The Energy-Aware Matrix Completion-Based Data Gathering Scheme for Wireless Sensor Networks

MANEL KORTAS[1,2]**, OUSSAMA HABACHI**[2]**, AMMAR BOUALLEGUE**[1]**, VAHID MEGHDADI**[2]**, TAHAR EZZEDINE**[1]**, AND JEAN PIERRE CANCES**[2]**, (Senior Member, IEEE)**

[1]Communications Systems Laboratory (SysCom), National Engineering School of Tunis, University of Tunis El Manar, Tunis 1002, Tunisia
[2]XLIM, University of Limoges, 87060 Limoges Cedex, France

Corresponding author: Manel Kortas (manel.kortas@unilim.fr)

ABSTRACT In the Wireless Sensor Networks (WSNs), ensuring long-term survival of the sensor devices is crucial, especially for non-energy harvesting networks where the sensors have to deal with the available limited power. Thus, there is a huge need to efficiently select, in each time-slot, a small set of source nodes to monitor the network area and deliver their data to the sink. Note that there is a trade-off between energy efficiency, achieved through data-compression, and the informative quality received by the sink. Moreover, although applying a high data compression ratio extremely reduces the overall network energy consumption, the network lifetime is not necessarily extended due to the uneven energy depletion of the nodes' batteries. To this end, in this paper, we propose the Energy-Aware Matrix Completion based data gathering approach (EAMC), which designates the active nodes according to their residual energy levels. To collect data readings, the proposed EAMC relies on a nodes clustering phase and a MC based data sampling. Then, the interpolation of all the missing data is performed by the sink thanks to a Three-stage MC based recovery framework. Since we are interested in high data loss scenarios, the limited amount of delivered data must be sufficient in terms of informative quality it holds in order to reach a satisfactory recovery accuracy for the entire data. Hence, the EAMC selects the nodes depending on their inter-correlation as well as the network energy efficiency, with the use of a combined energy-aware and correlation-based metric. This introduced active node cost function changes with the type of application one wants to perform with the intention to reach a longer lifespan for the network. Therewith, the numerical results show that the EAMC achieves an attractive and competitive trade-off between the data reconstruction quality and the network lifetime for all the investigated scenarios.

INDEX TERMS Wireless sensor networks, matrix completion, data gathering, spatial data interpolation, energy efficiency.

I. INTRODUCTION

With the rapid progress achieved in the information technology fields, the Internet of Things (IoT) has been emerging and promising to revolutionize the life quality by deploying massively the sensor nodes. According to [1], it is estimated that, by 2025, around 41.6 billion sensor-based devices, generating 79.4 *ZB* of Data, will be connected to the Internet as part of the IoT. On the other hand, we expect that the global IoT market will grow from US $190 billion in 2018 to US $1.1 trillion in 2026 [2]. The current need for Machine-Type Communications (MTC) has provided a vari-

ety of communication technologies in order to satisfy the heterogeneous IoT requirements [3]. Recently, the massive IoT access has been considered as a part of the 5th generation cellular systems (5G). Nevertheless, researchers, scientists, and engineers are facing emerging challenges to effectively incorporate the IoT based systems in the fifth generation 5G wireless communications [4]. Wireless sensor networks (WSNs), which is a key pillar of IoT, can gather data from harsh environments at a low cost. However, these devices are characterized by limited capabilities in terms of energy capacity and computation. In addition, due to the explosive growth of the number of sensor nodes and the amount of the generated data, it is vital for the network to preserve a high energy efficiency and improve its utilization during the

The associate editor coordinating the review of this manuscript and approving it for publication was Gongbo Zhou.

data gathering process. Indeed, since the sensors' batteries are generally not rechargeable, when the first node uses up all its energy, in most of cases, the entire network is partitioned then paralyzed causing its death. Usually, the activities for which a sensor node consumes its energy are sensing, processing and communication. Most of the existing energy management strategies assume that radio transmission and reception are the most energy-consuming operations [5], [6]. Therefrom, several papers have mainly focused on data compression in order to minimize the amount of data readings to be forwarded. Without the need of any extra communication and computational overheads, the Matrix Completion (MC) technique, viewed as an extension of Compressive Sensing (CS), reduces the number of active agents, in each time slot, while all the missing data can still be recovered [7]–[11].

In this paper we focus on the twofold data compression scenario that had been addressed in our previous work [7] with the Cluster-based MC data gathering approach (CBMC), where the wireless network gets denser. This atypical data sampling scenario consists on choosing a significant number of sensor nodes to remain inactive during the whole sensing period, whilst the rest of nodes serve as the representative of the entire network. Withal, these active nodes do not send, every time slot, their raw data to the sink. Instead, they deliver a part of their data readings under a sampling ratio guaranteed by the MC theory. Indeed, each time slot, a subset of the representative nodes is picked to get the role of source nodes. In [7], relying on a MC-based approach, the sink is able to recover all the missing readings, even the fully empty data rows that correspond to the inactive nodes, thanks to a complementary minimization problem based on an interpolation technique that is annexed to the MC one. Different from [7], where the representative nodes are chosen randomly, in this paper, we focus on the node selection process taking into account the reconstruction quality as well as the energy efficiency. In order to neatly represent the entire network and reach an interesting data reconstruction quality, despite the challenging addressed scenarios, correlation among sensors has been calculated and those holding the greatest informative values are better ranked to be chosen as representative nodes.

Reducing the amount of sensing data can indeed minimize the power consumption of the network and save its energy. Nevertheless, it is not sufficient since it does not necessarily alleviate the problem of energy load imbalance between nodes. Indeed, depending on the events to be monitored, even though the representative sensors may change from one detection period to another, the signal in most WSNs is time-stationary. Hence, the set of selected representative nodes can remain the same for many successive sensing periods. To avoid the overcharge that may occur over some sensor nodes, and thus their fast death, the representative nodes should change from a detection period to another. In addition, in the multi-hop WSNs, data packets that are generated from the source nodes should be relayed via intermediate nodes to be routed to the sink. Accordingly, nodes around the sink would exhaust their batteries faster as they carry heavier

traffic loads than the border nodes, which causes the problem of energy hole. In this case, even if the rest of nodes still holds sufficient energy, communication with the sink will be cut off leading to the end the network lifespan. To overcome the issue of uneven energy depletion phenomenon that was noticed in WSNs, in addition to the correlation, we have incorporated the sensors' residual energies in the representative node selection function with the proposed Energy-Aware MC-based data gathering approach (EAMC). It is noteworthy that taking into account the residual energy in the node selection process is related to the type of application one wants to perform. To this end, we have evaluated our selection strategy under different scenarios and network topologies while presenting for each one the adequate energy-aware metric.

More specifically, our main contributions in this paper are given as follows:

- As a sequel of [7], the selection process of the representative sensor nodes is no longer random. It is rather deterministic. Both energy and correlation are considered in the cost function of the EAMC in order to select the sensor nodes that can well represent the network and at the same time afford the sustainability as long as possible for the network lifetime.
- Different topologies and scenarios have been assessed under the adequate energy-aware proposed metrics. Indeed, in the star topology networks, where communication with the sink is direct, choosing a node to be a representative one according to its residual energy in order to improve the network energy utilization is sufficient. Whereas, in the mesh topology networks, where routing schemes must be applied and data is forwarded via relaying nodes, the entire route should be assessed. In addition to the correlation, a node can be chosen to be a representative one if there is no depleted relaying node in its route.
- In this paper, we target to minimize the energy consumption and extend the network lifespan through nodes energy load balancing, while, at the same time, ensuring a sufficiently good quality of data reconstruction. In the numerical results section, we have studied the trade-off between the data recovery error and the network lifetime for all the investigated scenarios.
- The assumption that the energy consumed in the data acquisition is much lower than that consumed in radio communications does not hold in a number of practical applications, such as the gas sensors which are considered as power hungry sensors [5]. Therefore, in this paper, we have assessed our approach under both sensor nodes types, the ordinary sensors (low sensing power sensors) and the power hungry ones.

The paper is organized as follows. The next section discusses the related works. Section III is devoted to state the used data reconstruction framework and the energy consumption system model. In section IV, we present the proposed energy-aware data gathering strategy under different scenarios. Then,

in order to evaluate the performance of the proposed scheme, we carry out, in section V, with various simulations, where we vary the cost selection function, the addressed scenario and the type of the deployed sensor nodes. Finally, we conclude the work in section VI.

II. RELATED WORKS

In the state-of-art of the energy-efficient based algorithms for data gathering, reducing the amount of collected data readings or reducing the packets size are two well investigated methods that are closely related to the minimization of the energy consumption [12], [13]. CS and MC, known as two popular sparsity representation techniques, take benefits from the redundancy that occurs in the environmental WSN signals to drastically reduce the number of transmitted measurements without any on-sensors computation [5], [7]–[11], [14]–[18].

In [14], a state-of-the-art of MC-based scheme for data gathering has been proposed in order to introduce the short-term stability with the low-rank feature. The presented MC-based approach requires a fewer number of sensor nodes' readings compared to the algorithms of comparison and hence achieves a longer lifespan for the network. In [17], a Routing-Aware Space-Time Compressive Sensing approach has been proposed in order to improve the trade-off between the network energy saving and the data reconstruction accuracy. However, the issue of energy load balancing between nodes was not addressed in [17] and has been left as a perspective. Likewise, Li et al., in [18], have combined the CS and the routing scheme and proposed a Multi-Strip Data Gathering approach for Green data collecting. Using this approach, the network is partitioned into multiple strips, where nodes around each strip forward data to the center with data fusion technique. The amount of data readings undertaken by sensors is relatively balanced since the transmitting nodes are changing. Yet, according to [19], this scheme doesn't use an adaptive distributed technique to minimize the complexity in data gathering. Authors of [15] have developed a novel energy-efficient and CS-based data gathering approach for clustered WSNs to reduce the number of required measurements using a graph theory. Moreover, they have provided mathematical foundations to the introduced approach, which leads to a nonuniform collection of data readings via a random walk based manner. In [20], a Reducing Delay and Maximizing Lifetime data gathering method has been introduced, where the data transmission radius is no longer the same for all nodes. Instead, some nodes that hold energy surplus are using a larger data transmission radius than the others in order to reduce the amount of forwarded data and hence extend the network lifetime. On the heels of MC theory, Tan et al., in [11], have targeted to enhance the network energy efficiency and proposed a low redundancy data collection scheme. This MC-based approach serves to quickly compensate the set of collected data in cases of packet loss. In order to not affect the network lifetime, this approach takes advantage of the energy surplus, remaining away from the sink area, and conceives the backup data set in order to satisfy

the minimum number of measurements required by the MC theory. Different from the compression-based aspect that the aforementioned schemes have proceeded, the authors of [21] have addressed the network lifetime issue by reducing the number of nodes state transitions, pointing out that the processor consumes energy through state transition. Likewise, in [22], authors have proposed a Rotating Random Sparse Sampling technique, where the data compression ratio varies over time slots, during data gathering, and nodes are sleeping most of the time. These techniques bear a resemblance to ours in the sense that, in our scenario, a set of sensor nodes is scheduled to not sense the environment for a large number of consecutive time slots in order to reduce their power consumption.

In line with the consideration of the transmission path to increase the network lifetime, Yao et al., in [23], have developed an Energy-efficient Delay-aware Lifetime-balancing data collection algorithm for the heterogeneous WSNs, in which nodes holding poor communication links and less remaining energy have a lower chance to be chosen as forwarders. However, in this paper, at the beginning of each collection period, a set of nodes is selected to be the sources. However, in our approach, the source nodes differ from a time slot to another ensuring the diversity in the reported data and thus a better monitoring quality and energy balancing. Similarly, the paper [24] has proposed two algorithms, where the current sensor node always chooses, as next hop, the node that has the highest residual energy. Yet, the proposed techniques have been proved in [25] to be unable to manage the problem of void hole. To overcome the energy hole problem, authors in [26] have introduced a new layer, referred to as the charging layer, into the basic node network protocol stack. As soon as the battery level of a node goes done, it is charged wirelessly using witricity (wireless electricity). In this context of rechargeable nodes' batteries, the Energy-Harvesting Wireless Sensor Networks (EHWSNs), where nodes can replenish their batteries with energy from the surrounding environment, have been getting attention from researchers. Among the in-network processing-based schemes, an *m*-hop averaging data compression technique, with energy harvesting, has been proposed in [12] in order to deal with the unevenness of the energy levels among the nodes. In this algorithm, each node has to continuously assemble the usable energy levels of other nodes then makes a decision about how much it needs to compress the forwarded data packets after comparing its own energy level with those assembled from the next *m* nodes within *m* hops. As the packet is relayed towards the sink, the data packet length becomes smaller leading to a gradual decrease in the energy cost. Different from [12], in [10], authors have presented an Adaptive Collection scheme-based Matrix Completion, which adjusts the amount of data to be gathered at each moment depending on the residual usable energy absorbed from solar radiation. This scheme has been designed to improve the network energy utilization, increase the duty cycle of sensors far away from the sink and gather as less data readings as possible, when there is no sufficient

usable energy and vice versa. Yet, this is not the case with our scenario since our deployed sensor nodes can neither charge their batteries nor renew them. Making use of only the available energy, our proposed scheme targets to extend the network lifetime and prevent it from being prematurely partitioned or dead by considering the residual energy of the entire multi-hop path that links the source node with the sink. Furthermore, at the end of the network lifetime, the remaining energy of the border nodes (i.e. nodes far away from the sink) is almost close to the average remaining energy thanks to the introduced energy-aware cost functions of the EAMC that select the representative sensor nodes. Without any extra communication between nodes, the proposed metrics of the EAMC scheme aim not only for achieving energy efficiency but also for preserving a sufficiently good quality of data reconstruction as they take into account correlation among sensors to select those who can report more information about the network.

III. PRELIMINARY AND ENERGY CONSUMPTION MODEL

In this section, we firstly present the developed techniques that the sink node applies in order to recover the entire data matrix $X \in IR^{N \times T}$, after receiving a partly empty matrix, where N denotes the number of deployed sensor nodes in a square observation area, whereas T designates the number of time slots t composing the detection period T. Precisely, the entry $x_{i,t}$ of X holds the tth data reading that can be sampled by the ith node. Then, in the second part, we describe the adopted energy consumption model that we used to calculate the network lifetime.

A. THE THREE-STAGE MC-BASED RECONSTRUCTION APPROACH

In this subsection, we present in details the three-stage MC-based reconstruction approach that we have introduced in [7] in order to deal with the high data loss ratios, even with the existence of a significant number of completely missing data rows in the received data matrix $M \in IR^{N \times T}$. These empty rows result from the inactive nodes, denoted also by absent nodes, that do not participate in the data sensing during the entire detection period T. Note that an inactive sensor node i leads to an empty data row i in M. This MC-based framework guarantees the recovery of all these rows relying on the partially received data from the active nodes. Indeed, in the beginning of each detection period T, the sink selects a set of active nodes $\mathcal{N}_{rep} = \{1, \ldots, N_{rep} \ll N\}$ to be the representative of the network and send a part of their data readings under a fixed compression ratio. To model the data sensing schedule, we use a binary matrix $\Omega_M \in IR^{N \times T}$, where $\Omega_M(i, t) = 1$ if $x_{i,t}$ will be sensed then delivered to the sink and 0 otherwise. Here, the incomplete received data matrix M can be expressed as a Hadamard product between X and M. Hence, to replace any missing entry in M, we set a "zero" as a placeholder. Following this kind of data sampling, M consists in the combination of N_{rep} partially empty data rows and $(N - N_{rep})$ completely empty ones.

1) OVERVIEW OF MATRIX COMPLETION

Recently, MC technique has emerged to benefit from the signal low-rank feature in order to fill the missing data using a limited number of the matrix entries [27]. That is, the compressed data, which is a partly empty matrix $M \in IR^{N \times T}$ of rank $r \ll min\{N, T\}$, can be entirely recovered, if a sub-set of its entries M_{ij} as well as their indices $(i, j) \in \Omega$ are known by the receiver. The entry-wise partial observation operator $P_\Omega : IR^{N \times T} arrow IR^{N \times T}$ is defined as follows:

$$[P_\Omega(X)]_{ij} = \begin{cases} X_{ij} & (i, j) \in \Omega \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

Here, P_Ω denotes the projection onto the subspace, constituted by sparse matrices, with non-zeros elements restricted only to the index subset Ω [28]. According to [27], if Ω holds enough information and if X is a low rank or approximately a low-rank matrix, we can recover the unknown entries by solving the following rank minimization problem:

$$\text{minimize } rank(X) \quad s.t \ P_\Omega(X) = P_\Omega(M). \quad (2)$$

However, problem (2) is not convex. Fortunately, the nuclear norm $\parallel X \parallel_*$ minimization problem, which is a convex relaxation, can be solved. Indeed, it is used as an alternative to the NP-hard rank minimization problem. Hence, we have:

$$\text{minimize } \parallel X \parallel_* \quad s.t \ P_\Omega(X) = P_\Omega(M). \quad (3)$$

In the literature, several solvers for this type of systems have been proposed. For example, the Singular Value Thresholding (SVT) optimizes an approximation of (3) by adding a Frobenius-norm term to the objective function [29]:

$$\text{minimize } \tau \parallel X \parallel_* + \frac{1}{2} \parallel X \parallel_F^2 \quad s.t \ P_\Omega(X) = P_\Omega(M). \quad (4)$$

Low rank matrix fitting (LMaFit) [28], Sparsity Regularized SVD (SRSVD) and Sparsity Regularized Matrix Factorization (SRMF) [30], among other schemes, have used the matrix factorization method. Different from (3), matrix factorization technique has been suggested to change (2) rather than using the nuclear norm.

2) THE THREE-STAGE RECONSTRUCTION PATTERN

The existence of $(N - N_{rep})$ totally empty rows in the received data matrix M impedes the MC technique and makes it completely unable to recover the original matrix. Indeed, since MC schemes are based on the minimization of the matrix rank, they become useless when there is any empty column or empty row in the matrix. In this case, the use of other complementary interpolation techniques becomes thoroughly needed. Thereupon and in this context, we have developed the following three-stage MC-based recovery scheme: **Stage 1:** Obviously, it is not possible to directly apply the MC method with the existence of the empty rows. Thus, we remove these rows from M. We denote the resultant matrix as $M_{MC} \in IR^{N_{rep} \times T}$, which contains the partially received readings of the active sensor nodes. We carry on with the same removal from Ω_M to obtain $\Omega_{MC} \in IR^{N_{rep} \times T}$. Then, using (4) or any other

method for the MC resolution, we fill the missing entries of M_{MC} that correspond to the non-delivered readings of the N_{rep} nodes. As it has been introduced in [29], τ roughly equals 100 times the largest singular value of M_{MC}. We denote $X' \in IR^{N_{rep} \times T}$ as the MC based estimation data. Finally, we update $X' \in IR^{N \times T}$ by adding the $(N - N_{rep})$ empty rows and placing them in their corresponding positions in M.

Stage 2: In this stage, we carried on with the spatial pre-interpolation method of [8], which estimates the data of an empty row using the available data of the neighboring nodes. However, for the considered/investigated high loss scenario, the number N_{rep} is very small compared to the total number N, which means that the $(N - N_{rep})$ inactive nodes represent the preponderant portion of the network. As the number of active nodes decreases, we face absent nodes having absent neighbor nodes as well. These nodes are denoted by Isolated nodes (IS). Using this interpolation technique, we can reconstruct data only for the absent nodes, whose neighbors belong to the set $\mathcal{N}_{rep}$, and the data of the Isolated nodes can't be successfully recovered, since the data of their neighbors, which are the absent nodes, is already unavailable in the sink. We assume that the network distribution contains N_{Is} Isolated nodes. Hence, the resulting data matrix $X'' \in IR^{N \times T}$, obtained following this stage, still holds N_{Is} empty rows to be estimated (N_{Is} all-zeros rows). For the detailed steps of the pre-interpolation technique, the reader may refer to [8, Section. VI].

Stage 3: In this stage, we resort to a complementary spatial interpolation method to recover the remaining part of the empty rows (Isolated nodes). Taking advantage of the spatial dependency among the sensors, we fill the N_{Is} empty rows by minimizing the following problem:

$$\text{minimize} (fac_1 \times \|\widehat{X} - X''\|_F^2 + fac_2 \times \|S \times \widehat{X}\|_F^2), \quad (5)$$

where $S \in IR^{N \times N}$ represents the spatial constraint matrix since it expresses the inter-nodes' readings similarities, fac_1 and fac_2 are two tuning parameters and $\widehat{X} \in IR^{N \times T}$ is the final interpolated data matrix. In (5), the matrix S reflects our knowledge about the spatial structure inherent in the data, as it is computed based on a data matrix $X_{lp} = [x_{lp\,1}^{tr}, x_{lp\,2}^{tr}, \ldots, x_{lp\,N}^{tr}]^{tr} \in IR^{N \times T_{lp}}$ that is delivered during a short learning period $T_{lp} \ll T$, where all sensor nodes report their information to the sink.

The resolution of this minimization problem (5) can be easily achieved using the semidefinite programming , and to solve it and get $\widehat{X}$, we opted for the CVX package [31], which is implemented in Matlab, as an advanced convex programming solver.

For detailed information about the computation steps of S and the adjustment of the tuning parameters fac_1 and fac_2, the reader may refer to [7, Section. V].

The numerical results of [7] have shown how interesting and important is the integration of the proposed minimization (5) to deal with the data recovery problem, appearing with the twofold data compression scenario. Indeed, for the high data compression ratios, the number N_{Is} of the IS becomes

significant. Hence, adding a third interpolation technique is deeply needed. Otherwise, we are susceptible to end up with a data matrix $\widehat{X}$, which is almost half built, even less.

Note that the presented three-stage MC-based data reconstruction approach is considered as building block of all the introduced data gathering schemes of section IV in order to recover the entire data matrix X.

B. THE ENERGY CONSUMPTION MODEL

Generally, a sensor node consumes the energy of its battery in three operations that are communications (i.e. both data transmission and reception), data sensing and data processing.

Since with the MC method, there is no on-sensor computation, and data is directly sub-sampled in the compressed form (i.e. the data $x_{i,t}$ is available only if a location i is chosen to be sensed in the time slot t), we assume here that there is no energy consumed in data processing. Moreover, the high energy-intensive reconstruction algorithm is executed at the sink node,[1] which is free of energy constraint and whose energy consumption does not be included in the network overall energy consumption.

Regarding the transmission and reception activities, we consider that two sensor nodes j and i are able to directly communicate with each other, only if the geographic distance $dst_{i,j}$ between them is lower than some transmission range (r) that scales with $\Theta(\sqrt{logN/N})$ [32]. To measure the energy consumption during the data forwarding, we have opted for the following model [33]:

$$\begin{cases} E_{Tx}(L, dst_{i,j}) = E_{elec-tr} \times L + \varepsilon_{amp} \times L \times dst_{i,j}^2 \\ E_{Rx}(L) = E_{elec-rc} \times L, \end{cases} \quad (6)$$

where $E_{Tx}(L, dst_{i,j})$ denotes the energy consumed by a specific node i when forwarding an L-bit packet to a node j through a distance of length $dst_{i,j}$, whereas, $E_{Rx}(L)$ represents the amount of energy that is consumed when receiving the L-bit packet. In the above model, ε_{amp} represents the energy required by the transmitter's amplifier, whereas, $E_{elec-tr}$ and $E_{elec-rc}$ are the amounts of energy that are consumed by the transceiver circuitry respectively at the sender and at the receiver. Here, a representative sensor node can participate in data forwarding, even if it is not chosen to be a transmitting source node in the time slot t.

To monitor the network area and sense the data field, we have used the following expression to compute the energy dissipation by a sensor node when performing the sensing operation for L bit packet [6]:

$$E_{sens}(L) = L \times V_{sup} \times I_{sens} \times T_{sens}. \quad (7)$$

In (7), V_{sup} is the supply voltage, I_{sens} is the total current required for the data sensing operation, and T_{sens} denotes the time duration allowed to a sensor node for data sensing.

IV. OUR PROPOSED DATA GATHERING SCHEME

In this section, we present how the energy constraint can be jointly considered with the correlation criteria in the active

---

[1]We suppose that the sink is located in the center of the network area.

node selection process in order to maintain a load balancing among nodes and maximize the network lifetime, while still achieving a low data reconstruction error. Since the performance usually vary with the network configurations, we differentiate, in this section, the proposed energy-aware cost functions for the representative node selection according to the given network topologies.

To monitor the environment and collect measurements, we carry on with the sampling strategy proposed in [7] that is the CBMC, where, first of all, the network is portioned into J clusters, according to the sensors data readings, using the Normalized Spectral Clustering. We consider the whole network as the combination of J clusters, $(CL_j)_{j=1,...,J}$, where $N = \sum_{j=1}^{J} cl_j$ and cl_j is the number of sensor nodes belonging to CL_j. This preliminary phase has been accomplished in order to not disregard sensor nodes that belong to the small clusters, a deficiency or a slip that can occurs with high probability in the purely random sampling, usually used in the conventional MC. In order to equitably involve all the detected clusters in the sensing schedule, in each time slot t, using the same percentage and according to a given sampling ratio, a subset of nodes is picked from each cluster to monitor the network area. Precisely and firstly, the set $\mathcal{N}_{rep}$ of the representative sensor nodes consists of the combination of J subsets, $(\mathcal{N}_{rep_j})_{j=1,...,J}$, where $\mathcal{N}_{rep_j}$ includes N_{rep_j} representative sensor nodes picked from cluster CL_j using the same shared percentage pct_{Nrep}. That is:

$$N_{rep} = \sum_{j=1}^{J} N_{rep_j}, \quad \text{where } N_{rep_j} = pct_{Nrep}\% \times cl_j. \quad (8)$$

Then, in each time slot t, using the same shared percentage pct_m, m_j source nodes are designated and randomly chosen from the set $\mathcal{N}_{rep_j}$ to sense the field and transmit their data readings to the sink. That is:

$$m = \sum_{j=1}^{J} m_j, \quad \text{where } m_j = pct_m\% \times N_{rep_j}. \quad (9)$$

Numerical results of [7] have shown significant amelioration in terms of data recovery performance, when performing this kind of twofold data sampling mechanism. The main focus of this section is the selection of the representative sensor nodes since these nodes represent the set from which the source nodes are afterwards picked. Different from the CBMC of [7], where for each cluster j the subset of representative nodes $\mathcal{N}_{rep_j}$ is randomly selected from CL_j, in this paper, these nodes are chosen in a specific way using a cost function that selects nodes being able to report more information about their clusters and hence ensuring a better reported data quality. This correlation-based metric, firstly, measures correlation between nodes belonging to CL_j, then selects the node g^* possessing the maximum informative value m', computed according to the following equation:

$$g^* = arg \max_{g \in S_1^j}(m'_g), \quad (10)$$

where

$$m'_g = \left(\sum_{i \in S_1^j} \frac{\sigma_{ig}^2}{\sigma_g^2} \right). \quad (11)$$

In (11), σ_{ig} presents the covariance between the reading variable x_i of sensor node i and the reading variable x_g of sensor node g, and σ_g^2 presents the variance of x_g. S_1^j represents the set of nodes that are still not yet selected and the members with whom the inter-correlation of node g is calculated in order to select the most qualified representative member g^*. The use of the set S_1^j makes the nodes selection process iterative, where, in each iteration $n \in \{1, \ldots, N_{rep_j}\}$, as the sensor $g^*(n)$ is selected, it is removed from S_1^j to prepare the selection of the next representative node $g^*(n+1)$. Here, the removal of $g^*(n)$ from S_1^j cancels its impact on the rest of the sensor nodes that are still in S_1^j, which makes this algorithm effective and advantageous. Subsequently, to correctly select $g^*(n+1)$ without any distortion, the metrics m' should be recomputed following this removal. To proceed with the representative nodes selection procedure, we make use of the data matrix $X_{lp} = [x_{lp\,1}^{tr}, x_{lp\,2}^{tr}, \ldots, x_{lp\,N}^{tr}]^{tr} \in IR^{N \times T_{lp}}$ that we divide into J sub-matrices $X_{lp}^j \in IR^{cl_j \times T_{lp}}$, where X_{lp}^j contains data sent by nodes belonging to CL_j. Without loss of generality, for each cluster j and its corresponding data matrix X_{lp}^j, we perform the steps of the nodes selection process that have been outlined in [17, Section. VI] in order to get the set $\mathcal{N}_{rep_j}$. For each cluster CL_j, based on the correlation among the nodes belonging to CL_j and using (11), the sink keeps selecting and removing nodes from S_1^j to $\mathcal{N}_{rep_j}$, which is initially an empty set, until reaching the required number N_{repj}. Here, we denote the updated version of the CBMC as the Optimized Cluster-based MC data gathering approach (OCBMC). The only and unique difference here is that, with the OCBMC scheme, for each cluster j, the set $\mathcal{N}_{rep_j}$ of the representative nodes is neatly chosen from CL_j according to the correlation-based metrics (10). Whereas, with the CBMC, this set is randomly chosen from CL_j.

Usually, nodes are randomly scattered in the area to be monitored, without any infrastructure, leading to the existence of different network topologies, which are determined according to the nodes' locations and the connections between them and the sink node. Different topologies may exist in the WSNs, and they vary with the kind of application one wants to proceed. In the sequel, we consider the frequently used topologies, which are the star and the tree/mesh topologies with the twofold addressed scenario.

A. SINGLE-HOP STAR TOPOLOGY (SCENARIO ONE)
The star topology networks are single-hop systems [34] since all nodes operate as terminal devices and directly communicate with a centralized communication server, which is the sink. This type of architecture is generally used in wireless micro sensor networks as the covered area is, most of the time, small and limited by the communication range of the

end nodes. As we have previously stated, the first step in the network sampling proceeding is to partition nodes into J disjoint clusters. Performing this step is of prime importance to reach an adaptive and overall representation for the whole monitored area, and thus a more efficient data sampling. Benefiting from the dependency among nodes, the aforementioned node selection strategy targets to achieve a better data sampling quality and hence a much lower data reconstruction error at the sink node, despite the limited number of reported data readings with the addressed twofold data compression scenario. However, there is still a crucial factor that cannot be overlooked at all, and must be cautiously taken into consideration, which is the network lifespan and energy load balancing between nodes. Indeed, depending on the events to be monitored, even though the set of representative sensors may change from one detection period T to another, the signal in most WSNs is time-stationary. Hence, using (10), the set $\mathcal{N}_{rep}$ of selected representative nodes can remain the same for many successive detection periods. To avoid the overcharge that may occur over some continuously operating sensor nodes and thus the fast depletion of their batteries, the active node selection process should take into account not only correlation between sensor nodes but also their residual energies. Accordingly, we incorporate in (10) the fraction of the sensor residual energy as a complementary factor in order to choose the sensor node that can well represent the network and at the same time holds the higher residual energy. Precisely, for a given sensor node $g \in S_1^j$, the trade-off between its informative value m'_g, computed in (11), and its residual energy with regard to the other sensors' residual energies, Ef_{resd_g}, is achieved through a multiplication of the two considered factors. Thereby, (10) is replaced by (12) for our EAMC approach:

$$g^* = arg \max_{g \in S_1^j} \left(m'_g \times Ef_{resd_g} \right), \qquad (12)$$

where

$$Ef_{resd_g} = \frac{E_g}{\sum_{i \in S_1^j} E_i}. \qquad (13)$$

Thus, the EAMC represents an improvement of the OCBMC. The unique difference here is that, with the OCBMC scheme, the set $\mathcal{N}_{rep_j}$ is selected from cluster CL_j passing through the correlation-based cost function (10), whereas, with the EAMC for the first scenario, this set is selected from CL_j according to the combined energy-aware and correlation-based metric (12).

Performing (12) means that we attempt to choose the sensor node g that maximizes the combined metric $(m'_g \times Ef_{resd_g})$. Here, multiplying the two addressed factors unites them into a one single entity, and it is analogous to computing the needed correlation per unit of energy. In other words, this operation makes the relation between the two factors fusional. If one of them is weak it will automatically weaken the other, and the carrier sensor node will not be chosen. Since the residual energy of the operating nodes decreases from one detection

period to another, the metrics $(m'_g \times Ef_{resd_g})_{g \in S_1^j}$ vary and the representative nodes will be selected efficiently, according to the available energy in their batteries.

In order to get the set $\mathcal{N}_{rep_j}$ of the EAMC, we perform the same steps of the nodes selection process as before , while replacing only the metric (10) by the metric (12).

B. MULTI-HOP MESH TOPOLOGY

Compared to the star topologies, the mesh network does not suffer from the limited scalability. Thus much wider area can be covered and monitored thanks to the multi-hop transmissions. In this type of networks, several routes may exist between sensor nodes and the sink, and most of the time the network software chooses the shortest one for data delivery. To forward the data towards the sink, we opted for the shortest path tree, implemented with Dijkstra algorithm. According to [35], paths that are established by Dijkstra algorithm usually present a lower number of connections, hence, the average delay of message dissemination decreases and the energy management is reinforced. Note that the routing protocol to use is not the main focus of this paper since our aim is to achieve energy load balancing between nodes and reach a higher lifetime of the network with the already established routes. Updating the paths systematically according to the remaining energy level in order to further prolong the network lifetime is left as a perspective for future works.

In light of the importance of energy utilization enhancement, as far as the size of these networks gets bigger and the diameter of the covered area gets larger, the problem of uneven energy depletion aggravates and gets worse. In fact, data packets, which are generated by the source nodes, have to be relayed via intermediate nodes to be finally routed to the sink. Accordingly, nodes close to the sink are susceptible to carry much heavier traffic loads than nodes of outer-regions. Consequently, they would speedily run out of power, leading to the problem of energy hole around the sink. In this case, even if the rest of nodes, specially the border ones, still hold sufficient energy, communication with the sink would be cut off, causing probably the end of the network lifespan.

1) THE TWOFOLD COMPRESSION PATTERN (SCENARIO TWO)

To alleviate the overwhelming issue of energy hole, nodes' residual energies should be considered when selecting the set of representative nodes $\mathcal{N}_{rep}$. When all the nodes are directly connected to the sink, as in the star network topology, performing the selection cost function (12) is effective enough to attain the purpose of this paper. Yet, when the data have to be forwarded via relaying nodes to reach the destination, taking into account only the source node residual energy is completely insufficient. Instead, the residual energy level of all the relaying nodes that would participate in the data forwarding should be assessed. Indeed, considering the entire route cannot be addressed using (12) in this kind of networks, otherwise we will select sensor nodes with the higher residual

energy, while ignoring the continuity ability of the entire route. Suppose a sensor node g^*, holding the maximum value of the combined metric (i.e. correlation-energy), is selected and there is a relaying node with a used up battery in its route towards the sink. In this case, the path will be cut off announcing probably the end of the network lifetime. Therefore, in addition to the correlation, a node is chosen to be a representative one under the condition that there is no depleted relaying node in its route. That is, the metric (14) is chosen for our EAMC for the second scenario:

$$g^* = arg \max_{g \in S_1^j} \left(m_g'^2 \times \frac{E_g \times \min_{hp_g \in \mathcal{HP}_g} \left(E_{hp_g} \right)}{(\sum_{i \in \mathcal{N}} E_i)^2} \right), \quad (14)$$

where $\mathcal{HP}_g$ represents the set of nodes composing the route of the source node g towards the sink.[2] In (14), adding the term $(\min_{hp_g \in \mathcal{HP}_g}(E_{hp_g}))$ means that we take into account also the relaying node with the lowest residual energy in the representative node selection process in order to avoid the fast depletion of the routes and hence the network partition, while there are still nodes with sufficient remaining energy that can forward data. Here, if the energy level of the relaying node hp_g that is belonging to the route of node g towards the sink is very low compared to other nodes, the combined entity value will be weakened, and the node g won't be chosen as a representative node for the current detection period T. As we can notice, in this energy-aware cost function, we have strengthened the weight of the factor m_g', which reflects how much the sensor g can represent the network, in order to maintain a good/efficient recovery quality. It will be shown in the simulations section that the introduced cost function is able to achieve an interesting and satisfactory trade-off between the data recovery quality and the network lifespan.

2) THE SINGLE-LEVEL COMPRESSION SCENARIO (SCENARIO THREE)

Mainly, the above investigated twofold scenario was designed to overcome the continually growing dense WSNs. Nevertheless, in the case of extended networks, without being too much dense, there is no need to make a significant number of sensor nodes completely inactive, for the entire current detection period, when executing data sensing. Accordingly, in this part, we won't pass through the selection of a set of representative sensor nodes. Instead, we proceed directly for the source nodes schedule. Furthermore, we want to evaluate our approach under the ordinary data sampling scenario as well in order to provide an overall work, where nodes can participate at least once during one detection period T. To do so, in each time slot t, using the same shared percentage pct_m, m_j source nodes are directly selected from the set CL_j of nodes composing the cluster j, according to (14), to sense the field and transmit their data readings to the sink. That is, instead

---

[2]Note that $\mathcal{HP}_g$ contains only the relaying nodes and neither the source node g nor the sink belongs to it.

of (9) we have:

$$m = \sum_{j=1}^{J} m_j, \quad \text{where } m_j = pct_m\% \times cl_j. \quad (15)$$

Here, we proceed as if we set $pct_{Nrep} = 100$ and all the nodes are representative for the network. Certainly, there will be more computation than the twofold scenario, where the active node selection process, via (14), is effectuated once for the entire detection period T. Fortunately, the one that is responsible for all that calculation is the sink, which is free of energy constraint. In fact, we assume that the sink node has all the information regarding the sensor nodes' locations. Thus, it can compute, in advance, the energy to be consumed by the nodes for data sensing and forwarding. Thereupon, it is able to schedule beforehand the participation of the nodes during the entire detection period T. In order to not increase again the communication overhead, the sink informs the concerned nodes about their data sensing schedule at the beginning of the detection period T, i.e. we designate only one-shot scheduling transmission for the entire detection period T.

Since, in each time slot t, energy consumption is uneven between nodes due to the multi-hop systems configuration, over a period of time, we outface some sensor nodes whose routes hold relaying nodes with much lower residual energy than other nodes. Performing (14) will keep these nodes out of the selection range for several successive time slots, until other nodes take their places. The fact of not being selected as a source node for successive time slots and not reporting data to the sink leads to the existence of successive missing data in the received data matrix M. This sequence of missing entries that may exist in the rows, referred to as a row structure fault in [8], impedes the MC resolution and highly increases the data reconstruction error. Therefore, for this single-level compression scenario, an extra step is added to the three-stage reconstruction pattern and set at the beginning of the recovery process, in order to detect the rows that hold structure faults and consider them as completely empty rows. This step consists simply in finding the sequence of successive zero entries holding a length larger than a given fixed size, which represents the minimum size of successive data missing from which that sequence is considered as a structure fault. That is:

$$StrFault_{min} = pct_{strF}\% \times T, \quad (16)$$

where pct_{strF} represents the parameter that fixes the minimum size of successive data missing from which that sequence is considered as a structure fault in accordance with the duration T of the detection period. It will be shown in the simulation part that treating separately the rows that hold structure faults significantly improves and refines the data reconstruction accuracy.

V. NUMERICAL RESUTLS

In [7], we had compared the performance differences between our proposed approach CBMC and the closest scheme of [8]
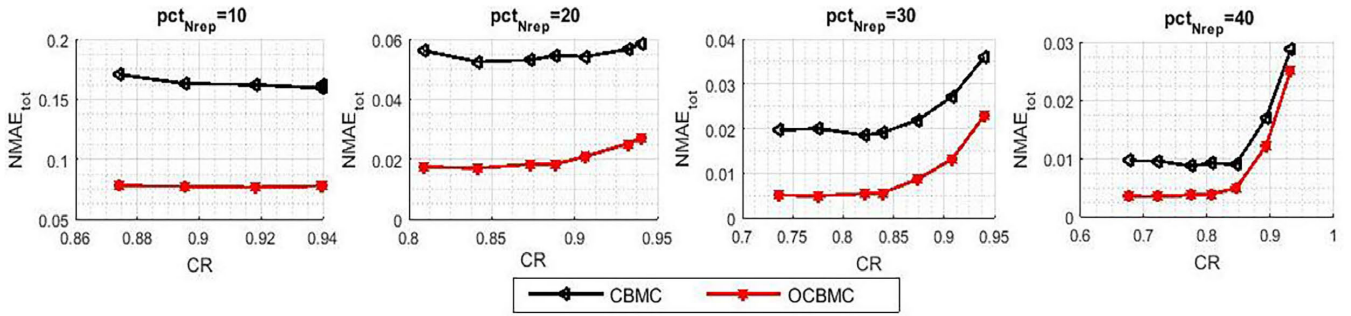
FIGURE 1. *NMAE_tot* for the CBMC and the OCBMC.

that had treated a relatively similar scenario to our twofold data loss one. We had found that our structured approach outperforms the closest one in terms of both data reconstruction error and overall network energy consumption. For that reason, in this work, we have been based on this comparison to carry on with our structured scheme and improve its design and techniques. Firstly, we simulate the OCBMC, where a cost selection function (10) based on the correlation among nodes has been introduced, with its previous version, the CBMC of [7], where the representative nodes are randomly chosen according only to the clusters consideration strategy. This simulation will highlight the enhancement of the data reconstruction accuracy brought by the OCBMC and prepare for the next simulation, which is the main focus of this paper. Secondly, the developed version EAMC, where energy is jointly taken into account with the correlation criteria, is evaluated facing its previous version, the OCBMC scheme. This simulation will reveal the impact of the updated selection cost function on the trade-off between the data recovery accuracy and the network lifetime under all the investigated scenarios and for the two types of sensor nodes. Finally, we summarize and confirm our results by measuring differently this trade-off in such a way that the compared approaches are pushed towards their limits.

To estimate the data reconstruction accuracy for the implemented schemes, we opted for the following metrics:

- *NMAE_tot*: The Normalized Mean Absolute Error on all the missing entries [14], [30], [36]:

$$NMAE_{tot} = \frac{\sum_{i,t:\Omega_M(i,t)=0} |X(i,t) - \widehat{X}(i,t)|}{\sum_{i,t:\Omega_M(i,t)=0} |X(i,t)|}, \quad (17)$$

where, X and $\widehat{X}(i,t)$ represent respectively the complete raw data matrix before sampling and the finally recovered one. The *NMAE_tot* calculates the error ratio on all the missing data values that are interpolated by the sink, including the partially missing data of the representative nodes and the fully missing data of the inactive nodes. Namely, the *NMAE_tot* measures the data recovery errors only on the non transmitted data entries $X(i,t)$, where $\Omega_M(i,t) = 0$.

- CR: The Compression Ratio:

$$CR = \frac{N - m}{N}. \quad (18)$$

To simulate the implemented schemes and evaluate their performance under different *CRs*, we vary pct_{Nrep} from 10 to 60, and for each given pct_{Nrep}, we vary pct_m from 10 to 80. Regarding the network parameters setting, we consider that $N = 50$ sensor nodes are randomly deployed in a square observation area of size $100m \times 100m$, and we monitor the WSN throughout a detection period of length $T = 100$ time slots.

The first simulation targets to prove the benefits brought by introducing the correlation-based deterministic cost selection function (10) on the data recovery accuracy. Moreover, since this work in an extension to [7], we compare in Figure 1 the updated OCBMC with the original version CBMC for the twofold data loss scenario with the multi-hop mesh network topology (scenario two). For both schemes, the J detected clusters are considered when choosing the representative nodes as well as when assigning the source ones. Besides, both of them perform the three-stage data reconstruction pattern of section III to recover the missing data. The unique difference is that in the OCBMC, the set $\mathcal{N}_{rep_j}$ is selected from cluster CL_j passing through the cost function (10), whereas in the CBMC, this set is selected randomly from CL_j. As expected, the data reconstruction error $NMAE_{tot}$ is notably reduced, especially for the small values of pct_{Nrep}. This confirms the fact that the introduced correlation based-metric (10) neatly selects the sensor nodes with the higher informative data values with respect to the other non selected ones, which nicely improves the data reconstruction quality.

We focus now on the principal purpose of this paper, which is the network lifetime improvement. In the simulations hereafter, we analyze the performance of the EAMC that takes into account the sensors' residual energies in the selection cost function. For each scenario and network topology, the adequate energy-aware cost function is considered for the EAMC. Namely, we evaluate the metric (12) for the single-hop star topology and the metric (14) for the multi-hop mesh topology. Moreover, we assess the proposed approach under both of sensor nodes types; the ordinary sensors, where

TABLE 1. Simulation Parameters for energy consumption.

Parameter	Ordinary sensor node	Hungry power sensor node
E_{init}	0.8 J	40 J
I_{sens}	50 μA [38]	25 mA [6]
T_{sens}	0.5 mS [6]	0.5 mS [6]
V_{sup}	2.25 V [38]	2.7 V [6]
$E_{elec-tr}$	50 nJ/bit [39]	
$E_{elec-rc}$	5 nJ/bit [39]	
ε_{amp}	100 $pJ/bit/m^2$ [6]	
L	1024 $bits$	

the energy consumed in data sensing is quite low, and the specific power hungry ones, where the acquisition energy costs can even be greater than that of the radio costs [37]. The parameters' setting values for the used energy consumption model, detailed in section III, are outlined in the table 1.

To begin, we consider the single-hop star network and we compare the EAMC approach to the OCBMC. Figures 2 and 3 depict the trade-off between the $NMAE_{tot}$ and the network lifetime. The network lifetime denotes the number of detection period T that a scheme can achieve without causing the death of any sensor node in the network, i.e. Nb_{rounds}. Indeed, the first node that exhausts all its battery energy announces the death of the network and determines its lifetime Nb_{rounds}. Note that for each case when we vary the compression ratios pct_{Nrep} and pct_m, the $NMAE_{tot}$ and the Nb_{rounds} are simultaneously calculated then depicted in Figures 2 and 3 for both compared approaches. Moreover, the final depicted $NMAE_{tot}$ represents the average of all the resulting $NMAE_{tot}$ during the ensured Nb_{rounds}.[3] Obviously, the impact of the correlation on the data recovery quality will be lightened, when we jointly take into account the residual energy, as a second weighty factor, with the correlation criteria in the cost selection function. However, in this paper, we target to reach a robust and equitable compromise between the two addressed factors. As we can note, we still achieve a sufficiently good data recovery accuracy even for the small values of pct_{Nrep} (i.e. when there is a significant number of completely empty data rows in M), while at the same time the network lifetime is highly improved. As an example, with the ordinary sensor nodes, for ($pct_{Nrep} = 20, pct_m = 10$), the $NMAE_{tot}$ passes from 0.023 to 0.037 with the EAMC, while the network lifetime is expanded with a percentage of 80, 652% (i.e. Nb_{rounds} passes from 43.34 to 224.04 rounds). On the other hand, with the hungry power sensor nodes, for ($pct_{Nrep} = 20, pct_m = 10$), the $NMAE_{tot}$ passes from 0.029 to 0.038 with the EAMC, while the network lifetime is expanded with a percentage of 75, 22% (i.e. Nb_{rounds} passes from 22.52 to 90.8 rounds). Moreover, we can notice that as the number of active nodes is increased, the gap of $NMAE_{tot}$ between the two

[3]The $NMAE_{tot}$ and the Nb_{rounds} of the simulations of Figures 4, 5, 6, 7 and 8 have been calculated following the same manner.

compared algorithms is significantly reduced, whereas that of the network lifetime is still clearly noteworthy. Precisely, with the hungry power sensor nodes, starting from $pct_{Nrep} = 40$, we start to reach gains on the network lifetime almost without losing costs in counterpart on the $NMAE_{tot}$.

Moving now to the second scenario, which is the twofold data compression in the multi-hop mesh network topology. For the ordinary sensor nodes, where the consumed energy during data detection is quite low compared to that used for data transmission, the studied trade-off between the $NMAE_{tot}$ and the Nb_{rounds} is illustrated in Figure 4. Particularly, as we can see, we keep intentionally considering the performance comparison of the $NMAE_{tot}$ that has been performed in the first simulation, i.e. Figure 1. Interestingly, it is noteworthy that even though the $NMAE_{tot}$ is slightly increased when considering the sensor residual energy in the metrics (12) and (14), it is still quite inferior to that given by the original proposed CBMC, when the correlation criteria is not taken into account. Moving to the performance comparison between the metrics (12) and (14), expectedly and as we have already explained in the previous section, the rending of (12), when it is performed in this scenario with this type of sensor node, is limited in terms of Nb_{rounds}, even though it gives the closest $NMAE_{tot}$ to the OCBMC. Since the entire route is considered with (14), sensor nodes having depleted relaying nodes in their paths towards the sink are less susceptible to be selected as representative for successive detection periods which restricts the degree of freedom in the choice of the set $\mathcal{N}_{rep}$ compared to the metric (12) and thus degrades the averaging of the $NMAE_{tot}$. Here, it is worth mentioning that as long as we keep reaching a sufficiently good recovery quality (i.e. a low $NMAE_{tot}$), we privilege the second crucial factor that is the network lifetime expanding. As we can see in Figure 4, for pct_{Nrep} equals to 60, the resulting data recovery error, when we perform the metric (14), is almost the same as that, when we use (12), whereas the Nb_{rounds} ensured by the retained metric (14) is higher than that given by (12). For example for ($pct_{Nrep} = 60, pct_m = 10$), with the same $NMAE_{tot}$, performing the EAMC using the cost function (14) can prolong the network lifetime with a percentage of 35.69% compared to the OCBMC, whereas, the rending of (12) is limited to 8.7%. This is due to the fact that the metric (14) focuses on the sensor node while taking into account the entire route through the value of ($\min_{hp_g \in \mathcal{HP}_g}(E_{hp_g})$). Hence, this technique is able to ensure a much longer lifetime for the network, when the sensor nodes are ordinary ones. Another example, for $pct_{Nrep} = 40$ and $pct_m = 10$, by increasing the $NMAE_{tot}$ from 0.025 to only 0.029, the metric (14) can prolong the network lifetime with a percentage of 47.28%.

Nevertheless, when the deployed sensor nodes are hungry power ones, as it has been depicted in Figure 5, the performance of both metrics (12) and (14) become very close. Indeed, with this type of sensor nodes, the amount of energy that is consumed in data forwarding by the relaying nodes becomes much less than that consumed in sensing by the assigned source nodes. Consequently, both metrics tend to
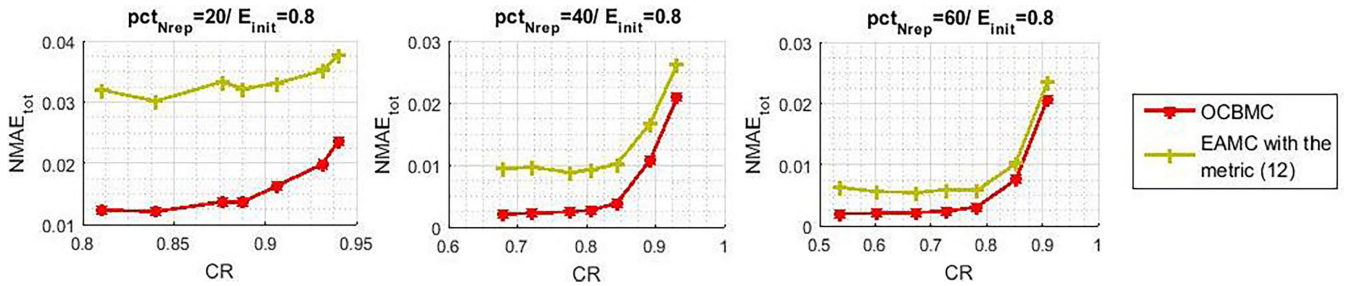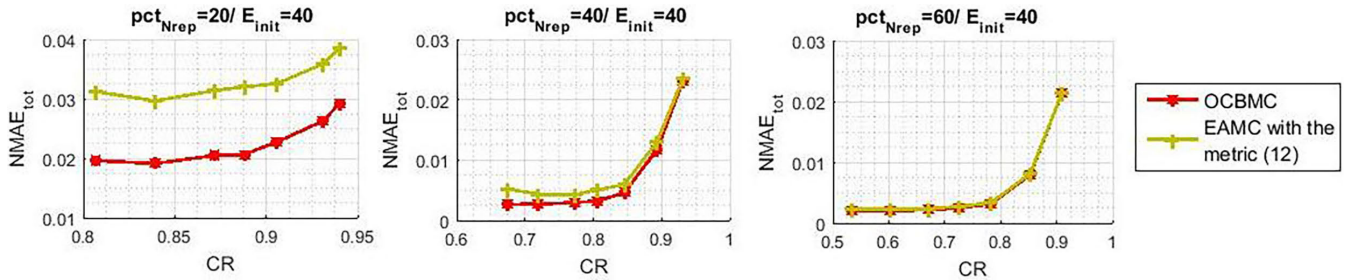
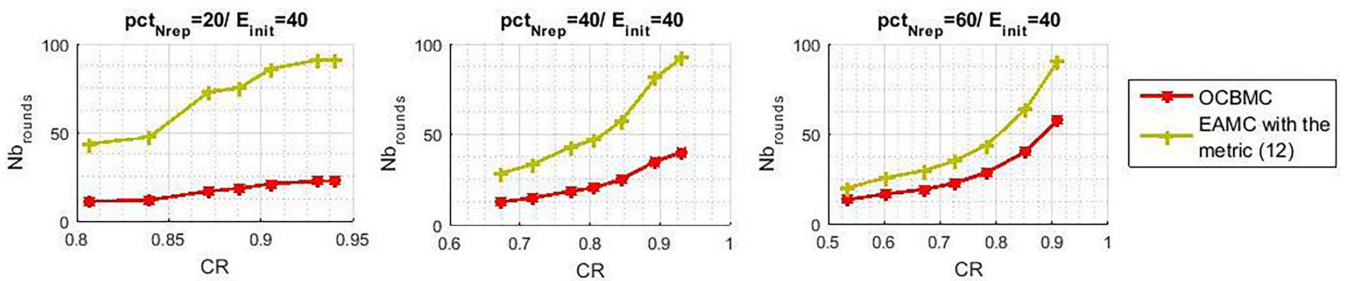(a) The $NMAE_{tot}$ for both OCBMC and EAMC approaches.
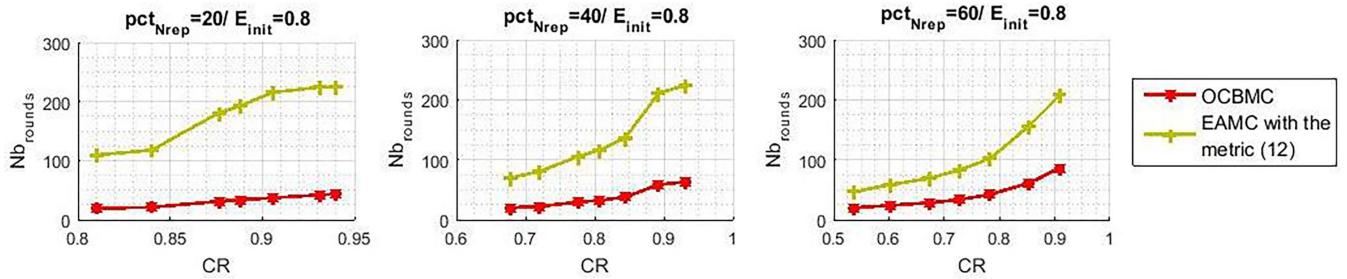
(b) Nb_{rounds} for both OCBMC and EAMC approaches.

FIGURE 2. Performance trade-off between the data reconstruction error and the network lifetime for both OCBMC and EAMC approaches in the single-hop star topology with ordinary sensors.

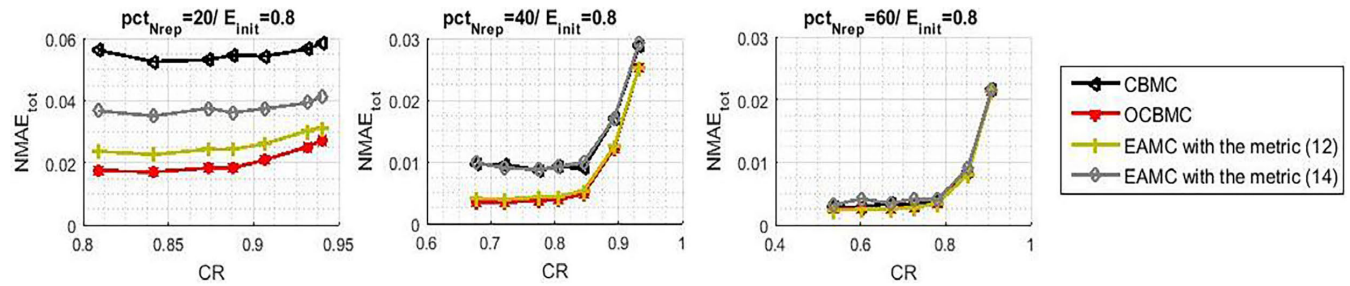(a) The $NMAE_{tot}$ for both OCBMC and EAMC approaches.

(b) Nb_{rounds} for both OCBMC and EAMC approaches.

FIGURE 3. Performance trade-off between the data reconstruction error and the network lifetime for both OCBMC and EAMC approaches in the single-hop star topology with the hungry power sensors.

choose the same set of representative nodes since the amount E_g (i.e. the residual energy of the source node of interest g) represents, with this type of nodes, the most important and the dominant component that heads the active node selection process. We can distinctly note the very significant improvement brought by the EAMC with (12) and (14) compared to the OCBMC in terms of network lifetime, especially for the

high CRs. As an example, for ($pct_{Nrep} = 40, pct_m = 10$), we can reach an amelioration of 124.4% in terms of Nb_{rounds} with both metrics, while still maintaining nearly the same $NMAE_{tot}$ compared to the OCBMC.

In order evaluate the scalability of the proposed solution in Figure 6, we have compared the performance of the implemented schemes, for the scenario of Figure 4, with respect
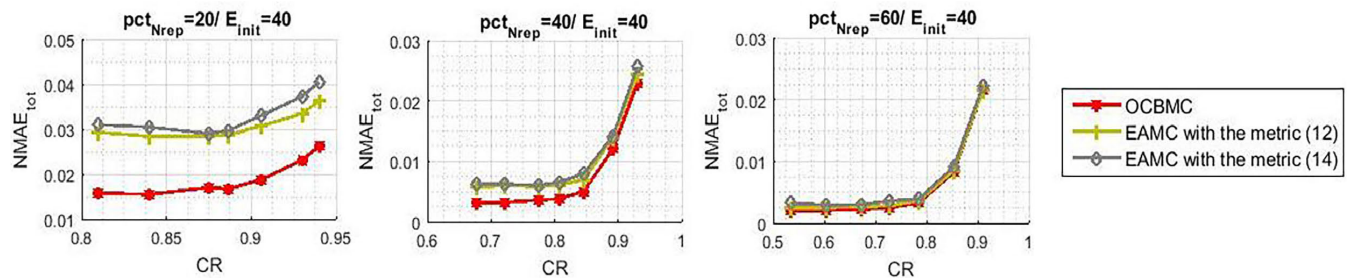
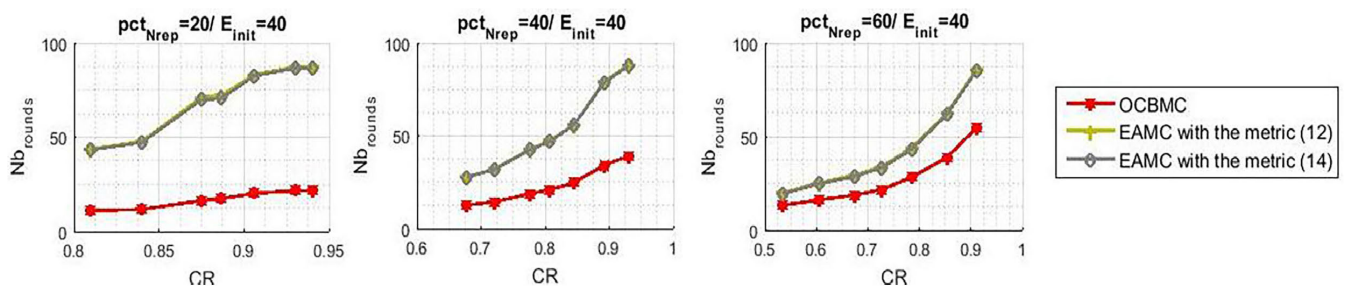(a) The $NMAE_{tot}$ for the CBMC, the OCBMC and the EAMC.
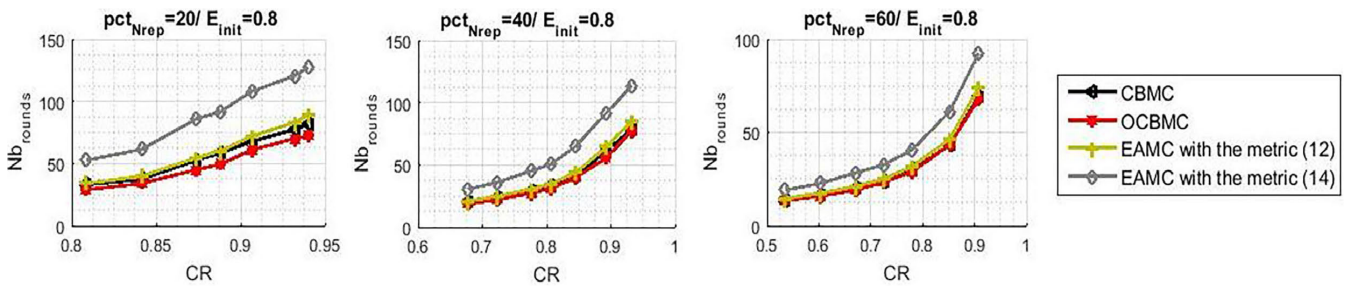
(b) Nb_{rounds} for the CBMC, the OCBMC and the EAMC.

FIGURE 4. Performance trade-off between the data reconstruction error and the network lifetime for the compared approaches in the twofold compression scenario and multi-hop mesh topology with ordinary sensors.

(a) The $NMAE_{tot}$ for both OCBMC and EAMC approaches.

(b) Nb_{rounds} for both OCBMC and EAMC approaches.

FIGURE 5. Performance trade-off between the data reconstruction error and the network lifetime for the compared approaches in the twofold compression scenario and multi-hop mesh topology with hungry power sensors.

to the number of sensor nodes. To this end, for a given *CR* with $pct_{Nrep} = 50$ and $pct_m = 10$, we have varied the number N of the deployed sensor nodes, and for each case we have measured the $NMAE_{tot}$ and the Nb_{rounds}. As we can note from this simulation, the EAMC with the metric (14) keeps outperforming both the EAMC with the metric (12)
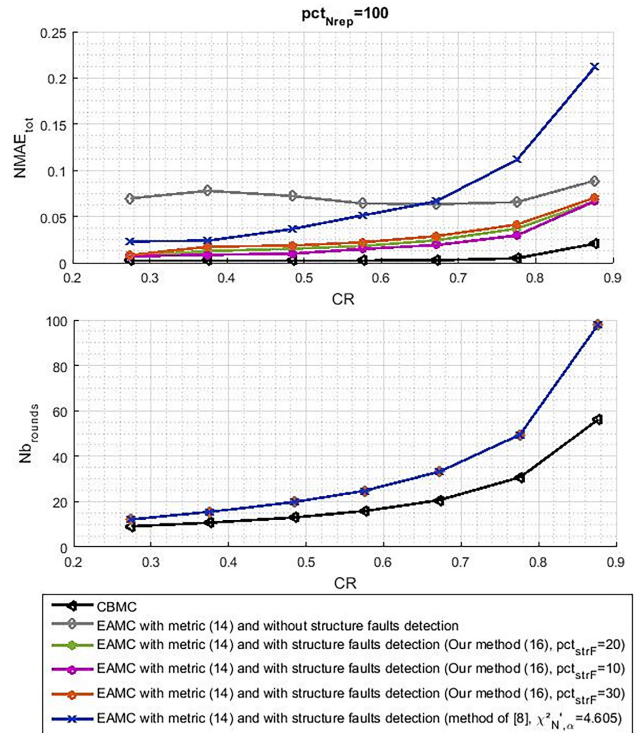
and the OCBMC in terms of Nb_{rounds} for the different values of N, although it increases a little bit the $NMAE_{tot}$. These simulations confirm the scalability of the proposed scheme and show that it keeps the same behaviors, even when we increase or decrease the number of sensor nodes N. Note that the data recovery error is reduced, for all the compared
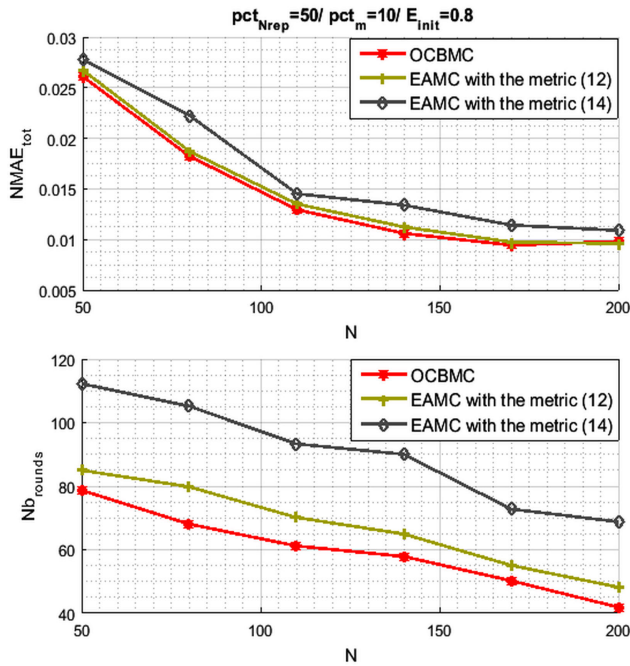
FIGURE 6. Performance trade-off between the data reconstruction error and the network lifetime for the compared approaches in the twofold compression scenario and multi-hop mesh topology with ordinary sensors and with respect to the number of sensor nodes.

FIGURE 7. Performance trade-off between the data reconstruction error and the network lifetime for the compared approaches in the single-level compression scenario and multi-hop mesh topology with the ordinary sensors.

schemes, as N is raised since the MC-based reconstruction methods work very well for the large-scale data estimation problems. Regarding the network lifetime, we can note that it is reduced as N is raised since the overall energy consumption is increased due to packet relaying.

In scenario three, whose results are depicted in Figures 7 and 8, we have compared the EAMC scheme using the metric (14) with its original version, the CBMC, for $pct_{Nrep} = 100$. Indeed, since in the WSNs, most of the time, the signal is time-stationary, using only the correlation criteria via the OCBMC to search for the m source nodes in each time slot t leads to probably having the same source nodes during all the detection period T. The resulting configuration entails a schedule, where the same chosen source nodes will transmit their data during all the time slots t composing the detection period T, while the rest of nodes remain completely inactive. The OCBMC is not suitable for this scenario since, in this part, we aim to address an ordinary data sampling scenario, where nodes can participate on data sensing and transmission at least once during one detection period T. As we can note from Figure 7, the $NMAE_{tot}$ brought by the EAMC (i.e. the gray curve) gets worse compared to the original CBMC, despite the amelioration that is achieved in terms of Nb_{rounds}. This is due to the existence of the row structure faults, which appear in the data vectors corresponding to the nodes that are remaining outside the range of selection for several successive time slots. In addition to the fully empty data rows, the structured data losses as the structure faults are among the serious obstacles that not only impede the MC

resolution but also pollute the received data [8]. For that reason, before applying the MC method, the rows holding these structure faults should be removed from M then recovered through stage two if the node corresponding to this row is an absent node or recovered through stage three if the node is an Isolated one. In [8, Section. V], authors have proposed an algorithm that detects rows holding structure faults. This technique is implemented here and the resulting performance are depicted with the dark blue curve. As we can note from the figure legend, this technique is dependent to two different parameters N' and α. Altogether, these parameters give $\chi^2_{N',\alpha}$, which represents the upper α percentage point of the chi-square distribution with the degree of freedom N'. Although they had shown how to choose N', the selection of α value has been done without any explanation, and according to our several simulations, it should be noted that the slightest variation of any of these parameters makes an important difference in the $NMAE_{tot}$ performance. Here, the value of α has been determined empirically.[4] As for our proposed structure faults detector (16), it has been evaluated with respect to a meaning parameter pct_{strF} that, in accordance with the detection period duration T, fixes the minimum size of successive data missing from which that sequence is considered as a structure fault. Surprisingly, we can clearly note from Figure 7 that, despite its simplicity, our proposed method can significantly
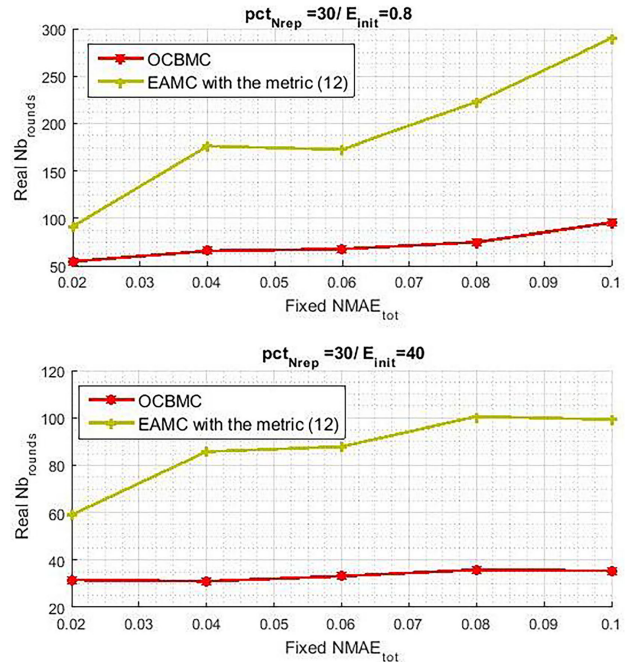
---

[4] We report here the simulation of only the most performing value.
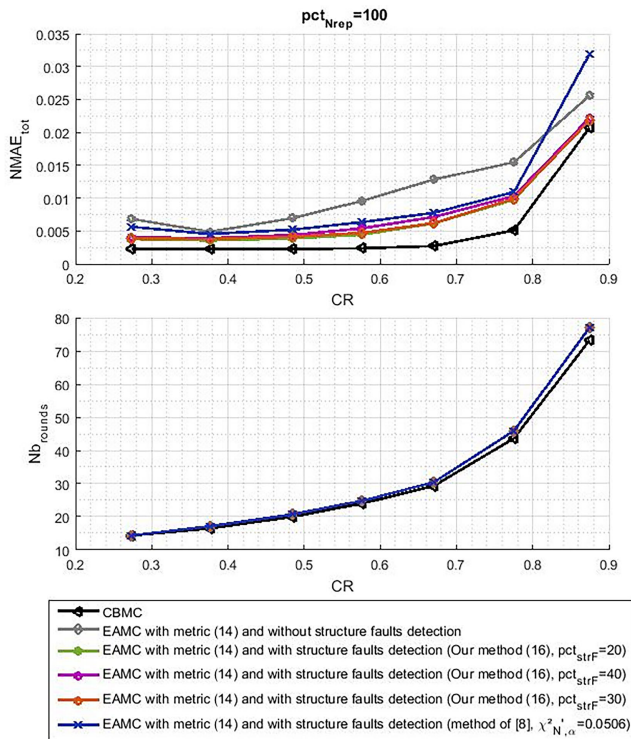
FIGURE 8. Performance trade-off between the data reconstruction error and the network lifetime for the compared approaches in the single-level compression scenario and multi-hop mesh topology with the hungry power sensors.
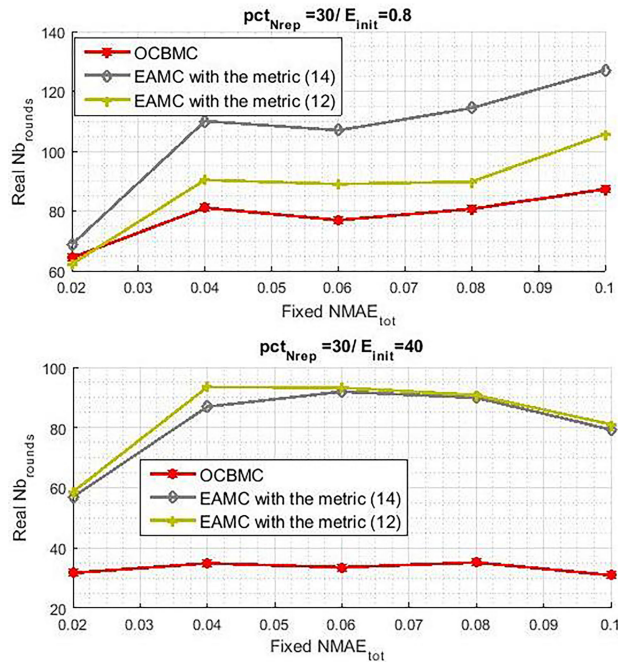

FIGURE 9. Real network lifetime vs. fixed upper bound data recovery error ratio for the compared approaches in the single-hop star topology with both types of sensor nodes.

reduce the data recovery error of our EAMC for the entire range of CR, while still keeping the same Nb_{rounds}, whereas, the improvement brought by the technique of [8] is limited by a restricted range of CR. Moreover, the $NMAE_{tot}$ given by the EAMC with our structure fault detection method is not only lower than that given by the EAMC without structure fault detection, but also it is lower than the $NMAE_{tot}$ resulting from the EAMC with the method of [8], which unfortunately makes the data recovery accuracy worse for the high CRs (i.e. $CR > 0.66$). In fact, using the same parameter value α for both high and low CRs impedes the imposed threshold for the structure faults detection from maintaining a low error ratio across the entire range of CR. These simulations show that our technique is not only simpler than that proposed in [8] but also it is more efficient.

Figure 8, which depicts the simulations obtained with the hungry power sensors, shows a significant decrease of the $NMAE_{tot}$ in the curves corresponding to EAMC compared to those of the ordinary low power sensors of Figure 7. Since the amount of consumed energy in data forwarding by the relaying nodes becomes much less than that consumed in sensing by the assigned source nodes, we do not have paths that run out quickly and the choice of active source node becomes directed by the energy of the node in question not the relaying nodes composing its path towards the sink. As a result, not only is the number of the appeared structure faults is reduced, so are their sizes. For that reason, we had been

obliged to vary again the parameter α in this simulation as the previous one of Figure 7 was not suitable for the present case. As for the network lifetime performance, unexpectedly, for this kind of data compression scenario and when the deployed sensors are hungry power ones, the energy consumption is too great that we can make significant improvements in terms of Nb_{rounds}.

In the final part of the simulations section, the trade-off between the data recovery error $NMAE_{tot}$ and the network lifetime, measured with Nb_{rounds}, has been investigated in other respects in such a way that it may reveal a closer performance to the reality. For a given $pct_{Nrep} = 30$, an error ratio upper bound is fixed. Here, the *Fixed NMAE*$_{tot}$ is varied from 0.02 to 0.1, and for each value we compute the maximum number of detection periods that the performed scheme can ensure, despite the possible existence of dead sensor nodes. As long as the implemented scheme can achieve an $NMAE_{tot}$ lower or equal to the upper bound fixed one (i.e. *Fixed NMAE*$_{tot}$), the *Real Nb*$_{rounds}$ keeps growing and the moment where the data recovery error ratio of this scheme exceeds the *Fixed NMAE*$_{tot}$ value we consider the network as dead. Doubtless, if a sensor node runs out of energy, it can no longer participate in data sensing and forwarding.

Likewise the previous simulations, we start with the single-hop star topology. We can note from Figure 9 that the EAMC, using the metric (12), can highly prolong the network lifetime compared to the OCBMC, while still able to maintain a data reconstruction quality better than the imposed level. Another interesting point that we can note from this comparison is as follows; as far as the upper bound error ratio is raised, the area

FIGURE 10. Real network lifetime vs. fixed upper bound data recovery error ratio for the compared approaches in twofold compression scenario and multi-hop mesh topology with both types of sensor nodes.

of improvement that the EAMC can provide for the network lifetime is larger than that provided by the OCBMC. In other words, the *Real Nb$_{rounds}$* achieved by the EAMC increases faster than that achieved by the OCBMC, when the upper bound error ratio is expanded.

Figure 10 depicts the obtained results for the twofold data compression in the multi-hop mesh network topology. As we can notice, the obtained performance results confirm those of Figure 4 and Figure 5 and prove the efficiency of the re-updated EAMC with both types of sensor nodes. Particularly, we can perceive the existence of drop points in the plots, namely, *Fixed NMAE$_{tot}$* = 0.06 with the ordinary sensor nodes and *Fixed NMAE$_{tot}$* = 0.1 with the hungry power sensor nodes. Since we perform a discretized *CR* to reach an *NMAE$_{tot}$* that is less than or equal to the fixed one, i.e. *pct$_m$* varies from 10 to 70 with step of 10, there is possibility that a *CR* under the needed one is used (*pct$_m$* above the needed one), and some specific nodes may untimely die. Consequently, the achieved *NMAE$_{tot}$* increases and exceeds the upper bound *Fixed NMAE$_{tot}$*, which causes the death of the network. The latter explanation denotes that the piecewise effect appears in these points. As we can note, the appearance of these particular points is clearer with the case of the hungry power sensor nodes, where the data sensing activities carried out by the source nodes consume the bulk of the total power consumption and the selection of the active nodes is mainly based on the residual energy of the sensor of interest. Nevertheless, note that what mostly interests us is the difference, in terms of network lifetime performance, between the curves that still exists even with the drop points.

VI. CONCLUSION

In light of the importance of the energy saving and lifetime for wireless sensor nodes that are suffering from a limited power capacity, in this paper, we have presented an adaptive data collecting scheme, called the EAMC. Since the all-node-active condition is completely impractical for the WSNs with the energy constraints, the proposed approach can dynamically designate the source sensor nodes that can afford the sustainability as long as possible of the network lifetime and alleviate the problem of energy load imbalacing, according to an energy-aware cost selection function. Additionally, the proposed EAMC scheme aims not only for achieving the energy efficiency for the network but also for preserving a sufficiently good quality of data reconstruction as it takes into account the correlation criteria among sensors in order to select those who can best represent the network. These nodes are qualified to report more information to the sink, despite the addressed high compression ratio. Furthermore, we have evaluated our approach under different network topologies and scenarios, while performing, in each time, the adequate energy-aware metric. Moreover, to recover the entire data matrix, despite the existence of a significant number of completely missing rows corresponding to the inactive nodes, we have relied on the three-stage data reconstruction framework of our previous work [7]. For the third addressed scenario, to refine and enhance the data recovery quality, we have added a simple step to the data recovery techniques, which efficiently detect the structure faults that may appear in the received data matrix. Simulations have proven that the proposed scheme can achieve an interesting trade-off between the data reconstruction accuracy and the network lifetime compared to the approach of comparison. Accordingly, the EAMC scheme can be considered as significant for research in the field of energy saving since, particularly, it is able to efficiently overcome the twofold data loss scenarios.

REFERENCES

[1] Business Wire. (Jun. 2019). *The Growth in Connected IoT Devices is Expected to Generate 79.4zb of Data in 2025, According to a New IDC Forecast.* [Online]. Available: https://www.businesswire.com/news/home/20190618005012/en/

[2] Fortune Business Insights. (Nov. 2019). *Market Research Report.* [Online]. Available: https://www.fortunebusinessinsights.com/industry-reports/internet-of-things-iot-market-100307

[3] G. A. Akpakwu, B. J. Silva, G. P. Hancke, and A. M. Abu-Mahfouz, "A survey on 5G networks for the Internet of Things: Communication technologies and challenges," *IEEE Access*, vol. 6, pp. 3619–3647, 2018.

[4] W. Ejaz, A. Anpalagan, M. A. Imran, M. Jo, M. Naeem, S. B. Qaisar, and W. Wang, "Internet of Things (IoT) in 5G wireless communications," *IEEE Access*, vol. 4, pp. 10310–10314, 2016.

[5] M. Razzaque and S. Dobson, "Energy-efficient sensing in wireless sensor networks using compressed sensing," *Sensors*, vol. 14, no. 2, pp. 2822–2859, Feb. 2014.

[6] M. N. Halgamuge, M. Zukerman, K. Ramamohanarao, and H. L. Vu, "An estimation of sensor energy consumption," *Prog. Electromagn. Res. B*, vol. 12, pp. 259–295, Sep. 2009.

[7] M. Kortas, O. Habachi, A. Bouallegue, V. Meghdadi, T. Ezzedine, and J.-P. Cances, "Energy efficient data gathering schema for wireless sensor network: A matrix completion based approach," in *Proc. Int. Conf. Softw., Telecommun. Comput. Netw. (SoftCOM)*, Sep. 2019, pp. 1–6.

[8] K. Xie, X. Ning, X. Wang, D. Xie, J. Cao, G. Xie, and J. Wen, "Recover corrupted data in sensor networks: A matrix completion solution," *IEEE Trans. Mobile Comput.*, vol. 16, no. 5, pp. 1434–1448, May 2017.

[9] M. Kortas, A. Bouallegue, T. Ezzeddine, V. Meghdadi, O. Habachi, and J.-P. Cances, "Compressive sensing and matrix completion in wireless sensor networks," in *Proc. Int. Conf. Internet Things, Embedded Syst. Commun. (IINTEC)*, Oct. 2017, pp. 9–14.

[10] J. Tan, W. Liu, T. Wang, N. N. Xiong, H. Song, A. Liu, and Z. Zeng, "An adaptive collection scheme-based matrix completion for data gathering in energy-harvesting wireless sensor networks," *IEEE Access*, vol. 7, pp. 6703–6723, 2019.

[11] J. Tan, W. Liu, M. Xie, H. Song, A. Liu, M. Zhao, and G. Zhang, "A low redundancy data collection scheme to maximize lifetime using matrix completion technique," *EURASIP J. Wireless Commun. Netw.*, vol. 2019, no. 1, p. 5, 2019.

[12] S. Kim, C. Cho, K.-J. Park, and H. Lim, "Increasing network lifetime using data compression in wireless sensor networks with energy harvesting," *Int. J. Distrib. Sensor Netw.*, vol. 13, no. 1, Jan. 2017, Art. no. 155014771668968.

[13] M. Hooshmand, M. Rossi, D. Zordan, and M. Zorzi, "Covariogram-based compressive sensing for environmental wireless sensor networks," *IEEE Sensors J.*, vol. 16, no. 6, pp. 1716–1729, Mar. 2016.

[14] J. Cheng, Q. Ye, H. Jiang, D. Wang, and C. Wang, "STCDG: An efficient data gathering algorithm based on matrix completion for wireless sensor networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 2, pp. 850–861, Feb. 2013.

[15] X. Li, X. Tao, and G. Mao, "Unbalanced expander based compressive data gathering in clustered wireless sensor networks," *IEEE Access*, vol. 5, pp. 7553–7566, 2017.

[16] Z. Chen, C. Chen, G. Hu, W. Ye, J. Zhang, and G. Yang, "Data reconstruction in wireless sensor networks from incomplete and erroneous observations," *IEEE Access*, vol. 6, pp. 45493–45503, 2018.

[17] M. Kortas, V. Meghdadi, A. Bouallegue, T. Ezzeddine, O. Habachi, and J.-P. Cances, "Routing aware space-time compressive sensing for wireless sensor networks," in *Proc. IEEE 28th Annu. Int. Symp. Pers., Indoor, Mobile Radio Commun. (PIMRC)*, Oct. 2017, pp. 1–6.

[18] Z. Li, Y. Liu, M. Ma, A. Liu, X. Zhang, and G. Luo, "MSDG: A novel green data gathering scheme for wireless sensor networks," *Comput. Netw.*, vol. 142, pp. 223–239, Sep. 2018.

[19] S. J and V. Mayil, "Fuzzy gene optimized reweight boosting classification for energy efficient data gathering in WSN," *Int. J. Comput. Netw. Comput.*, vol. 11, no. 2, pp. 95–112, Mar. 2019.

[20] Q. Wang, W. Liu, T. Wang, M. Zhao, X. Li, M. Xie, M. Ma, G. Zhang, and A. Liu, "Reducing delay and maximizing lifetime for wireless sensor networks with dynamic traffic patterns," *IEEE Access*, vol. 7, pp. 70212–70236, 2019.

[21] Z. Li, Y. Liu, A. Liu, S. Wang, and H. Liu, "Minimizing convergecast time and energy consumption in green Internet of Things," *IEEE Trans. Emerg. Topics Comput.*, to be published.

[22] Y. Xu, G. Sun, T. Geng, and J. He, "Low-energy data collection in wireless sensor networks based on matrix completion," *Sensors*, vol. 19, no. 4, p. 945, Feb. 2019.

[23] Y. Yao, Q. Cao, and A. V. Vasilakos, "EDAL: An energy-efficient, delay-aware, and lifetime-balancing data collection protocol for heterogeneous wireless sensor networks," *IEEE/ACM Trans. Netw.*, vol. 23, no. 3, pp. 810–823, Jun. 2015.

[24] N. Jan, N. Javaid, Q. Javaid, N. Alrajeh, M. Alam, Z. A. Khan, and I. A. Niaz, "A balanced energy-consuming and hole-alleviating algorithm for wireless sensor networks," *IEEE Access*, vol. 5, pp. 6134–6150, 2017.

[25] Z. Wadud, N. Javaid, M. Khan, N. Alrajeh, M. Alabed, and N. Guizani, "Lifetime maximization via hole alleviation in IoT enabling heterogeneous wireless sensor networks," *Sensors*, vol. 17, no. 7, p. 1677, Jul. 2017.

[26] K. Mahendrababu and K. L. Joshitha, "A solution to energy hole problem in Wireless Sensor Networks using WITRICITY," in *Proc. Int. Conf. Inf. Commun. Embedded Syst. (ICICES)*, Feb. 2014, pp. 1–6.

[27] E. Candès and B. Recht, "Exact matrix completion via convex optimization," *Commun. ACM*, vol. 55, no. 6, p. 111, Jun. 2012.

[28] Z. Wen, W. Yin, and Y. Zhang, "Solving a low-rank factorization model for matrix completion by a nonlinear successive over-relaxation algorithm," *Math. Prog. Comp.*, vol. 4, no. 4, pp. 333–361, Dec. 2012.

[29] J.-F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM J. Optim.*, vol. 20, no. 4, pp. 1956–1982, Jan. 2010.

[30] Y. Zhang, M. Roughan, W. Willinger, and L. Qiu, "Spatio-temporal compressive sensing and Internet traffic matrices," *SIGCOMM Comput. Commun. Rev.*, vol. 39, no. 4, p. 267, Aug. 2009.

[31] M. Grant and S. Boyd. *CVX: MATLAB Software for Disciplined Convex Programming, Version 2.1. 2014 MAR*. Accessed: Mar. 2014. [Online]. Available: http://www.cvxr.com/cvx

[32] H. Zheng, F. Yang, X. Tian, X. Gan, X. Wang, and S. Xiao, "Data gathering with compressive sensing in wireless sensor networks: A random walk based approach," *IEEE Trans. Parallel Distrib. Syst.*, vol. 26, no. 1, pp. 35–44, Jan. 2015.

[33] J.-H. Chang and L. Tassiulas, "Maximum lifetime routing in wireless sensor networks," *IEEE/ACM Trans. Netw.*, vol. 12, no. 4, pp. 609–619, Aug. 2004.

[34] D. Sharma, S. Verma, and K. Sharma, "Network topologies in wireless sensor networks: A review 1," *Indian J. Extra-Corporeal Technol.*, vol. 4, no. Spl-3, Apr./Jun. 2013. [Online]. Available: http://www.iject.org/vol4/spl3/c0116.pdf

[35] B. Musznicki, M. Tomczak, and P. Zwierzykowski, "Dijkstra-based localized multicast routing in Wireless Sensor Networks," in *Proc. 8th Int. Symp. Commun. Syst., Netw. Digit. Signal Process. (CSNDSP)*, Jul. 2012, pp. 1–6.

[36] H. Zhou, D. Zhang, and K. Xie, "Accurate traffic matrix completion based on multi-Gaussian models," *Comput. Commun.*, vol. 102, pp. 165–176, Apr. 2017.

[37] G. Anastasi, M. Conti, M. Di Francesco, and A. Passarella, "Energy conservation in wireless sensor networks: A survey," *Ad Hoc Netw.*, vol. 7, no. 3, pp. 537–568, May 2009.

[38] *Stts751 Description*. Accessed: Sep. 30, 2019. [Online]. Available: https://www.st.com/content/st_com/en/products/mems-and-sensors/temperature-sensors/stts751.html

[39] S. Chen, Z. Wang, and K. Sha, "Cluster-aware kronecker supported data collection for sensory data," in *Proc. 27th Int. Conf. Comput. Commun. Netw. (ICCCN)*, Jul. 2018, pp. 1–6.

MANEL KORTAS received the Engineering degree in telecommunications from the National Engineering School of Tunis, Tunisia, in 2015. She is currently pursuing the joint Ph.D. degree with the XLIM Research Institute, University of Limoges, France, and the SysCom Laboratory, National Engineering School of Tunis. Her research interests include compressive sensing, matrix completion, and sensor networks.

OUSSAMA HABACHI received the Engineering degree in computer sciences from the National School of Computer Sciences (ENSI), in September 2008, and the M.Sc. degree in network and communications from the University of Pierre and Marie Curie (UPMC), Paris, France, in 2009. From October 2009 to September 2012, he received his Ph.D. thesis in computer sciences with the University of Avignon. He is currently an Assistant Professor with the University of Limoges.

AMMAR BOUALLEGUE received the degree in electrical engineering and the Ph.D. degree in engineer from ENSERG, Grenoble, France, in 1971 and 1976, respectively, and the Ph.D. degree from ENSEEIHT, INP of Toulouse, France, in 1984. In 1976, he joined the National Engineering School of Tunis (ENIT), Tunisia. From 1984 to 1992, he was the Head of the Electrical Department, ENIT, and from 1993 to 1995, he was the Director of telecommunication with the High School of Tunis, Tunisia. He was also the Head of the Communication Systems (Sys'Com) Laboratory, Engineering School of Tunis, till 2012. Since 2012, he has been a Professor Emeritus. His research interests include passive and active microwave structures, signal coding theory, and digital communications.

VAHID MEGHDADI received the B.Sc. and M.Sc. degrees from the Sharif University of Technology, Tehran, Iran, in 1988 and 1991, respectively, and the Ph.D. degree from the University of Limoges, France, in 1998.

He was invited Researcher for eight months at the University of California Irvine (UCI), in 2007. He works as a Researcher with the XLIM-CNRS Laboratory, SRI (Intelligent systems and networks) Department, Limoges, France. He has been working with the Department of Electronic and Telecommunication, ENSIL/University of Limoges, as an Assistant Professor, since 2000, and as a Full Professor, since 2014. He has coauthored of more than 100 publications in scientific journals and conferences. His main research interests include telecommunication systems including (massive) MIMO systems, compressive sensing, sensor networks, coding, network coding, cooperative communications, and smart grids. He has served as a TPC Member in several international conferences. He received the award of excellence for scientific research, in 2004, 2008, and 2012, from the French Ministry of Research and Higher Education. He has served as the Scientific Manager and/or a Scientific Researcher for more than 12 research projects in the field of ICT (information and communications technology), including European projects, ANRs (The French National Research Agency) projects, Regional projects, and PHC (Hubert Curien Partnerships).

TAHAR EZZEDINE received the M.S., Ph.D., and HDR degrees in telecommunications from the National Engineering School of Tunis (ENIT), Tunisia. He is currently a Full Professor with the Department of Information and Communication Technology and the Founder/Leader of the Research and Development Team for research and development of wireless sensor networks in different fields with the SysCom Research Laboratory. His interests also include smart objects and the IoT.

JEAN PIERRE CANCES (Senior Member, IEEE) graduated from the Ecole Nationale Supérieure des Télécommunications de Bretagne, in 1990. He received the Agregation teaching degree in physics, in 1993, and the CAPES degree in mathematics. He is currently a Professor with ENSIL, Limoges, where he also teaches signal processing. His research activities concern the IoT multiple access schemes, including PDMA and SCMA-based NOMA techniques.

• • •