system to each query is to establish a routing (spanning) tree as the query arrives. Unlike the previous assumption that each relay node only transmits the same volume of data to its parent [9], [17], [25], we assume that the volume of the data transmitted by each relay node is various, which depends on not only the volume of sensed data by the node itself, but also the volume of the data received from its children. The objective is to maximize the network lifetime without any knowledge of future query arrivals and generation rates.

1.2 Related Work

To prolong the network lifetime, energy-aware optimization in wireless ad hoc networks has been paid significant attention in recent years. Energy efficient routing protocols for wireless ad hoc networks have been addressed in [2], [3], [4], [13], [16], [27], [28]. In particular, energy efficient broadcast or multicast routing in ad hoc networks, which aims to minimize the total transmission energy consumption, has been extensively studied [2], [16], [27], [28]. To solve this problem, an energy efficient broadcast or multicast tree rooted at the source is built, and the same message is broadcast from the tree root to every other node in the tree.

The data gathering problem in sensor networks is to build a routing tree rooted at a sink node too. The difference between the broadcast tree and the routing tree is that the latter is an *inverted* tree, in which the sensed data by every sensor must be relayed to the root. During the relay period, each relay node in the tree may or may not have aggregation ability. The data gathering problem aiming at the minimization of the total energy consumption has been studied in the literature [9], [12], [17], [25]. For example, Heinzelman et al. [9] initialized the study of this problem by proposing a clustering protocol called `LEACH`. The nodes in `LEACH` are grouped into a number of clusters in a self-organizing manner, and a clusterhead serves as a local "base station" to aggregate the gathered messages from its members and forward the result to the sink directly. Lindsey and Raghavendra [17] studied the problem by providing an improved protocol called `PEGASIS`, in which all the nodes in the network form a chain and one of the nodes in the chain is chosen as the head that will be responsible for reporting the aggregated result to the base station. Tan and Körpeoğlu [25] studied the problem as well by proposing another protocol called `PEDAP`, based on the above two solutions, which uses a heuristic to assign weights to links and finds a minimum spanning tree rooted at the sink node in terms of total transmission energy consumption. Kalpakis et al. [12] considered this problem by proposing an integer program solution and a heuristic solution. It should be mentioned that, although each of the above approaches for data gathering does consider the energy consumption issue, none of them provides an energy consumption metric explicitly as the optimization objective.

It is well-known that most existing data gathering approaches based on routing trees assume that the length of the message transmitted by each relay node is independent of the lengths of its children messages, i.e., each node transmits the same volume of data no matter how much data it received from its children. Such queries in databases include `AVG`, `MIN`, `MAX`, `COUNT`, etc. However, there are a number of data gathering applications in which the length of the message transmitted by a relay node depends on not only the length of its sensed message, but also the lengths of the messages received from its children. Thus, the length of the message transmitted by each relay node varies, depending on whether or not the data (its own sensed data and the collected data from its children) is aggregated before transmitted. There are several studies that deal with this latter data gathering problem [5], [7], [21], [23] with different optimization metrics of energy consumption. For

example, Goel and Estrin [7] addressed the problem by minimizing the total transmission energy consumption, assuming that the aggregation function at each relay node is modeled as a given concave, nondecreasing cost function. They also proposed a hierarchical matching algorithm for the problem, which delivers an approximate solution that is up to a logarithmic factor of the optimum. Cristescu et al. [5] studied a variant of the problem—the data correlation problem with an objective to minimize the total transmission energy consumption. Under the assumption that each node knows which node to be merged to generate a merged message with the minimum length and each relay node has data aggregation and compression ability, they showed that the data correlation problem is NP-complete and provided an integer program solution using the Slepian-Wolf coding approach. von Rickenbach and Wattenhofer [21] recently also studied the data correlation problem by providing an approximation algorithm using the *shallow light tree* concept [15]. The solution delivered by their algorithm is within $2(1+\sqrt{2})$ times of the optimum. Intanagonwiwat et al. [10], [11] studied the general data gathering issue by incorporating the semantics of aggregation query into building an energy efficient routing tree. The tree, however, is not necessarily a spanning tree. For example, in [11], they proposed a data dissemination schema called *directed diffusion with opportunistic aggregation*, where data is opportunistically aggregated at intermediate nodes on a low-latency tree. In [10], they explored the greedy aggregation by providing a novel approach that adjusts aggregation points to increase the amount of path sharing, reducing the energy consumption. In this paper, the proposed heuristics `MMRE`, `MML`, `BT`, `MDST`, and `SPT` will also trade off different energy optimization metrics to prolong the network lifetime.

1.3 Contributions

Our major contributions in this paper are as follows: We first present a generic cost model of energy consumption for data gathering in sensor networks if a routing tree is used for query evaluation. This generic cost model unifies several well-known cost models. We then show the online data gathering problem to be NP-complete, and instead propose several heuristic algorithms for the problem. We finally conduct extensive experiments by simulation to evaluate the performance of the proposed algorithms. The experimental results show that, among the proposed algorithms, algorithm `MNL`, which takes into account both the residual energy and the volume of data at each sensor node, significantly outperforms the others. To the best of our knowledge, we are not aware of any other discussions of this problem in the literature to date.

1.4 Paper Organization

The rest of the paper is organized as follows: In Section 2, the system model is given. In Section 3, a generic cost model of energy consumption for data gathering is proposed and the online data gathering problem is defined. In Section 4, the problem is shown to be NP-complete and several heuristic algorithms are proposed. In Section 5, extensive experiments by simulation are conducted to evaluate the performance of the proposed algorithms. The conclusion is given in Section 6.

2 System Model

We consider a wireless sensor network consisting of n stationary *sensor nodes* and a base station s (also referred to as the sink node) distributed over a region. The location of each sensor and the base station are fixed and known a priori. Each sensor node is equipped with an omnidirectional antenna and able to vary its transmission power dynamically. In other words, the wireless sensor network can be modeled by a directed graph $M = (N, A)$, where N is the set of nodes with $|N| = n + 1$ and there is a directed edge $\langle u, v\rangle$ in A if node v is within the transmission range of u when u uses its maximum power level to broadcast a message. For two nodes u and v with distance $d_{u,v}$, the transmission energy at node u is modeled to be proportional to $d_{u,v}^{\alpha}$ if a unit of message is transferred from u to v directly, where α is a path-loss exponent parameter that typically takes on a value between 2 and 4, depending on the characteristics of the communication medium. Unless otherwise specified, for the sake of simplicity, in this paper, we take into account the transmission energy consumption only and assume that the other energy consumptions such as reception are negligible, as it is well-known that the radio frequency (RF) transmission is the dominant energy consumption in wireless communications.

3 Generic Cost Model of Energy Consumption

In this section, we introduce a generic cost model of energy consumption for data gathering queries in sensor networks if a routing tree will be used for such a purpose. This cost model will unify several known cost models of energy consumption.

3.1 A Generic Cost Model

Given a data gathering query, we aim to build a routing tree T rooted at the sink node and spanning all the other nodes in the network. For each node v, let $p(v)$ be the parent of v in T and m_v be the length of the sensed message by v itself. Given a relay node v in T with t children ($t \geq 1$), assume that the lengths of messages that v received from its children are $l_1, l_2, \ldots, \text{and } l_t$, respectively. v may or may not aggregate these messages and its own sensed message before transmitting them as a single message to its parent $p(v)$ during a data gathering session. In other words, the length of the message transmitted by v to its parent $p(v)$ is a function f with parameters $l_1, l_2, \ldots, l_t$ and m_v. The definition of f may vary, depending on application domains. In practice, most known approaches of data gathering in the literature make use of one of the following two different definitions of function f:

1. The length of the message transmitted by a relay node is independent of the message lengths of its children and itself, i.e., the length of the message transmitted by each relay node in the tree is identical. One typical application background of this definition is to measure the average temperature of a given region. To do so, each sensor node v has a pair of variables $(Temp_v, N_v)$, where $Temp_v$ is the sum of

sensed temperatures by all the sensors in the subtree rooted at v and N_v is the number of sensors in the subtree including v itself. Each relay node will transmit its pair of the data to its parent. In the end, the sink node s calculates the average temperature of the network, which is $Temp_s/N_s$.

2. The length of the message transmitted by a relay node depends on the message lengths of its children and itself, which is linear or sublinear to the sum of the lengths of its children messages and its own sensed one.

In the following, we assume that, for a data gathering query, a routing tree T rooted at the sink node will be built. Assume that V_T is the set of the nodes in T and $v \in V_T$. Let r_e and r_s be the amounts of energy consumption of receiving and sensing a unit of message by a sensor.

If v is a relay node with children $u_1, u_2, \ldots,$ and u_t, assume that the message length transmitted by u_i is l_i, $1 \leq i \leq t$. The cost $c(v)$ of v is thus

$$c(v) = f(l_1, l_2, \ldots, l_t, m_v) d^{\alpha}_{v,p(v)} + \sum_{i=1}^{t} l_i * r_e + m_v * r_s,$$

where the value of f is the length of the message transmitted by v, depending on the lengths of the messages transmitted by its child nodes $u_1, u_2, \ldots, u_t$ and the length m_v of the sensed message by v itself. Otherwise, v is a leaf node, f is a function of m_v only, and the cost $c(v)$ of v is

$$c(v) = f(m_v)(d^{\alpha}_{v,p(v)} + r_s),$$

which is the energy consumption of sensing and transmitting a m_v-unit message to the parent of v in T. It is obvious that $c(v)$ is the total energy consumption at node v for the current data gathering query.

Having function f defined above, to prolong the network lifetime when dealing with data gathering by using different energy optimization metrics, there are two types of energy optimization metrics that have been widely used, as follows:

- One is to find a routing tree T such that the total transmission energy consumption $\sum_{v \in V_T} c(v)$ in the tree is minimized, which aims to prolong the network lifetime through minimizing the total energy consumption per data gathering query. However, this optimization does not take into account the energy consumption at each individual node. Thus, a relay node near the tree root (the sink) may run out of energy and fail very quickly, which leads to the network being partitioned.
- Another is to find a routing tree T such that the minimum residual energy among the nodes is maximized, i.e., *maximize* $\min_{v \in V_T}\{re'(v)\}$, where $re(v)$ and $re'(v) = re(v) - c(v)$ are the residual energy at v before and after the realization of the current data gathering query. This optimization aims to prolong the network lifetime through extending the lifetime of individual nodes by maximizing the minimum residual energy among the nodes.

3.2 Several Known Cost Models of Energy Consumption

In the following, we show that three well-known cost models of energy consumption can be derived from the above generic cost model.

Case 1. Assume that $f(l_1, l_2, \ldots, l_t, m_v) = k$ for all $v \in V_T$, $r_s = 0$, and $r_e = 0$, i.e., the length of the message transmitted by v is independent of the message lengths of its children and itself.

1. If the total transmission energy consumption is considered as the optimization objective, then *the minimum energy data gathering problem* is to find a spanning tree T rooted at the sink node such that its cost $C(T)$ is minimized, where

$$C(T) = \sum_{v \in V_T} f(l_1, l_2, \ldots, l_t, m_v) d^{\alpha}_{v,p(v)} = k \sum_{v \in V_T} d^{\alpha}_{v,p(v)}. \tag{1}$$

   This cost model of energy consumption for data gathering has been widely adopted in the literature [9], [17].
2. If the minimum residual energy among the nodes is taken as the optimization objective, then *the max-min energy data gathering problem* is to find a spanning tree T rooted at the sink node such that its cost $C(T)$ is maximized, where

$$C(T) = \min_{v \in V_T}\{re(v) - k d^{\alpha}_{v,p(v)}\}. \tag{2}$$

   This cost model has been used in the literature as well [12], [25].

Case 2. Assume that $f(l_1, l_2, \ldots, l_t, m_v) = kD(v)$ and $m_v = k$ for all $v \in V_T$, $r_s = 0$, and $r_e = 0$, where $D(v)$ is the number of descendants of v in T including v itself, which means that each node just relays its descendant data to its parent without any aggregation on the data.

1. If the total transmission energy consumption is taken as the optimization objective, then *the minimum energy precise data gathering problem* is to find a spanning tree T rooted at the sink node such that its cost $C(T)$ is minimized, where

$$\begin{aligned} C(T) &= \sum_{v \in V_T} f(l_1, l_2, \ldots, l_t, m_v) d^{\alpha}_{v,p(v)} \\ &= k \sum_{v \in V_T} D(v) d^{\alpha}_{v,p(v)}. \end{aligned} \tag{3}$$

   A similar cost model has been used in the literature [5], [7], [21].
2. If the minimum residual energy among the nodes is taken as the optimization objective, then *the max-min energy precise data gathering problem* (MMEPD) is to find a spanning tree T rooted at the sink node such that the cost $C(T)$ is maximized, where

$$C(T) = \min_{v \in V_T}\{re(v) - kD(v) d^{\alpha}_{v,p(v)}\}. \tag{4}$$

Case 3. It is the mixture of Cases 1 and 2 that takes into account both optimization objectives simultaneously, which can be further divided into two subcases.

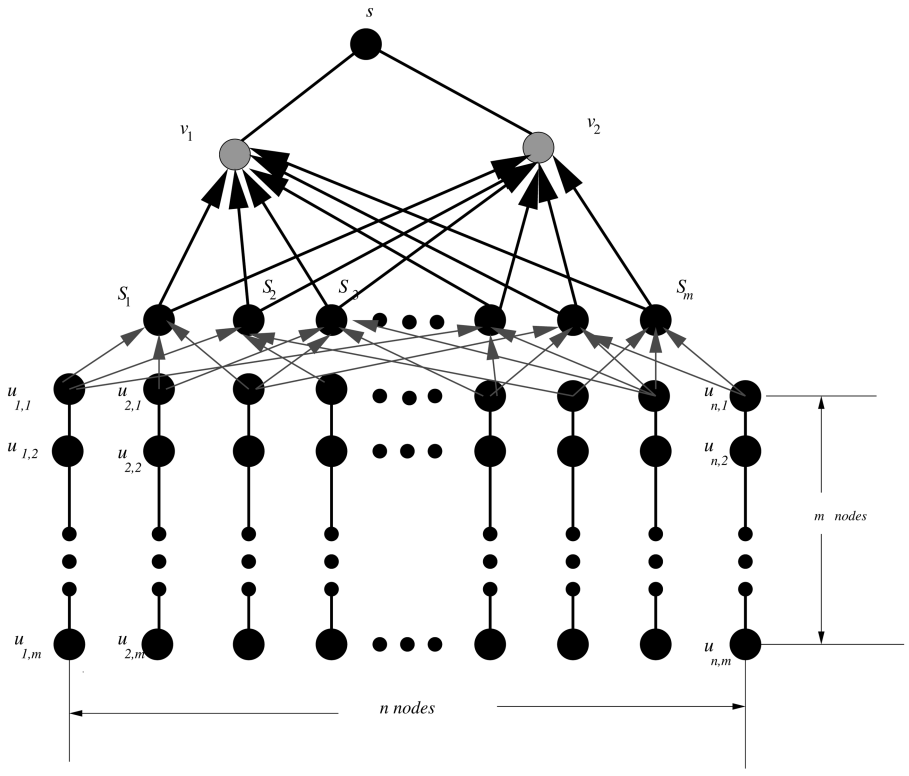


Fig. 1. An instance of a sensor network derived from an instance of SC.

When $f(l_1, l_2, \ldots, l_t, m_v) = k$ for all $v \in V_T$ and $r_s = r_e = 0$, *the balanced minimum energy data gathering problem* is to find a spanning tree T rooted at the sink node such that 1) the total transmission energy consumption in T is minimized, i.e., *minimize* $k\sum_{v \in V_T} d^{\alpha}_{v,p(v)}$, and 2) the minimum residual energy among the nodes is maximized, i.e., *maximize* $\min_{v \in V_T}\{re(v) - kd^{\alpha}_{v,p(v)}\}$.

When $f(l_1, l_2, \ldots, l_t, m_v) = kD(v)$ for all $v \in V_T$ and $r_s = r_e = 0$, *the balanced minimum energy precise data gathering problem* is to find a spanning tree rooted at the sink node such that 1) the total transmission energy consumption in T is minimized, i.e., *minimize* $k\sum_{v \in V_T} D(v)d^{\alpha}_{v,p(v)}$, and 2) the minimum residual energy among the nodes is maximized, i.e., *maximize* $\min_{v \in V_T}\{re(v) - kD(v)d^{\alpha}_{v,p(v)}\}$.

3.3 The Problem Definition

Given a wireless sensor network $M = (N, A)$ with a sink node, we assume that there is an unknown sequence of data gathering queries which arrive one by one. As a query arrives, the response by the system to the query is to build a routing tree rooted at the sink and spanning the other nodes for it. We further assume that the length of the message transmitted by each relay node in the routing tree is the sum of the lengths of its children messages and its own sensed message. The *online data gathering problem* is to maximize the network lifetime without knowledge of future query arrivals and generation rates. In other words, the problem is to maximize the number of queries answered until the first node in the network fails. Note that, although for a given query, the length of the sensed message by each sensor node is identical, it is various for different queries.

4 Algorithms for Online Data Gathering

In this section, we first show a special case of the online data gathering problem where the sequence of data gathering queries contains only one query—the decision version of MMEPD, is NP-complete through a reduction of the *minimum set cover problem* (SC for short) to it. We then propose heuristic algorithms for the online data gathering problem.

4.1 NP-Hardness of MMEPD

Given an n-element set $S = \{a_1, a_2, \ldots, a_n\}$ and a collection of subset $\mathcal{C} = \{S_1, S_2, \ldots, S_m\}$ of S, the decision version of the set cover problem is to determine whether there is a set cover $\{S_{i_1}, S_{i_2}, \ldots, S_{i_K}\}$ of cardinality K covering the elements in S, where $S_{i_j} \subseteq S$ for all i_j, $1 \le i_j \le m$, $1 \le j \le K$, and $K \le m$. We state the decision version of MMEPD is NP-complete as follows:

Theorem 1. *Given a sensor network $M = (N, A)$ with a sink node s, the decision version of* MMEPD *in M is NP-Complete.*

Proof. Given an instance of SC, an instance of MMEPD in a sensor network can be constructed as follows (see Fig. 1).

For each element a_i in S, there are m corresponding nodes $u_{i,1}, u_{i,2} \ldots, u_{i,m}$ in the network, and there is a directed edge $\langle u_{i,j+1}, u_{i,j}\rangle$ from $u_{i,j+1}$ to $u_{i,j}$ for every j, where a directed edge $\langle u, v\rangle$ from u to v means that v is within the transmission range of u, $1 \le i \le n$ and $1 \le j \le m-1$. In addition, there are m set cover nodes $S_1, S_2, \ldots, S_m$ in the network, corresponding to the m subsets of S. For each set cover node S_j, there is a directed edge $\langle u_{i,1}, S_j\rangle$ from node $u_{i,1}$ to node S_j if an element $a_i \in S_j$, $1 \le i \le n$, and $1 \le j \le m$. There are two special nodes, v_1 and v_2. There is a directed edge $\langle S_j, v_i\rangle$

from every set cover node S_j to v_i, $i = 1, 2$, and $1 \le j \le m$. There is a sink node s (the base station) equipped with unlimited energy supply. Both v_1 and v_2 are connected to the sink node.

Having constructed the sensor network, we now assume that each node has a 1-bit sensed message to transmit and consumes a unit of energy for transmitting a 1-bit message to its neighbors. We further assume that every other node except v_1 and v_2 in the network has enough residual energy, while the amounts of residual energies of v_1 and v_2 are assumed to be $mn + K + 2$ and $m - K + 2$, respectively, just before the current data gathering query is considered.

It is obvious that the decision version of an instance of SC can be reduced to an instance of MMEPD within polynomial time, since the construction of the sensor network takes polynomial time of its input size $|S| + |\mathcal{C}| = n + m$.

Given the sensor network constructed as above, MMEPD is to find a spanning tree in the network rooted at s such that the minimum residual energy among the nodes is maximized after the realization of the data gathering query. Assume that there is no aggregation at each relay node, and the length of the message transmitted by each relay node is the sum of the message lengths of its children and the node itself. It can be seen that the bottlenecks of the residual energy among the nodes are nodes v_1 and v_2. There are in total $mn + m + 2$ bits of data for the current query (assuming each sensor node senses 1-bit message) to be forwarded to s using the relay nodes v_1 and v_2. Thus, the maximum value of the minimum residual energy at v_1 or v_2 is one unit if there exists such a spanning tree T that 1) the $m - K + 1$ set cover nodes among the m set cover nodes are the children of v_2 and the leaves of T and 2) the remaining K set cover nodes are the children of v_1 and the relay nodes of T.

Given the spanning tree T constructed for MMEPD, it is easy to construct a set cover for the SC instance. That is, the union of the subsets corresponding to the K nonleaf set cover nodes in T contains all the elements in S. □

4.2 Algorithm MNL

In the following, we propose a heuristic algorithm for the online data gathering problem. For convenience, we treat the wireless sensor network $M(N, A)$ as a directed graph $G(V, E)$, where the set of nodes V consisting of sensors and $\langle u, v\rangle \in E$ if and only if u and v are within the transmission ranges of each other. The basic idea behind the proposed algorithm is that, once a data gathering query arrives, a data gathering tree for the query is constructed using a greedy policy that maximizes the minimum residual energy among the nodes. Specifically, the nodes are included into the tree one by one. Initially, only the sink node is included. Each time a node v is included into the tree, either the network lifetime derived from the current tree is at least as long as that without the inclusion of v to the tree or the amount of reduction of the network lifetime is minimized. In other words, a node v is chosen to be included into the tree if it leads to maximizing the minimum residual energy among the tree nodes including itself.

Denote by T and V_T the tree and the set of nodes included in T so far. Initially, the set of nodes in T contains the sink node only, i.e., $V_T = \{s\}$. Each time the algorithm picks up a node v from the set $V - V_T$ such that $\min\{re'(u) \mid u \in V_T \cup \{v\} - \{s\} \text{ and } \langle v, u\rangle \in E\}$ is maximized, $re'(u)$ is the residual energy at node u after the addition of v to T. The algorithm continues until $V - V_T = \emptyset$. In what follows, we detail the choice of v. Assume that $P_{u,s}$ is the unique path in T from node u to node s and $p(u)$ is the parent of u in T. Let node $v \in V - V_T$ be the considered node.

1. If there are l edges from v to the nodes in V_T, denoted by $\langle v, u_1\rangle, \langle v, u_2\rangle, \ldots, \langle v, u_l\rangle$, where $u_i \in V_T$ for all i, $1 \le i \le l$, define

$$g(v, u_i) = \min\{re(v) - kd^{\alpha}_{v,u_i}, re(u) - kd^{\alpha}_{u,p(u)} \mid u \in P_{u_i,s},\ u \neq s\}, \tag{5}$$

   which is the minimum residual energy among the nodes in the path $P_{v,s}$ if the edge $\langle v, u_i\rangle$ is included as a tree edge, $1 \le i \le l$. Now, let $g(v, u_{l_0}) = \max\{g(v, u_i) \mid 1 \le i \le l\}$ with $1 \le l_o \le l$. Then, define

$$g_{\max}(v) = g(v, u_{l_0}) = \max\{g(v, u_i) \mid 1 \le i \le l\} \tag{6}$$

   and

$$temp\_parent(v) = u_{l_0}. \tag{7}$$

   It is obvious that the edge $\langle v, u_{l_0}\rangle$ is the best choice if v is included into T after the current iteration.
2. Otherwise (there is not any edge from v to the nodes in V_T), define $g_{\max}(v) = 0$.

   A node $v_0 \in V - V_T$ is finally chosen to be added to T by setting $p(v_0) = temp\_parent(v_0)$ if

$$g_{\max}(v_0) = \max\{g_{\max}(v) \mid v \in V - V_T\}.$$

   This means that the inclusion of v_0 will result in the minimum amount or no reduction of the network lifetime.

The detailed algorithm is presented as follows. Once a data gathering query arrives, to respond to the query, the algorithm is executed immediately.

Algorithm Maximum_Network_Lifetime(G, re, k)
/* G is the current sensor network and re is an array of the residual energy at each node */
begin

1. $V_T \leftarrow \{s\}$; $terminate \leftarrow$ "false";
   /* V_T is the set of nodes in the tree, $terminate$ is a boolean variable, */
   /* and k is the size of the sensed data by node $added\_node$. */
2. $Q \leftarrow V - V_T$; /* the set of nodes that are not in the tree */
3. $re(s) \leftarrow \infty$; /* the sink node has unlimited energy supply */
4. repeat
5. $g_{\max} \leftarrow 0$;
   /* the maximal minimum residual energy at nodes in the tree */
6. for each $v \in Q$ do

7. compute $g_{\max}(v)$ and $temp\_parent(v)$;
8. if $g_{\max}(v) > g_{\max}$ then
9. $g_{max} \leftarrow g_{max}(v)$;
10. $added\_node \leftarrow v$;
/* the node that will be added to the tree */
endif;
endfor;
11. if $g_{max} > 0$ then
12. $p(added\_node) \leftarrow temp\_parent(added\_node)$;
13. for each node $u \in P_{added\_node,s}$ do
14. $re(u) \leftarrow re(u) - kd^{\alpha}_{u,p(u)}$;
endfor;
15. $V_T \leftarrow V_T \cup \{added\_node\}$;
16. $Q \leftarrow Q - \{added\_node\}$;
17. else $terminate \leftarrow$ 'true';
18. endif;
19. until ($Q = \emptyset$) or $terminate$;

end.

For the sake of convenience, we refer to algorithm `Maximum_Network_Lifetime` as `MNL` and have the following theorem:

Theorem 2. *Given a sensor network $M = (N, A)$ with a sink node s, $n = |N|$ sensor nodes, and $m = |A|$ links, there is a heuristic algorithm for the online data gathering problem in M which takes $O(mn^2)$ time for each data gathering query.*

Proof. The time complexity of `MNL` is analyzed as follows: The number of iterations from Step 4 to Step 19 is n. Within each iteration, the calculation of $g_{\max}(v)$ takes $O(ndeg_v)$ time because every node in the path $P_{v,s}$ from v to s needs to be visited at least once, where deg_v is the degree of v in the network. Thus, within an iteration, the sum of the running time of these steps is $O(n\sum_{v\in Q} deg_v) = O(mn)$, where m is the number of links in the network. Therefore, the proposed algorithm takes $O(mn^2)$ time. □

4.3 Other Heuristic Algorithms

In the following, we present the other four heuristics for the problem, which will be used as benchmarks to evaluate the performance of algorithm `MNL`. We start by introducing two simple heuristic algorithms: algorithm `MMRE` and algorithm `SPT` based on single-source shortest path trees. We then present two more involved heuristics: algorithm `BT` and algorithm `MDST`.

4.3.1 Algorithm `MMRE`

Algorithm `MMRE` aims to prolong the network lifetime through the maximization of the minimum residual energy (MMRE) among the nodes in the network, which is similar to an algorithm given in [13]. The only difference between them is that, in this case, an inverted spanning tree instead of a broadcast tree is constructed. Given a data gathering query with the length k of the sensed message by each node, algorithm `MMRE` proceeds as follows.

Let T be the tree and V_T be the set of nodes in T. The sink node s is included in T and $V_T = \{s\}$ initially. Each time it picks up a node $v \in V - V_T$ if v satisfies

$$re'(v) = \max_{\langle v',u'\rangle \in E} \{re(v') - kd^{\alpha}_{v',u'} \mid v' \in V - V_T,\ u' \in V_T\}, \quad (8)$$

where $re(v)$ and $re'(v)$ are the residual energies at v before and after the current data gathering query is realized. Add node v and the edge $\langle v, u\rangle$ into T, where $u = p(v)$ is the parent of v in T. The algorithm continues until $V - V_T = \emptyset$.

4.3.2 Algorithm `SPT`

Algorithm `SPT` aims to prolong the network lifetime through the minimization of the total transmission energy consumption of relaying the sensed message from a sensor to the sink node. Given a data gathering query with the length k of sensed message by each node, algorithm `SPT` proceeds as follows.

An energy graph $G(V, E)$ is derived from the sensor network, where V is the set of sensor nodes and the sink node s. There is a directed edge $\langle u, v\rangle$ in E from u to v if the residual energy at u is at least $kd^{\alpha}_{u,v}$. The weight assigned to the edge is $d^{\alpha}_{u,v}$, which is the energy consumption of transmitting a unit message between the two nodes. A single-source shortest path tree rooted at the sink node is constructed. Clearly, for each node v, the minimum transmission energy consumption to send its k-unit sensed message to the sink node is $kW(P_{v,s})$, where $P_{v,s}$ is the unique path in the tree from v to s and $W(P)$ is the weighted sum of the edges in path P. Thus, the total transmission energy consumption for realizing a data gathering query is $\sum_{v\in V} kW(P_{v,s})$.

4.3.3 Algorithm `BT`

Given a data gathering query, an undirected, energy graph $G(V, E, \omega)$ for the sensor network is defined, where V is the set of sensor nodes and E is the set of undirected links. A link $(u, v) \in E$ if 1) u and v are within the transmission range of each other when they use their maximum power to transmit a message and 2) the residual energy $re(u)$ and $re(v)$ at u and v are at least $kd^{\alpha}_{u,v}$. The weight assigned to (u, v) is $d^{\alpha}_{u,v}$, which is the amount of energy consumption for transmitting a unit-length message between u and v.

Now, we aim to prolong the network lifetime by dealing with two opposite optimization objectives. One is to minimize the total energy consumption of all nodes by transmitting their sensed data to the sink node. Thus, a minimum spanning tree in G rooted at the sink node is desirable. Another is to minimize the total energy consumption by each node to forward its sensed data to the sink node. Thus, a shortest path tree in G rooted at the sink node is expected. However, constructing a spanning tree that meets these two optimization objectives is NP-hard [15]. Instead, an approximation algorithm that balances these two optimization objectives is available, and a solution delivered by the proposed algorithm is within (α', β') times of the optimum [15], i.e., the solution is no greater than α' times of the cost of a minimum spanning tree and no greater than β' times of the cost of a single source shortest path tree (SSP). Clearly, $\alpha' \geq 1$ and $\beta' \geq 1$. In our choice, we set $\alpha' = \beta'$, which means $\alpha' = \beta' = 1 + \sqrt{2}$. The final tree is referred to as a *balanced tree*.

The proposed algorithm `BT` proceeds as follows: It constructs a balanced tree in the energy graph G which balances the cost of the minimum spanning tree and the cost of the shortest path tree with $\alpha' = \beta' = 1 + \sqrt{2}$, using an

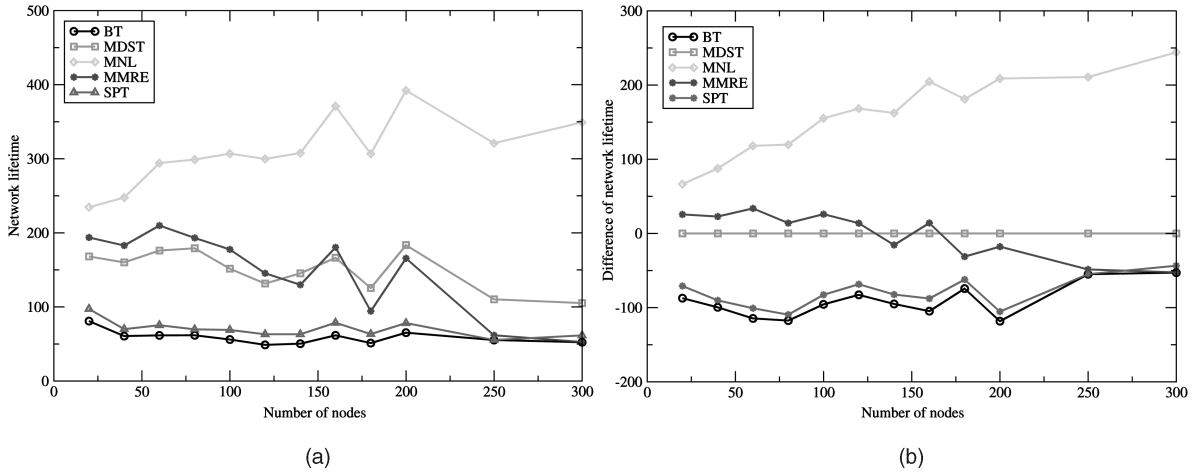


Fig. 2. Network lifetime delivered by MNL, MDST, MMRE, SPT, and BT with initial energy 2×10^6 units and path-loss exponent $\alpha = 2$.

algorithm due to Khuller et al. [15]. The tree is then used to realize the data gathering query.

Despite algorithm BT taking the total energy consumption for realizing a data gathering query into consideration, it does not take into account the residual energy at each individual node. This results in the nodes near the tree root (the sink node) running out of their batteries quickly after a number of data gathering queries are realized, since those nodes always relay messages for the other nodes. Note that von Richenbach and Wattenhofer [21] developed an algorithm similar to BT to deal with the data correlation problem, which is essentially different from ours.

4.3.4 Algorithm MDST

Given a data gathering query with length k of the sensed message by each node, the idea behind algorithm MDST is similar to the one of algorithm BT, but a different weight function is used. Specifically, algorithm MDST proceeds as follows:

A directed energy graph $G(V, E, \omega_1)$ for the sensor network is defined, where V is the set of sensor nodes and E is the set of directed links. A link $\langle u, v\rangle \in E$ if 1) v is within the transmission range of u when u uses its maximum power level to transmit a message and 2) the residual energy $re(u)$ at u is at least $kd^{\alpha}_{u,v}$. The weight assigned to $\langle u, v\rangle$ is given below.

It is generally difficult to construct a routing tree that balances the residual energy at each node and the total energy consumption due to the involvement of two optimization objectives. However, a heuristic, based on an exponential function of the network resource utilization for other routing problems, has been shown to be very efficient to cope with two optimization objectives simultaneously [1], [14]. Here, we adopt a heuristic based on an exponential function of energy utilization at each node. To incorporate the residual energy at each node into the optimization objectives, a weight function $\omega_1 : E \mapsto \mathcal{R}$ for links in G is defined. The weight assigned to a link $\langle u, v\rangle \in E$ is

$$\omega_1(u, v) = d^{\alpha}_{u,v}(\lambda^{\beta(u)} - 1), \tag{9}$$

where $\beta(u) = \frac{C(u)-re(u)}{C(u)} = 1 - \frac{re(u)}{C(u)}$ is the energy utilization ratio at node u between its consumed energy $C(u) - re(u)$ and its initial capacity $C(u)$, and $\lambda > 1$ is constant.

Having the weighted directed graph, find a minimum directed spanning tree (also called an optimal branching) rooted at the sink node by an algorithm due to Gabow et al. [6]. The algorithm will provide a feasible solution to the problem if the residual energy at each node is sufficient for transmitting its message.

5 Performance Evaluation

In this section, we evaluate the performance of the proposed algorithms through experimental simulations. We assume that network instances comprise 20, 40, 60, 80, 100, 120, 140, 160, 180, 200, 250, and 300 nodes, respectively. We further assume that the nodes in each instance are distributed in a $100 \times 100\ m^2$ region, and one of the nodes is the *sink node*, which has unlimited energy supply. The initial energy assigned to each sensor node except the sink node is identical. For a given number of nodes n, each network instance of size n is generated using the NS-2 simulator.

In all our experiments, we assume that data gathering queries arrive one by one. The length k of the sensed message by each sensor node for a given data gathering query is identical, which is a value that is drawn from a uniformly distributed interval between 1 and 7. However, for different queries, the lengths of sensed messages by each sensor may be different. The network lifetime, which is the time of the first node failure due to expiration of its energy, will be used as the metric to evaluate the performance of different algorithms. In our simulations, the network lifetime L of a sensor network for a given query sequence is measured by the first L queries that have been answered and L is maximized. Specifically, for every network size n, the network lifetime shown in each figure is the mean of 300 individual network lifetimes. These 300 values are derived from 10 different query sequences that are randomly generated, and each query sequence will be run on 30 different network topologies of the same network size. Note that, for each of the 10 query sequences, we will run it on the 30 different network topologies. We assume that the

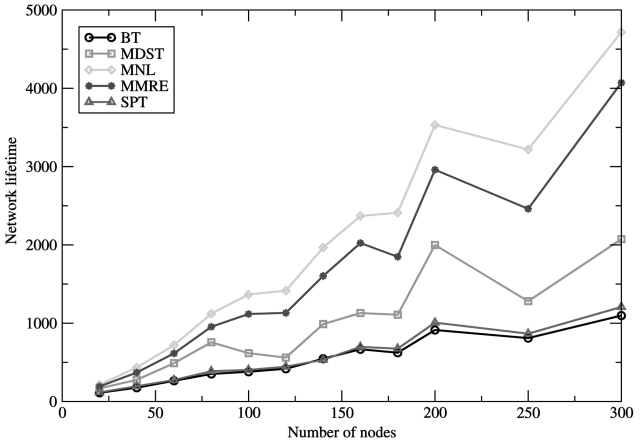


Fig. 3. Network lifetime delivered by MNL, MMRE, MDST, SPT, and BT with initial energy 2×10^9 units and path-loss exponent $\alpha = 4$.

length of the message transmitted by a relay node in a routing tree is the sum of the lengths of its children messages and its own sensed message.

The input parameters of each algorithm include the network size, query sequences, network topologies, and initial energy assignment at each node. In all our experiments, we assume that the initial energy at every node except the sink node is identical, which is 2×10^6 or 2×10^9 units when the path-loss exponent α is set to be 2 or 4, respectively. The parameters α' and β' in algorithm BT are set to be $1 + \sqrt{2}$.

5.1 Performance Evaluation of Various Algorithms

We first evaluate the performance of algorithm MNL against the other algorithms MMRE, SPT, BT, and MDST. Fig. 2a clearly shows that algorithm MNL significantly outperforms all the other algorithms in terms of network lifetime delivered, where the value of λ in algorithm MDST in Fig. 2 is set to be 100. If we put the performance of algorithm MDST as the baseline, Fig. 2b illustrated the relative difference of performance among the heuristics in terms of network lifetime, from which we can see that there is no significant difference in the performance among all the other heuristics except MNL and MDST when the problem size approaches over 250, given the initial energy and the length of the monitored square region.

We then study the impact of different values of path-loss exponent on the performance of various heuristics. Fig. 3 indicates that the performance of algorithms MNL, MMRE, MDST, BT, and SPT is essentially consistent with their corresponding ones, shown in Fig. 2a, and algorithm MNL is still the best. Meanwhile, it can seen that algorithm MMRE outperforms algorithm MDST for each network size when the path-loss exponent is 4. The reason behind it is that the amount of transmission energy required between a pair of nodes is now proportional to the power 4 rather than power 2 of their distance. Thus, although the total weight in the routing tree delivered by algorithm MDST is small, this does not prevent that a large weighted edge $e = \langle u, v \rangle$ is also contained by the tree, and u will consume its residual energy faster, compared with its energy consumption in the case where $\alpha = 2$, when transmitting the same message length to its parent.

We finally analyze the impact of various values of λ on the performance of algorithm MDST through experimental simulations, where the value of λ is ranged from 10^1 to 10^{10}. Recall that the weight assigned to each link $\langle u, v \rangle$ in the energy graph is $d_{u,v}^{\alpha}(\lambda^{\beta(u)} - 1)$, which is proportional to the ratio $\beta(u)$ of energy consumption at node u. Since $\lambda > 1$, different values of λ will result in different routing trees, thereby different network lifetime. Figs. 4a and 4b shows that the network lifetime delivered by algorithm MDST depends on not only the value of λ but also the network size n when the path-loss exponents are 2 and 4, respectively. In both cases, the network lifetime is maximized when $\lambda \approx 100$.

6 Conclusions

In this paper, we first presented a generic cost model of energy consumption for data gathering in sensor networks. We then showed that the online data gathering problem is NP-complete if the length of the message transmitted by each relay node varies, and instead proposed heuristic algorithms for the problem. We finally conducted extensive

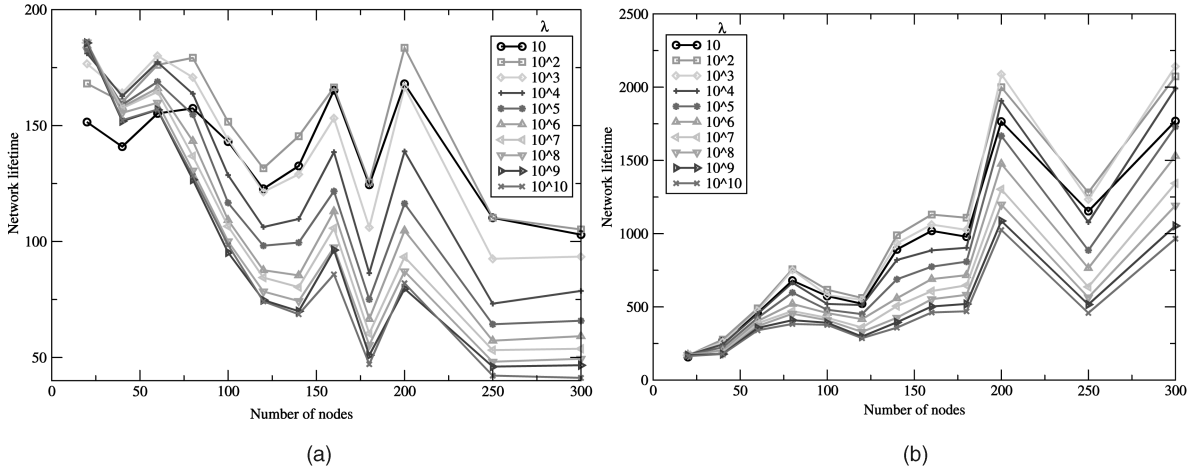


Fig. 4. The impact of λ in algorithm MDST on the network lifetime.

experiments by simulation to evaluate the performance of the proposed algorithms. The experimental results showed that algorithm MNL significantly outperforms the other algorithms including MDST, MMRE, SPT, and BT.

ACKNOWLEDGMENTS

The authors appreciate the anonymous referees for their constructive comments and suggestions which have helped improve the quality and presentation of this paper.

REFERENCES

[1] J. Aspnes, Y. Azar, A. Fait, S. Plotkin, and O. Warrts, "On-Line Routing of Virtual Circuits with Applications to Load Balancing and Machine Scheduling," *J. ACM,* vol. 44, pp. 486-504, 1997.

[2] M. Cagalj, J.-P. Hubaux, and C. Enz, "Minimum-Energy Broadcast in All-Wireless Networks: NP-Completeness and Distribution Issues," *Proc. ACM MobiCom '02,* 2002.

[3] J.-H. Chang and L. Tassiulas, "Energy Conserving Routing in Wireless Ad Hoc Networks," *Proc. INFOCOM '00,* 2000.

[4] J.-H. Chang and L. Tassiulas, "Fast Approximate Algorithms for Maximum Lifetime Routing in Wireless Ad Hoc Networks," *Proc. Int'l Federation for Information Processing TC6/European Commission Int'l Conf.,* pp. 702-713, 2000.

[5] R. Cristescu, B. Beferull-Lonzano, and M. Vetterli, "On Network Correlated Data Gathering," *Proc. INFOCOM '04,* 2004.

[6] H.N. Gabow, Z. Galil, T. Sencer, and R.E. Tarjan, "Efficient Algorithms for Finding Minimum Spanning Trees in Undirected and Directed Graphs," *Combinatorica,* vol. 6, pp. 109-122, 1986.

[7] A. Goel and D. Estrin, "Simultaneous Optimization for Concave Costs: Single Sink Aggregation or Single Source Buy-at-Bulk," *Proc. ACM/SIAM Symp. Discrete Algorithms,* pp. 499-505, 2003.

[8] R. Govindan, J.M. Hellerstein, W. Hong, S. Madden, M. Franklin, and S. Shenker, "The Sensor Network as a Database," Technical Report 02-771, Computer Science Dept., Univ. of Southern California, Sept. 2002.

[9] W.R. Heinzelman, A. Chandrakasan, and H. Balakrishnan, "Energy-Efficient Communication Protocol for Wireless Microsensor Networks," *Proc. IEEE Hawaii Int'l Conf. System Sciences,* 2000.

[10] C. Intanagonwiwat, D. Estrin, R. Govindan, and J. Heidemann, "Impact of Network Density on Data Aggregation in Wireless Sensor Networking," *Proc. 22nd IEEE Int'l Conf Distributed Computing Systems,* pp. 457-458, 2002.

[11] C. Intanagonwiwat, R. Govindan, D. Estrin, J. Heidemann, and F. Silva, "Directed Diffusion for Wireless Sensor Networking," *IEEE/ACM Trans. Networking,* vol. 11, pp. 2-16, 2003.

[12] K. Kalpakis, K. Dasgupta, P. Namjoshi, "Efficient Algorithms for Maximum Lifetime Data Gathering and Aggregation in Wireless Sensor Networks," *Computer Networks,* vol. 42, pp. 697-716, 2003.

[13] I. Kang and R. Poovendran, "Maximizing Static Network Lifetime of Wireless Broadcast Ad Hoc Networks," *Proc. IEEE Int'l Conf. Comm. (ICC '03),* 2003.

[14] K. Kar, M. Kodialam, T.V. Lakshman, and L. Tassiulas, "Routing for Network Capacity Maximization in Energy-Constrained Ad-Hoc Networks," *Proc. INFOCOM '03,* 2003.

[15] S. Khuller, B. Raghavachar, N. Young, "Balancing Minimum Spanning and Shortest Path Trees," *Proc. Fourth ACM-SIAM Symp. Discrete Math.,* 1993.

[16] W. Liang, "Constructing Minimum-Energy Broadcast Trees in Wireless Ad Hoc Networks," *Proc. MobiHoc '02,* 2002.

[17] S. Lindsey and C.S. Raghavendra, "PEGASIS: Power-Efficient Gathering in Sensor Information Systems," *Proc. IEEE Aerospace Conf.,* pp. 1125-1130, 2002.

[18] S. Madden, M.J. Franklin, J.M. Hellerstein, and W. Hong, "TAG: A Tiny Aggregation Service for Ad Hoc Sensor Networks," *ACM SIGOPS Operating Systems Rev.,* vol. 36, pp. 131-146, 2002.

[19] S. Madden, M.J. Franklin, J.M. Hellerstein, and W. Hong, "The Design of an Acquisitional Query Processor for Sensor Networks," *Proc. ACM SIGMOD '03,* pp. 491-502, 2003.

[20] S. Madden, R. Szewczyk, M.J. Franklin, and D. Culler, "Supporting Aggregate Queries Over Ad Hoc Wireless Sensor Networks," *Proc. Fourth IEEE Workshop Mobile Computing and System Applications,* 2002.

[21] P. von Richenbach and R. Wattenhofer, "Gathering Correlated Data in Sensor Networks," *Proc. Second ACM DIALM-POMC Joint Workshop Foundations of Mobile Computing,* Oct. 2004.

[22] A. Mainwaring, J. Polastre, R. Szewczyk, D. Culler, and J. Anderson, "Wireless Sensor Networks for Habitat Monitoring," *Proc. First ACM Int'l Workshop Wireless Sensor Networks and Applications,* pp. 88-97, 2002.

[23] M.A. Sharaf, J. Beaver, A. Labrinidis, and P.K. Chrysanthis, "Balancing Energy Efficiency and Quality of Aggregate Data in Sensor Networks," *J. Very Large Data Bases,* 2004.

[24] A. Singh, M. Woo, and C.S. Raghavendra, "Power-Aware Routing in Mobile Ad Hoc Networks," *Proc. MobiCom '98,* pp. 181-190, 1998.

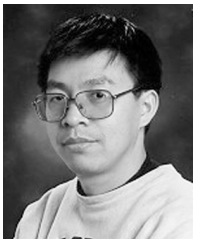
[25] H.Ö. Tan and İ. Körpeoğlu, "Power Efficient Data Gathering and Aggregation in Wireless Sensor Networks," *ACM SIGMOD Record,* vol. 32, pp. 66-71, 2003.

[26] D.B. Terry, D. Goldberg, D. Nichols, and B.M. Oki, "Continuous Queries over Append-Only Databases," *Proc. ACM SIGMOD '02,* 2002.

[27] J.E. Wieselthier, G.D. Nguyen, and A. Ephremides, "On the Construction of Energy-Efficient Broadcast and Multicast Trees in Wireless Networks," *Proc. INFOCOM '00,* 2000.

[28] J.E. Wieselthier, G.D. Nguyen, and A. Ephremides, "Resource Management in Energy-Limited, Bandwidth-Limited, Transceiver-Limited Wireless Networks for Session-Based Multicasting," *Computer Networks,* vol. 39, pp. 113-131, 2002.

[29] Y. Yao and J. Gehrke, "The Cougar Approach to In-Network Query Processing in Sensor Networks," *ACM SIGMOD Record,* vol. 31, pp. 9-18, 2002.



Weifa Liang (M'99–SM'01) received the BSc degree from Wuhan University, China, in 1984, the MEng degree from the University of Science and Technology of China in 1989, and the PhD degree from the Australian National University in 1998, all in computer science. He is currently a senior lecturer in the Department of Computer Science at the Australian National University. His research interests include the design of energy-efficient routing protocols for wireless ad hoc and sensor networks, routing protocol design for WDM optical networks, design and analysis of parallel and distributed algorithms, data warehousing and OLAP, query optimization, and graph theory. He is a senior member of the IEEE and the IEEE Computer Society.

Yuzhen Liu received the BSc and MEng degrees from Wuhan University in China, both in computer science. She is currently pursuing the PhD degree in the Department of Computer Science at the Australian National University. Her research interests include the design and analysis of routing algorithms for wireless ad hoc and sensor networks, design and analysis of network security protocols, trusted computing, and embedded systems.