Peer-to-Peer Networking and Applications manuscript No.
(will be inserted by the editor)

An Energy-Efficient Data Prediction and Processing Approach for the Internet of Things and Sensing based Applications

Hassan Harb* · **Chady Abou Jaoude** ·
Abdallah Makhoul

* ***Corresponding author***

Received: Jan 16, 2019 / Accepted: Oct 09, 2019

Abstract The Internet of Things (IoT) is a vision in which billions of smart objects are linked together. In the IoT, “things” are expected to become active and enabled to interact and communicate among themselves and with the environment by exchanging data and information sensed about the environment. In this future interconnected world, multiple sensors join the internet dynamically and use it to exchange information all over the world in semantically interoperable ways. Therefore, huge amounts of data are generated and transmitted over the network. Thus, these applications require massive storage, huge computation power to enable real-time processing, and high-speed network. In this paper, we propose a data prediction and processing approach aiming to reduce the size of data collected and transmitted over the network while guaranteeing data integrity. This approach is dedicated to devices/sensors with low energy and computing resources. Our proposed technique is composed of two stages: on-node prediction model and in-network aggregation algorithm. The first stage uses the Lagrange interpolation polynomial model to reduce the amount of data generated by sensor nodes while, the second stage uses a statistical test, i.e. Kolmogorov-Smirnov, and aims to reduce the redundancy between data generated by neighbouring nodes. Simulation on real sensed data reveals that the proposed approach significantly reduces the amount of data generated and transmitted over the network thus, conserving sensors' energies and extending the network lifetime.

Keywords IoT · WSN · Data prediction · Lagrange interpolation · Kolmogorov-Smirnov test · In-network computing · Real sensor data

H. HARB and C. ABOU JAOUDE
TICKET Lab, Faculty of Engineering, Antonine University, Baabda, Lebanon
E-mail: hassan.harb@ua.edu.lb E-mail: chady.aboujaoude@ua.edu.lb

A. MAKHOUL
FEMTO-ST Institute/CNRS, the DISC department, Univ. Bourgogne Franche-Comté, Belfort, France
E-mail: abdallah.makhoul@univ-fcomte.fr

1 Introduction

In our modern era, the number of smart connected devices (Smartphones, smartwatches, smart glasses, smart TVs, smart cars, etc.) is increasing significantly; all of which collect and share data in an IoT based architecture [43]. Thus, the accumulated volume of data has reached, in many applications, the order of petabyte and, sometimes, the zettabytes. This amount is expected, by the data scientists, to be doubled every two years. Hence, new techniques and mechanisms need to be proposed in order to analyze and discover meaningful knowledge in IoT and sensing based applications.

As mentioned earlier, data acquisition is the first step in the life cycle of data. Hence, sensors and wireless sensor networks (WSNs) are used in IoT applications to collect and process data before sending them to the end-user. Generally, WSN consists of a large number of sensor nodes to monitor a phenomenon or a physical condition in environmental or industrial process. The primary practice of WSN is in the area where humans are not able to fetch the data and can get support from sensor nodes. Consequently, the efficiency of a WSN totally relies on the minimum dis-chargeable battery in the nodes. However, sensing the massive volume of data with transmission operation cost consumes a significant amount of the sensor energy. Hence, on-node and in-network data mechanisms are becoming mandatory in IoT and sensing based applications to consume the energy in an intelligent way so that the network can run for a long period.

In this paper, our efforts are devoted to present a novel big data prediction and statistic mechanism encompasses two main stages to maintain the network and minimize the energy consumption in sensor nodes or things. The first stage is a data prediction scheme for the IoT network. It is applied at the nodes/things themselves. At this stage, we consider that sensors are collecting and broadcasting a huge volume of data packets periodically to a specific intermediate node called aggregator. In order to save energy and reduce the amount of the transmitted data, we propose that each node sends the coefficients of Lagrange interpolation polynomial, instead of sending the whole collected data. Thus, the sink will be able to recover data based on received Lagrange coefficients. At the second stage, the aggregator uses a statistical-based data model to search neighbouring nodes that periodically generate similar data then to eliminate redundancy and reduce data transmission. Thus, it will provide a great reduction in power consumption in the aggregator node. The proposed model is mainly based on the Kolmogorov-Smirnov test. Furthermore, to show the efficiency of our proposal we conducted several simulations on real collected data. The obtained results show clearly that our approach saves nodes energies and reduces the size of the collected and transmitted data while preserving data integrity.

The rest of the paper is organized as follows. In Section 2, we briefly present literature on on-node and in-network approaches. Section 3 defines terminologies and network design that are related to the paper. Section 4 details the data prediction model proposed at the sensor node level while in Section 5, a data statistical model based on Kolmogorov-Smirnov test is developed and detailed at the CH node. A set of simulations are conducted to evaluate the proposed mechanism in Section 6. Section 7 concludes this paper.

2 Related Work

WSNs are self-awareness networks and they face many benefactions like dense deployment of nodes, sever energy, computation power and massive data collection. Hence, on-node and in-network data processing are beneficial to big data applications in WSN [34]. Researchers on [1–4] have presented a review article to identify the importance of such approaches as well as to briefly describe and compare most of the proposed techniques existing in the literature.

The on-node data approach is used at each sensor node itself and aims to accumulate data in an energy efficient manner by reducing data redundancy, before sending data to the parent node [5–13, 32, 33]. In [5], the authors propose a data prediction model applied at both sensors and sink. The prediction model is based on line equation trough two n-dimensional vectors and aims to predict the future readings of the sensors based on the previous one. Another data prediction model has been proposed in [6]. The authors use the concept of time series analysis in order to analyze the variations in sensed data so as it can be interpreted based on an autoregressive model of order p. The authors in [7] propose an Adams-Bashforth-Moulton algorithm aims to optimize the accuracy of prediction obtained with Milne Simpson scheme proposed in [8]. Both algorithms are simulated on real data sensor and an optimization level of energy and accuracy is noticed. In [9], the author propose a control scheme based on data compression and sensing rate in order to reduce the amount of data collected at the sink node. The idea behind this scheme is that every parent node sends a threshold, called data quota, to all its node children. According to the received quota and its remaining energy, the children node selects its suitable compression method and its sensing rate during the next period. In [10], a coding provenance scheme (CBP) has been proposed. Compared to traditional compression techniques, CBP ensures a high provenance compression rate as well as it encodes and decodes incrementally the compression ratio at the base station depending on the condition observed. Finally, the authors in [13] propose an efficient and robust compression method named Sequential Lossless Entropy Compression (S-LEC). S-LEC uses a differential predictor that arranges the alphabet of integer residues into a number of groups. For each group, two codes are assigned: entropy and binary codes. The first code specifies the group where the second one represents the index inside the group. In [45–47], the authors tackle a new area of IoT by integrating with cloud computing and big data technologies. First, they propose an efficient algorithm for advanced scalable media-based smart big data, i.e. 3D Ultra HD, on intelligent cloud computing systems [46]. 3D Ultra HD is based on encoding methods and can outperform the traditional HEVC standard. Then, an architecture relying on the security of the network has been proposed to improve the privacy of data transmitted between IoT and cloud [47]. The idea is to install a security wall between the cloud server and the different users on the Internet.

The in-network data approach is used at an intermediate node, mostly called aggregator or Cluster-Head (CH), and aims to a find correlation between neighboring nodes so as to transfer valuable data to the sink [14–23, 35]. The authors in [14] propose a structure fidelity data collection (SFDC) technique dedicated for cluster-based periodic applications in WSNs. SFDC searches both spatial and temporal correlation between nodes, using distance functions and similarity metrics respectively. Another spatiotemporal node correlation, based on the Pearson Product-Moment Coefficient (PPMC) metric, has been proposed in [15]. PPMC aims to conserve sensor energies by switching high correlated nodes into sleep mode. Compared to other existing methods, PPMC has been evaluated based on experiments on real sensors. In [16], the authors propose a polynomial regression-based data aggregation protocol that conserves network energies as well as the privacy of sensed data. Instead of sending its raw data, each sensor uses coefficient regression polynomials to represent their data while the aggregation is made on such secret coefficients. In [17], a two-level node mechanism has been proposed which is dedicated to periodic sensor applications. First, the authors propose an on-node aggregation method to remove redundant data collected by the sensor. Then, an in-network data reduction called prefix frequency filtering (PFF) is introduced at the CH level. PFF allows CHs to find similarities between data collected by neighboring nodes in the same cluster, using Jaccard similarity function. The authors in [18] propose a Semi Distributed Heuristic Energy efficient Aggregation Tree (SDHEAT) algorithm for WSN. Mainly, SDHEAT is based on three concepts: heuristic tree formation, sensing priority and distributed nature and aims to reduce the overall network consumption while conserving information integrity. Finally, the authors in [19] propose an energy-efficient communication method that is dedicated to periodic underwater sensor applications. on the basis of the proposed technique, each node cleans its collected data before transmitting to the appropriate CH. When receiving datasets, the CH applies Kmeans algorithm adopted to the ANOVA model with statistical tests in order to eliminate inter-node correlation.

3 Network Design and Preliminaries

3.1 Network Design

In this paper, we consider the scenario where several sensor nodes are grouped together to form a WSN network in view of sensing data and ahead of this data to the sink [24, 25]. In order to reduce the energy consumption, an important approach seeks to divide the whole network into subareas termed as clusters within which a head or a leader is elected. This particular node is called a cluster head (CH) and has a mission to gather data from its subordinate cluster members and sends them to the sink node. Obviously, the selection of CHs is itself a trending research challenge that considers several factors like node, battery life of node, and distance to/from the sink. Fortunately, the significance of selecting cluster-based network design is to ensure network scalability, reduce the distance between source nodes and sink, perform data fusion and aggregation, and consumes less network energy.

Fig. 1 presents the cluster-based network design that we consider in our study. Our proposed mechanism consists of two stages: on-node and in-network. The on-node operation is performed by the sensor nodes themselves while the in-network is applied at the CH level. After data being periodically collected, the sensor nodes use a data prediction model in order to reduce the data size sent to its corresponding CH. Upon receiving the prediction model for all its sensors, the CH uses a statistical model in order to prevent sending similar prediction models generating by neighbouring to the sink node. Finally, the sink receives the subset of prediction models and try to recover data of the whole sensors.

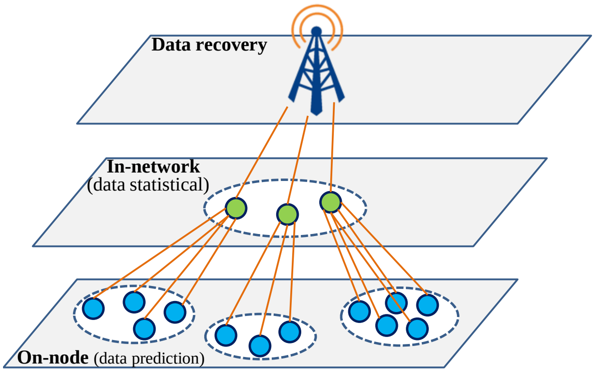


Fig. 1 A cluster-based network architecture

Indeed, dividing the network into clusters is not an easy task and it faces many challenges. Hence, one can find a lot of works in the literature that are interested in issues related to cluster network like the selection of cluster heads [36–38], optimization of cluster size [39, 40], communication between sensors/CHs and CHs/sink [41, 42], etc. However, our concern in this paper is to study the variation of data collected by the sensors and not the formation of clusters themselves. Therefore, we consider a geographical clustering scheme in which near sensors are already assigned to the same cluster.

3.2 Problem Description and Notations

WSN is represented as a connected graph G =(N,E), where $N = \{N_1, N_2, \ldots, N_n\}$ is a set of n (sensor) nodes and E is a set of edges. A sensor node collects data over a period of time (P_k) and subsequently transmits all the sensed data to the next level of hierarchy (e.g. CH). Sensor networks supporting this kind of applications are known as periodic wireless sensor networks (PWSNs). Each period P_k of a sensor node N_i is divided into a finite number of time slots as follows: $P_k = [s_1, s_2, \ldots, s_F]$. At each slot s_j, each sensor node N_i captures a new data value v_{ik_j}, and eventually forms a vector of sensed data during the period P_k as follows: $V_{ik} = [v_{ik_1}, v_{ik_2}, \ldots, v_{ik_F}]$. Fig. 2 depicts the data collection scenario of a node N_i which takes four measures ($= F$) during each period P_k ($k \in [1, 3]$) and transmits the resulted vector of data $V_{ik} = [v_{ik_1}, v_{ik_2}, v_{ik_3}, v_{ik_4}]$ to the next hierarchical node. The data collected by sensor nodes are highly dependent on the rate of change of the environmental condition or the sensed phenomena. The sensed data is more correlated and redundant when monitored condition changes slowly or short slot is taken. A data vector V_{ik} created by the sensed data of node N_i may contain redundant data (or similar data), especially when the monitored condition varies slowly or when the frequency of sensing is high or the time slots are short.

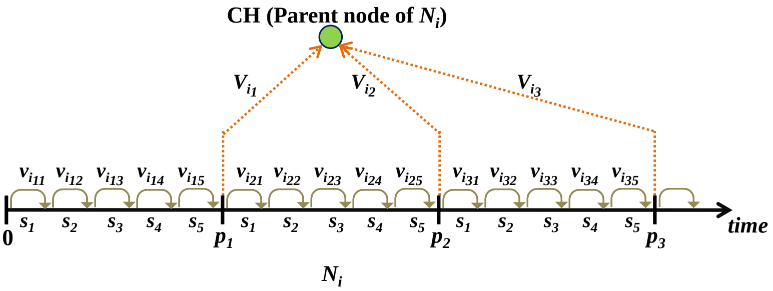

CH (Parent node of N_i)
V_{i1}
V_{i2}
V_{i3}
v_{i11} v_{i12} v_{i13} v_{i14} v_{i15}
v_{i21} v_{i22} v_{i23} v_{i24} v_{i25}
v_{i31} v_{i32} v_{i33} v_{i34} v_{i35}
s_1 s_2 s_3 s_4 s_5
0
p_1
p_2
p_3
time
N_i

Fig. 2 Data collection at node N_i in PWSN.

4 On-Node Data Prediction

In IoT and WSN applications, sensors are able to collect all kinds of data like ecological conditions (relative humidity, pressure, temperature, sound pollution, etc.), motion monitoring (traffic, animals, enemy), and disaster phenomenon (tsunami, volcano, forest fire, etc.) [44]. However, the issue of the huge bulk of datasets produced by these sensors forms a very serious challenge. Furthermore, data transmission in WSN consumes a significant amount of sensor energy and occupies a large volume of its memory. In this section, we present the first stage, e.g. on-node , of our mechanism which is applied at each sensor node. The on-node stage aims to prevent sending similar data points sensed by each sensor at each period P_k, based on a prediction model that uses the Lagrange interpolation polynomial.

4.1 Lagrange Interpolation Polynomial Model

In numerical analysis, Lagrange polynomials are used for polynomial interpolation. For a given set of points (x_j, y_j), the Lagrange polynomial is the polynomial of lowest degree that assumes at each value x_j the corresponding value y_j.

Definition 1 Given a set of k data points, $(x_1, y_1), \ldots, (x_j, y_j), \ldots, (x_k, y_k)$, the interpolation polynomial in the Lagrange form is a linear combination of the form:

$$L(x) = \sum_{j=1}^{k} y_j l_j(x) \tag{1}$$

$$\text{where } l_j(x) = \prod_{1 \le m \le k} \frac{x - x_m}{x_j - x_m} = \frac{(x - x_0)}{(x_j - x_0)} \times \ldots$$

$$\times \frac{(x - x_{j-1})}{(x_j - x_{j-1})} \times \frac{(x - x_{j+1})}{(x_j - x_{j+1})} \times \cdots \times \frac{(x - x_k)}{(x_j - x_k)}, \tag{2}$$

4.2 Illustrative Example

Given the three points $A(-2, -5), B(-1, -1)$ and $C(1, 1)$. The interpolating polynomial of Lagrange degree $d = 2$ can be calculated as follows:

$$L(x) = (-5) \times \frac{x+1}{-2+1} \times \frac{x-1}{-2-1} + (-1) \times \frac{x+2}{-1+2} \times \frac{x-1}{-1-1}$$

$$+ (1) \times \frac{x+2}{1+2} \times \frac{x+1}{1+1} = -x^2 + x + 1. \tag{3}$$

with the coefficients $a_2 = -1, a_1 = 1, a_0 = 1$.

4.3 Data Prediction Model

As mentioned before, every data sets collected by each sensor at the end of each period, e.g. V_{ik}, can contain redundant values. In order to reduce the size of data sent to the CH, we propose to find the Lagrange interpolating polynomial of the data vector V_{ik} and send only the coefficients of the Lagrange polynomial instead of the whole vector V_{ik}. As a result, the data transmission between sensor/CH will be reduced while the sink can, at any time, retrieve/recover all the collected data based on the coefficients of the received equation.

Indeed, we need $d + 1$ points in order to calculate a Lagrange polynomial of degree d. For instance, in the previous subsection, Lagrange polynomial of degree 2 was needing 3 data points to find the Lagrange coefficients. However, the period size in our case contains F readings where F is much greater than Lagrange degree, i.e. $F >> d$. Hence, in order to overcome this problem, we propose to select d readings among V_{ik} in order to calculate the set of Lagrange coefficients of each sensor N_i, e.g. L_{ik}, based on the following equation:

$$L_{ik} = \{v_{1+j\times\lfloor F/d\rfloor}, v_F\} \tag{4}$$

where $d_{1+j\times\lfloor F/d\rfloor}$ are all readings collected at slot numbers $s_{1+j\times\lfloor F/d\rfloor}$ (such that $j \in [0, F]$ and $1 + j \times \lfloor F/d\rfloor < F$) and d_F is the last reading in V_{ik}.

Finally, each sensor node will calculate its set of Lagrange coefficients, e.g. $C_{ik} = \{a_d, a_{d-1}, \ldots, a_0\}$, that will send toward the CH at the end of each period. Obviously, C_{ik} contains the coefficients of equation computed based on the Lagrange interpolation model mentioned in equation 1.

Fig. 3 shows an illustrative example of data prediction model proposed at sensor node level. We consider a period of 10 slots ($F = 10$) where a data vector V_i is formed at the period p_1. Suppose we want to find a Lagrange polynomial of degree 4, so the readings in L_i can be selected as follows: $\{v_{1+0\times\lfloor 10/4\rfloor} = v_1, v_{1+1\times\lfloor 10/4\rfloor} = v_3, v_{1+2\times\lfloor 10/4\rfloor} = v_5, v_{1+3\times\lfloor 10/4\rfloor} = v_7, v_{i_{10}}\}$, where $v_{i_{10}}$ is added because it represents the last reading in V_i. Then, we find the Lagrange coefficient set C_i by applying the Lagrange interpolation model as shown in equation 1.

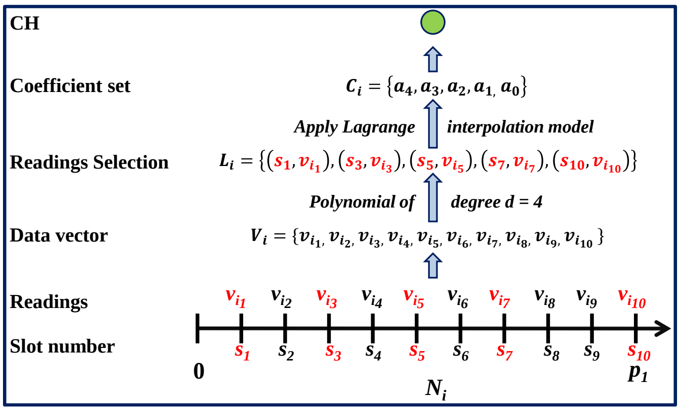

CH
Coefficient set
$C_i = \{a_4, a_3, a_2, a_1, a_0\}$
Apply Lagrange
interpolation model
Readings Selection
$L_i = \{(s_1, v_{i_1}), (s_3, v_{i_3}), (s_5, v_{i_5}), (s_7, v_{i_7}), (s_{10}, v_{i_{10}})\}$
Polynomial of
degree d = 4
Data vector
$V_i = \{v_{i_1}, v_{i_2}, v_{i_3}, v_{i_4}, v_{i_5}, v_{i_6}, v_{i_7}, v_{i_8}, v_{i_9}, v_{i_{10}}\}$
Readings
v_{i1} v_{i2} v_{i3} v_{i4} v_{i5} v_{i6} v_{i7} v_{i8} v_{i9} v_{i10}
Slot number
s_1 s_2 s_3 s_4 s_5 s_6 s_7 s_8 s_9 s_{10}
0
p_1
N_i

Fig. 3 Illustration of data prediction model at the sensor node.

In order to be more formally, Algorithm 1 describes the on-node data model that is applied at each sensor node. As an input, the algorithm takes the period size F and the desired Lagrange degree d and it returns the Lagrange coefficient set that will send to its CH (lines 22-23). First, the sensor collects the vector of readings during a period (lines 1-5) then, it selects readings at indexes determined by equation 4 (lines 6-10). Finally, the sensor calculates its Lagrange coefficient set that to be sent to its CH based on the Lagrange model (lines 11-21).

Algorithm 1 On-Node Stage Algorithm.

```
Require: Node: $N_i$, period size: $F$, Lagrange degree: $d$.
Ensure: Coefficient set: $C_i$.
 1: $V_i \leftarrow \emptyset$
 2: for $k = 1$ to $F$ do
 3:     take reading value $v_k$
 4:     $V_i \leftarrow V_i \cup \{v_k\}$
 5: end for
 6: $L_i \leftarrow \emptyset$
 7: for $k = 1$ to $F/d$ do
 8:     $L_i \leftarrow L_i \cup \{v_{1+k \times \lfloor F/d \rfloor}\}$
 9: end for
10: $L_i \leftarrow L_i \cup \{v_F\}$
11: $y = 0$
12: for each $l_k \in L_i$ do
13:     $prodfunc = 1$
14:     for each $l_t \in L_i$ do
15:         if $t \neq k$ then
16:             $prodfunc = prodfunc \times (x - x_t)/(x_k - x_t)$
17:         end if
18:     end for
19:     $y = y + s_t \times prodfunc$
20: end for
21: simplify $y$
22: $C_i \leftarrow \{y_{a_d}, \ldots, y_{a_0}\}$
23: return $C_i$
```

5 In-Network Data Reduction

In this stage, our objective is to allow CH to reduce the total number of data set coefficients sent to the sink node. Indeed, statistical analysis approach is an important primitive that aims to search correlation between data sets so that overall communication bandwidth and energy consumption of CH is reduced. The motivation behind applying the statistical approach is that data of multiple sensor nodes can be summarized by the CH so that only the useful information is sent to the sink.

5.1 Kolmogorov-Smirnov Test

In our work, we are interested in the Kolmogorov-Smirnov (K-S) test which is one of the most tests used in statistical analysis. Recently, K-S has been used to detect changes in stationarity in big data [26], to check if the phylogeny molecular is clock-like [27], to verify homogeneity of data measured at certain high energy physics [28] and so on.

Generally, the K-S test is used to quantify the distance between empirical distribution functions of two data samples. The null distribution of this statistic is calculated under the null hypothesis that the two data samples are drawn from the same distribution. First, it calculates the cumulative distribution functions of the two samples, then, it computes the maximal deviation between them. If the maximal deviation exceeds a given threshold, then the test fails and the distribution empirical function of the two data samples are not theoretically drawn from the same distribution function.

Let consider two empirical distribution functions DF_i and DF_j of the same size F. DF_i (respectively DF_j) is in the form $\{(s_i, v_i)$ such that $i \in [1, F]\}$. Then, assume that the corresponding cumulative distribution functions of DF_i and DF_j are DF_{ci} and DF_{cj} respectively. The Kolmogorov-Smirnov statistic test between DF_i and DF_j is given as follows:

$$\mathcal{D}_{i,j}(DF_i, DF_j) = \sup_x |DF_{ci}(x) - DF_{cj}(x)|, \tag{5}$$

where sup is the supremum function between DF_{ci} and DF_{cj}.

The null hypothesis is rejected at level α if:

$$\mathcal{D}_{i,j}(DF_i, DF_j) > \mathcal{C}(\alpha)\sqrt{\frac{2 \times F}{F^2}}. \tag{6}$$

Where the value of $\mathcal{C}(\alpha)$ is given in the table below for the most common levels of α:

α	0.10	0.05	0.025	0.01	0.005	0.001
$\mathcal{C}$	1.22	1.36	1.48	1.63	1.73	1.95

5.2 Illustrative Example

Table 1 shows an illustrative example for the computation of the difference between two empirical distribution functions DF_1 and DF_2 based on the K-S test. The two functions have the same size of 25. First, K-S computes the cumulative distribution functions of DF_1 and DF_2 represented by DF_{c1} and DF_{c_2} respectively. Subsequently, at the first slot, the value of $DF_{c1}(1)$ is the same as that of $DF_1(1)$ whilst, for the next slots, the value of $DF_{c1}(j)$ can be calculated as the sum of the value taken at slot j in addition to all values taken at previous slots ($< j$) and so on. For instance, at slot #1, $DF_{c1}(1) = DF_1(1) = 0.09$ whilst, in the slot #2, $DF_{c1}(2) = DF_1(1) + DF_1(2) = 0.06 +$

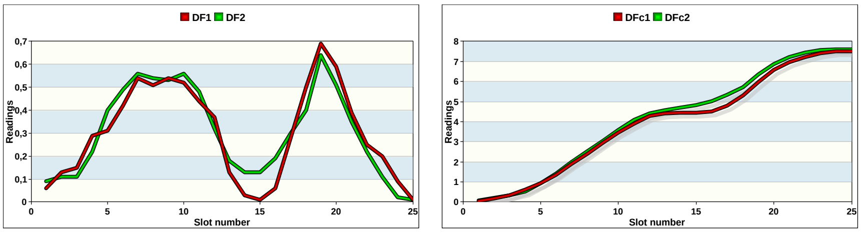
$0.13 = 0.19$. After calculating the cumulative values of DF_1 and DF_2, K-S finds the difference, e.g. absolute value, for every pair of values of DF_{c1} and DF_{c2} at the same slot. For instance, at slot #1, the difference equals to $|DF_{c1}(1) - DF_{c2}(1)| = |0.06 - 0.09| = 0.03$ and so on. Finally, K-S finds the maximum value between all differences as the dissimilarity between DF_1 and DF_2, e.g. $\sup_x |DF_{c1}(x) - DF_{c2}(x)| = 0.52998$.

Table 1 Difference computation between two distribution function based on K-S test.

Slot #	DF_1	DF_2	DF_{c1}	DF_{c2}	Difference
1	0.06	0.09	0.06	0.09	0.03
2	0.13	0.11	0.19	0.21	0.01
3	0.15	0.11	0.34	0.32	0.02
4	0.29	0.22	0.63	0.54	0.09
5	0.31	0.4	0.94	0.94	0.0
6	0.42	0.49	1.36	1.42	0.07
7	0.54	0.56	1.9	1.98	0.08
8	0.51	0.54	2.41	2.52	0.11
9	0.54	0.53	2.95	3.04	0.09
10	0.52	0.56	3.47	3.6	0.13
10	0.44	0.48	3.91	4.08	0.17
12	0.37	0.32	4.28	4.4	0.12
13	0.13	0.18	4.41	4.58	0.17
14	0.03	0.13	4.44	4.71	0.27
15	0.01	0.13	4.46	4.84	0.38
16	0.06	0.19	4.51	5.03	0.51
17	0.28	0.3	4.8	5.33	0.53
18	0.5	0.4	5.3	5.73	0.43
19	0.69	0.64	5.99	6.37	0.38
20	0.59	0.51	6.58	6.88	0.3
21	0.39	0.35	6.97	7.23	0.26
22	0.25	0.22	7.22	7.45	0.23
23	0.2	0.11	7.42	7.56	0.15
24	0.09	0.02	7.51	7.59	0.08
25	0.01	0.01	7.52	7.59	0.08
				Dissimilarity=	0.52998

Now, assume that $\alpha = 0.01$, then $\mathcal{C}(\alpha)\sqrt{\frac{2\times F}{F^2}} = 1.63 \times \sqrt{\frac{2\times 25}{25^2}} = 0.46103$. Thus, $\mathcal{D}(DF_1, DF_2) = 0.52998 > 0.46103$ then the null hypothesis is rejected and DF_1 and DF_2 are not drawn from the same distribution function.

Fig. 4 presents the curves of both empirical distribution functions DF_1 and DF_2 (Fig. 4(a)) and those of their corresponding cumulative functions DF_{c1} and DF_{c2} (Fig. 4(b)), according to slot number in x-axis and reading values in y-axis. We can observe that the curves of both empirical functions are not closer to each other (see Fig. 4(a)) due to the difference between their reading values. As a result, their cumulative functions notice a notable difference between the curves, especially at slot #15. This confirms the behaviour of the K-S test that considers both functions as dissimilarly distributed.



(a) empirical

(b) cumulative

Fig. 4 Curves of empirical and cumulative distribution functions.

5.3 Data Reduction Algorithm Based on K-S Test

In this section, we present the in-network stage based on K-S test which is applied at each CH node (Algorithm 2). After receiving list of coefficient sets sent from sensors at each period, the CH reconstructs first the empirical distribution function for each sensor, then it calculates its corresponding cumulative function (lines 4-8). After that, for every pair of empirical functions where the maximal distance between their corresponding cumulative functions is less than the permitted threshold, the CH adds this pair to the list of a redundant sets (lines 10-11). Finally, for each pair of redundant set, the CH adds one of them to the final list of coefficient sets sent to the sink while removing removing all pairs of redundant sets that contain either one of them from the set of pairs (which means it will not check them again) (lines 16-20). Subsequently, in order to save the information integrity, the CH assigns to each set its weight (line 21) when sending it to the sink.

Algorithm 2 In-Network Stage Algorithm.

```
Require: Number of slots per period: $F$, Set of coefficient sets: $C = \{C_1, C_2, \ldots, C_n\}$, Rejected level: $\alpha$.
Ensure: List of sent coefficient sets among $C$: $\mathcal{S}$.
1:  $S \leftarrow \emptyset$ // list of pairwise similar distributed functions
2:  $E = \mathcal{C}(\alpha) \times \sqrt{\frac{2\times F}{F^2}}$
3:  for each set $C_i \in C$ do
4:      compute $DF_i$ of $C_i$
5:      compute $DF_{ci}$ of $DF_i$
6:      for each set $C_j \in C$ such that $j > i$ do
7:          compute $DF_j$ of $C_j$
8:          compute $DF_{cj}$ of $DF_j$
9:          calculate $\mathcal{D}_{i,j}(DF_i, DF_j)$ based on Eq. 6
10:         if $\mathcal{D}_{i,j} \leq E$ then
11:             $S \leftarrow S \cup \{(C_i, C_j)\}$
12:         end if
13:     end for
```

```
end for
$\mathcal{S} \leftarrow \emptyset$
for each pair of sets $(C_i, C_j) \in S$ do
    $\mathcal{S} \leftarrow \mathcal{S} \cup \{C_i\}$
    Remove all pairs of sets containing one of the two sets $C_i$ and $C_j$
    $wgt(C_i)$ = number of removed pairs + 1
end for
return $\mathcal{S}$
```

6 Experimental Results

This section conducts a set of simulations to investigate the effectiveness of the proposed technique in respecting to several parameters. To address this issue, the scalar dataset were picked up from sensors that they were deployed in the Intel Berkeley Research lab [29]. Table II shows the information about the deployed network. In the simulation, the results are obtained after implementing our technique in Java-based simulator. Furthermore, our technique has been evaluated in comparison with two recent techniques: S-LEC [13] and PFF [17]. S-LEC and PFF are two data reduction techniques where the first one is used to reduce the data transmission at sensor level while the second one is used to reduce that at CH level.

Table 2 Simulation parameters and their values.

Parameter	Value
Year	2004
duration	February 28th to April 5th
Dimension of area	42 × 33 meters
Number of sensors	46
Observed conditions	temperature, humidity, light
Collected readings	2.3 million
Slot interval	31 seconds

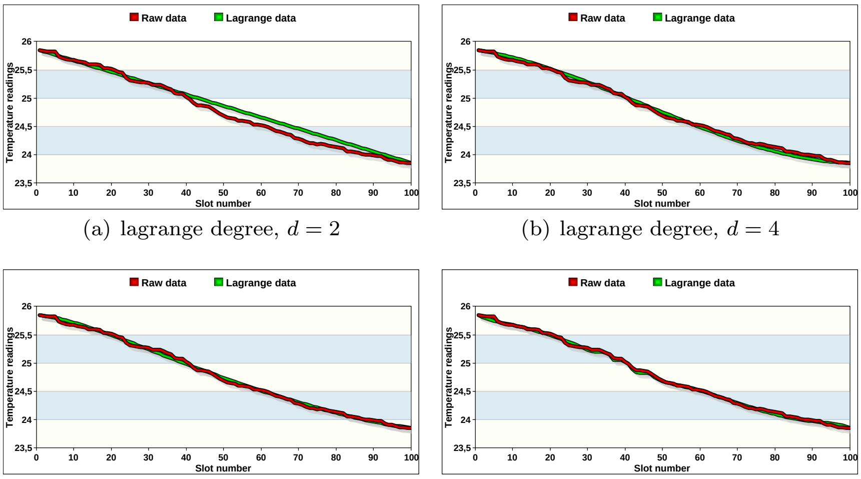
6.1 Data Prediction at the Sensor Nodes Level

In this section, we aim to study and analyze the performance of our mechanism at the sensor node level according to the following metrics:

6.1.1 Effect of Lagrange Degree Variation

In the on-node stage, each sensor will find the polynomial model of the data collected at each period. Thus, the performance of this stage is highly dependent on Lagrange degree d which we choose to vary, in our simulation, from

2 to 8. Indeed, selecting an appropriate value of d is subject to the experts which can decide based on two factors: application requirements and capacity of sensor resources; Greater value of d is selected makes the computation more complex and leads to reduce the integrity of information when recovering data at the sink node. Fig. 5 shows the effect of varying the Lagrange degree in comparison between raw data (data collected by sensors) and Lagrange data (data recovered at the sink). We fixed the observed condition to the temperature readings and the period size to 100 slots. As expected, when the Lagrange degree increases the curves of raw and Lagrange data becomes more closer. This is because, when d increases, more points are taken to find the Lagrange coefficients that increase the accuracy of the recovered data. In addition, we observe that the accuracy of the information and the convergence between raw data collected by the sensors and those recovered at the sink node fully occurs when d arrives at 6.



(a) lagrange degree, $d = 2$

(b) lagrange degree, $d = 4$

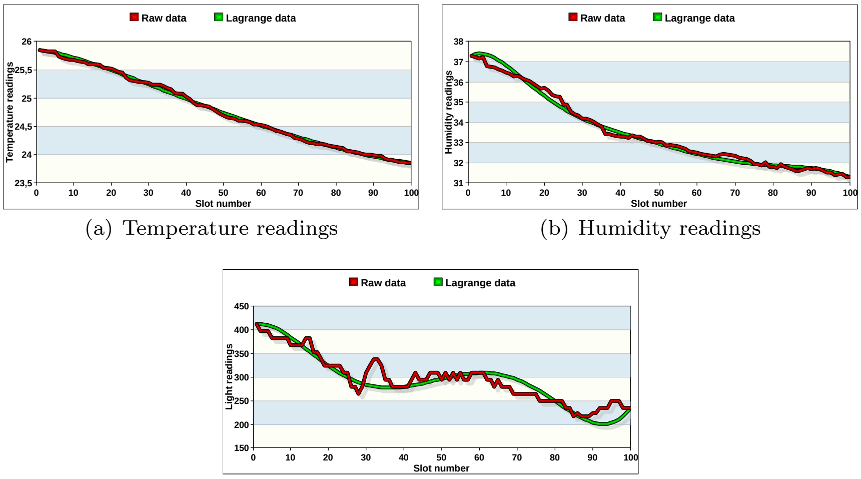
(c) lagrange degree, $d = 6$

(d) lagrange degree, $d = 8$

Fig. 5 Comparison between raw and Lagrange data, $period = 100\ slots$.

6.1.2 Effect of Condition Variation

Fig. 6 shows the efficiency of the on-node stage when varying the observed condition to temperature, humidity and light respectively. We fixed the Lagrange degree to 6 and the period size to 100 slots. The obtained results reveal two facts: first, the Lagrange data recovered at the sink node can differ from one condition to another even though the Lagrange degree is fixed. Second, the accuracy of the Lagrange data is dependent on the dynamic of the monitored condition as well as the Lagrange degree; the more the monitored condition varies slow, i.e. in temperature, Lagrange data will be much closer to the raw data thus, the accuracy is more conserved. Otherwise, i.e. condition speeds up, Lagrange and raw data will diverge then, the accuracy will decrease (i.e. light condition).



(a) Temperature readings

(b) Humidity readings

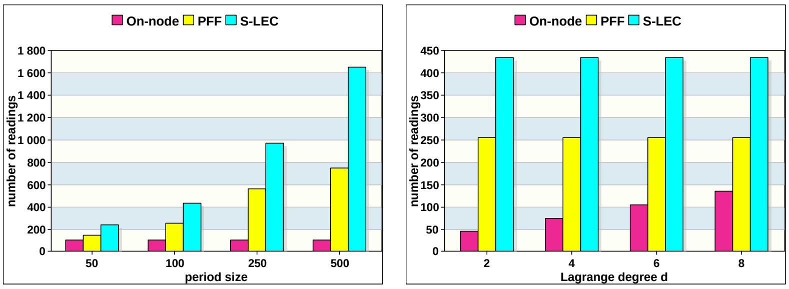
(c) Light readings

Fig. 6 Effect of Lagrange degree when varying condition, $period = 100\ slots$, $d = 6$.

6.1.3 Periodic Data Transmission Ratio

In this section, we show the average number of readings sent from each sensor node to the CH (Fig. 7). In Fig. 7(a), we fixed the Lagrange degree to 6 and we varied the period size from 50 to 500 slots while in Fig. 7(b) the period size is fixed to 100 slots and the degree is changed from 2 to 8. In addition, at the sensor level, we compared the results of on-node stage to the aggregation phase used in PFF technique [17] and a data reduction technique called S-LEC [13]. As shown, the on-node stage proposed in our technique reduces more data transmission compared to other techniques. Subsequently, it reduces up to 86% and 93% compared to PFF and S-LEC when fixing d (Fig. 7(a)), and up to 82% and 89% when fixing period size (Fig. 7(b)). This is because, the sensor node only sends, using on-node stage, the Lagrange coefficients to the CH while in the PFF and S-LEC, it uses aggregation and compression methods to send a portion of collected data instead of the whole raw data. Furthermore, we notice that the data transmission ratio at the sensor node in PFF and S-LEC is based on the similarity between the collected data; more data are similar then less data is sent. Otherwise, the transmission ratio using on-node stage is fixed and only dependent on the Lagrange degree. Thus, the transmission ratio at each sensor can be calculated as follows: $T_r = (d + 1) \times P$, where P is the total number of periods before the sensor loses its available energy.

Therefore, as energy consumption is highly related to data transmission ratio, on-node stage proposed in our technique can be considered as an energy-efficient method that allows the sensor node to reduce its energy consumption and extend its lifetime.



(a) lagrange degree, $d = 6$

(b) period size = 100 slots

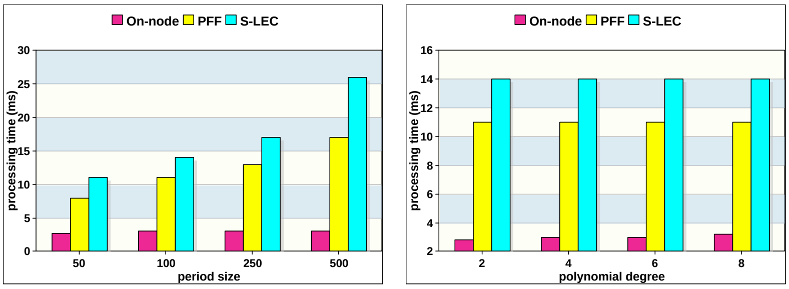
Fig. 7 Number of readings periodically sent to the CH.

6.1.4 Processing Time

In this section, we discuss the complexity of the on-node stage as well as that of PFF and S-LEC. Indeed, the complexity is an important metric that must be considered in WSN due, in one hand, to the limited sensor resources and, another hand, it can affect the data delivery delay to the sink node. In our technique, the complexity of on-node stage is dependent on equation 4 that is used to select a set of readings in order to find the Lagrange coefficients. Therefore, the complexity of on-node stage is a constant and it is computed as $O(d + 1)$, where d is the Lagrange degree. However, in PFF and S-LEC, each reading should be compared to all other collected readings in the same period in order to find its similar ones. Thus, The complexity can be considered as $O(F)$, where F is the total number of readings collected in each period. Furthermore, Fig. 8 shows the processing time needed to apply each of the three compared techniques. The processing time is a good indicator to the complexity of any technique. Similar to Fig. 7, we fixed the Lagrange degree in Fig. 8(a) and varied period size while in Fig. 8(b) we varied d and fixed period slots. The obtained results show that on-node stage accelerates time processing at the sensor from 3 to 5 times compared to PFF and from 4 to 8 compared to S-LEC. We can also observe that the processing time of on-node stage is almost fixed in all cases while that required for PFF and S-LEC increases with the increasing of period size.

6.2 CH Study: Results and Analysis

This section presents and analyzes the results obtained at the CH nodes while respecting to the following metrics:



(a) Lagrange degree, $d = 6$

(b) period size = 100 slots

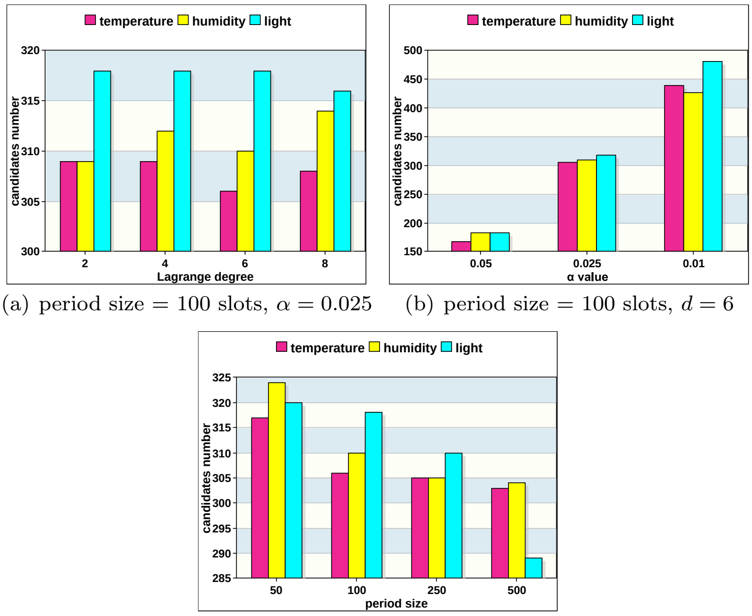
Fig. 8 Processing time of each sensor at the end of the simulation.

6.2.1 Periodic Number of Candidates

The in-network stage allows CH to find all similar coefficient set candidates for a later redundancy elimination before sending to the final destination (sink). Fig. 9 shows the average number of candidates found by the CH at each period. We studied the candidate number in terms of three variables: Lagrange degree (d), rejected level (α) and period size (F). As shown in the figure, we varied d from 2 to 8 and F from 50 to 500 slots while α takes the values 0.05, 0.025 and 0.01. The obtained results reflect a huge amount of redundancy among the neighbouring nodes where the similarity differs from condition to another. Subsequently, we can notice that the light condition varies slowly followed by humidity than temperature conditions. Hence, light condition supports more similarity between data generated between neighbouring sensor nodes than the other conditions.

Furthermore, the following observations can be noticed:

– by varying the Lagrange degree, the number of candidates still almost fix. This is because the similarity between data sets is dependent on the dynamic of the condition and not the degree of Lagrange.
– by decreasing the α value, the number of candidates increases. This is because when the rejection level decreases the similarity constraint becomes more flexible thus, more data sets will be similar. For instance, by decreasing α from 0.05 to 0.01 the candidate's number increases by approximately 3 times.
– by increasing the period size from 50 to 500, the candidate's number decreases. Indeed, more the period size increases more the difference between data sets increases thus, the number of candidates decreases.



(a) period size = 100 slots, $\alpha = 0.025$ (b) period size = 100 slots, $d = 6$

(c) $d = 6$, $\alpha = 0.025$

Fig. 9 Periodic number of candidates obtained by applying in-network stage.

6.2.2 Illustrative Example of Neighboring Correlation

Fig. 10 shows an illustrative example to which sensor nodes are correlated based on the K-S test. We take the sensor node 1 and we present its correlation with neighbouring nodes for temperature, humidity and light conditions. The results show that the temperature sensor in node 1 has correlation with the set of temperature sensors $[2, 13, 24, 26, 27, 30, 31, 33, 43]$ while the humidity and light sensors in node 1 have correlation with $[13, 24, 26, 27, 28, 29, 30, 31, 33]$ and $[2, 3, 15, 21, 22, 23, 26, 32, 33, 35, 36, 41, 43, 46]$ respectively. Thus, we can deduce the following: 1) the sensors in the same node do not have the same number of correlated sensors; the light sensor has more correlations than temperature and humidity sensors. 2) the node is more correlated to its nearest nodes than the other nodes.

6.2.3 Energy Consumption

In our simulation, we used the energy model in [30,31] in order to evaluate the performance of the in-network stage of our mechanism. This model considers that energy consumption is highly dependent on the data transmission and receiving while negligent the other factors (sensing and processing). The energy consumed, during each period, by a CH for receiving data from N sensors is

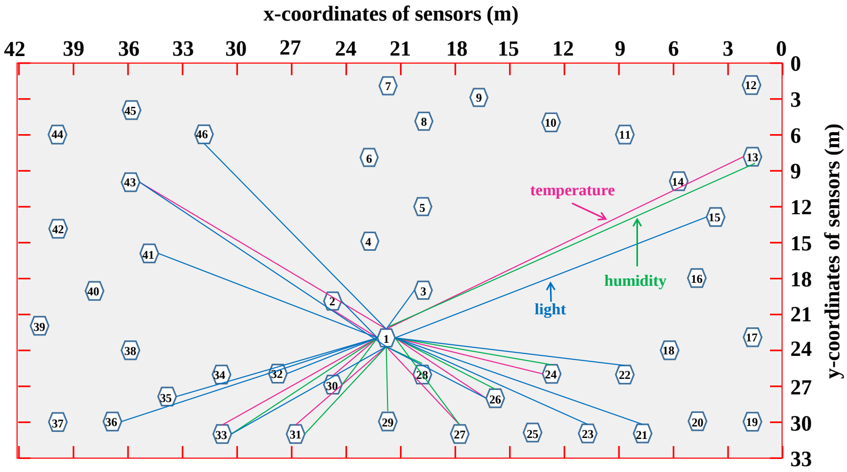


Fig. 10 Example of neighboring correlation between sensor nodes, period size= 100 slots, $d = 6$, $\alpha = 0.025$.

only dependent on the amount of data and can be calculated as follows:

$$E_{RX} = N \times (d+1) \times 64 \times E_{elec} \tag{7}$$

where $(d+1)$ is the size of coefficient set sent from each sensor, 64 indicates the bit representation of each value, and E_{elec} is the energy consumption of a CH in its electronic circuitry (usually $E_{elec} = 50nJ/bit$).

However, the energy consumption of a CH for transmitting $\mathcal{N}$ coefficient sets to the sink node, which is at distance '$dist$' is:

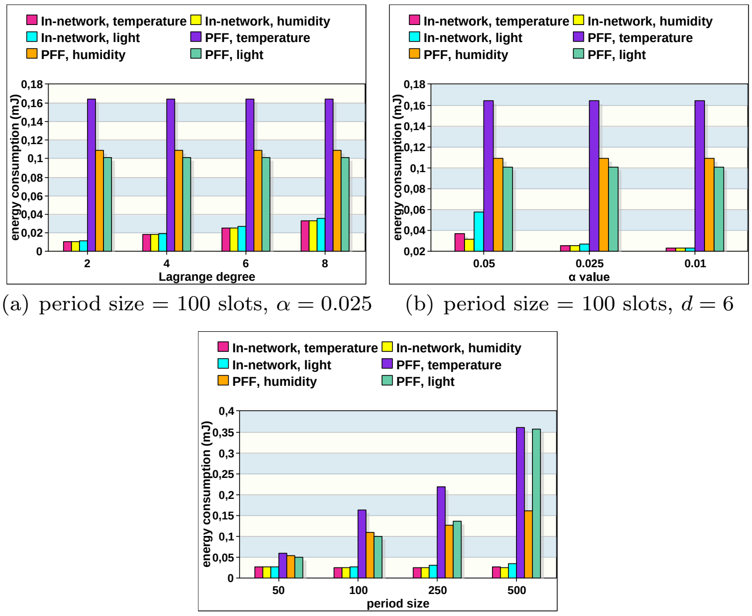
$$E_{TX} = E_{elec} \times \mathcal{N} \times (d+1) \times 64 + \beta_{amp} \times \mathcal{N} \times (d+1) \times 64 \times dist^2 \tag{8}$$

where β_{amp} represents the energy consumption in RF amplifiers for propagation loss (usually $\beta_{amp} = 100pJ/bit$).

Fig. 11 shows the energy consumption in CH by applying in-network stage and PFF while varying the tested parameters to different values. Obviously, the energy consumed in CH is dependent on the amount of data received from sensors (eq. 7) and the number of coefficient sets after eliminating redundant sets (Fig. 9). The results show that in-network stage can efficiently reduce the amount of data transmitted to the sink compared to PFF. Subsequently, our stage conserves the energy of CH by at least 45% and up to 90%. In addition, we can also observe that the energy consumption in CH decreases with the increasing Lagrange degree d or decreasing value of α.

6.2.4 Processing Time

This section discusses the complexity of the in-network stage proposed in our technique at the CH level, compared to that required in PFF. At each period,



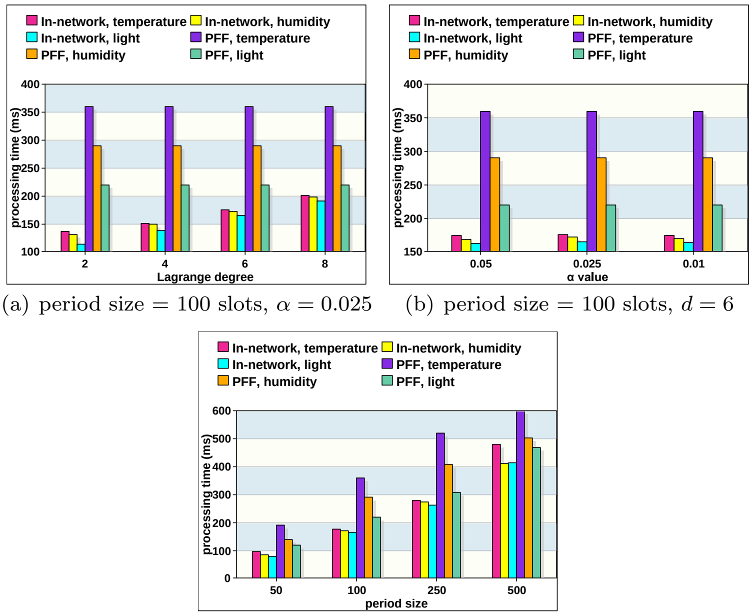
(a) period size = 100 slots, $\alpha = 0.025$ (b) period size = 100 slots, $d = 6$

(c) $d = 6$, $\alpha = 0.025$

Fig. 11 Energy consumption in CH.

the complexity of our stage can be computed based on two steps: first, the CH has to reconstruct raw data for each sensor based on its coefficient set; Second, it has to apply K-S test in order to find the maximum difference between every pair of data sensors. Therefore, the complexity of the in-network stage is considered as $O(F \times N^2)$, where N is the number of total sensors. On the other hand, e.g. using PFF, the CH has to calculate the similarity between every pair of sets. Unfortunately, such calculation requires to check the similarity between every reading in the first set and all readings in the second one which requires $O(F^2)$, where F is the total number of readings in each set. In addition, checking all pair of data will require $O(N^2)$ thus, the PFF has $O(F^2 \times N^2)$ as complexity.

Fig. 12 shows the processing time required at the CH when applying in-network stage and PFF. As shown in the figure, we varied the value of variables (d, α and period size) like those used in Fig. 11. As expected, our stage outperforms PFF in terms of execution time in all cases. In addition, the results show that in-network stage reduces up to 61%, 54% and 48% of the execution time at the CH for temperature, humidity and light respectively, compared to PFF. We can also notice the following observations: 1) the processing time using our stage is almost fix independently of the monitored condition. This is because K-S test compares sets of equal size (e.g. F), unlike PFF that assumes a similarity threshold between the collected data. 2) the processing time of our stage increases with the increasing value of Lagrange degree. This is due to the number of readings taken when computing the Lagrange function that increases when d increase. 3) the processing time in our stage is almost fix when varying α level. 4) our stage and PFF require more execution time when the period size increases. This is due to the computation process that requires more processing when increasing the number of readings in the sets.



(a) period size = 100 slots, $\alpha = 0.025$ (b) period size = 100 slots, $d = 6$

(c) $d = 6$, $\alpha = 0.025$

Fig. 12 Processing time at CH.

6.3 Further Discussion

In this section, we give further consideration to our proposed technique with studying the feasibility of applying it under which conditions and circumstances of the application.

First, the data accuracy is preserved in our technique depending on the chosen value of Lagrange degree d. Therefore, for the critical applications where a high level of accuracy is needed, like in military and healthcare applications, the value of d can be increased. Otherwise, i.e. for low critical applications like weather monitoring, Lagrange degree can be decreased.

Second, the data latency in our technique is dependent on the period size and the Lagrange degree. Therefore, for the real-time applications where a fast decision is necessary to take, like in disaster applications, the values of such parameters should be decreased in order to decrease the execution time of our technique.

Finally, for applications where the energy of the network is the most important factor to save, like in hostile and remote zones, our technique must decrease the values of Lagrange degree d or increase the false rejection probability α at the CH node.

7 Conclusion and Perspectives

As the number of connected devices will continue to rise every day, the IoT will take more attention from both industries and governments. Thus, data reduction and prediction algorithms will remain at the heart of data management in IoT. In this paper, we proposed a novel big data prediction and statistic mechanism dedicated to large-scale sensor networks. Our mechanism is based on the cluster-based network and it works in two stages: on-node prediction and in-network aggregation stages. In the first stage, we focus on reducing data transmitted by sensors using the Lagrange interpolation polynomial model. The second stage focuses on reducing data generated by neighbouring nodes using a statistical test called Kolmogorov-Smirnov. Through simulation on real sensor data, we demonstrate that the proposed mechanism is better than existing techniques in terms of network energy consumption and network lifetime.

Many enhancements can be done on our mechanism in future work. First, we plan to apply our mechanism in real-world scenarios by conducting experimentations in environment and healthcare applications. Second, we seek to improve the computational algorithm at the CH level in order to reduce the data latency of our mechanism, especially in a dense network. Third, we want to add a scheduling strategy in order to put sensors generating redundant data in sleep/active mode. Thus, the energy of the sensor will be more minimized and the network lifetime will be more enhanced.

8 Conflict of Interest

The authors declare that they have no conflict of interest.

References

1. Akila Cse and Uma Maheswari, *A Survey on Recent Techniques for Energy Efficient Routing in WSN*, International Journal of Sensors and Sensor Networks, Vol. 6, Iss.1, pp. 8-15, 2018.
2. Gabriel Martins Dias, Boris Bellalta, and Simon Oechsner, *A Survey About Prediction-Based Data Reduction in Wireless Sensor Networks*, Journal of ACM Computing Surveys (CSUR), Vol. 49, Iss. 3, Article No. 58, 2016.
3. Vishal Krishna Singh, Vivek Kumar Singh, and Manish Kumar, *In-Network Data Processing Based on Compressed Sensing in WSN: A Survey*, An International Journal Wireless Personal Communications, Vol. 96, No. 2, pp. 2087-2124, September 2017.
4. S.G. Shivaprasad Yadav, A. Chitra, and C. Lakshmi Deepika, *Reviewing the process of data fusion in wireless sensor network: a brief survey*, International Journal of Wireless and Mobile Computing, Vol. 8, Iss. 2, pp. 130-140, April 2015.

5. Samer Samarah, *Vector-based data prediction model for wireless sensor networks*, International Journal of High Performance Computing and Networking (IJHPCN), Vol. 9, No. 4, pp. 310-315, 2016.
6. Gopal Krishna, Sunil Kumar Singh, Jyoti Prakash Singh and Prabhat Kumar, *Energy Conservation Through Data Prediction in Wireless Sensor Networks*, Proceedings of 3rd International Conference on Internet of Things and Connected Technologies (ICIoTCT), Jaipur, India, pp. 986-992, March 26-27, 2018.
7. Md Monirul Islam, Zabir Al Nazi, A. B. M. Aowlad Hossain and Md Masud Rana, *Data Prediction in Distributed Sensor Networks Using Adam Bashforth Moulton Method*, Journal of Sensor Technology, Vol. 8, pp. 48-57, 2018.
8. Islam, M.M., Nazi, Z.A, Rana, M.M. and Hossain, A.A.B.M., *Information Prediction in Sensor Networks Using Milne-Simpson's Scheme*, Proceedings of the International Conference on Advances in Electrical Engineering, pp. 494-498, September 2017.
9. Ikjune Yoon, and Dong Kun Noh, *Energy-Aware Control of Data Compression and Sensing Rate for Wireless Rechargeable Sensor Networks*, Journal of Sensors, Vol. 18, Iss. 8, Article 2609, 2018.
10. Qinbao Xu, Rizwan Akhtar, Xing Zhang, and Changda Wang, *Cluster-Based Arithmetic Coding for Data Provenance Compression in Wireless Sensor Networks*, Journal of Wireless Communications and Mobile Computing, Vol. 2018, Article ID 9576978, 15 pages, 2018.
11. Sunyong Kim, Chiwoo Cho, Kyung-Joon Park, and Hyuk Lim, *Increasing network lifetime using data compression in wireless sensor networks with energy harvesting*, International Journal of Distributed Sensor Networks, Vol. 13, Iss. 1, 2017.
12. Tarek Sheltami, Muhammad Musaddiq, and Elhadi Shakshuki, *Data compression techniques in Wireless Sensor Networks*, Journal of Future Generation Computer Systems, Vol. 64, Iss. C, pp. 151-162, 2016.
13. Y. Liang and Y. Li, *An efficient and robust data compression algorithm in wireless sensor networks*, Journal of Future Generation Computer Systems, IEEE Communications Letters, Vol. 18, No. 3, pp. 439-442, 2014.
14. Mou Wu, Liansheng Tan, and Naixue Xiong, *A Structure Fidelity Approach for Big Data Collection in Wireless Sensor Networks*, Sensors journal, Vol. 15, pp. 248-273, 2015.
15. Sunil Dhimal, and Kalpana Sharma, *Energy conservation in wireless sensor networks by exploiting inter-node data similarity metrics*, International Journal of Energy, Information and Communications, Vol. 6, No. 2, pp. 23-32, 2015.
16. Suat Ozdemir, Miao Peng and Yang Xiao, *PRDA: polynomial regression-based privacy-preserving data aggregation for wireless sensor networks*, Wireless communications and mobile computing journal, Vol. 15, pp. 615-628, 2015.
17. Jacques Bahi, Abdallah Makhoul and Maguy Medlej, *A two tiers data aggregation scheme for periodic sensor networks*, Ad Hoc & Sensor Wireless Networks, Vol. 21, Iss. (1-2), pp. 77-100, 2014.
18. Shahinaz M. Al-Tabbakh, *Novel technique for data aggregation in wireless sensor networks*, International Conference on Internet of Things, Embedded Systems and Communications (IINTEC), Gafsa, Tunisia, October 20-22, pp. 1-8, 2017.
19. Hassan Harb, Abdallah Makhoul and Raphael Couturier, *An enhanced k-means and anova-based clustering approach for similarity aggregation in underwater wireless sensor networks*, IEEE Sensors Journal, Vol. 15, No. 10, pp. 5483-5493, 2015.
20. Degan Zhang, Ting Zhang, Jie Zhang, Yue Dong and Xiao-dan Zhang, *A kind of effective data aggregating method based on compressive sensing for wireless sensor network*, EURASIP Journal on Wireless Communications and Networking, Vol. 2018, No. 159, 15 pages, 2018.
21. Guorui Li, Haobo Chen, Sancheng Peng, Xinguang Li, Cong Wang, Shui Yu, and Pengfei Yin, *A Collaborative Data Collection Scheme Based on Optimal Clustering for Wireless Sensor Networks*, Sensors (Basel), Vol. 18, No. 8, pages 2487, 2018.
22. Nandoori Srikanth, M.S. Ganga Prasad, *Efficient Clustering Protocol Using Fuzzy K-means and Midpoint Algorithm for Lifetime Improvement in WSNs*, International Journal of Intelligent Engineering and Systems, Vol. 11, No. 4, pp. 61-71, 2018.

23. Adil Khan, Chandra P. Gupta, Iti Sharma, *Addressing Data Aggregation Using Polynomial Regression in WSNs*, International Journal of Sensors, Wireless Communications and Control, Vol. 5, Iss. 2, pp. 114-120, 2015.
24. Sreya Ghosh, and Iti Saha Misra, *Design and testbed implementation of an energy efficient clustering protocol for WSN*, IEEE International Conference on Innovations in Electronics, Signal Processing and Communication (IESC), Shillong, India, April 6-7, pp. 1-6, 2017.
25. Shaweta Mahajan, and Vijay Kumar Banga, *Inter cluster data aggregation balanced energy efficient network integrated super heterogeneous protocol for wireless sensor networks*, Twelfth International Conference on Wireless and Optical Communications Networks (WOCN), Bangalore, India, Sept. 9-11, pp. 1-6, 2015.
26. Dongbin Zhao, Li Bu, Cesare Alippi, Qinglai Wei, *A Kolmogorov-Smirnov Test to Detect Changes in Stationarity in Big Data*, IFAC-PapersOnLine, Vol. 50, Iss. 1, pp. 14260-14265, 2017.
27. Fernando Antoneli, Fernando M. Passos, Luciano R. Lopes, Marcelo R. S. Briones, *A Kolmogorov-Smirnov test for the molecular clock based on Bayesian ensembles of phylogenies*, PLOS ONE journal, Vol. 13, No. 1, pp. 1-22, 2018.
28. Trusina J, Franc J, and Kus V, *Statistical homogeneity tests applied to large data sets from high energy physics experiments*, Journal of Physics: Conference Series, Vol. 936, conference 1, pp. 1-6, 2017.
29. Samuel Madden, *Intel Berkeley Research lab*, http://db.csail.mit.edu/labdata/labdata.html, 2004.
30. Wendi Beth Heinzelman, *Application Specific Protocol Architectures for Wireless Networks*, PhD thesis, Massachusetts Institute of Technology, June 2000.
31. Wendi Beth Heinzelman, A. Chandrakasan, H. Balakrishnan, *Energy-Efficient Communication Protocol for Wireless Microsensor Networks*, Proceedings of the 33rd Hawaii International Conference on System Sciences, January 2000.
32. Hassan Harb, Abdallah Makhoul, Chady Abou Jaoude *A Real-Time Massive Data Processing Technique for Densely Distributed Sensor Networks.* IEEE Access 6: 56551-56561 (2018)
33. Gaby Bou Tayeh, Abdallah Makhoul, David Laiymani, Jacques Demerjian *A distributed real-time data prediction and adaptive sensing approach for wireless sensor networks.* Pervasive and Mobile Computing 49: 62-75 (2018)
34. Hassan Harb, Abdallah Makhoul *Energy-Efficient Sensor Data Collection Approach for Industrial Process Monitoring.* IEEE Trans. Industrial Informatics 14(2): 661-672 (2018)
35. Hassan Harb, Abdallah Makhoul, David Laiymani, Ali Jaber *A Distance-Based Data Aggregation Technique for Periodic Sensor Networks.* ACM Transactions on Sensor Networks13(4): 32:1-32:40 (2017)
36. Trupti Mayee Behera, Sushanta Kumar Mohapatra, Umesh Chandra Samal, Mohammad S Khan, Mahmoud Daneshmand, Amir H Gandomi *Residual Energy Based Cluster-head Selection in WSNs for IoT Application.* IEEE Internet of Things Journal, 1(1): 1-8 (2019)
37. Suparna Biswas, Jayita Saha, Tanumoy Nag, Chandreyee Chowdhury, Sarmistha Neogy *A novel cluster head selection algorithm for energy-efficient routing in wireless sensor network.* 2016 IEEE 6th International Conference on Advanced Computing (IACC), 588-593 (2016)
38. R Raj Priyadarshini, N Sivakumar *Cluster head selection based on Minimum Connected Dominating Set and Bi-Partite inspired methodology for energy conservation in WSNs.* Journal of King Saud University-Computer and Information Sciences, 1-20 (2018)
39. Yousif Khalid Yousif, R Badlishah, N Yaakob, A Amir *An energy efficient and load balancing clustering scheme for wireless sensor network (WSN) based on distributed approach.* Journal of Physics: Conference Series 1019(1):012007 (2018)
40. Sang Kang *Energy Optimization in Cluster-Based Routing Protocols for Large-Area Wireless Sensor Networks.* Journal of Symmetry 11(1):37 (2019)
41. Govind P Gupta *Improved Cuckoo Search-based Clustering Protocol for Wireless Sensor Networks.* Procedia Computer Science 125:234-240 (2018)

42. Amine Rais, Khalid Bouragba, Mohammed Ouzzif *Routing and Clustering of Sensor Nodes in the Honeycomb Architecture*. Journal of Computer Networks and Communications 2019: (2019)
43. Andreas P Plageras, Kostas E Psannis, Christos Stergiou, Haoxiang Wang, B Brij Gupta *Efficient IoT-based sensor BIG Data collection–processing and analysis in smart buildings*. Future Generation Computer Systems Journal 82:349-357: (2018)
44. Christos Stergiou, Kostas E Psannis, Byung-Gyu Kim, Brij Gupta *Secure integration of IoT and cloud computing*. Future Generation Computer Systems Journal 78:964–975: (2018)
45. Christos Stergiou, Kostas E Psannis *Recent advances delivered by Mobile Cloud Computing and Internet of Things for Big Data applications: a survey*. International Journal of Network Management 27(3):e1930: (2017)
46. Kostas E Psannis, Christos Stergiou, BB Gupta *Advanced media-based smart big data on intelligent cloud systems*. IEEE Transactions on Sustainable Computing 4(1):77-87: (2019)
47. Christos Stergiou, Kostas E Psannis, Brij B Gupta, Yutaka Ishibashi *Security, privacy & efficiency of sustainable Cloud Computing for Big Data & IoT*. Sustainable Computing: Informatics and Systems 19:174-184: (2018)