Journal of Computer and System Sciences 79 (2013) 349–368

Contents lists available at SciVerse ScienceDirect

Journal of Computer and System Sciences

www.elsevier.com/locate/jcss

A novel distributed framework for optimizing query routing trees in wireless sensor networks via optimal operator placement

Georgios Chatzimilioudis [a], Alfredo Cuzzocrea [b,c], Dimitrios Gunopulos [d,*], Nikos Mamoulis [e]

[a] *Dept. of Computer Science, University of Cyprus, 1678 Nicosia, Cyprus*
[b] *ICAR-CNR, Italy*
[c] *University of Calabria, 87036 Cosenza, Italy*
[d] *Dept. of Informatics and Telecommunications, University of Athens, 15784 Ilisia, Greece*
[e] *Computer Science Department, Hong Kong University, Hong Kong*

ARTICLE INFO

Article history:
Received 28 March 2011
Received in revised form 30 January 2012
Accepted 11 September 2012
Available online 23 October 2012

Keywords:
In-network query processing over wireless sensor networks
Query optimization in wireless sensor networks
Operator placement in wireless sensor networks
Query algorithms over wireless sensor networks

ABSTRACT

In this paper, we focus the attention on the *operator placement problem* in *Wireless Sensor Networks* (WSN). This problem is very relevant for *in-network query processing* over WSN, where *query routing trees* are decomposed into three sub-components that must be processed at query time, namely *operator tree*, *operator placement assignment scheme* and *routing scheme*. In particular, the operator placement assignment defines on which node of the network each (query) operator will be hosted and executed. Hence, operator placement plays a key role in the context of query optimization issues in WSN research. In line with this main motivation, in this paper we present an optimal distributed algorithm to adapt the placement of a single operator in high communication cost networks, such as a wireless sensor network. Our parameter-free algorithm finds the optimal node to host the operator with minimum communication cost overhead. Three techniques, proposed here, make this feature possible: (1) identifying the special, and most frequent case, where no flooding is needed, otherwise (2) limitation of the neighborhood to be flooded and (3) variable speed flooding and eves-dropping. When no flooding is needed the communication cost overhead for adapting the operator placement is negligible. In addition, our algorithm does not require any extra communication cost while the query is executed. In our experiments we show that for the rest of cases our algorithm saves 30%–85% of the energy compared to previously proposed techniques. To our knowledge this is the first optimal and distributed algorithm to solve the 1-median (*Fermat node*) problem. A comprehensive experimental evaluation and the proposal of two solutions that are capable of dealing with *adaptive properties* of the operator placement problem, which is an innovative perspective of research in this scientific field, represent two further contributions of our research.

1. Introduction

1.1. Overview and motivations

Information is valuable and most technology applications invest in extracting valuable information from detailed readings or input data. This is done using queries on data. The optimization of query processing has been a significant topic of the

* Corresponding author.
E-mail addresses: gchatzim@cs.ucr.edu (G. Chatzimilioudis), cuzzocrea@si.deis.unical.it (A. Cuzzocrea), dg@di.uoa.gr (D. Gunopulos), nikos@cs.hku.hk (N. Mamoulis).

0022-0000/$ – see front matter © 2012 Published by Elsevier Inc.
http://dx.doi.org/10.1016/j.jcss.2012.09.013

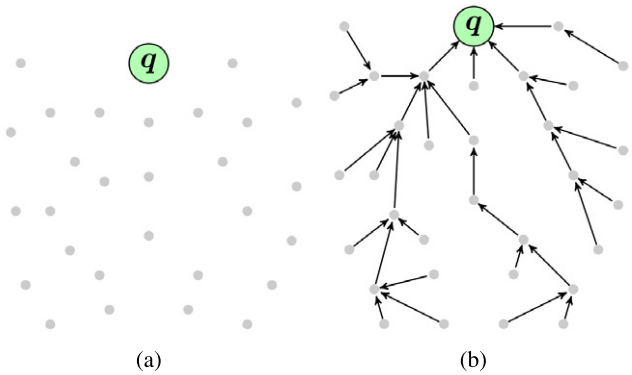

Fig. 1. (a) A wireless sensor network with a querying node and (b) a possible Query Routing Tree.

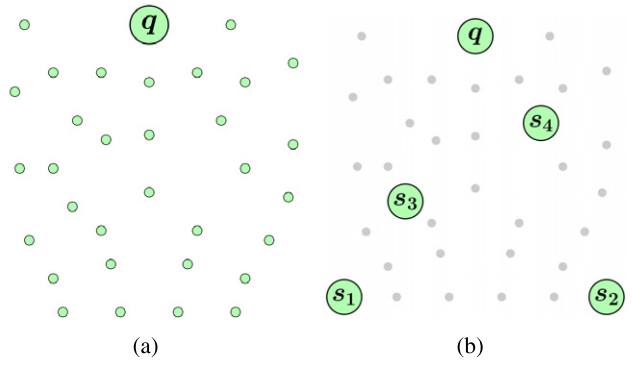

Fig. 2. Green nodes represent nodes that store data needed to answer a query Q. (a) An example of a universal query and (b) an example of a subset query. (For interpretation of the reference to color in this figure legend, the reader is referred to the web version of this article.)

research community throughout the years as it allows to optimize objective functions that are vital to the efficiency of applications.

Wireless Sensor Networks (WSN) are deployed to understand the physical world at a high fidelity using the numerous readings from the sensors. Monitoring and controlling high performance office buildings is an example of a WSN application of great interest to the industry (Fig. 3). A wireless sensor network is deployed throughout the building using various sensor nodes, including light, particle, motion, temperature, humidity, RFID, camera and other sensors. The data collected can be used to improve the Indoor Environmental Quality of the office space while minimizing the energy consumption of the building. This WSN application will be used throughout the paper as a motivational example where various queries need to be optimized.

An important objective function to optimize in WSN is the network lifetime. In WSN the transmission of data consumes the most energy [26]. The foremost goal when optimizing query processing is to minimize the communication cost. We propose techniques specifically designed for various query types that will minimize data transfers during query execution.

Various queries can be injected to the wireless sensor network of our high performance building application. Queries are answered by constructing a Query Routing Tree (QRT) that determines how data will be routed through the network and where it will be processed. Using sophisticated query routing trees the query execution can be optimized according to the given objective function. Two are the main tasks for a sophisticated QRT: constructing a QRT and adapting a QRT to changes in the network. We will present existing work and new techniques that efficiently complete both tasks. An example of a query routing tree can be seen in Fig. 1.

To better study the possible optimization of a query routing tree we distinguish between two major types of queries: universal and subset queries. Universal queries need data from all the nodes of the network to be answered and usually involve a single aggregational operator that is applied to all the data sources. Subset queries only need data from a subset of nodes and can involve various operators applied on different pairs/groups of data sources. A visual example of such queries inside a sensor network can be seen in Fig. 2.

Consider the high performance building application (Fig. 3) and the need to look for association rules between the different reading of the sensor nodes regarding possible light energy savings. In this example query we would like to stream data from all the nodes in the network to a sink and process the readings there. This would be a universal query. It would be a snapshot universal query if we just need the current readings that are stored at each node, i.e. take a snapshot of the current readings in the network. A continuous query is if we need the readings over several epochs of new readings and

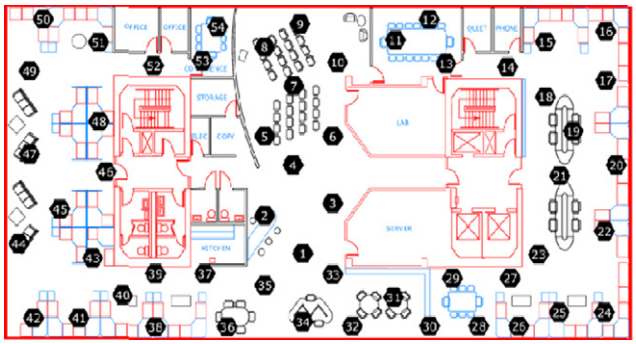

Fig. 3. An application example: Office building wireless sensor setup for monitoring Indoor Environmental Quality in respect to energy consumed by the building.

transmission. In such case there can be some universal aggregation or compression operator that can be applied on the data at every node where two data streams meet.

Universal queries are optimized mainly by balancing the data load or the node degree among the QRT nodes. Balancing data load lets nodes share energy consumption equally and thus increases network lifetime. Balancing the degree among the QRT nodes minimizes the chances of collision between data packet which requires the retransmission of the collided packets.

An instance of subset queries is “Find the employees that were at room A, B and C”. Such a query is needed to identify employee habits in order to make the workplace more productive by optimizing equipment and workstation arrangement and floor planning. Similarly this query could be a snapshot query or a continuous, long-running query. In this example the operator defined by the query is the intersection operator on the employee ID readings of the sensors responsible for rooms A, B and C.

Subset queries have a greater potential for optimization. A query routing tree can be broken down into an operator tree, an operator placement assignment and a routing scheme. The operator tree defines the order in which the operators will be applied to the data and how the results will be pipelined from one operator to the other. The operator placement assignment defines on which node of the network each operator will be hosted and executed. The routing defines how data will be routed from a data source to an operator node, from an operator node to another operator node or to the sink. Depending on the operators defined by a subset query all of the above can be optimized resulting in significant energy savings.

The duration of a query also plays a significant role in the optimization techniques that can be used. Snapshot queries, that are only executed instantaneously, need a good query tree construction algorithm since the QRT is going to be used only once. Continuous queries are executed over several epochs, during which data and the network can change resulting in degradation of the QRT efficiency. The longer the query needs to be executed the more query tree adaptation is more important than a good initial tree construction algorithm.

Among the mentioned sub-components of a query routing tree, the operator placement assignment problem is, without doubts, the most important one, and a lot of attention has been devoted to so-called operator placement operators for WSN. This is due to the fact that, very often, network applications perform in-network query processing for efficiency purposes. Sensor networks are being deployed in the physical or urban environment to benefit scientific research or security surveillance. Another example of a query in a network (similar to the previous one on the high performance building application), which is monitoring traffic in a busy downtown area, could be “How many cars took the same route of passing through intersections A, B and C?”. To avoid the cost of communicating all the data lists from the nodes in regions A, B and C to the querying node, the query must be executed in-network. Data lists generated on the source nodes are fed into operators on intermediate nodes that combine several lists from different sources. The amount of data is reduced due to the selectivity of the operators and the data that reaches the querying node is the final answer.

An operator, that is involved in the in-network processing, can be placed on a node of the network. It takes in elements from source nodes, processes them, and sends the output to either another operator node or to the sink. Shipping elements over an edge in the graph imposes a cost that is dependent on the weight of the elements. Therefore, the placement of an operator can greatly affect the cost of answering a query since it affects the number of edges the elements have to travel over and the weight of the elements, since usually the output weight is not the sum of the input weights.

It is typical to have continuous queries that require an answer over a continuous period of epochs. In most applications the sources and operators are not producing the same weight of elements in every epoch. Similarly, nodes in the network might be mobile resulting in different hop-distances between nodes in every epoch. Therefore, the initial operator placement might not be good enough for future epochs. It is a large overhead to re-run the algorithms for finding a good placement for the operators of the query. Instead, the technique followed in literature is to *update* the placement of just the operators that are affected by the weight change in order to keep the cost of query execution in the next epoch to a minimum. This operator placement update needs to be done with the least amount of communication cost overhead possible.

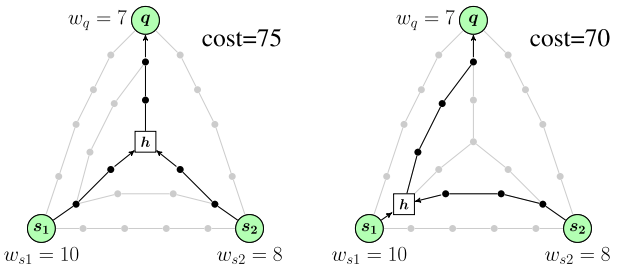

Fig. 4. Example of optimal operator placement: (a) Data flow during query execution before the operator placement is optimized, and (b) data flow during query execution after the operator placement is optimized. The *Fermat* node is an *external* node. The cost represents the cost of our objective function, not the actual communication cost. (For interpretation of the colors in this figure, the reader is referred to the web version of this article.)

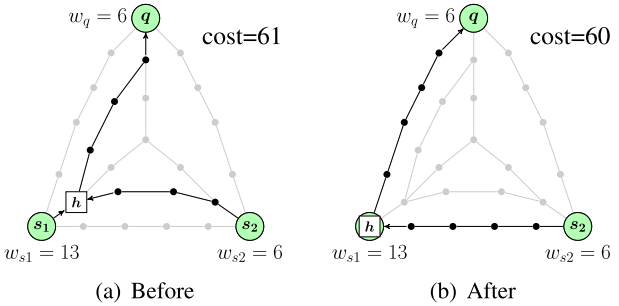

Fig. 5. Example of optimal operator placement: (a) Data flow during query execution before the operator placement is optimized, and (b) data flow during query execution after the operator placement is optimized. The *Fermat* node is a *datanode*. The cost represents the cost of our objective function, not the actual communication cost.

In Figs. 4 and 5 let the nodes $s1$, $s2$ and q be the sources and the sink (henceforth called altogether *datanodes*) that send/receive data from the binary operator hosted at node h. Let w_i be the weight of the data to be sent from node n_i. We can see that by picking the right node to host the operator with the right distances from the *datanodes*, we can reduce the objective function for the communication cost of executing the query (difference between Figs. 4(a), 5(a) and 4(b), 5(b)). Depending on the data loads and the path lengths, the optimal node to place the operator can be either an *external* node (Fig. 4), or one of the *datanodes*, i.e. a source or sink (Fig. 5). Our algorithm finds the optimal new placement for an operator while creating far less communication cost overhead than previous work. Note that we do not assume that the communication cost can be computed by summing the data-load sent over each link. We just use this as an objective function to estimate the actual communication cost. In our experiments we use a more accurate model for the communication cost.

Especially in high communication-cost applications minimizing the communication cost is the key issue. High communication cost networks play an important role in real world applications, as much as they do in research. The communication cost can be posed by monetary, temporal, resource or energy demands. As an example of a high communication cost network we will use a wireless sensor network throughout the paper. Wireless sensors have very limited energy resources. The task that has by far the highest demand in energy on a wireless sensor is the transmission and reception of data. Thus, the cost to pay for communicating is in form of energy. Minimizing the total energy consumed makes the whole network more energy sufficient, and minimizing the maximum energy consumption per node increases the network's lifetime.

1.2. Contributions of our research

Our *distributed Fermat node search* algorithm (*dFNS*) achieves two goals: finding the best node to place an operator and minimize communication cost doing so. To achieve this it (1) identifies the special case where no flooding is needed, (2) if flooding is needed, it minimizes the flooding radius, and (3) uses variable speed flooding and eves-dropping. Our algorithm is parameter-free, decentralized, optimal and outperforms previously proposed methods in minimizing communication cost overhead.

As shown in our experiments, there is a high chance (56%–85%) that the optimal node to place the operator is a *datanode* (source or sink), like in the example of Fig. 5. Such a case can be identified by our algorithm and the operator is simply placed on the optimal *datanode* without any further communication cost to find the optimal operator node.

In any other case, *dFNS* finds the optimal operator node (*Fermat* node) by extending a flood from each of the *datanodes*. We generate a set of possible distance combinations that the new hosting node can have to produce a smaller hosting cost.

Using these candidate distance combinations *dFNS* calculates the minimum possible radius for each flood, guaranteeing that the nodes that participate are kept to a minimum without compromising the optimality of the algorithm.

We adapt our proposed algorithm to existing work in WSN. Using an existing framework for answering multi-predicate snapshot queries, we extend the framework to deal with continuous queries. The framework answers continuous queries in epochs and adapts the operator placement to data load changes. In our experimental evaluation we compare against the only other existing distributed algorithm for operator placement updates and show that using our proposed algorithm we can save 30%–80% of the communication cost overhead.

In addition to this, we further extend the proposed framework as to deal with other important aspects of the operator placement problem in adaptive environments. In particular, we find there is still better optimization techniques needed for adapting the placement of multiple operators in a continuous subset query and propose a better algorithm. There is also no technique proposed so far to adapt an operator tree of commutative operators. We propose the first technique to adapt an operator tree in a distributed fashion that also incorporates operator placement. These extensions are significant contributions of our research as well.

1.3. Paper organization

The remaining part of the paper is organized as follows. In Section 2, we formalize our system model and the basic terminology that will be utilized in the subsequent sections, and we give formal definitions and proposition that we use in our algorithms. In Section 3, we present previous work related to our research. We formulate our problem definition and preliminary annotation in Section 4 to be able to describe our algorithm in detail in Section 5. In Section 6 the framework in which our algorithm is implemented is described and in Section 7 we present our thorough experimental evaluation that shows the efficiency of our algorithm. Furthermore, in Section 8 we provide meaningfully extensions of our framework, mainly devoted to deal with adaptive properties of the operator placement problems. Finally, in Section 9 we provide conclusions and future work of our research.

2. Formal definitions and system model

Let V denote a set of n sensing devices $\{v_1, v_2, \ldots, v_n\}$. Assume that v_i $(i \leqslant n)$ is able to acquire m physical attributes $\{a_1, a_2, \ldots, a_m\}$ from its environment at every discrete time instance t. This generates at each t and for each v_i $(i \leqslant n)$ one tuple of the form $\{t, a_1, a_2, \ldots, a_m\}$. This scenario conceptually yields an $n \times m$ matrix of readings $X := (v_{ij})_{n \times m}$ for each timestamp. This matrix is *horizontally fragmented* across the n sensing devices (i.e., row i contains the readings of sensor v_i and $X = \bigcup_{i \in n} X_i$). Now let $G = (V, E)$ denote the network graph that represents the implicit network edges E of the sensors in V. The edges in E are implicit, because there is no explicit connection between adjacent nodes, but nodes are considered neighbors if they are within communication range R (i.e., a fundamental assumption underlying the operation of a radio network).

A user can run *queries* on a WSN. WSN can be viewed as a network of tiny distributed databases and *queries* can be posed through a node of the network. To answer a query, data generated by the sensors needs to be collected and processed. The processing is done centralized on the basestation or distributed on the nodes of the network (in-network processing). We assume that nodes can pose queries over the sensor network. The sensor node that issues the query Q is called querying node and is denoted as q. We will use the terms querying node and sink interchangeably throughout this work.

For simplicity let us adopt a declarative SQL-like syntax (similarly to [21,33]) to express the ideas presented in this work in brevity. For instance, the following query declares that each sensing device should recursively collect the node identifier and the temperature from its children every 31 seconds and communicate the results to the sink.

```
SELECT nodeid, temp
FROM sensors
EPOCH DURATION 31 seconds
```

Note that our model also supports continuous aggregate queries. For instance, the following query declares that each sensing device should aggregate the average light measurement for each room from its children every 31 seconds and communicate the results to the sink.

```
SELECT roomid, AVG(light)
FROM sensors
GROUP BY roomid
EPOCH DURATION 31 seconds
```

Continuous queries are answered in consecutive data acquisition rounds called *epochs*. An epoch is a small time period in which the query is answered once, like a snapshot query. A user specifies a continuous query Q to be evaluated once during the interval of an *epoch* (denoted as e), which is the time interval after which each s_i $(i \leqslant n)$ will re-compute Q.

Assuming that the nodes have a restricted communication range data will have to travel over a multi-hop path toward the querying node. In this case a routing tree T is created connecting every node over a multi-hop path with the sink. The querying node q is the root of this tree and receives the information needed to answer the query. We will denote as d_v the depth of v in this tree.

In order to process queries efficiently over a WSN, sensors need to be organized in a *query routing tree* T. A *query routing tree* is an acyclic subset of the communication graph G (i.e., a spanning tree) which is denoted as $T = (V', E')$, where $V' \subseteq V$ and $E' \subseteq E$. T can be constructed based on query semantics, power consumption, remaining energy and others. A *query routing tree* provides each sensor with a path over which query answers can be transmitted to the querying node and allows for waking window and data reduction techniques to optimize the energy consumption in the network. An epoch can use the same query tree T as the previous epoch and save the overhead of tree construction. Otherwise, it can be deemed more efficient to reconstruct or adapt the query tree for the new epoch.

Energy expenditure in data communication is far greater compared to data processing. The example described in [26], effectively illustrates this disparity. Mixers, frequency synthesizers, voltage control oscillators, phase locked loops (PLL) and power amplifiers, all consume valuable power in the transceiver circuitry. This involves both data transmission and reception. It can be shown that for short-range communication with low radiation power (~ 0 dbm), transmission and reception energy costs are nearly the same.

In [27], the authors present a formulation for the radio power consumption (Pc) as

$$P_c = N_T\big(P_T(T_{on} + T_{st}) + P_{out} * T_{on}\big) + N_R\big(P_R(R_{on} + R_{st})\big)$$

where P_T/P_R is the power consumed by the transmitter/receiver; P_{out}, the output power of the transmitter; T_{on}/R_{on}, the transmitter/receiver on time; T_{st}/R_{st}, the transmitter/receiver start-up time; and N_T/N_R, the number of times transmitter/receiver is switched on per unit time, which depends on the task and medium access control (MAC) scheme used. We can further say that $T_{on} = L/R$, where L is the packet size and R the data rate.

3. Related work

There is existing work that can be used or be adapted to solve most of the query types presented (e.g., [9]). We will present previous work categorized according to the types of queries they optimize. First we present techniques that can be used to optimize universal queries and expose their strengths and drawbacks. Then we present techniques that optimize snapshot and continuous subset queries. Finally, we will focus the attention on operator placement literature, the main context of our research.

To answer universal queries we need to acquire data from all the nodes of the network in order to compute the answer. These are queries that perform a common operation on all the data. Optimizing universal queries has been tackled with two different methods. There are works that aim at balancing the workload among nodes in an effort to minimizing the maximum energy consumption per node [32]. These works create a query routing tree that balances the workload among nodes in order to distributed the energy consumption evenly. For continuous universal queries that need adaptation, the works propose some simple mechanisms to maintain a good workload balance in the query routing tree. These works also have a simple way of adapting to data/network change using information acquired during the tree construction. The candidate parents are stored and whenever a change occurs it is easy to connect to a different parent keeping the objective function optimized.

The other method for optimizing universal queries used in literature is balancing the degree among nodes in an effort to minimize the packet collisions during query execution [3,7]. These works construct a query routing tree that balances the children among parents. It has been shown by Andreou et al. [3] that this leads to reduced collisions in the wireless channel. Similarly, when adaptation is needed due to network or data changes both works use the information acquired during tree construction to keep their objective function optimized. Other approaches propose to provide topology control over the WSN for query optimization purposes (e.g., [10]).

To answer subset queries we need to acquire data from a specific regions of the network. These queries are also known as multi-predicate queries [16]. The computation of the answer usually involves specific operators among different region data. The operators to be used are defined by the query. Queries can have commutative and non-commutative k-ary operators. If all operators are commutative, then query execution can be optimized by defining the right operator tree. An operator tree is a tree that determines the sequence of the operators, and how their inputs and outputs are pipelined. If a query has non-commutative operators then the operator tree is defined by the query.

Data acquisition inside a region can be seen as a local universal query and region data can be collected on one representative node for each region. Advantages for this technique are shown in works like [30,35]. These works propose techniques on how to choose representative and adapt in order to maximize lifetime. Doing the data acquisition inside a region can be optimized using the universal query optimization literature presented above. Subset queries can be seen as queries that need data from a subset of nodes that are representatives of their region.

A query routing tree for subset queries can be broken down into an operator tree, an operator placement assignment, and a routing algorithm. Optimizations can be done in all three parts. There is work optimizing one or two parts at a time. As far as routing goes there is work like [17,13] that focus on optimizing routing in order to achieve energy savings while answering subset queries and placing operators opportunistically, when there is a need for adaptation due to network or data changes.

Works that do not focus on routing, including this one, regard routing as a separate part from query tree construction or assume optimal shortest-path routing. Work that regards routing as a separate part does not take into consideration the cost for maintaining and acquiring routes for the routing algorithm are not included in the experimental evaluations.

For queries with commutative operators works like [8,6] propose algorithms for constructing sophisticated operator trees. Both algorithms are centralized. The work of [8] is also using joins to reduce transmitted data. Their divide and conquer heuristic algorithm is developed for uniform networks and requires queries that involve both 'data reducing' and 'data increasing' join operators on different attributes in order to work. In our case we deal only with intersections which is always 'data-reducing' and which is defined between the same common attribute over all relations. Using their algorithm in our problem would return a random query tree. [6] uses a heuristic algorithm to come up with a near-optimal tree solution.

Now, let us focus the attention on operator placement literature in greater detail. The vast majority of the literature on operator placement in WSN focuses on finding a good operator placement at query initialization as described in the introduction. Those algorithms are centralized; i.e., the base-station knows the location of the sensors or has complete knowledge about the network [28,18,2,23,6].

Ying et al. [34] propose a distributed algorithm to do the same task as above, namely static operator placement. Nodes exchange information with their neighbors iteratively until they find the optimal placement for all given operators. Any node that has found a better cost for routing data or placing the operator, broadcasts this information to its neighbors. This algorithm is suited only for initial operator placement for queries with many operators, since it involves every node inside the network. Further, using this technique, it is hard to guarantee convergence, optimality, and low communication cost overhead.

Instead of sticking to a static plan, dynamic environments require adaptive query processing. A comprehensive survey on adaptive query processing is presented by Deshpande et al. [13]. They categorize all techniques proposed that focus on using runtime feedback to modify query processing in a way that provides better response time, more efficient CPU utilization or network utilization. Our work would fall under the category of adaptive join processing with non-pipelined execution.

Next, we cite literature that deals specifically with operator placement adaptation, picking a new hosting node for one of the operators. There are two categories here: algorithms that pick the best neighboring node as the new host and converge to the optimal operator placement with time, and algorithms that find the best hosting node immediately. The former method is also called operator migration and we will call the later method *placement update*.

An alternative to operator placement update is operator *migration*, where the operator is moved gradually from one node to the next node towards the optimal placement. Algorithms following this principle are simple and their decision making is only local. On the other hand, it takes several epochs of query execution to reach the optimal operator placement. For the same reason, these methods suffer greatly from oscillating changes, that might force it to migrate an operator to a different direction before even reaching the optimal placement. Further, they are prone to local minima and impose extra cost during query execution in order to probe for a better operator host on every neighbor; [22,5,24] are works in this category.

Finding directly the optimal hosting node is the approach adopted in this paper. This problem is the same as the 1-median problem or single facility location problem in graphs. There is extensive literature on centralized algorithms for this problem [25], but not on distributed algorithms. In a distributed environment we cannot adapt any of the centralized algorithms, since they all require that a central authority knows the topology of the whole network.

Zoe Abrams and Jie Liu in their paper named "Greedy is Good" (GIG) [1] propose a decentralized solution for the 1-median problem in graphs. They try to find the optimal hosting node of a single operator by flooding a small neighborhood around each *datanode*. It follows the intuition that the optimal hosting node will be somewhere close to all the datanodes. Their algorithm, *GIG*, aims to minimize the nodes involved in the flood by making use of some parameters set by the user. Surprisingly, they do not aim to minimize the number of messages exchanged by those nodes and thus the communication cost overhead is not minimized. Further, their algorithm does not guarantee to find the optimal operator node as we will see in the example in Fig. 6.

We propose a parameter-free algorithm based on the same principles as *GIG*, but show how using the right techniques the right heuristics we can achieve a 30%–100% energy reduction compared to *GIG*. Some extra points that distinguishes our work from previous work are the following:

- our algorithm is distributed and we only collect a negligible amount of network information;
- we do not assume any location awareness for the nodes – it follows that we cannot use geographical routing to our advantage;
- our algorithm does not impose any overhead during the query execution phase;
- our algorithm is parameter-free, thus its efficiency is independent of any user input;
- our algorithm guarantees optimality;
- our algorithm supports adaptive properties of WSN;

4. Distributed Fermat node search: Preliminaries

Assume that in the network seen in Fig. 4(a) the colored nodes are 3 customers s_1, s_2 and q. Each customer i needs quantity w_i from a commodity produced by a service that is currently hosted in node h. The cost of servicing customer i is the cost of sending weight w_i over the shortest path from node h to i. Find the node, that minimizes the cost of servicing the customers, to host the service. This is also known as the 1-median problem and can be extended to an arbitrary number of customers. Equivalently in WSN we have an operator that collects data from a number of sources and sends the result of

the operation to a sink. In Figs. 4 and 5 we are dealing with binary operators (two sources s_1 and s_2, one sink q). Note that there are no restrictions in the relation between the quantities w_i, thus we can use any kind of operator.

We assume that sending data of weight w from node i to the operator host h and sending the same amount of data from operator host h to node i imposes the same cost. This is why we generalize and call both, sources and sinks, *datanodes*. Now the problem of finding the optimal operator placement is similar to the Fermat point problem [31], the three factory problem [15], and to the 1-median problem or single facility location problem. We call the optimal node to place the operator *Fermat* node and formulate our problem as follows:

Fermat node (or 1-median) problem definition. *Given a weighted graph $G(N, L)$ and a set of datanodes $D \subset N$, find the Fermat node f in the graph that minimizes the cost of shipping data from the nodes in D to node f.*

For the objective function that we use in our algorithm we assume that the cost of shipping data from node u to node v is proportional to the data load w_u to be shipped and the weight of the path used. The path weight $W(u, v)$ is equal to the sum of the weights of all links $l \in L$ that make up path (u, v): $W(u, v) = \sum_{l \in links(u,v)} w_l$, where w_l is the weight of the link $l \in L$. The cost of shipping data from node u to node v is defined as

$$t(u, v) = w_u * W(u, v)$$

This simplified version is used only as an objective function in our algorithm to estimate communication cost. Note that the computation of the actual energy consumed by the network when transmitting a message over a path is more complicated. In the network simulator we used to run our experiments the communication cost model is much more realistic. It also takes into account the energy consumed by the neighbors of u and v since they also receive the packet.

To estimate the energy needed to answer a query Q over a query tree T once, we define $T(i, j)$ as the function that returns 1 whenever the edge (i, j) is used in the query evaluation:

$$T(i, j) = 1 \quad \text{if communication edge } (i, j) \text{ is used} \tag{1}$$

The total communication cost C_Q of answering query Q will be:

$$C_Q = \sum_{i=1,\, j=1}^{m,m} c * T(i, j) * B_i \tag{2}$$

where nodes $i, j \in V$. This does not include the cost of disseminating the query or constructing the tree T.

Hosting cost, c, is the cost of sending data from the nodes in D to the hosting node h. It is equal to

$$c = \sum_{d \in D} t(d, f) \tag{3}$$

To minimize this cost we need to find the *Fermat* node and place the operator there. Finding the *Fermat* node involves a number of nodes that need to exchange messages. This imposes a communication overhead. The problem we solve in this paper is the following:

Our problem definition. *Given a weighted graph $G(N, L)$, with identical link weights, and a set of weighted datanodes $D \subset N$, solve the Fermat node problem with minimum overhead.*

The communication cost in a wireless sensor network is the energy consumed for performing communication. The total communication cost is the sum of the energy consumed by each node in the network. The maximum communication cost is the maximum energy consumed by a single node. By minimizing the number of nodes involved and the messages exchange between them, we keep the total communication cost and the maximum communication cost per node to a minimum.

Networks are inherently distributed, thus no node has global knowledge about the network topology. This rules out the application of one of many proposed algorithms in literature (Section 3), that solve the *Fermat node problem*. We propose a fully distributed algorithm, that does not require the gathering of network information in order to compute the *Fermat* node. In the rest of the paper we will make extensive use of the following notions, that are formally defined here.

Shortest path length is the length of the shortest path between two nodes, i.e. u and v, and is denoted as $|(u, v)|$. We assume that the graph has bidirectional links, thus $|(u, v)| = |(v, u)|$.

Datanodes is the set of nodes D that either transmit data (*source nodes*) or receive data (*sink node*) to/from the node that hosts the m-ary operator (*hosting node*). The opposite of the datanode set is the *external nodes* set $X = (N - D)$. *Leader* node is the node that decides on initiating and terminating the *dFNS* algorithm.

Note that we assume error-free readings, otherwise we would need specialized techniques for probabilistic or model-base query execution [12]. We also do not assume any correlation between data that could assist us in saving energy during query execution [11]. The only information we need is what nodes the data is coming from/going to and the size in bytes.

Our framework operates independently of how an operator placement update is triggered or oscillating updates, due to the rapid changes in the network, are avoided.

Distance combination, α, denotes the k-ary set of shortest path lengths from all datanodes in D to the *hosting* node h.

$$\alpha = [\alpha_1, \alpha_2, \ldots, \alpha_k] = \left[\left|(d_1, h)\right|, \left|(d_2, h)\right|, \ldots, \left|(d_k, h)\right|\right]$$

where $d_i \in D$ and $k = |D|$. Each distance combination α has its hosting cost c_α. Note, when we have an m-ary operator it means we have m inputs and one output. It follows that the number of datanodes is $m+1$ and thus $m+1 = k$.

Flooding is the task of broadcasting data from one node to all its neighbors and repeating this for each neighbor. Each node broadcasts the data only once. By setting a restriction to the flooding *radius*, the broadcast message travels only *radius* hops away (Hops-To-Live = *radius*). This limits the nodes in the network that are flooded.

5. A novel distributed Fermat node search algorithm

We assume that a node h, that hosts an operator with datanodes D, knows the shortest path distances between any pair of datanodes in D. Note that the datanodes D of a single operator are only a very small subset of the nodes in the network ($|D| \ll |N|$). This information can be piggy-backed from each datanode $d \in D$ to node h, since there is direct unicast communication between them. The task of retrieving this information for each datanode d can be performed with efficient algorithms proposed in literature, such as doubling broadcast distance. Other than the datanodes of an operator, no other node in the network need to know their distance to any other node.

Each datanode d has its own hosting cost c_d. We call *best* datanode b the datanode with the minimum hosting cost $c_b = min\{c_d\}_{\forall d \in D}$. Using b as the solution to the *Fermat* problem is called *datanode solution*. There are cases where it is impossible for an *external* node to have better hosting cost than datanode b. Identifying those cases is simple and imposes no communication cost. All our techniques make use of hosting cost c_b of the best node.

5.1. Candidate nodes

Candidate nodes are called the nodes in the network that have a hosting cost less than the hosting cost c_b of the best datanode. We need to compare all candidate nodes in order to find the actual *Fermat* node. Minimizing the number of candidate nodes is one of the key features of our algorithm. Note that there can be several nodes with the same minimum hosting cost, thus there can be several *Fermat* nodes. We just need to pick one of them.

To be able to calculate the hosting cost of an *external* node, we need to know its distance to the datanodes. Although external nodes might serve as relay nodes, they never communicate directly with any datanode, thus we cannot assume that they know their distance to each datanode in advance. To find the distance from an *external* node to each datanode we can initiate a flood from each datanode counting hops.

Nodes inside the intersection of all floods know the distance to all datanodes. This is true since we assume that the flood reaches a node over the shortest path from the initiator. These nodes can now calculate their hosting cost and, if it is smaller than c_b, they become candidate *Fermat* nodes. Candidate nodes report their hosting cost to the *leader* node, that decides what node is the actual *Fermat* node.

By reducing the number of candidate nodes, and therefore the messages (reports) sent to the *leader* node, we can save on communication cost. *dFNS* includes the hosting cost c_b of the best datanode in the initial flooding message as a cost threshold. Nodes, that have a hosting cost higher than this cost threshold, are not considered candidate nodes. Nodes that have a better hosting cost designate themselves as *Fermat* candidates and update the cost threshold inside the flooding message before it gets forwarded. We also let candidate nodes eves-drop messages sent by their neighbors in order to increase the probability that a message with a lower cost threshold is received to minimize the number of candidates.

5.2. Calculating all candidate distance combinations

Before looking for the actual candidate nodes in the network we calculate all possible distance combinations that would qualify a node as a candidate node. These *candidate distance combinations* are calculated at the datanodes without any communication with neighbors. This is done in order to be able to restrict the communication cost while searching for the actual candidate nodes inside the network. Most of the notation used here is defined in Section 4.

The datanode computes all candidate distance combinations A and their respective hosting cost c_α, $\alpha \in A$. This is the basic building block for our algorithm. To efficiently compute this set we use information about the shortest path lengths between the datanodes D. The distance combinations that violate the shortest path length between the datanodes (triangle inequality) and have a greater hosting cost than c_b, the hosting cost of the best datanode b, are discarded. Formally the restrictions for each distance combination α are:

$$\begin{aligned} &\left|(d_i, d_j)\right| \leqslant a_i + a_j, \quad \forall i, j \in D \\ &c_\alpha < c_b \end{aligned} \tag{4}$$

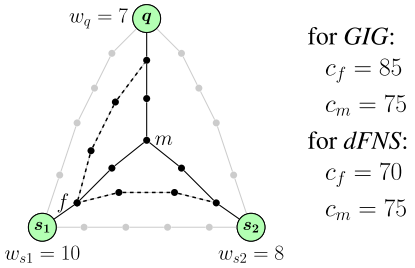

Fig. 6. *GIG* [1] is not an optimal algorithm. An example where *GIG* misses the optimal operator placement (f). This happens because the distances from candidate nodes to the datanodes are overestimated.

The distance combination with the minimum hosting cost is called *ideal* distance combination and is denoted as ϵ. Depending on the network, a node with the distance combination ϵ might exist or not. If a candidate node has the ideal hosting cost c_ϵ, then no further action is needed to distinguish it as the *Fermat* node.

The algorithm we propose to compute the distance combinations is optimized to find the set fast and effectively, pruning combinations that do not satisfy the constraints in Eq. (4) early. We start from the distance combination that corresponds to picking the best datanode b as the *Fermat* node. In this distance combination the value for $|(b, f)|$ will be 0. The other distances start from the minimum value possible that satisfies the constraints. We recursively increment each distance by 1. The pseudo-code is shown in Algorithm 1. This algorithm returns the set of all possible distance combinations that would result in a smaller hosting cost than c_b. It also designates the *ideal* distance combination ϵ.

Algorithm 1 CDCGenerator(*distanceList*, i)

Require: list of datanodes and their loads, distance between every pair of datanodes
1: $l_{current}$ = minimumDistance(*distanceList*, i)
2: *distanceList* $\leftarrow l_{current}$
3: **while** (*distanceList* satisfies constraints AND $l_{current}$ <maximumDistance(*distanceList*, i)) **do**
4: **if** *distanceList.size* == number of datanodes **then**
5: C $\leftarrow$ *distanceList*
6: update *bestCombination*
7: **else**
8: C $\leftarrow$ CDCGenerator(*distanceList*, $i + 1$)
9: **end if**
10: $l_{current} = l_{current} + 1$
11: remove last entry of *distanceList*
12: *distanceList* $\leftarrow l_{current}$
13: **end while**
14: **return** C

The function *maximumDistance*(*distanceList*, i) returns the maximum distance that a node can have from datanode d_i so that it satisfies the constraints of Eq. (4) and does not exceed the maximum distance between d_i and d_j, where $j > i$.

5.3. No flooding cases

If we get a distance combination set, that is empty when running the *CDCGenerator*() algorithm, it means that there cannot exist an *external* datanode with better hosting cost than the best datanode b. In those special cases, no flooding is needed to look for external *candidate* nodes. Node b is the optimal new operator host and our algorithm terminates by placing the operator there. Contrary to their characterization as *special*, these cases comprise 56%–85% of the cases as shown by experiments.

5.4. Flooding radius

Flooding the whole network from each datanode in D poses a very big communication cost. Our algorithm efficiently restricts the flooding radius, guaranteeing at the same time that the *Fermat* node will be found. For this it uses the *candidate distance combinations*.

The same intuition is used in the *GIG* algorithm [1], only they use a suboptimal method to restrict the flooding. In addition, *GIG* cannot guarantee optimality since the distance from an external node to a datanode can be overestimated. This can be seen in Fig. 6. According to *GIG* flooding is extended until all floods intersect, in this case node m. Then m broadcasts a message to every node inside the flooding union, which would be every node in this example, counting hops from m. This way the distance $|(x, d)|$ from a node x to a datanode d is calculated as $|(x, d)| = |(x, m)| + |(m, d)|$, which is clearly an overestimation. In our example the distance between node f and q is incorrectly estimated as $|(f, q)| = 5$ by *GIG* (following the solid edges) and correctly as $|(f, q)| = 4$ by *dFNS* (following the dashed edges). As a result the *GIG* algorithm

would choose node m as the new operator node, although the actual *Fermat* node and optimal new operator node is f. The hosting costs estimated by *GIG* and our algorithm can also be seen in Fig. 6.

There is a maximum radius that each datanode has to flood in order to be able to reach every candidate distance combination. The maximum radius is set to guarantee completeness, i.e. if there is an *external* node with hosting cost better that c_b of the best datanode it will be found. For the maximum radius of datanode d_i we use the maximum value of α_i $\forall \alpha \in A$.

5.5. Flooding speed

Increasing the likelihood that the floods will intersect at the *Fermat* node first, increases the likelihood that more nodes inside the flood intersection will receive a lower cost threshold. This in turn leads to fewer nodes reporting to the lead node as *Fermat* candidates, thus saving on communication cost. We define a primary speed for each flood in order for them to reach the *ideal* distance combination ϵ at the same time point.

After the floods reach the *ideal* distance combination ϵ they will keep expanding until they reach the maximum radius. We define a secondary speed for the foods that will make their intersection grow faster toward the distance combinations with the lower hosting costs.

The flooding speed is implemented by delaying the relay (broadcast) of the flooding message at every node. More specifically, a timeout is defined at flood initialization, that each node should obey before re-broadcasting the flooding message. This timeout is defined by multiplying the estimated time it takes for the message to travel over one hop by a delay factor. Based on the *ideal* distance combination ϵ, we compute the primary delay factor pf of the flood for each datanode $d_i \in D$ as follows:

$$pf(i) = max\{e_i\}_{\forall i}/\epsilon_i - 1$$

When the delay factor is 0 then the flooding message gets forwarded immediately.

To calculate the secondary delay factor sf we reverse the order of the pf. The datanode d_i with the maximum pf will have a secondary delay factor equal to the minimum pf. The datanode d_j with the minimum pf will have a secondary delay factor equal to the maximum pf and so on. The intuition about this is that the cheapest distance combinations will be the ones with minimum values satisfying the triangle inequality between the datanodes.

5.6. The dFNS algorithm

Here we describe the *dFNS* algorithm as general steps taken inside the network. Assume each operator node has some pre-specified criteria that decide whether an operator placement update is needed or not. These criteria could involve monitoring the change in the data loads of the datanodes, the change in the location of the datanodes, the remaining energy on the operator node, and estimations on whether an operator placement update would be worth the cost overhead for a cheaper query execution in the next epoch. What happens after this decision is taken is described next and shown in pseudo-code in Algorithm 2.

Assume there is an operator placed on node h, that sends/receives data from datanodes D. Thus, it knows the data loads for every node in D. When the criteria of node h to update the operator are met, node h becomes the *leader* node and initiates the *dFNS* algorithm (Algorithm 2). It calculates all candidate distance combinations using Algorithm 1. If the candidate distance combination set is empty, the *leader* informs all datanodes that the new operator placement has changed to b. Otherwise, if there are candidate distance combinations for external nodes, the *leader* computes the time-point to initiate flooding for synchronization. Without synchronization variable speed flooding would not have the desired effect. The *leader* sends a message to all datanodes in D containing the time point to initiate the flooding and the candidate distance combinations.

Using the candidate distance combination set, datanode d_i can calculate the hosting cost c_b of the *best* datanode. It also can compute the minimum and maximum radius, and the primary and secondary speed of its flood, described in Section 5.5 and Section 5.4 respectively. d_i prepares a flooding message that contains the cost threshold set to c_b, the timeout needed to realize the primary speed, the timeout needed to realize the secondary speed, the minimum radius, and the maximum radius of the food. d_i initiates its flooding at the given time-point broadcasting its flooding message. All the candidate nodes send their report to the *leader*. After all reports are received, the *leader* calculates the best candidate node, informs the datanodes about the new operator host and passes on any information regarding the operator to the new host node.

When an external node n receives a flooding message from datanode d_i for the first time it stores it and performs a series of checks. If n is not beyond the minimum radius then it just forwards the message. Any consecutive receptions of the same message are ignored. Otherwise, if n has received a message from all the datanodes in D it can calculate its hosting cost c_n. If c_n is smaller than the cost threshold contained in any of the flooding messages, node n updates the cost threshold inside every flooding message that was not yet forwarded and stores c_n. Also, n sets a timeout to report to the *leader* node as a candidate node. A final check that node n performs when receiving a message from datanode d_i is whether it is not on the maximum radius so it can forward the message it received.

Algorithm 2 The general *dFNS* steps

Steps taken by *leader* node:

```
1: A = CDCGenerator(∅, 0)
2: if A = ∅ then
3:     Place operator on b
4: else
5:     Set timepoint t for initiating flood
6:     m ← t, A
7:     send message m to D
8: end if
9: timeout until all candidate nodes have reported
10: choose best candidate as new operator host
11: inform datanodes about new operator host
12: send operator information to new operator host
```

Steps taken by each datanode $d_i \in D$:

```
1: compute minimum and maximum radius
2: compute primary and secondary speed
3: initiate flood at timepoint t
```

Every candidate node that has not reported yet to the *leader* performs eves-dropping on its neighbors. When such a candidate node receives a message containing a lower cost threshold than its hosting cost, it cancels the timeout for reporting to the *leader* node and withdraws its designation as a candidate node. This way the number of candidate node reports sent to the *leader* node are minimized. The pseudo-code is omitted due to lack of space.

5.7. *Optimality of* dFNS

Our algorithm always finds the node in the network that minimizes the hosting cost as defined in objective function (3).

The dFNS *algorithm is optimal.*

Proof. First, all possible distance combinations A, that have a better hosting cost than the *best* datanode, are found using Algorithm 1. This is true since the algorithm is exhaustive. The radius for the flood of datanode d_i is set to the maximum value of α_i $\forall \alpha \in A$. This guarantees that all possible nodes with distance combinations equal to the ones in the candidate set A will be inside the intersection of all floods. This allows them to calculate their hosting cost and become candidate nodes. □

6. Initial operator placement

Our distributed Fermat Node Search algorithm (*dFNS*) can be used in any framework that optimizes continuous queries, to always keep the operator placement optimal. In addition, frameworks that are made to optimize snapshot queries can be adapted for answering continuous queries by using *dFNS*. The query execution is divided in epochs. As soon as the query execution terminates, an operator node checks whether its operator meets the placement update criteria. Details of these criteria are orthogonal to this work. If those are met, *dFNS* is triggered and the operator placement is optimized before the next epoch. Note, that *dFNS* has no overhead whatsoever during query execution. The overhead is most of the time negligible even when an operator placement update takes place. We have a communication cost overhead only in the less frequent cases, where a flood is needed to find the new optimal operator host.

Most of the previously proposed algorithms for initial operator placement (Section 3) are centralized, assuming global knowledge of the network. When answering snapshot queries, the initial placement should be as optimized as possible, which cannot be achieved without collecting network information. For continuous queries, however, the quality of the initial operator placement is less of an issue, as the query executes for several epochs. A rough initial placement is calculated with the information that is locally present or is collected locally without significant overhead, avoiding the collection of global network knowledge. After the query execution in the first epoch is over, we can check the criteria for each operator and, if they are met, run *dFNS* to optimize the operator placement for consecutive epochs.

We use the algorithm proposed in Chatzimilioudis et al. [6] for finding an initial operator placement. We exploit the mandatory query dissemination to collect some information about the network with minimum overhead. Every node, that receives the query dissemination and has data needed for answering the query, sends to the querying node its position and a summary of its data. Techniques for building a summary of small size and high information have been previously proposed in the literature [14,20,29,4]. Using this information the query node can roughly estimate the hop-distance between the datanodes and the selectivity of each operator.

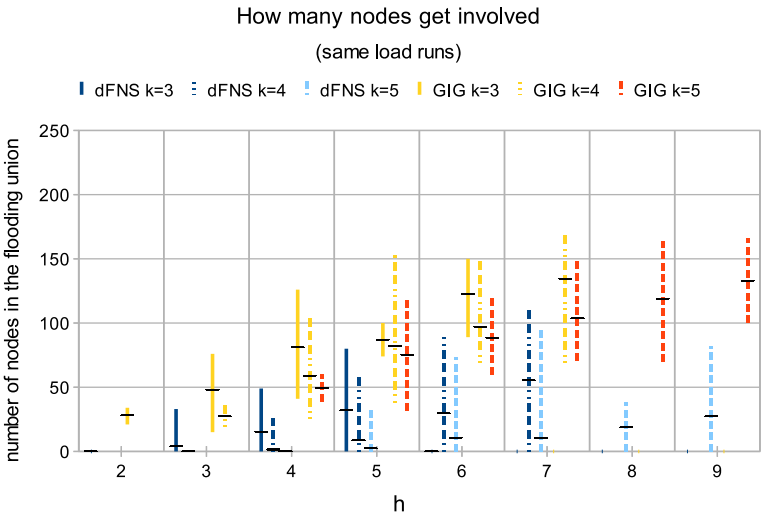

Fig. 7. Number of nodes involved in the flooding process (minimum, average and maximum value) using the same load for each datanode. When lines are missing it means there were no simulation runs possible for the combination of k and h values.

7. Experimental evaluation and analysis

The experiments were run on an Intel Core2 Duo CPU at 2.5 GHz with a 4GB RAM. We implemented the algorithms in Java and used J-Sim [19] as our network simulator. We used the energy model of J-Sim to account for the energy spent by the network when transmitting data.

Comparison is done against the algorithm proposed by Zoe Abrams and Jie Liu in [1], noted as *GIG*. The authors' implementation was not available, so we re-implemented *GIG* with clarifications from the authors. This algorithm needs two parameters from the user in order to run: the radius α of the initial flood and a function $G()$ defining how the radius is increased for each consecutive flood. Its performance heavily depends on these parameters. For our experimental setup we use the optimal values $\alpha = 1$ and $G(r_{new}) = r_{previous} + 1$, which are the same as in the original paper. Those choices are optimal since most of the resulting networks have the datanodes in close proximity and only a small flooding radius of 1 or 2 hops is needed. We also implemented the variable speed flooding function that is only suggested as a future optimization in [1]. This function sets the flooding speed of datanode i to be inversely proportional to the data load of datanode i.

For the experiments we create a network with 512 nodes randomly scattered in a physical space of size 1000×1000. We randomly place the datanodes in a square region of size 200×200 at the center of the space. This is done so that the flooding process can reach a large number of nodes without hitting the edge of the network, and we get more accurate results regarding the efficiency of the algorithms. This is the same network setup as in [1].

We run experiments for m-ary operators with $m = 2, 3, 4$, thus we have $k = 3, 4, 5$ datanodes respectively, showing the efficiency of the compared algorithms. For each value of k we run 80 simulations. The amount of the communication cost overhead is immediately dependent on how far the datanodes are from each other, since the further apart the bigger the floods will have to be. Therefore, our experiments are grouped by metric h. We sum the distances from the datanodes to the *Fermat* node returned by the algorithm in the equation:

$$h = \sum_{\forall d_i \in D} \left|(d_i, f)\right|$$

7.1. Reproducing experiments of GIG

For the first set of experiments we copied the operator migration experiments conducted in [1]. The authors used the same data load w for all the datanodes and used three metrics: number of nodes involved in the flooding process (Fig. 7), number of candidate nodes (Fig. 9) and quality of the first candidate (Fig. 10). All figures denoted as "same load runs" belong to the first set of experiments. We got approximately the same results for the *GIG* algorithm as in [1]. Our proposed algorithm (*dFNS*) outperforms *GIG* in this first set of experiments.

In Fig. 7 the minimum, average and maximum number of nodes involved in the flooding process is shown when running each algorithm for $k = 3, 4, 5$ datanodes. The simulation runs are grouped by h, how far apart the datanodes are. For each value of h the leftmost three lines belong to *dFNS*, whereas the rightmost three lines belong to *GIG*. Some lines are missing, since not all combinations of k and h are possible. For example, when we have $k = 5$ distinct datanodes it is impossible to find an operator node whose sum of distances to the datanodes is less than 4, $h < 4$. We can have $h = k - 1$ only if the *Fermat* node returned by the algorithm is one of the datanodes itself and every other datanode is only 1 hop away from the *Fermat* datanode.

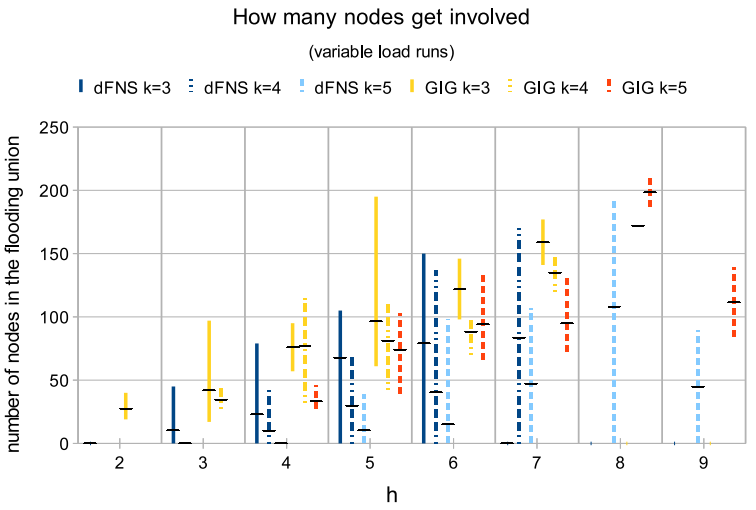

Fig. 8. Number of nodes involved in the flooding process (minimum, average and maximum value) using the variable load for each datanode. When lines are missing it means there were no simulation runs possible for the combination of k and h values.

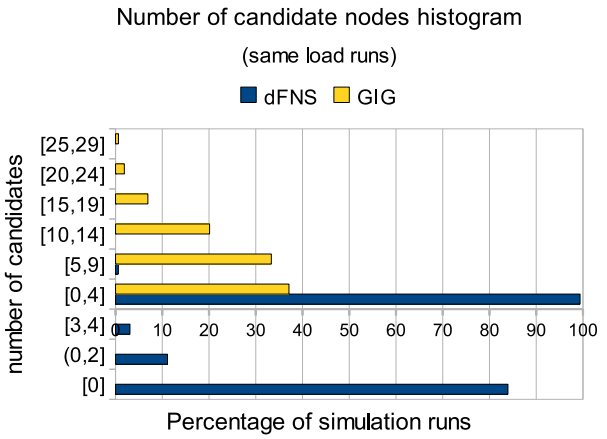

Fig. 9. Number of candidate nodes that report to the leader node. The less candidates the less communication needed. Beneath the x-axis the group $[0, 4]$ is divided into more detailed groups to show the distribution for *dFNS*.

One would expect that *GIG* always involves less nodes in its flooding than *dFNS*, since it stops flooding as soon as the floods intersect for the first time. As Fig. 7 shows, though, *dFNS* has a far smaller mean value of the number of nodes involved in flooding than *GIG*. All this can be attributed to the fact that *dFNS* identifies the special cases where a datanode is the *Fermat* node, and avoids flooding. Those are the frequent cases where the number of nodes involved is zero.

For the second metric, we plot the number of candidate nodes produced by each algorithm in a histogram in Fig. 9. The less candidate nodes, the fewer candidate node messages have to be sent to the leader node to decide on the best candidate. We can see that *dFNS* has never more than 4 candidates. This happens because our algorithm looks for candidates only in the intersection of its extended floods, whereas *GIG* looks for candidates inside the whole union of its floods. Below the x-axis the group of [0–4] candidate is broken down just to show the distribution for our algorithm.

In Fig. 10 we can see the quality of the first candidate. The quality is expressed by the ratio λ equal to the hosting cost of the actual *Fermat* node over the hosting cost of the first candidate node found. If $\lambda = 1$ it means the first candidate node is the actual *Fermat* node. The first time all floods intersect, there are one or more candidate nodes inside the intersection, which will all report to the leader. The one with the best hosting cost is called the first candidate node. This is how the first candidate is defined for both algorithms. The quality of the first candidate depends solely on the variable speed flooding function used. We can see that our variable speed flooding function has always a better first candidate.

The simulations described so far were conducted solely for the purpose of matching the experiments in [1], in order to compare our algorithm head on against *GIG*. The second set of experiments is a fairer comparison between the two algorithms, since in real world applications the datanodes usually have different data loads.

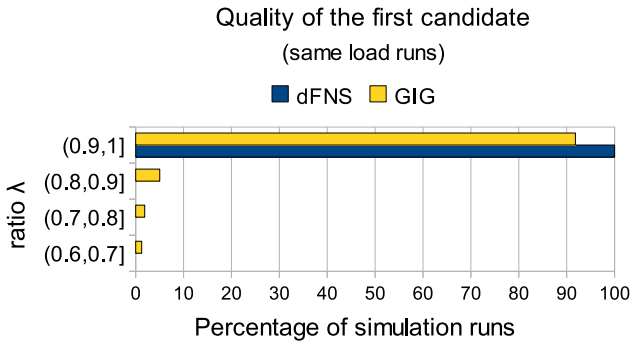

Fig. 10. Quality of the first candidate node encountered while flooding (or the best of the first set). Quality is expressed by dividing the hosting cost of the actual *Fermat* node to the first candidate. The closer the ratio is to 1 the better the variable speed flooding function of the algorithm used.

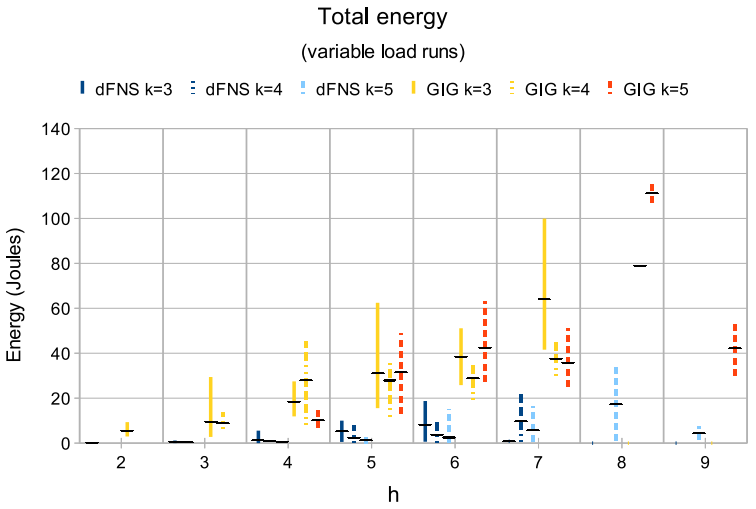

Fig. 11. Total energy consumed (minimum, average and maximum value). When lines are missing it means there were no simulation runs possible for the combination of k and h values.

We run all of the above experiments again with variable data loads. We varied the loads of the datanodes slightly with a Gaussian distribution around weight w used in the previous experiments. A greater variation in the loads would leave us with only a few runs where no datanode is a Fermat node, which is the only case where we can study these heuristics. The results regarding the first metric can be seen in Fig. 8. We excluded the results for the other two metrics because the result were identical to the "same load" simulations seen in Figs. 9 and 10.

Apart from better efficiency, *dFNS* also finishes faster since it does not use incremental flooding, where the network is flooded repeatedly until an intersection is found. *dFNS* floods once to a predefined restricted neighborhood.

7.2. Actual communication cost overhead

We also conducted our own experiments using as metrics the total energy and the maximum energy per node consumed for finding the new *Fermat* node. After all, this is what we are trying to minimize with our algorithm. These are more important metrics compared to the above and the ones that actually define the performance of the algorithms.

Figs. 11 and 12 show that *dFNS* clearly has a smaller energy overhead for determining the optimal hosting node. *GIG*'s limited cost flooding results in reflooding the neighborhood incrementally, thus yields a very big total energy cost. In addition it refloods the whole flooding union to look for candidate nodes. In our algorithm very often we do not need to flood in the first place. This keeps the mean value of total energy low. In the case where flooding is needed, our algorithm might use slightly larger flood radii, but it floods only once, and no further communication between nodes is needed to find any candidate nodes.

To simulate the energy in the previous experiments, we used the following parameters for our sensors: power consumption for transmission 0.660 Watts and power transmission for reception 0.395 Watts. The data rate of the radio is set to 19.2 kbps and the load of each transmission is 1Kb.

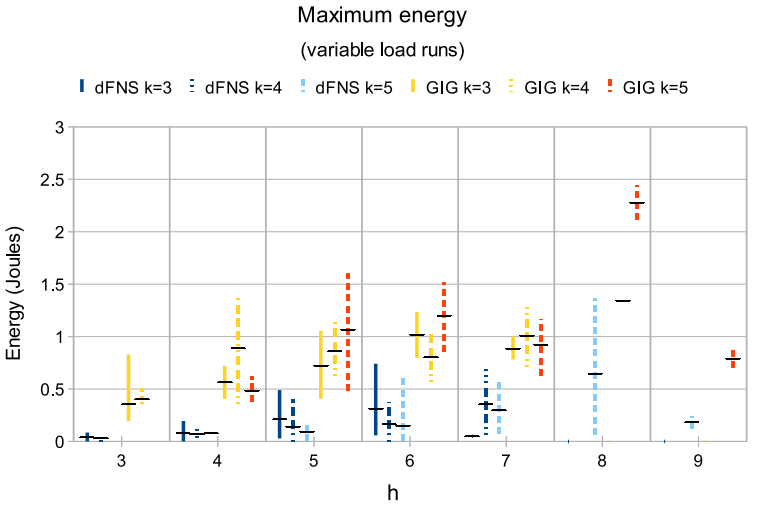

Fig. 12. The maximum energy consumed by a single node (minimum, average and maximum value). When lines are missing it means there were no simulation runs possible for the combination of k and h values.

Table 1
Percentage of simulation runs where a datanode is the optimal node to place the operator and thus no flooding is needed.

	$k=3$	$k=4$	$k=5$
same load	85%	84%	83%
variable load	68%	66%	56%

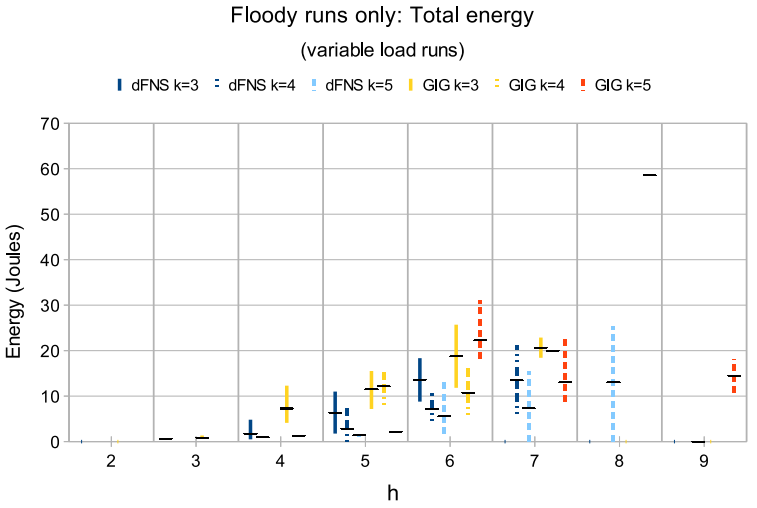

Fig. 13. Total energy consumed (minimum, average and maximum value) only for simulation runs that needed to use flooding in order to find the *Fermat* node. When lines are missing it means there were no simulation runs possible for the combination of k and h values.

These differences are affected by the fact that *dFNS* takes advantage of the cases where a datanode is a *Fermat* node in order to save energy by avoiding flooding. We can see in Table 1 that the majority of cases have a datanode that is the actual *Fermat* node and thus we do not need to look any further for the optimal operator placement.

7.3. *How good is* dFNS *in finding external* Fermat *nodes*

It is clear that our algorithm successfully identifies the special cases where flooding can be avoided. Here we evaluate our algorithm for the other case by collecting information only from the simulation runs where the *Fermat* node is an external node and flooding is needed (*floody runs*). Figs. 13 and 14 show that *dFNS* still outperforms *GIG*, although the savings are less significant compared to the cases where no flooding is needed. Fig. 13 shows the minimum, mean and maximum values of

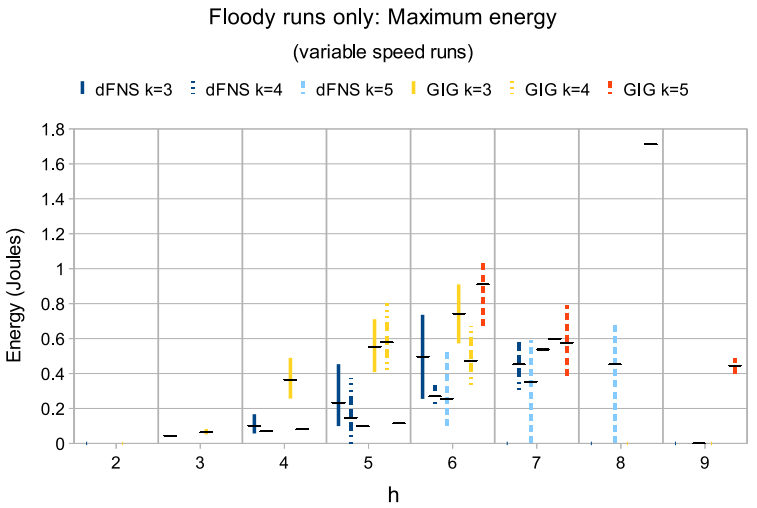

Fig. 14. The maximum energy consumed by a single node (minimum, average and maximum value) only for simulation runs that needed to use flooding in order to find the *Fermat* node. When lines are missing it means there were no simulation runs possible for the combination of k and h values.

the total energy consumed for finding the external *Fermat* node. Similarly, Fig. 14 shows the minimum, mean and maximum values of the maximum energy per node.

8. Dealing with adaptive properties of the operator placement problem

In this section, we provide two additional contributions on the operator placement problem, which extends the baseline framework *dFNS*. These contributions deal with the adaptive properties of the main problem investigated in our research, and, namely, treat the *operator placement adaptation* for multiple operators (*OPA*) (Section 8.1), and the *operator tree adaption* (*OTA*) (Section 8.2), which is possible when commutative query operators over WSN are considered.

8.1. Operator placement adaptation

We present a novel placement adaptation framework for multiple operators (*OPA*). The algorithm's efficiency is based on the fact that the optimal operator host is usually a source/sink node itself as shown by the proposed framework *dFNS*. Recall that *dFNS* achieves two goals: finding the best node to place an operator and minimize communication cost doing so. To achieve this it (1) identifies the special case where no flooding is needed, (2) if flooding is needed, it minimizes the flooding radius, and (3) uses variable speed flooding and eves-dropping. Our algorithm is parameter-free, decentralized, optimal and outperforms previously proposed methods in minimizing communication cost overhead.

In any other case, *dFNS* finds the optimal operator node (*Fermat* node) by extending a flood from each of the *datanodes*. We generate a set of possible distance combinations that the new hosting node can have to produce a smaller hosting cost. Using these candidate distance combinations *dFNS* calculates the minimum possible radius for each flood, guaranteeing that the nodes that participate are kept to a minimum without compromising the optimality of the algorithm.

The adaptation algorithm gets initiated at the lowest possible operator nodes that might need to be adapted. Those operator nodes are the nodes that satisfy the following condition:

condition A = operator i has a change in its output data load AND there is no change its input.

The execution of the adaptation algorithm travels up the query routing tree as long the following conditions apply:

condition B = operator i has a change in its output data load OR changed its placement.

As the lower operator nodes complete the algorithm they send a relay message up to the next operator node up the tree.

condition C = operator i has a change in its input data loads AND i received a completion message from all its children.

When we reach an operator node that has no change in its output or when we reach the sink then we repeat the same process top-down. This time we also complete all placement updates that need to search for the optimal operator host inside the network.

The *CDCGenerator*() takes as input the distances between the datanodes. There are cases where we do not need to run *CDCGenerator* in order to determine whether the optimal operator host is a data node. When we have two sources u and v

Algorithm 3 OPA-upward

Event: Condition A or C holds
Actions:
1: compute *best* datanode *b*
2: *cdc* ← *CDCGenerator*(*distancebetweendatanodes*)
3: **if** *cdc* = ∅ **then**
4: update operator placement
5: inform datanodes
6: **end if**
7: **if** Condition B holds **then**
8: send completion message to parent operator
9: **end if**

Algorithm 4 OPA-downward

Event: node *i* finished executing OPA-upward AND Condition B holds OR completion message received from parent operator
Actions:
1: compute *best* datanode *b*
2: *cdc* ← *CDCGenerator*(*distancebetweendatanodes*)
3: **if** *cdc* = ∅ **then**
4: update operator placement
5: **else**
6: optimal host ← *FNS*()
7: **if** optimal host = ∅ **then**
8: update operator placement
9: **end if**
10: **end if**
11: **if** received a completion message from child operator during OPA-upward **then**
12: send completion message to children operators
13: **end if**

and it holds that $2 * w_u \leqslant w_v$ then the optimal operator host is always data node u. Identifying those cases saves energy from the adaptation process and there is no need to assume knowledge over the distances between the data nodes (source and sink nodes of an operator).

8.2. Operator tree adaptation

Whenever a query involves only commutative operators we can also adapt the operator tree to network and data changes. Even for queries with non-commutative operators, if there is an operator subtree that involves only commutative operators the same technique can be used to adapt that subtree. To the best of our knowledge this is the first work to propose a distributed algorithm for adapting an operator tree.

Our operator tree adaptation (*OTA*) algorithm is based on our placement adaption algorithm. Whenever two operators are placed on the same node then we can initiate the tree adaptation algorithm. Until we get two operators on the same node we just run our placement adaptation algorithm. We show that when a tree adaptation is needed then the operator adaptation algorithm is guaranteed to place two operators on the same node.

Whenever a node hosts two operators it checks whether switching the operators' inputs and outputs with each other can yield a better communication cost during query execution. If a change is needed then it is performed on the operator tree immediately. The placement adaptation algorithm is then called for both operators on this node to find their optimal placement using their new datanodes. Placement update is performed immediately. The placement update naturally triggers more changes in the tree. These changes are administered by the placement adaptation algorithm presented in the previous section.

All of our proposed algorithms can have been presented using binary operators. This does not prohibit our algorithms to optimize queries with k-ary operators. Any k-ary operator can be transformed into a sequence of $(k-1)$ binary operators where their outputs are pipelined. Transforming k-ary operators to binary and optimizing the query routing tree using the binary version of the operators give us more flexibility for optimization. Whenever two or more binary parts of the same operator are placed on the same node, we can say that we place a k-ary operator on that node.

8.3. Remarks

Upon the same experimental framework presented in Section 7, the two proposed *dFNS* extensions *OPA* and *OTA* can be easily implemented and tested as to observe principal variations. However, since both extensions make use of *dFNS* as baseline algorithm, it is reasonable enough to assert that behaviors follow the ones presented in Section 7, which is a matter of (further) contribution in addition to the (more relevant) capability of dealing with adaptive properties of the operator placement problem ensured by the *OPA* and *OTA* solutions.

9. Conclusions and future work

In this paper, we have presented an optimal distributed algorithm to update the placement of a single operator achieving minimum cost for executing continuous queries. Our algorithm imposes minimal communication cost overhead for finding the optimal node to host the operator. Previous work in WSN has only proposed approximate/heuristic algorithms. Besides the advantage of optimality, our experiments show that the cost overhead of our algorithm is reduced by 50%–100% compared to previously proposed techniques. Our distributed Fermat Node Search algorithm (*dFNS*) can be used in any framework that optimizes continuous queries and has specific criteria for triggering operator placement updates. *dFNS* can be seamlessly integrated to keep the operator placement optimal. Experiments have clearly proved the feasibility and the reliability of our proposed framework. In addition to previous contributions, we have also provided two novel solutions, namely *OPA* and *OTA*, which are capable of taking into account adaptive properties of the operator placement problems.

Future work is oriented towards developing novel optimization solutions for in-network query processing over WSN. Sophisticated techniques to answer subset queries over a subset of regions should be optimized holistically. Constructing a query routing tree that acquires data from inside a region to representative nodes and calculates in-network the answer to the query while routing the data towards the sink will make even more optimization possible. It would also be of great interest to optimize a query routing tree as a whole including all three of its subparts: operator tree, operator placement and routing. Work so far, including this one, focuses either on one or two parts at a time.

Acknowledgment

This research was partially funded by the DISFER project.

References

[1] Z. Abrams, J. Liu, Greedy is good: On service tree placement for in-network stream processing, in: International Conference on Distributed Computing Systems, 2006, p. 72.
[2] L. Amini, N. Jain, A. Sehgal, J. Silber, O. Verscheure, Adaptive control of extreme-scale stream processing systems, in: ICDCS '06: Proceedings of the 26th IEEE International Conference on Distributed Computing Systems, IEEE Computer Society, Washington, DC, USA, 2006, p. 71.
[3] P. Andreou, A. Pamboris, D. Zeinalipour-Yazti, P.K. Chrysanthis, G. Samaras, Etc: Energy-driven tree construction in wireless sensor networks, in: Mobile Data Management, 2009, pp. 513–518.
[4] B.H. Bloom, Space/time tradeoffs in hash coding with allowable errors, Commun. ACM 13 (7) (July 1970) 422.
[5] B.J. Bonfils, P. Bonnet, Adaptive and decentralized operator placement for in-network query processing, Telecommun. Syst. 26 (2–4) (2004) 389–409.
[6] G. Chatzimilioudis, H. Hakkoymaz, N. Mamoulis, D. Gunopulos, Operator placement for snapshot multi-predicate queries in wireless sensor networks, in: MDM 2009: Proceedings of the 10th International Conference on Mobile Data Management, May 2009.
[7] G. Chatzimilioudis, D. Zeinalipour-Yazti, D. Gunopulos, Minimum-hot-spot query trees for wireless sensor networks, in: MobiDE, 2010, pp. 33–40.
[8] M.-S. Chen, P.S. Yu, A graph theoretical approach to determine a join reducer sequence in distributed query processing, IEEE Trans. Knowledge Data Eng. 6 (1) (1994) 152–165.
[9] A. Cuzzocrea, Intelligent algorithms for data-centric sensor networks, J. Network Comput. Appl. 35 (4) (2012) 1175–1176.
[10] A. Cuzzocrea, A. Papadimitriou, D. Katsaros, Y. Manolopoulos, Edge betweenness centrality: A novel algorithm for qos-based topology control over wireless sensor networks, J. Network Comput. Appl. 35 (4) (2012) 1210–1217.
[11] A. Deshpande, Prdb: Managing large-scale correlated probabilistic databases (abstract), in: L. Godo, A. Pugliese (Eds.), SUM, in: Lecture Notes in Comput. Sci., vol. 5785, Springer, 2009, p. 1.
[12] A. Deshpande, C. Guestrin, S. Madden, Model-based querying in sensor networks, in: L. Liu, M.T. Özsu (Eds.), Encyclopedia of Database Systems, Springer, US, 2009, pp. 1764–1768.
[13] A. Deshpande, Z.G. Ives, V. Raman, Adaptive query processing, Found. Trends Databases 1 (1) (2007) 1–140.
[14] C. Estan, J. Naughton, End-biased samples for join cardinality estimation, in: Proceedings of the 22nd International Conference on Data Engineering, 2006. ICDE '06, April 2006, p. 20.
[15] I. Greenberg, R.A. Robertello, The Three Factory Problem, Mathematical Association of America, 1965.
[16] M. Hadjieleftheriou, G. Kollios, P. Bakalov, V.J. Tsotras, Complex spatio-temporal pattern queries, in: Proceedings of the 31st International Conference on Very Large Data Bases, Trondheim, Norway, August 30–September 2, 2005, ACM, 2005, pp. 877–888.
[17] C. Intanagonwiwat, R. Govindan, D. Estrin, J.S. Heidemann, F. Silva, Directed diffusion for wireless sensor networking, IEEE/ACM Trans. Netw. 11 (1) (2003) 2–16.
[18] N. Jain, R. Biswas, N. Nandiraju, D. Agrawal, Energy aware routing for spatio-temporal queries in sensor networks, in: Wireless Communications and Networking Conference, 2005 IEEE, vol. 3, March 2005, pp. 1860–1866.
[19] J. Kacer, Discrete event simulations with J-Sim, in: Intermediate Representation Engineering for Virtual Machines, Dublin, Ireland, 2002, pp. 13–18.
[20] R.J. Lipton, J.F. Naughton, D.A. Schneider, Practical selectivity estimation through adaptive sampling, in: Proceedings of the 1990 ACM SIGMOD International Conference on Management of Data, 1990, pp. 1–11.
[21] S. Madden, M.J. Franklin, J.M. Hellerstein, W. Hong, Tinydb: an acquisitional query processing system for sensor networks, ACM Trans. Database Syst. 30 (1) (2005) 122–173.
[22] K. Oikonomou, I. Stavrakakis, A. Xydias, Scalable service migration in general topologies, in: International Symposium on a World of Wireless, Mobile and Multimedia Networks, 2008, pp. 1–6.
[23] A. Pathak, V.K. Prasanna, Energy-efficient task mapping for data-driven sensor network macroprogramming, in: DCOSS '08: Proceedings of the 4th IEEE international conference on Distributed Computing in Sensor Systems, Springer-Verlag, Berlin, Heidelberg, 2008, pp. 516–524.
[24] P.R. Pietzuch, J. Ledlie, J. Shneidman, M. Roussopoulos, M. Welsh, M.I. Seltzer, Network-aware operator placement for stream-processing systems, in: ICDE, 2006, p. 49.
[25] R.L.F. Pitu, B. Mirchandani, Discrete Location Theory, Wiley, New York, NY, USA, July 1990.
[26] G.J. Pottie, W.J. Kaiser, Wireless integrated network sensors, Commun. ACM 43 (5) (2000) 51–58.
[27] E. Shih, S.-H. Cho, N. Ickes, R. Min, A. Sinha, A. Wang, A. Chandrakasan, Physical layer driven protocol and algorithm design for energy-efficient wireless sensor networks, in: MOBICOM, 2001, pp. 272–287.

[28] U. Srivastava, K. Munagala, J. Widom, Operator placement for in-network stream query processing, in: PODS '05: Proceedings of the Twenty-Fourth Symposium on Principles of Database Systems, ACM, New York, NY, USA, 2005, pp. 250–258.
[29] Y. Tao, G. Kollios, J. Considine, F. Li, D. Papadias, Spatio-temporal aggregation using sketches, in: ICDE, IEEE Computer Society, 2004, pp. 214–226.
[30] S. Tilak, N.B. Abu-Ghazaleh, W.R. Heinzelman, A taxonomy of wireless micro-sensor network models, Mobile Comput. Commun. Rev. 6 (2) (2002) 28–36.
[31] E. Weiszfeld, Sur le point pour lequel la somme des distances de n points donnes est minimum, Tohoku Math. J. 43 (1937) 355–386.
[32] T. Yan, Y. Bi, L. Sun, H. Zhu, Probability based dynamic load-balancing tree algorithm for wireless sensor networks, in: Int. Conf. Networking and Mobile Computing, 2005.
[33] Y. Yao, J. Gehrke, Query processing in sensor networks, in: CIDR, Conf. on Innovative Data Systems Research, 2003.
[34] L. Ying, Z. Liu, D.F. Towsley, C.H. Xia, Distributed operator placement and data caching in large-scale sensor networks, in: INFOCOM, IEEE, 2008, pp. 977–985.
[35] O. Younis, S. Fahmy, Heed: A hybrid, energy-efficient, distributed clustering approach for ad hoc sensor networks, IEEE Trans. Mob. Comput. 3 (4) (2004) 366–379.