Energy-Aware Data Processing Techniques for Wireless Sensor Networks: A Review

Suan Khai Chong[1], Mohamed Medhat Gaber[2],
Shonali Krishnaswamy[1], and Seng Wai Loke[3]

[1] Centre for Distributed Systems and Software Engineering,
Monash University, 900 Dandenong Rd, Caulfield East, Victoria 3145
Australia
[2] School of Computing
University of Portsmouth
Portsmouth, Hampshire, England, PO1 3HE
UK
[3] Department of Computer Science and Computer Engineering
La Trobe University
Victoria 3086
Australia

Abstract. Extensive data generated by peers of nodes in wireless sensor networks (WSNs) needs to be analysed and processed in order to extract information that is meaningful to the user. Data processing techniques that achieve this goal on sensor nodes are required to operate while meeting resource constraints such as memory and power to prolong a sensor network's lifetime. This survey serves to provide a comprehensive examination of such techniques, enabling developers of WSN applications to select and implement data processing techniques that perform efficiently for their intended WSN application. It presents a general analysis of the issue of energy conservation in sensor networks and an up-to-date classification and evaluation of data processing techniques that have factored in energy constraints of sensors.

1 Introduction

Data processing engines in wireless sensor networks (WSNs) are typical energy and processing power constrained P2P systems. There are several challenges that need to be addressed in order to promote the wider adoption and application of WSNs. These challenges relate to both individual sensor hardware and operations of the sensor network as a whole. Individual sensors have limitations in terms of sensing and processing capabilities [48] while on the level of the whole network, the issues extend to finding a suitable communication protocol to deal with frequent topology changes, routing protocols to maximise sensor lifetime and dealing with extensive data generated from sensor nodes [1]. In order to deal with the extensive data generated from sensor nodes, the main strategies that have been proposed in the domain of WSNs involve:

A. Hameurlain, J. Küng, and R. Wagner (Eds.): TLDKS III , LNCS 6790, pp. 117–137, 2011.

- **How frequently to sense and transmit data?** the frequency of transmission is important because radio communication consumes the most energy in a sensor network [2]. Energy consumption is significantly reduced when sensors exchange only data that is necessary for the application, for example, sending data on user demand only.
- **How much data has to be processed?** data exchanged between sensor nodes can be either raw (i.e. sensed readings) or processed (e.g. averaged sensed readings). Processing sensor data enables essential information to be filtered out from all data collected on sensors and communication of only data that is important for the application.
- **How is the data to be communicated?** this refers to the communication and routing protocols to transport sensor data from one sensor to another sensor or base station in the network. For instance, the decision to communicate the data from a sensor directly to the base station or via neighbouring sensors to base station. Communication of data from a sensor node directly to a base station may drain the node's energy significantly due to transmission over a long distance whereas routing via peer nodes may prolong the node's energy but decrease overall network lifetime.

Generally, approaches that deal with decisions about the amount of data to be processed or communicated can be classified as data processing approaches. Approaches that deal with the underlying mechanism to communicate the data are referred to as communication protocols for wireless sensor networks. These two types of approaches can work together to conserve energy. We focus on data processing approaches that efficiently reduce the amount of sensor data exchanged, and thereby prolong sensor network lifetime. These can be divided into approaches that operate at the network level or at the node level, as explained in section 2. Following this, section 2 then further discusses the techniques that work at the network level, while section 3 elaborates on the techniques that work at the node level. Lastly, section 4 draws some conclusions about these techniques.

2 Network-Based Data Processing Approaches

As sensor data communication operation is significantly more costly in terms of energy use than sensor data computation, it is logical to process sensor data at or close to the source to reduce the amount of data to be transmitted [48]. This data originates from the sensing capabilities of sensor nodes and can be either stored to be processed at a later time or treated as a continuous stream to be processed immediately [14].

In this section, we introduce the approaches that process such sensor data with a focus on reducing overall energy consumption. These approaches can be broadly classified to be either: (1) network-based, which refers to approaches that involve processing sensor data at the base station; or (2) node-based, which refers to approaches that involve processing sensor data locally at sensor nodes. The former category, network-based, is discussed in details in this section.

Figure 1 illustrates the data processing at network and node levels. The taxonomy of network and node-based data processing approaches is presented in figure 2.

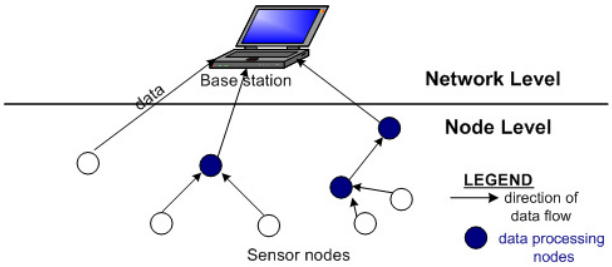

Fig. 1. Processing at network and node levels

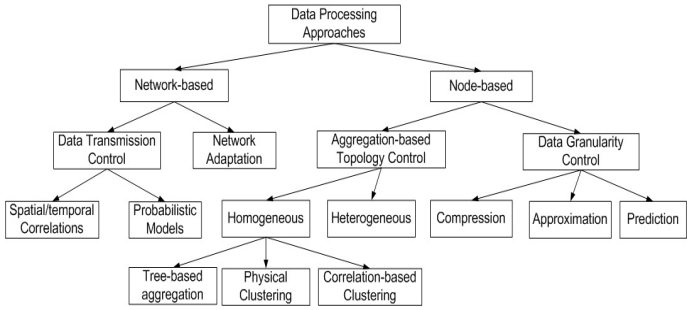

Fig. 2. Data processing approaches taxonomy

Network-based approaches focus on processing of sensory data collected at a resource-rich central location (for instance, a base station) in a sensor network in order to conserve energy. This rationale has been adopted in sensor data sampling techniques, WSN monitoring applications and sensor data querying systems with respect to *data transmission control* and *network adaptation*.

2.1 Data Transmission Control

In this category of approaches, the main idea is that if sensor data correlations or probabilistic models about sensor data is derived at the base station, the base station selectively choose nodes in the sensor network to send their data samples during data gathering. As a consequence, while sampling sensor nodes, the energy consumption of nodes is minimised through reducing the number of data samples that needs to be collected at the base station. In the following section, we discuss existing works that have predicted sensor correlations and sensory values at the base station as a means to efficiently control data collection in this manner.

2.1.1 Spatial/Temporal Correlations

Spatial and temporal correlations have been utilised to more efficiently perform sampling. In [46], the blue noise sensor selection algorithm is presented to selectively activate only a subset of sensors in densely populated sensor networks while sampling. This selection of nodes is derived through building a statistical model of the sensor network nodes known as blue noise pattern (typically used in image processing applications). This model is used to deselect sensor nodes that are active whilst maintaining sufficient accuracy in sensing. The algorithm used also ensures that sensor nodes with the least residual energy are less likely to be selected through incorporating of an energy cost function in the sensor selection algorithm. This approach is an improvement over existing coverage-preserving approaches such as PEAS [60] and node scheduling [55] that do not consider load balancing in node selection.

Another study that utilises spatial correlation to improve sampling is [57] that reduces the number of nodes reporting data to the base station (also referred to as sink). At the sink, an algorithm (known as Iterative Node Selection (INS)) is utilised to derive the number and location of representative nodes in the WSN. These values are derived by using the spatial correlations between sensor observations to choose the representative nodes in the event area whilst, minimising the resulting distortion in sensor data received at the sink. In these approaches, energy is conserved as only a subset of nodes are selected to send data while other sensor nodes can be switched off. A similar idea has also been explored in [59] whereby the authors propose a technique termed 'backcasting' to periodically select only a small subset of sensors with high spatial densities to communicate their information to the base station, whilst ensuring that there is minimal accuracy loss in the data collected.

2.1.2 Probabilistic Models

Probabilistic models have also been used as a means to reduce the amount of communication from sensor nodes to the network's base station [11,56]. In [11], this is achieved through utilising a probabilistic model based on time-varying multivariate Gaussians to derive the probability density function over the set of sensor attributes present in a stored collection of data. The resulting probabilities can then be used to derive spatial and temporal correlations between sensing attributes, which are consequently utilised to form the answer to any user query. Their approach conserves energy as these correlations can be used to develop a query plan that chooses the most cost effective plan to query the sensors. For instance, in answering a query, the model might suggest a plan to sample a voltage sensor (voltage reading used to predict temperature reading) rather than the temperature sensor as the cost to sample a temperature sensor is higher.

In [12], the authors propose a database abstraction known as model-based views to presents users with a consistent view of sensor data queried irrespective of the time and space. This is because the proposed system, MauveDB, applies a predictive model to the static data as well as keeping the output of the model used consistent with changes to the underlying raw data. This allows a conceptual view of sensor data to be presented to the user on demand. The same notion

of providing a consistent view of data has also been presented in [28] where a dynamic probabilistic model (DPM) is exploited to create a probabilistic database view.

In another study, [8] propose to conserve energy by efficiently maintaining probabilistic models without sacrifice to sensor communication cost and data quality. More specifically, in their approach, two dynamic probabilistic models are maintained over sensor network attributes: one distributed across and the other at the base station. By keeping these two models in sync at all times, the approach allows data to be communicated only when the predicted values at the base station are not within bounds. Consequently, energy is conserved as the spatial correlations calculated at sensors node can be used to reduce the amount of data communicated from nodes to the base station when answering a query. Similarly, in [56], their approach involves enforcing the building of autoregressive models locally on the node to predict local readings and collectively at the base station. In this case, time series forecasting is employed for prediction.

Separately, in [26], the authors propose an architecture that focus on predicting data reliably at the base station mainly without sensor involvement to minimise communication overhead. This has been achieved through the use of Kalman Filter to filter arriving sensor streams and predict future sensory values. The novelty of the Kalman Filter in their approach lies its ability to predict effectively the internal state of the data stream system in a wide range of wireless sensor applications, for instance in object tracking or network monitoring.

2.2 Network Adaptation

Alternatively, a network-based approach can conserve energy through adapting sensor network operations to physical parameters in the network such as energy levels or to adapt sensing to specific requirements of the WSN monitoring application (for instance, perform sensing only when an expected event is likely to occur). Approaches such as [42,49] have explored these aspects for energy conservation.

In [44], a system level dynamic power management scheme is used, that adapts sensing and communication operations to the expected occurrence (in terms of probability measure) of the event being monitored. This approach is proposed as traditional power management schemes only focus on hardware idleness identification to reduce energy consumption. Their results have shown a gain of 266.92% in energy savings compared to the naive approach that does not apply such adaptation behaviour.

Separately, in [9], the authors propose the use of low-power sensors as tools for enhancing the operating-system-based energy management. From readings obtained by low-powered sensors, the system would infer the user intent and use it to match user needs to conserve energy. For example, energy-consuming cameras that capture user faces would turn off power until low-power thermal sensors indicate a user's presence. They have evaluated this approach by an experimental setting consisting of a camera that periodically captures images and a face detection algorithm that determines the presence or absence of a user. When a user is present at the screen, the display on the computer will be

turned on and vice-versa. An average energy saving of 12% has been reported respectively for their prototype using low-power sensors.

In [42], sensor energy is conserved through the use of *energy maps*. The energy maps provide global information about the energy status of all sensors. This information enables an algorithm to make intelligent energy-saving decisions such as selecting packet routes that preserves energy of sensors with low energy reserves. The building of energy maps is, however, an energy-consuming task. The method proposed is to allow each node to calculate its energy dissipation rate from its past readings and forwards the information to monitoring nodes, which can then update this information locally by calculations. They have evaluated, using simulations, their approach by comparing the amount of energy saved using their proposed approach to the approach where each node sends periodically to a monitoring node only its available energy.

Also, in [49], the authors describe a power-efficient system architecture that exploits the characteristics of sensor networks. This architecture manipulates access points with more resources within one hop transmission of sensor nodes to coordinate sensor nodes activities. Three operating phases for a set of sensor nodes and their access points are evaluated: (1) the topology learning phase; (2) the topology collection phase; and (3) the scheduling phase. Both topology learning and collection phases aid the access points in determining a complete topology information of cooperating sensor nodes. With this information, in the scheduling phase, the access points will then determine packet-sensing behaviour of connected sensor nodes. The authors compared their scheme to a random access scheme. Their results have an extended sensor network lifetime of 1 to 2 years relative to 10 days if the random access scheme is applied, transmission of packets are optimally scheduled to avoid retransmissions, while in the random access scheme, energy is lost due to the large number of packet retransmissions that can occur in collisions.

Figure 3 lists the approaches, their main feature and limitation as well as some representative techniques for each application category and the energy savings that have been reported.

In this section, we have discussed network-based data processing approaches that target energy conservation in WSNs. The following section describes

Approach Type	Feature	Limitation	Techniques	Energy Saved
Data transmission control	+ Generic energy conservation techniques.	- Assumes correlated data.	Blue noise sensor selection	85% energy saved compared to random node selection.
			Iterative node selection	Up to 50% energy saved in ns-2 simulation.
			Kalman Filter	10% improvement in saving communication resource.
Network adaptation	+ Techniques optimised to application.	- Application-specific.	Sensing user intention	12% of energy saved for system overall.
			System-level DPM	266.92% energy saved compared to using a random scheme.

Fig. 3. Comparison of network-based approaches

approaches that focus on processing the data on sensor nodes to further reduce network data transmissions and extend network lifetime.

3 Node-Based Data Processing Approaches

Node-based approaches are those that focus on data processing on sensor nodes locally. As illustrated in figure 2, this section discusses two main aspects of energy conservation at the node level:

How data is to be communicated? this concerns data-centric routing techniques used to determine the structure of communication in the WSN in an energy-efficient manner. In figure 2, this is illustrated as *aggregation-based topology control*.

How much/frequently data is to be communicated? this refers to data processing techniques that have been proposed to conserve energy at the node level (*data granularity control* in figure 2).

The following sections describe the aforementioned two categories in greater detail.

3.1 Aggregation-Based Topology Control

A wireless sensor network can be either heterogeneous, whereby the network is made up of nodes with different data processing capability and energy capacity; or homogeneous whereby all nodes have equal data processing capability and energy capacity. In the heterogeneous network, nodes with more resources can automatically be used for data processing or as intermediate nodes for aggregation. For instance, in [22], resource-rich Stargate nodes are responsible for performing integer-only Fast Fourier Transforms and machine learning while other mica2 nodes in the system are only used to collect acoustic data samples. In [32], data processing nodes are iPaqs in their hierarchical network comprised of *macro-nodes* (iPaqs) and *micro-nodes* (mica motes).

However, in the homogeneous network, all nodes have equal capabilities and a data communication structure is necessary in order to balance the communication load in the network. This section focuses on existing literature that have studied how data from a sensor network can efficiently enable energy-efficient routing of network sensor data to the base station, otherwise known as data aggregation. As illustrated in figure 2, the approaches that enables data aggregation can be further divided into: (1) tree-based aggregation; (2) physical clustering; and (3) correlation-based clustering.

3.1.1 Tree-Based Aggregation

Tree-based aggregation describes data-centric routing techniques that apply the idea of a communication tree rooted at the base station; in which data packets from leaf nodes to the base station are aggregated progressively along the tree through the intermediate nodes. Additional data processing can be performed

on the intermediate nodes that route packets in the communication tree. Tree-based aggregation includes tree-based schemes [31,25,35] and variations to the tree-based schemes [43,38].

Early tree-based schemes have been initially discussed in [31] including:

1. The center at nearest source (CNS) scheme, in which data is aggregated at the node nearest to the base station.
2. The shortest paths tree (SPT) scheme, in which data is transmitted along the shortest path from the node to the base station and data is aggregated at common intermediate nodes.
3. The greedy incremental tree (GIT) scheme, in which the paths of the aggregation tree are iteratively combined to form more aggregation points (the GIT scheme has been further evaluated in [24]).

A well-known tree-based scheme is Directed Diffusion [25]. It is a data-centric routing protocol that allows an energy-efficient routing path to be constructed between the base station and the sensor node that answers the query. The protocol works in the following way. When a query is sent from the base station to the WSN, the base station propagates *interest* messages to surrounding nodes nearest to it. This interest message describes the type of sensory data relevant to answering the query. The nodes, upon receiving this interest message, perform sensing to collect information to answer the query and rebroadcasts the message to their neighbours. In this process, every node also sets up a data propagation gradient used to route answers back to the base station along the reverse path of the interest. The intermediate nodes involved in this propagation can perform data aggregation or processing. The energy expenditure from this technique comes from the frequency of the gradient setup between the nodes, which typically requires data exchanges between neighbourhood nodes.

The Tiny AGgregation approach (TAG) [35] is another approach that uses a tree-based scheme for aggregation. In TAG, the communication tree is constructed by first having the base station broadcast messages to all sensor nodes in order to organise the nodes with respect to their node and distance from the base station in terms of levels. For instance, any node without an assigned level that hears a broadcast message will set its own level to be the level in the message plus one. The base station broadcasts this message periodically to continually discover nodes that might be added at a later stage to the topology. Any messages from a sensor node are then first sent to its parent at the level above it and this process is repeated until it eventually reaches the base station. The main energy expenditure in TAG is in the requirement of having nodes in constant listening mode to receive broadcast messages from the base station. As a consequence, less running energy may be consumed in comparison to Directed Diffusion. However, Directed Diffusion has the advantage in saving energy in the long-term due to its use of cost-optimised routes to answer queries.

Other variations to the tree-based scheme include Synopsis Diffusion [43] and Tributaries and Delta [38]. In [43], a ring topology is formed when a node sends a query over the sensor network. In particular, as the query is distributed across to the nodes, the network nodes form a set of rings around the querying node

(i.e. the base station). Nevertheless, although the underlying communication is broadcast, the technique only requires each node to transmit exactly once, allowing it to generate equal optimal number of messages as tree-based schemes. However, in this technique, a node may receive duplicates of the same packets from other neighbouring nodes, which can affect the aggregation result. Improving over [43] and tree-based approaches, [38] have proposed an algorithm that alternates between using a tree structure for efficiency in low packet loss situations and the ring structure for robustness in cases of high packet loss.

Topology control protocols can also be clustering-based. In the clustering scheme, cluster heads are nominated to directly aggregate data from nodes within their cluster. The following sections describe the two main types of cluster schemes, namely physical clustering and correlation-based clustering.

3.1.2 Physical Clustering

In physical clustering, the clustering is performed on the basis of physical network parameters such as a node's residual energy. Physical clustering protocols discussed in this literature include Hybrid Energy Efficient Distributed Clustering: HEED [63], Low Energy Adaptive Clustering Hierarchy: LEACH [19] and their variants that run on sensor nodes. Firstly, [19] have proposed the LEACH protocol that allows nodes distributed in a sensor network to organise themselves into sets of clusters based on a similarity measure. Among these sets of clusters, sensors would then elect themselves as cluster-heads with a certain probability. The novelty of LEACH lies in the randomised rotation of high-energy cluster-head selection among sensors in its cluster to avoid energy drain on a single sensor. Finally, local data fusion is performed on cluster heads to allow only aggregated information to be transmitted to the source, thereby enhancing network lifetime. LEACH-centralised (LEACH-C) [20] improves over LEACH in terms of data delivery by the use of an overlooking base station to evenly distribute cluster head nodes throughout the sensor network. The base station does this by computing the average node energy, and that allows only nodes over the average node energy to become cluster heads for the current iteration.

Improving upon LEACH and its variant is the distributed clustering algorithm known as HEED (Hybrid Energy-Efficient Distributed Clustering) [63] for sensor networks. The goal of HEED is to identify a set of cluster heads, which can cover the areas that the sensor nodes monitor, on the basis of the residual energy of each node. This is achieved using a probability function to determine the likelihood a node will become a cluster head in order to select the node that attracts more nodes in terms of proximity and which has most residual energy left. HEED, however, has the drawback that additional sensor energy would be depleted in the changing of the cluster heads at each reclustering cycle. To address this limitation, it is essential to prolong the time between changing the cluster heads and running the HEED protocol over longer time intervals. In terms of evaluations, HEED has shown favourable results compared to other techniques such as LEACH, which selects a random cluster head and also has a successful implementation on Berkeley motes.

Adopting similar notions in HEED, several other techniques have tried to improve HEED/LEACH by developing ways to more efficiently select cluster heads. In [27], sensor nodes communicate among themselves through broadcasts messages to form tighter sensor clusters of closer proximity to one another. The sensors stop their broadcasts when the cluster becomes stable. Their technique has been shown to increase the number of sensor nodes alive over LEACH but a shorter time to first node death. Similarly, [61] improves over LEACH by favouring cluster heads with more residual energy and electing them based on local radio communication to balance load among the cluster heads.

Alternatively, [33] present a chain-based protocol, termed PEGASIS (Power-Efficient Gathering in Sensor Information Systems) that improves over LEACH. The main idea in PEGASIS is for nodes to receive from and transmit data to their close neighbours in a chain-like manner, taking turns as intermediate nodes that would transmit directly to the base station. This is done to reduce the number of nodes communicating directly to the base station. The chain can be created by randomly choosing nodes with favourable radio communication strength or created by the base station, which will broadcast the route that forms the chain to all sensor nodes.

3.1.3 Correlation-Based Clustering

More recently, existing work have shown that cluster heads can also be selected in favour of spatial or temporal data correlations between sensor nodes [62]. One such work is Clustered AGgregation (CAG) [62] in which the authors exploit

Approach Type	Features	Limitations
Tree-based aggregation	+ Only local knowledge of topology required. + Duty-cycling can be implemented on top of aggregation technique to save energy. + Robustness in ring topologies.	- Some tree-based protocols not dynamic in presence of node changes, resulting in energy loss from data retransmissions.
Physical clustering	+ Clustering can be on basis of node residual energy, thereby further prolonging overall network lifetime. + Dynamic to topology changes.	- Energy lost when nodes change their cluster memberships.
Correlation-based clustering	+ Clusters created are correlated in data similarity. Useful to answer approximate data queries.	- Does not adapt to node residual energy.

Fig. 4. Topology control techniques summary

spatial and temporal correlations in sensor-data to form clusters with similar node sensory values within a given threshold and that the clusters remain fixed until the sensory value threshold has changed over time. When the threshold values change, the related sensor nodes will then communicate with neighbouring nodes associated with other clusters to change their cluster memberships. CAG allows the user to derive approximate answers to a query. A similar approach can also be observed in [40] whereby spatial correlations are used to group sensor nodes when they have the same behaviour in movement.

Figure 4 illustrates the qualitative differences between the aforementioned aggregation-based topology control mechanisms. We discuss next how energy can be conserved through data processing regardless of the sensor network topology used.

3.2 Data Granularity Control

In this section, we describe existing approaches that reduce the amount of data communicated in-network and thus, prolong sensor network lifetime. The specific type of approach that can be applied to reduce communication is dependent upon the granularity of data required for the WSN application. For instance, a critical-sensing WSN application such as smart home health care systems [29] would require accurate values from sensor nodes at all times to monitor patient health conditions, whereas coarse-granularity data would suffice for certain event detection systems [17,18]. As a consequence, various data granularity control mechanisms have been proposed to cater to the data requirements of WSN applications. As illustrated in figure 2, these techniques include *compression*, *approximation* and *prediction*. In the following sections, we describe the approaches in each of these categories in detail, in the context of their purpose in energy conservation.

3.2.1 Compression

Data compression at sensor nodes serves to reduce the size of data packets that are to be transmitted through packet encoding at the sensor node and decoding at the base station. It is desirable to apply compression techniques on sensory data when highly accurate sensory data from the sensing application is required. However, existing data compression algorithms such as bzip2 [52] are not feasible to be directly implemented on sensors due to their large program sizes [30]. Furthermore, as discussed in [4], there is a net energy increase when an existing compression technique is applied before transmission. Compression algorithms for sensor networks are required to operate with a small memory footprint and low-complexity due to sensor hardware limitations. The compression schemes that serve to meet these requirements focus on either improving the efficiency of the coding algorithm (light-weight variants of existing compression algorithms) or utilising the data communication topology to reduce the amount of data to be compressed (for instance, performing compression mainly on intermediate nodes in a data aggregation tree).

Studies that have manipulated the data communication structure to more efficiently perform compression include [47,3]. In [47], a scheme known as 'Coding

by Ordering' has been proposed to compress sensory data by encoding information according to the arrival sequence of sensory data packets. The compression scheme merges packets routed from nodes to base station into one single large packet up an aggregation tree. The main idea used in compression is to disregard the order in which sensor data packets reach the base station in order to suppress some packets sent from intermediate aggregator nodes to the base station. In effect, this omits the encoding required for the suppressed packets and thus, improves the efficiency to perform compression. This idea to manipulate sensor communication behaviour to improve compression efficiency is also shared in [3], whereby the rationale used is to buffer sensor data for a specified time duration at the aggregator node's memory. This allows the aggregator node to combine data packets and reduce data packet redundancies prior to transmission. The proposed scheme termed 'Pipelined In-Network Compression' by [3] improves data compression efficiency by allowing sensor data packets to share a prefix system with regard to node IDs and timestamp definitions. In this scheme, the data compression efficiency depends on the length of the shared prefix, i.e. the longer the length of the shared prefix, the higher the data compression ratio.

Spatial and temporal correlations in sensor data have also been manipulated to reduce the compression load [7,21,45]. In [7], the approach is to use the base station to determine the correlation coefficients in sensor data and use the derived correlations to vary the level of compression required at individual sensor nodes. The advantage of this approach is that it reduces the communication load to calculate correlations in-network but assumes the availability of the base station to perform the computation. On the contrary, the approach proposed in [21] uses the idea of computing spatial correlations on sensors locally from data packets overheard on the broadcast transmission channel from neighbouring. This approach allows sensor nodes to collaborate on the data encoding process in order to reduce overall data compression ratio on the node, whilst still enabling the data packets to be decoded exactly at the base station. This in effect reduces the amount of transmission required. The benefits of using spatial correlations with compression have been studied more broadly in [45], whereby the authors explored the energy efficiency in the compression of correlated sensor data sources on sensor data given varying levels of correlations.

Apart from the using physical or data parameters to enhance the data compression technique, it is also important for the designed data compression algorithm to be tailored to resource constraints of sensor hardware. This has been studied in [51,39]. In [51], the authors propose a compression scheme with low memory usage to run on sensor nodes. Their proposed compression scheme, S-LZW improves over an existing compression scheme, LZW compression [58]. This has been achieved by setting desirable attributes for LZW compression on a sensor node with regard to the dictionary size, the data size to compress at one time and the protocol to use when the data dictionary fills up. To further improve the running of LZW algorithm on a sensor node, the authors also proposed the use of an in-memory data cache to filter repetitive sensor data. More recently, in [39], the authors proposed a compression algorithm for WSN that outperforms

the S-LZW in terms of compression ratio and computational complexity by exploiting high correlations existing between consecutive data samples collected on the sensor node.

3.2.2 Approximation

In general, the aforementioned data compression techniques focus on reducing the amount of data packets to be transmitted in-network when the accuracy of data collection is important. Alternatively, in applications where approximations in the collected data can be tolerated, approximations on sensory data can be performed instead to further reduce data transmissions. The application of approximation to reduce data transmissions in-network has been explored in [64,54,37,5,35,36,6].

These studies have explored the computation of aggregates in sensor network data. In [64], the authors propose protocols to continuously compute network digests. These digests are specifically digests defined by decomposable functions such as *min*, *max*, *average* and *count*. The novelty in their computation of network digests lies in the distributed computation of the digest function at each node in the network in order to reduce communication overhead and promote load-balanced computation. The partial results obtained on the nodes are then piggybacked on neighbour-to-neighbour communication and eventually propagated up to the root node (base station) in the communication tree. On the contrary, in [5], the amount of communications is reduced by having every node store an estimated value of the global aggregate and updates this estimate periodically depending on changes in locally sensed values. The locally stored global aggregate of a node is only exchanged with another neighbourhood node if the aggregate value changes significantly after a local update. A distributed approximation scheme has also been used in [6] in which they proposed a probabilistic grouping algorithm to run on local sensor nodes so that the computed local aggregates can progressively converge to the aggregated value in real-time. The proposed scheme has the additional benefit that it is robust to node link failures. Apart from common computing aggregates such as *min*, *max*, *average* and *count* in the discussed approaches, *median* (most frequent data values) is another aggregate function that can be used for gathering sensory data. In this regard, [54] have proposed a distributed data summarization technique, known as *q-digest* that can operate on limited memory. This facilitates computation of aggregates such as medians or histograms.

Similarly, in [37,36], the focus is on reducing the total number of messages required to compute an aggregate in a distributed manner. In [37], the authors propose a scheme termed *pipelined aggregate* in which the aggregates are propagated into the network in time divisions. Applying this scheme, at each time interval, the aggregate is broadcast to sensors one radio hop away. A sensor that hears the request transmit a partial aggregate to its parent by applying the aggregate function to its own readings and readings of its immediate child neighbour. As stated by the authors, the drawback to this scheme is in the number of messages that need to be exchanged to derive the first aggregates from all sensors in the network. In a related work, [36] discuss broader data

acquisition issues pertaining to the energy-efficiency of the query dissemination process and energy-optimised query optimisation. For instance, in [36], semantic routing trees are proposed for collecting aggregates from the sensor network.

3.2.3 Prediction

The third class of techniques that can be applied to reduce network data transmissions is prediction. Prediction techniques at the node level derive spatial and temporal relationships or probabilistic models from sensory data to estimate local data readings or readings of neighbouring nodes. When sensory data of particular sensor nodes can be predicted, these sensor nodes are then suppressed from transmitting the data to save communication costs. Similar to compression and approximation, prediction-based techniques are also required to run in a light-weight manner on sensor nodes. In the literature, prediction techniques have been proposed as algorithms for enhancing data acquisition in [34,15,53] and as generic light-weight learning techniques to reduce communication costs in transmissions.

For energy-efficient data acquisition, [34] proposed their Data Stream Management System (DSMS) architecture to optimise the data collection process in querying. In their architecture, sensor proxies are used as mediators between the query processing environment and the physical sensors, where proxies can adjust sensor sampling rates or request for some further operations before sending data by intelligently sampling sensors (for instance, sampling less frequently if the user query demands so) rather than just sampling data randomly. The energy efficiency in data collection has also been addressed in [15], whereby the idea is to select in a dynamic fashion the subset of nodes to perform sensing and transmitting data to the base station and data values for the rest of the nodes predicted using probabilistic models. Similarly, in [53], data to be communicated is reduced by controlling the number of sensor nodes communicating their sensory data. This is achieved by setting threshold values so that a sensor should only send its reading when its reading is outside the set threshold.

In [41], the prediction involves the building of local predictive models at sensor nodes and having sensors transmit the target class predictions to the base station. The models built at sensors can then be used to predict target classes such as to facilitate anomalies detection, which reduces the transmissions of sensor data necessary otherwise to the base station. Such a distributed approach has also been adopted by [13] who propose sensors that re-adjust their actions based on the analysis of information shared among neighbouring sensors. These sensors perform actions that could be more resource and time-efficient. In particular, they acquire spatio-temporal relationships by learning from a neighbourhood of sensor data and history data. Markov models are used to calculate probabilities of the required data fitting into different time intervals. Using the calculated probabilities, sensor readings with the highest confidence for missing sensor data are chosen. Once the correlations between sensor data are learnt and reused over time, the need to send prediction models from the base stations are omitted, thereby saving energy through reduced sensing and communication costs. Experimental results show the efficiency of the classifier based on simulations.

In another study, [10] propose a generalised Non-Parametric Expectation-Maximization (EM) algorithm for sensor networks. Conventionally, a parametric EM algorithm is a clustering algorithm that, from chosen initial values for specified parameters and a probability density function, computes cluster probabilities for each instance. Ultimately, in the maximization step, the algorithm will derive a convergence or local maximum after several iterations. However, for a sensor network, the algorithm is not applicable because sensors frequently report false values, requiring them to calculate the estimates as required by the algorithm, which is infeasible for resource-limited sensor nodes. The solution provided is one that uses the estimation step on nodes such that nodes will estimate their current value from the knowledge of other values from neighbouring nodes. The algorithm requires that sensor nodes report their discrete values. A more recent study is work done by [40] where the authors investigated correlations that can be formed when sensors in loading trucks experience similar vibrations when the trucks send out the same load. The correlation information of the sensor nodes then allowed them to group trucks carrying out the same load. The unique contribution in their work lies in the incremental calculation of the correlation matrix. The idea of exploiting correlations in sensor node readings to reduce transmissions has also been explored in [16].

Generally, the distributed nature of the algorithms proposed in [41,13,40] allow the computation load to be more balanced across the sensor network to achieve computational efficiency. Alternatively, a more centralised node processing model can be applied such as work in [50] whereby the authors propose delegating the base station to calculate the classification model and uploads the predicted model to sensor nodes. Sensor nodes then use the model to selectively report the 'interesting' data points to the base station as a means to reduce energy consumption. Similarly, in Prediction-based Monitoring in Sensor Networks: PREMON [23], the processing is shared by intermediate nodes in the network. In this work, the authors describe a video compression-based prediction technique, the block-matching function of MPEG, to predict spatio-temporal correlation models at

Approach Type	Features	Limitations
Compression-based	+ Preserves the accuracy of readings collected.	- A need for decoding at the base station. - Data compression techniques are computationally heavy.
Approximation-based	+ Low computational overhead relative to data compression.	- Some techniques only applicable to certain WSN applications such as applying aggregates for data querying.
Prediction-based	+ Generic light-weight algorithms have been proposed.	- Some prediction techniques are application-specific. - Prediction may sacrifice data collection accuracy when some nodes are not required to send.

Fig. 5. Data granularity control techniques summary

intermediate nodes. The prediction correlation models are then passed on to sensor nodes within the aggregation group and a sensor node will then send its reading when the readings have not been predicted. To save transmission costs, PREMON enables a sensor to only send its actual readings when the readings cannot be inferred from the predicted model. However, the approach has a heavy overhead as predictions often have to be continuously sent to the sensors.

Figure 5 illustrates the qualitative differences between the aforementioned data granularity control techniques.

4 Conclusions

Existing data processing approaches have explored how sensor communication can be improved through utilising processed data obtained from sensors. This has been achieved by utilising a suitable energy trade-off between computation and communication operations:

Network-based approaches. At the network data level, the data model is such that sensor data arrives at the application running at a central processing location, enabling the application to have a global view of data distribution within a sensor network. By utilisation of a resource-rich base station, network-based approaches can utilise information obtained from expert knowledge or prediction at network data level to obtain cues that would improve sensing and communication operations at a high level.

Node-based approaches. At the node data level, as the computation process is performed locally on sensor nodes, efforts towards energy conservation focus on developing data processing algorithms suitable for resource-constrained sensor nodes. This involves optimising the data processing algorithm to operate on minimal storage using limited processing power while being adaptive to the sensor's remaining energy. Through local data processing data on sensor nodes, the overall amount of data that needs to be transmitted in-network is then consequently reduced, i.e. applying either compression, approximation or prediction.

It can be observed that approaches at the network level have addressed energy conservation by controlling certain aspects of network operation. The work done thus far has focused on improving particular node operations, for instance in sampling, the aim is to reduce the number of nodes in a group that send data to the base station and consequently reduce energy consumption. Similarly, for node level approaches, energy conservation has focussed on reducing data (lossy or lossless) sent from nodes to the base station through building light-weight processing algorithms for sensor networks using compression, approximation or prediction.

Underlying these approaches, we can observe the notion that if some information about the sensor network can be obtained, then this information can

be used to drive sensor network operations efficiently. In the aforementioned approaches, this notion has not been explicated or formalised. Therefore, one direction of future work could be to explore using computed information (locally at sensor node or globally at the base station) to autonomously decide how a sensor should operate efficiently at any point during its sensing task. For example, if the computed information suggests that sensing is no longer required at a sensed region, then the energy-efficient operation to be carried out by sensors in that sensed region may be to sleep or to sense less frequently. In addition, this raises related research questions, namely, how then would streaming sensory information be obtained efficiently and how it would be used to control sensors in a scalable manner.

References

1. Akyildiz, I., Su, W., Sankarasubramaniam, Y., Cayirci, E.: Wireless sensor networks: A survey. Computer Networks: The International Journal of Computer and Telecommunications Networking 38(4), 393–422 (2002)
2. Anastasi, G., Conti, M., Falchi, A., Gregori, E., Passarella, A.: Performance measurements of mote sensor networks. In: ACM/IEEE International Symposium on Modeling, Analysis and Simulation of Wireless and Mobile Systems, Venice, Italy, pp. 174–181 (2004)
3. Arici, T., Gedik, B., Altunbasak, Y., Liu, L.: Pinco: A pipelined in-network compression scheme for data collection in wireless sensor networks. In: 12th International Conference on Computer Communications and Networks, Texas, USA, pp. 539–544 (2003)
4. Barr, K., Asanovic, K.: Energy-aware lossless data compression. ACM Transactions on Computer Systems 24(3), 250–291 (2006)
5. Boulis, A., Ganeriwal, S., Srivastava, M.: Aggregation in sensor networks: An energy-accuracy tradeoff. In: 1st IEEE International Workshop on Sensor Network Protocols and Applications, California, USA, pp. 128–138 (2003)
6. Chen, J., Pandurangan, G., Xu, D.: Robust computation of aggregates in wireless sensor networks: Distributed randomized algorithms and analysis. In: 4th International Symposium on Information Processing in Sensor Networks, California, USA, pp. 348–355 (2005)
7. Chou, J., Petrovic, D., Ramchandran, K.: A distributed and adaptive signal processing approach to reducing energy consumption in sensor networks. In: IEEE International Conference of the IEEE Computer and Communications Societies, San Francisco, USA, pp. 1054–1062 (2003)
8. Chu, D., Deshpande, A., Hellerstein, J., Hong, W.: Approximate data collection in sensor networks using probabilistic models. In: 22nd International Conference on Data Engineering, Atlanta, USA (2006)
9. Dalton, A., Ellis, C.: Sensing user intention and context for energy management. In: 9th Workshop on Hot Topics in Operating Systems, USENIX, Hawaii, USA, pp. 151–156 (2003)
10. Davidson, I., Ravi, S.: Distributed pre-processing of data on networks of berkeley motes using non-parametric em. In: 1st Internation Workshop on Data Mining in Sensor Networks, Los Angeles, USA, pp. 17–27 (2005)

11. Deshpande, A., Guestrin, C., Madden, S., Hellerstein, J., Hong, W.: Model-driven data acquisition in sensor networks. In: 30th Very Large Data Base Conference, Toronto, Canada, pp. 588–599 (2004)
12. Deshpande, A., Madden, S.: Mauvedb: Supporting model-based user views in database systems. In: Special Interest Group on Management of Data, Illonois, USA, pp. 73–84 (2006)
13. Elnahrawy, E., Nath, B.: Context-aware sensors. In: Karl, H., Wolisz, A., Willig, A. (eds.) EWSN 2004. LNCS, vol. 2920, pp. 77–93. Springer, Heidelberg (2004)
14. Gaber, M., Krishnaswamy, S., Zaslavsky, A.: Mining data streams: A review. ACM SIGMOD Record 34(2), 18–26 (2005)
15. Gedik, B., Liu, L., Yu, P.: Asap: An adaptive sampling approach to data collection in sensor networks. IEEE Transactions on Parallel and Distributed Systems 18(12), 1766–1782 (2007)
16. Goel, S., Passarella, A., Imielinski, T.: Using buddies to live longer in a boring world. In: 4th Annual IEEE International Conference on Pervasive Computing and Communications, Washington, USA, p. 342 (2006)
17. He, T., Krishnamurthy, S., Stankovic, J., Abdelzaher, T., Luo, L., Stoleru, R., Yan, T., Gu, L., Hui, J., Krogh, B.: Energy-efficient surveillance system using wireless sensor networks. In: 2nd International Conference on Mobile Systems, Applications and Services, Boston, USA, pp. 270–283 (2004)
18. Hefeeda, M., Bagheri, M.: Wireless sensor networks for early detection of forest fires. In: IEEE International Conference on Mobile Adhoc and Sensor Systems, Pisa, Italy, pp. 1–6 (2007)
19. Heinzelman, W., Chandrakasan, A., Balakrishnan, H.: Energy-efficient communication protocol for wireless microsensor networks. In: 33rd Annual Hawaii International Conference on System Sciences, Maui, USA, pp. 2–12 (2000)
20. Heinzelman, W., Chandrakasan, A., Balakrishnan, H.: An application-specific protocol architecture for wireless microsensor networks. IEEE Transactions on Wireless Communications 1(4), 660–670 (2002)
21. Hoang, A., Motani, M.: Exploiting wireless broadcast in spatially correlated sensor networks. In: International Conference on Communications, Seoul, Korea, pp. 2807–2811 (2005)
22. Hu, W., Tran, V., Bulusu, N., Chou, C., Jha, S., Taylor, A.: The design and evaluation of a hybrid sensor network for cane-toad monitoring. In: 4th International Symposium on Information Processing in Sensor Networks, California, USA, pp. 28–41 (2005)
23. Imielinski, T., Goel, S.: Prediction-based monitoring in sensor networks: Taking lessons from mpeg. ACM Computer Communication Review 31(5), 82–98 (2001)
24. Intanagonwiwat, C., Estrin, D., Govindan, R., Heidemann, J.: Impact of network density on data aggregation in wireless sensor networks. In: 22nd International Conference on Distributed Computing Systems, Vienna, Austria, p. 457 (2002)
25. Intanagonwiwat, C., Govindan, R., Estrin, D.: Directed diffusion: A scalable and robust communication paradigm for sensor networks. In: ACM/IEEE International Conference on Mobile Computing and Networking, Boston, USA, pp. 56–67 (2000)
26. Jain, A., Chang, E., Wang, Y.: Adaptive stream resource management using kalman filters. In: Special Interest Group on Management of Data, Paris, France, pp. 11–22 (2004)
27. Kamimura, J., Wakamiya, N., Murata, M.: A distributed clustering method for energy-efficient data gathering in sensor networks. International Journal on Wireless and Mobile Computing 1(2), 113–120 (2006)

28. Kanagal, B., Deshpande, A.: Online filtering, smoothing and probabilistic modeling of streaming data. In: 24th International Conference on Data Engineering, Cancun, Mexico, pp. 1160–1169 (2008)
29. Keshavarz, A., Tabar, A.M., Aghajan, H.: Distributed vision-based reasoning for smart home care. In: ACM SenSys Workshop on Distributed Smart Cameras, Boulder, USA, pp. 105–109 (2006)
30. Kimura, N., Latifi, S.: A survey on data compression in wireless sensor network. In: International Conference on Information Technology: Coding and Computing, Washington, USA, pp. 8–13 (2005)
31. Krishnamachari, B., Estrin, D., Wicker, S.: The impact of data aggregation in wireless sensor networks. In: 22nd International Conference on Distributed Computing Systems, Washington, USA, pp. 575–578 (2002)
32. Kumar, R., Tsiatsis, V., Srivastava, M.: Computation hierarchy for in-network processing. In: 2nd ACM International Conference on Wireless Sensor Networks and Applications, California, USA, pp. 68–77 (2003)
33. Lindsey, S., Raghavendra, C.: Pegasis: Power-efficient gathering in sensor information systems. In: Aerospace Conference Proceedings, Montana, USA, pp. 1125–1130 (2002)
34. Madden, S., Franklin, M.: Fjording the stream: An architecture for queries over streaming sensor data. In: 18th International Conference on Data Engineering, San Jose, USA, pp. 555–566 (2002)
35. Madden, S., Franklin, M., Hellerstein, J., Hong, W.: Tag: a tiny aggregation service for ad hoc sensor networks. In: 5th Annual Symposium on Operating Systems Design and Implementation, Boston, USA, pp. 131–146 (2002)
36. Madden, S., Franklin, M., Hellerstein, J., Hong, W.: The design of an acquisitional query processor for sensor networks. In: ACM Special Interest Group on Management of Data, Wisconsin, USA, pp. 491–502 (2003)
37. Madden, S., Szewczyk, R., Franklin, M., Culler, D.: Supporting aggregate queries over ad-hoc wireless sensor networks. In: 4th IEEE Workshop on Mobile Computing Systems and Applications, New York, USA, pp. 49–58 (2002)
38. Manjhi, A., Nath, S., Gibbons, P.: Tributaries and deltas: Efficient and robust aggregation in sensor network stream. In: Special Interest Group on Management of Data, Baltimore, USA, pp. 287–298 (2005)
39. Marcelloni, F., Vecchio, M.: A simple algorithm for data compression in wireless sensor networks. IEEE Communications letters 12(6), 411–413 (2008)
40. Marin-Perianu, R., Marin-Perianu, M., Havinga, P., Scholten, H.: Movement-based group awareness with wireless sensor networks. In: Pervasive, Toronto, Canada, pp. 298–315 (2007)
41. McConnell, S., Skillicorn, D.: A distributed approach for prediction in sensor networks. In: 1st International workshop on Data Mining in Sensor Networks, Newport Beach, USA, pp. 28–37 (2005)
42. Mini, R., Nath, B., Loureiro, A.: A probabilistic approach to predict the energy consumption in wireless sensor networks. In: IV Workshop de Comunicacao sem Fio e Computacao Movel, Sao Paulo, Brazil, pp. 23–25 (2002)
43. Nath, S., Gibbons, P., Seshan, S., Anderson, Z.: Synopsis diffusion for robust aggregation in sensor networks. In: ACM Conference on Embedded Networked Sensor Systems, Baltimore, USA, pp. 250–262 (2004)
44. Passos, R., Nacif, J., Mini, R., Loureiro, A., Fernandes, A., Coelho, C.: System-level dynamic power management techniques for communication intensive devices. In: International Conference on Very Large Scale integration, pp. 373–378. Nice, French Riviera (2006)

45. Pattem, S., Krishnamachari, B., Govindan, R.: The impact of spatial correlation on routing with compression in wireless sensor networks. ACM Transactions on Sensor Networks 4(4), 24–33 (2008)
46. Perillo, M., Ignjatovic, Z., Heinzelman, W.: An energy conservation method for wireless sensor networks employing a blue noise spatial sampling. In: International Symposium on Information Processing in Sensor Networks, California, USA, pp. 116–123 (2004)
47. Petrovic, D., Shah, R., Ramchandran, K., Rabaey, J.: Data funneling: routing with aggregation and compression for wireless sensor networks. In: 1st IEEE International Workshop on Sensor Network Protocols and Applications, California, USA, pp. 156–162 (2003)
48. Pottie, G., Kaiser, W.: Wireless integrated network sensors. Communications of the ACM 43(5), 51–58 (2000)
49. Puri, A., Coleri, S., Varaiya, P.: Power efficient system for sensor networks. In: 8th IEEE International Symposium on Computers and Communication, Kemer, Antalya, Turkey, vol. 2, pp. 837–842 (2003)
50. Radivojac, P., Korad, U., Sivalingam, K., Obradovic, Z.: Learning from class-imbalanced data in wireless sensor networks. In: 58th Vehicular Technology Conference, Florida, USA, vol. 5, pp. 3030–3034 (2003)
51. Sadler, C., Martonosi, M.: Data compression algorithms for energy-constrained devices in delay tolerant networks. In: ACM Conference on Embedded Networked Sensor Systems, Colorado, USA, pp. 265–278 (2006)
52. Seward, J.: bzip2 compression algorithm (2008), `http://www.bzip.org/index.html`
53. Sharaf, M., Beaver, J., Labrinidis, A., Chrysanthis, P.: Tina: A scheme for temporal coherency-aware in-network aggregation. In: ACM Workshop on Data Engineering for Wireless and Mobile Access, California, USA, pp. 69–76 (2003)
54. Shrivastava, N., Buragohain, C., Agrawal, D., Suri, S.: Medians and beyond: New aggregation techniques for sensor networks. In: ACM Conference on Embedded Networked Sensor Systems, Baltimore, USA, pp. 239–249 (2004)
55. Tian, D., Georganas, N.: A node scheduling scheme for large wireless sensor networks. Wireless Communications and Mobile Computing Journal 3(2), 271–290 (2003)
56. Tulone, D., Madden, S.: Paq: Time series forecasting for approximate query answering in sensor networks. In: 3rd European Conference on Wireless Sensor Networks, Zurich, Switzerland, pp. 21–37 (2006)
57. Vuran, M., Akyildiz, I.: Spatial correlation-based collaborative medium access control in wireless sensor networks. IEEE/ACM Transactions on Networking 14(2), 316–329 (2006)
58. Welch, T.: A technique for high-performance data compression. IEEE Computer 17(6), 8–19 (1984)
59. Willett, R., Martin, A., Nowak, R.: Backcasting: Adaptive sampling for sensor networks. In: International Symposium on Information Processing in Sensor Networks, California, USA, pp. 124–133 (2004)
60. Ye, F., Zhong, G., Cheng, J., Lu, S., Zhang, L.: Peas: A robust energy conserving protocol for long-lived sensor networks. In: 23rd International Conference on Distributed Computing Systems, Providence, Rhode Island, pp. 28–37 (2003)
61. Ye, M., Li, C., Chen, G., Wu, J.: Eecs: An energy efficient clustering scheme in wireless sensor networks. In: 24th IEEE International Performance Computing and Communications Conference, Arizona, USA, pp. 535–540 (2005)

62. Yoon, S., Shahabi, C.: The clustered aggregation (cag) technique leveraging spatial and temporal correlations in wireless sensor networks. ACM Transactions on Sensor Networks 3(1), 1–39 (2007)
63. Younis, O., Fahmy, S.: Heed: A hybrid, energy-efficient, distributed clustering approach for ad-hoc sensor networks. IEEE Transactions on Mobile Computing 3(4), 366–379 (2004)
64. Zhao, J., Govindan, R., Estrin, D.: Computing aggregates for monitoring wireless sensor networks. In: 1st IEEE International Workshop on Sensor Network Protocols and Applications, California, USA, pp. 139–148 (2003)