Data-Centric Node Selection for Machine-Type Communications with Lossy Links

Hung-Hsien Chen* and Hung-Yun Hsieh*†
*Graduate Institute of Communication Engineering
†Department of Electrical Engineering
National Taiwan University, Taipei, Taiwan 106
Email: hungyun@ntu.edu.tw

***Abstract*—While node selection has been popularly studied in the literature for wireless sensor networks, a majority of papers assume a simplistic wireless model without taking communication costs such as radio resource usage and link loss into consideration. In a lossy environment, since data sent back by the selected subset of sensors may suffer from random losses, it may become necessary to use more radio resource usage by either selecting more sensors than needed as backups or providing more transmission opportunities to the selected sensors. In this paper, we investigate how the limited radio resource can be effectively allocated to a selected subset of sensors using machine-type communications for minimizing the data reconstruction error in a data gathering application with lossy links. We first formulate a node selection problem and then investigate two algorithms as solutions. The first algorithm exploits meta-heuristic randomized search in the search space to find a near-optimal solution. The second one, on the other hand, incurs a much lower computation cost by greedily selecting most informative sensors one by one to represent the population. Through computer simulation, we show that providing more transmission opportunities to the selected subset of sensors can achieve a more desirable performance in terms of radio resource usage and energy conservation than selecting more sensors as backups for machine-type communications with lossy links.**

I. Introduction

Node selection is a technique investigated popularly in wireless sensor networks (WSNs) to select only a subset from all sensors deployed for reporting data back to the sink. Early research endeavors, however, either do not include the wireless communication model into the optimization problem or assume a simplistic one without taking communication cost such as radio resource usage into consideration. For example, the authors in [1] consider the problem of choosing a subset of sensor measurements to minimize the parameter estimation error under Gaussian noise. A heuristic algorithm is proposed such that the subset of sensors that minimizes the determinant of the estimator covariance matrix is selected. The communication model in such work is non-existent, where sensor transmission/reception can happen without requiring any radio resource or incurring any communication cost.

For related work that *does* take the wireless communication model into consideration, a majority of papers on node selection deals with the *energy cost* problem to prolong the lifetime of the sensor network. For example, the authors in [2] propose an optimal node selection algorithm to select a subset of camera sensors for estimating the location of a target while minimizing the energy cost. The work in [3] investigates the problem of reducing unnecessary energy costs for transmitting redundant data in multi-hop wireless sensor networks with multiple sinks through relay node selection. [4] proposes a data-driven active node selection algorithm for IoT that uses the current data readings, residual energy, and power costs for selecting the next active nodes.

In this paper, we investigate the use of machine-type communications (MTC) to support machine-to-machine (M2M) applications such as data gathering from sensors. While related work typically assumes multi-hop communications for WSNs, the scenario we consider involves direct communications from sensors (machines) to the base station (BS) based on the *wide-area cellular technology.* Since the BS is typically the communication bottleneck in MTC, we consider a resource-constrained scenario such that the amount of radio resource available at the BS is not sufficient to support data reporting from all sensors. In such a scenario, the question arises as to how the radio resource can be effectively utilized to support a selected subset of sensors while *minimizing data reconstruction errors* for the data gathering application.

Related wok has investigated the node selection problem for minimizing reconstruction errors through compressive sensing (CS). For example, the authors in [5] propose a correlation-based node selection algorithm to successively decide important nodes that bring the largest improvement in terms of reconstruction quality. Although these papers focus on reconstruction error similar to our work, one discrepancy is that they *assume ideal wireless channels without considering data loss*, an inherent and unavoidable problem in wireless environments, when deciding the subset of nodes to select. The authors in [6] do investigate the node selection problem under link loss similar to our work. Each link is associated with a packet loss probability and the authors combine the packet loss matrix into the objective of minimizing expected reconstruction error through node selection. Unlike [6] where selected nodes are allocated the same amount of radio resource for transmission, we consider a different approach by *allocating more radio resource to "important" nodes* to prevent losing informative data carried by these nodes. While such an approach could possibly not select as many nodes as that in [6], the benefit is that important data has a higher probability of being received. More importantly, from the perspective of energy conservation,

978-1-7281-4355-2/20/$31.00 ©2020 IEEE

Authorized licensed use limited to: Carleton University. Downloaded on November 05,2020 at 09:30:02 UTC from IEEE Xplore. Restrictions apply.

a smaller subset of active (selected) nodes can help reduce energy cost and prolong network lifetime since unselected nodes can stay in the sleep mode without being awakened to participate in the data transmission process.

To proceed, we start with the framework of Bayesian Compressive Sensing (BCS) [7] and formulate an optimization problem for node selection in Section II. The objective is to minimize the mean squared error (MSE) of reconstructed data from the subset of selected sensors without violating the radio resource constraint. Since each sensor may experience link loss in sending back its data to the base station, additional resource blocks are provided to selected sensors depending on the expected transmission counts (ETX) of individual links. To solve the formulated problem, we investigate two algorithms in Section III. The first is a meta-heuristic algorithm based on the cross-entropy algorithm. It aims to find a near-optimal solution of the problem through randomized search in iterations. To reduce the computation complexity of the meta-heuristic algorithm, we propose a greedy algorithm that iteratively selects the "most informative sensors" to represent the set of sensors not yet selected. We calculate the conditional variance and select sensors one by one until the radio resource is used up. The performance of the proposed node selection algorithms are compared against related work [5], [6], [8] in Section IV. Our evaluation results show that the proposed solution for node selection indeed has performance benefits compared to related work, thus motivating further investigation along this direction.

II. System Model and Problem Formulation

We consider an MTC scenario with N sensors (machines) randomly deployed in the serving area of a base station (BS). Each sensor periodically measures the data of interest from the data field and waits for the transmission opportunity to report the data back to the BS through one-hop communication. The BS allocates radio resource to sensors in the cell and it doubles up as a data aggregator (sink) to support the target application. Uplink transmissions from sensors to the BS are assumed to be orthogonal (e.g. TDMA or OFDMA) such that each sensor is provided with dedicated radio resource (e.g. time slots or resource blocks) for transmission. To optimize allocation of the radio resource, the BS determines for each allocation period the subset of sensors for uplink transmissions as follows.

A. System Model

We assume the data of interests forms a Gaussian random source field that is observed by the set of sensors $\mathcal{N} = \{1, 2, \cdots, N\}$. Each data source $X_i, \forall i \in \mathcal{N}$, is a Gaussian random variable with expected value μ_i and standard deviation σ_i. Specifically, $\mathbf{X} = \{X_1, X_2, \cdots, X_N\}$ is the collection of data at some time instant that follows a multivariate Gaussian distribution as

$$f(\mathbf{X}) = \frac{1}{(2\pi)^{\frac{N}{2}} |\mathbf{\Sigma}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{X}-\boldsymbol{\mu})^T \mathbf{\Sigma}^{-1}(\mathbf{X}-\boldsymbol{\mu})}, \quad (1)$$

where $\boldsymbol{\mu}$ is the expectation vector of $\mathbf{X}$, $\mathbf{\Sigma} = [\sigma_{ij}]_{N\times N}$ is the covariance matrix of $\mathbf{X}$. Covariance σ_{ij} between sensors i and j is calculated as follows:

$$\sigma_{ij} = \sigma_i \sigma_j e^{-d_{ij}^2/\kappa}, \quad (2)$$

where κ is the pre-determined correlation exponent and d_{ij} is the Euclidean distance between sensors i and j.

Since the Gaussian random variable is continuous, each sensor $i \in \mathcal{N}$ quantizes the data for packetization with a pre-configured quantization level Δ_i. With a sufficiently small quantization level, the entropy $H(i)$ (i.e., the minimum number of bits needed to encode/compress the information) of the quantized source at sensor i can be approximated as

$$H(i) = \frac{1}{2}\log_2\left[\frac{2\pi e}{\Delta_i^2}\sigma_i^2\right], \quad \forall i \in \mathcal{N}. \quad (3)$$

In general, the joint entropy of all data gathered by the subset $\mathcal{C} \subseteq \mathcal{N}$ of sensors can be written as follows [9]:

$$H(\mathcal{C}) = \frac{1}{2}\log_2\left[\frac{(2\pi e)^{|\mathcal{C}|}}{\sum_{i\in\mathcal{C}}\Delta_i^2}|\mathbf{\Sigma}_{\mathcal{C}}|\right], \quad (4)$$

where $|\mathbf{\Sigma}_{\mathcal{C}}|$ is the determinant of covariance matrix $\mathbf{\Sigma}_{\mathcal{C}}$. We define compression ratio

$$\eta = 1 - \frac{H(\mathcal{N})}{\sum_{i\in\mathcal{N}} H(X_i)} \quad (5)$$

to indicate the level of data correlation among sensors.

B. Problem Formulation

We assume that in each resource allocation period the BS has a total of T resource blocks (e.g. time slots) to allocate. Each sensor is allocated at least one unit of resource block for data transmission and we assume that the total number of sensors N is more than the number of resource blocks T. The goal is to select a subset of M sensors $(M \leq T < N)$ to transmit data to the BS such that the reconstruction errors of the gathered data is minimized. Note that since link loss could occur in wireless communications, a sensor may need more than one resource block for retransmissions to ensure its data being received with high probability. Therefore, while we can select a subset of $M = T$ sensors and allocate one resource block for each sensor, it is non-trivial to decide if we should select a smaller subset $M < T$ with *some sensors allocated additional resource blocks for minimizing reconstruction error.*

To proceed, denote $\mathbf{\Phi} \in \mathbb{R}^{M\times N}$ as the sampling matrix whose elements are all zeros except for M unity elements in different rows. If sensor i is selected, then there is a unity element in column i of $\mathbf{\Phi}$; otherwise all elements of column i are zeros. Let $\mathbf{x} = [x_1, x_2, \cdots, x_N] \in \mathbb{R}^N$, where x_i represents the data of sensor i, and $\mathbf{y} = \mathbf{\Phi}\mathbf{x} = [y_1, y_2, \cdots, y_M] \in \mathbb{R}^M$ be the data gathered by the BS. Since data measured by different sensors may exhibit high spatial correlations [10], in the context of compressive sensing, it is possible to project $\mathbf{x}$ onto a sparse domain using a set of orthonormal basis matrix $\mathbf{\Psi}$ such that $\mathbf{x} = \mathbf{\Psi}\mathbf{u}$. Using PCA (Principal Component Analysis), for example, the columns of $\mathbf{\Psi}$ can be obtained

Authorized licensed use limited to: Carleton University. Downloaded on November 05,2020 at 09:30:02 UTC from IEEE Xplore. Restrictions apply.

as unitary eigenvectors of the sample covariance matrix $\mathbf{\Sigma}$ placed according to the decreasing order of the corresponding eigenvalues. Once the sparsification matrix $\mathbf{\Psi}$ and the sampling matrix $\mathbf{\Phi}$ are known, $\mathbf{u}$ can be recovered from $\mathbf{y}$ by using techniques such as the $\ell 1$-magic [11].

Different from the least-square estimation scheme adopted in [6], in this paper, we start with the Bayesian estimation scheme for reconstructing $\mathbf{u}$ [7], [12]. Firstly, the measurement $\mathbf{y}$ is assumed to be Gaussian with $\mathbf{y} = \mathbf{\Phi x} + \mathbf{z} = \mathbf{\Phi \Psi u} + \mathbf{z}$, where $\mathbf{z}$ is independently identically distributed (*i.i.d*) additive white Gaussian noise with mean zero and variance σ^2. Hence,

$$p\left(\mathbf{y} \mid \mathbf{u}, \alpha_0\right) = \left(2\pi/\alpha_0\right)^{-\frac{M}{2}} e^{\left(-\frac{\alpha_0}{2}\|\mathbf{y}-\mathbf{\Phi\Psi u}\|_2^2\right)}, \tag{6}$$

where $\alpha_0 = 1/\sigma^2$. The element of the sparse vector $\mathbf{u}$ is also assumed as an independent Gaussian distribution with zero mean and variance α_i^{-1}. With such, the distribution of the sparse vector $\mathbf{u}$ can be written as follows:

$$p\left(\mathbf{u} \mid \boldsymbol{\alpha}\right) = \left(2\pi\right)^{-\frac{N}{2}} \left|\mathbf{\Lambda}\right|^{-\frac{1}{2}} e^{\left(-\frac{1}{2}\mathbf{u}^T\mathbf{\Lambda}^{-1}\mathbf{u}\right)}, \tag{7}$$

where $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \cdots, \alpha_N]$ and $\mathbf{\Lambda} = diag\left(\alpha_1, \alpha_2, \cdots, \alpha_N\right)$ is a diagonal matrix. Note that the three unknown parameters: $\mathbf{u}$, α_0, and $\mathbf{\Lambda}$ must be determined during the estimation process. Now, through Bayesian statistics, we can obtain the posterior probability $p\left(\mathbf{u} \mid \mathbf{y}, \alpha_0, \mathbf{\Lambda}\right)$ as a Gaussian distribution with mean $\boldsymbol{\mu} \in \mathbb{R}^N$ and covariance matrix $\mathbf{\Sigma} \in \mathbb{R}^{N\times N}$ as follows:

$$\begin{aligned} \boldsymbol{\mu} &= \alpha_0 \mathbf{\Sigma}\left(\mathbf{\Phi\Psi}\right)^T \mathbf{y}, \\ \mathbf{\Sigma} &= \left(\mathbf{\Lambda} + \alpha_0 \mathbf{\Psi}^T\mathbf{\Phi}^T\mathbf{\Phi\Psi}\right)^{-1}. \end{aligned} \tag{8}$$

Note that $\boldsymbol{\mu}$ and $\mathbf{\Sigma}$ can be calculated via an iterative update process [7], [12]. After the estimate $\hat{\mathbf{u}}$ is obtained, the estimate of the original signal $\hat{\mathbf{x}} = \mathbf{\Psi}\hat{\mathbf{u}}$ can also be obtained as a Gaussian distribution with mean $\mathbf{\Psi}\boldsymbol{\mu}$ and covariance $\mathbf{\Psi\Sigma\Psi^T}$. Now since the MSE of Bayesian estimation is the trace of the error covariance matrix, it can be expressed as follows:

$$\begin{aligned} Trace\left(\mathbf{\Psi\Sigma\Psi}^T\right) &= Trace\left(\mathbf{\Sigma}\right) \\ &= Trace\left(\mathbf{\Lambda} + \alpha_0\mathbf{\Psi}^T\mathbf{\Phi}^T\mathbf{\Phi\Psi}\right)^{-1}, \end{aligned} \tag{9}$$

where the first equation follows because the sparsification matrix $\mathbf{\Psi}$ is an orthogonal matrix and the trace is similarity-invariant. Note that $\tilde{\mathbf{\Phi}} = \mathbf{\Phi\Phi^T}$ is a diagonal matrix and $\tilde{\Phi}_{ii} = 1$, for $i \in \mathcal{N}$, if node i is selected; otherwise $\tilde{\Phi}_{ii} = 0$.

To take link loss into consideration, let q_i, $i \in \mathcal{N}$, be the probability that a packet is successfully received from sensor i to the BS. Denote $\mathbf{H} \in \mathbb{R}^{N\times N}$ as a diagonal matrix with the i^{th} element $h_{ii} = \lfloor 1/q_i \rfloor$ being the expected transmission count (ETX) of sensor i. We thus formulate the node selection problem for lossy WSNs as follows:

$$\min_{\tilde{\mathbf{\Phi}}} \; Trace\left(\mathbf{\Lambda} + \alpha_0\mathbf{\Psi}^T\tilde{\mathbf{\Phi}}\mathbf{\Psi}\right)^{-1} \tag{10}$$

subject to

$$\begin{aligned} &\tilde{\Phi}_{ii} \in \{0,1\}, \quad i = 1, 2, \cdots, N, \\ &Trace\left(\tilde{\mathbf{\Phi}}\mathbf{H}\right) \le T. \end{aligned} \tag{11}$$

Since link loss may occur, each selected node is allocated multiple resource blocks depending on the link quality to ensure that its data can be received at the BS with high probability. The second constraint means that the maximum amount of resource blocks that can be allocated is T. Under the resource constraint, our goal is to collect the most important information in this network to ensure the minimum reconstruction error.

III. Proposed Algorithms

Finding the optimal subset of sensors for the formulated problem is a combinatorial optimization problem. To solve the formulated problem, we first consider a solution based on a meta-heuristic search algorithm. Specifically, we extend the cross-entropy (CE) algorithm proposed in [9] for node selection. The concept of the CE algorithm is to translate a combinatorial optimization problem into a probability estimation problem. To proceed, note that determining $\mathbf{\Phi}$ is equivalent to determining the binary sensor selection vector $\mathbf{b} = [b_1, b_2, \cdots, b_N]$, where a sensor i is selected if and only if $b_i = 1$. The binary vector $\mathbf{b}$ can first be translated into a Bernoulli probability vector $\mathbf{p} = [p_1, p_2, \cdots, p_N]$, where each sensor i is selected $(b_i = 1)$ and allocated a time slot for transmission with probability p_i and not selected $(b_i = 0)$ with probability $1 - p_i$. The algorithm then runs in iterations and in each iteration a set of samples are randomly generated, where each sample consists of the set of sensors selected based on the current value of $\mathbf{p}$. The objective in (10) for each generate sample is evaluated, and the top few samples that can yield minimal reconstruction errors are used to update the probability vector for the next iteration (p_i is adjusted based on the counts sensor i is selected in these top samples). The algorithm can stop when the probability vector converges to a binary vector within the desired tolerance. We refer readers to [9] for details of the algorithm design due to lack of space.

While the CE algorithm can generally yield a reasonably good solution, it incurs non-negligible computation cost that depends on the number of samples generated and the number of iterations needed for algorithm convergence. To address this issue, in this paper we propose a greedy node selection algorithm that aims to find informative nodes directly based on correlation among sensor nodes. Such an algorithm is motivated by a node selection procedure called Enhanced Correlation Based Deterministic Node Selection (ECB-DNS) proposed in [5]. Specifically, ECB-DNS runs in iterations, and initially D, the set of sensors that are not yet selected, is equal to $\mathcal{N}$. In each iteration one sensor j^* with the maximum informative value (minimum conditional variance) $m_j^{'}$ with respect to the set of sensors that are not yet selected D is chosen as follows:

$$j^* = \arg\max_{j\in D}\left(m_j^{'}\right), \text{ where } \; m_j^{'} = \left(\sum_{i\in D}\frac{\sigma_{ij}^2}{\sigma_j^2}\right). \tag{12}$$

Every time such a sensor is found, the algorithm removes it from D and adds it to S, the set of selected sensors. The

Authorized licensed use limited to: Carleton University. Downloaded on November 05,2020 at 09:30:02 UTC from IEEE Xplore. Restrictions apply.

algorithm then recomputes relevant metrics conditioning on its removal from the main set until M sensors are selected.

The problem with ECB-DNS, however, is that it does not take into account of the relation of sensors in the selected set such that it is possible that *similar sensors with high data redundancy* are selected. To address this problem, we propose an algorithm based on *enhanced correlation* that selects the sensor j^* with the minimum conditional variance with respect to the sensors that are not selected conditioned on the selected set. To start with, let $\mathbf{X}$ be a multivariate normal distribution of a k-dimensional random vector, where $\mathbf{X} = (X_1, \cdots, X_k)$ and $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with k-dimensional mean vector $\boldsymbol{\mu} = \mathrm{E}[\mathbf{X}] = (\mathrm{E}[X_1], \mathrm{E}[X_2], \cdots, \mathrm{E}[X_k])$ and $k \times k$ covariance matrix $\Sigma_{ij} = \mathrm{E}[(X_i - \mu_i)(X_j - \mu_j)] = \mathrm{Cov}[X_i, X_j]$. Without loss of generality, we partition $\mathbf{X}$ as follows:

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix} \text{ with sizes } \begin{bmatrix} q \times 1 \\ (k-q) \times 1 \end{bmatrix}, \tag{13}$$

and accordingly $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are partitioned as follows:

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix} \text{ with sizes } \begin{bmatrix} q \times 1 \\ (k-q) \times 1 \end{bmatrix}, \tag{14}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix} \text{ with sizes } \begin{bmatrix} (q \times q) & q \times (k-q) \\ (k-q) \times q & (k-q) \times (k-q) \end{bmatrix}.$$

The conditional variance of $\mathbf{X}_1$, given that we have sample $\mathbf{X}_2$, is equal to $Var[\mathbf{X}_1 \mid \mathbf{X}_2 = x] = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$.

With these concepts in mind, we proceed to design the proposed greedy node-selection algorithm as follows. Assuming we have selected some sensor nodes S and let D be the set of sensor nodes which are not selected yet. We compute the overall conditional variance m_j with respect to the set D of sensors that are not selected conditioned on the selected set S as follows:

$$\begin{aligned} m_j &= \sum_{i \in D} Var[X_i \mid X_j \cup S] \\ &= \sum_{i \in D} \left(\sigma_i^2 - \Sigma_{i,j\cup S} \Sigma_{j\cup S, j\cup S}^{-1} \Sigma_{j\cup S, i} \right) \\ &= \sum_{i \in D} \sigma_i^2 - \sum_{i \in D} \Sigma_{i,j\cup S} \Sigma_{j\cup S, j\cup S}^{-1} \Sigma_{j\cup S, i}. \end{aligned} \tag{15}$$

Note that node j^* with the smallest conditional variance is found as:

$$\begin{aligned} j^* &= \arg\min_{j \in D} \left(\sum_{i \in D} \sigma_i^2 - \sum_{i \in D} \Sigma_{i,j\cup S} \Sigma_{j\cup S, j\cup S}^{-1} \Sigma_{j\cup S, i} \right) \\ &= \arg\max_{j \in D} \left(\sum_{i \in D} \Sigma_{i,j\cup S} \Sigma_{j\cup S, j\cup S}^{-1} \Sigma_{j\cup S, i} \right), \end{aligned} \tag{16}$$

where the second equality follows since the first sum in the first equation does not depend on the index j. In this case, m_j is related to the amount of uncertainty for readings from unselected sensors in D given the value sampled at sensor j union the selected set S. A smaller value of uncertainty m_j indicates that sensor j is a *better candidate to represent other unselected nodes*. The detailed procedure of the proposed greedy algorithm is presented in Algorithm 1. Initially, $D = \mathcal{N}$ is the whole node set and $S = \emptyset$ is an empty set. Every time a node is selected, the algorithm removes it from D and adds it to S. Then, the amount of resource allocated to the selected sensor is calculated from the expected transmission count to avoid important data being dropped due to link loss. The node selection procedure proceeds until all available T resource blocks are allocated.

Algorithm 1 Greedy Node Selection Algorithm

1: $D = \{1, 2, \cdots, N\}$, $S = \emptyset$, Z is the allocated time slots, T is the total time slots that can be allocated.
2: **repeat**
3: Select the node j^* with the highest value with respect to all the other sensors from D according to (16).
4: $D \leftarrow D \setminus \{j^*\}$
5: $S \leftarrow S \cup \{j^*\}$
6: $Z \leftarrow Z + \lfloor 1/q_{j^*} \rfloor$
7: **until** ($Z > T$ is reached)
8: **return** S

It has to be noted that [8] also proposes a node selection algorithm called STCS (Space-Time Compressive Sensing) based on ECB-DNS [5] for routing in a wireless mesh network. However, the algorithm in [8] selects in each iteration n a node $j(n)$ with the highest variance conditioned only on the previously selected node $j(n-1)$ in iteration $n-1$. This is different from the proposed approach that considers the variance conditioned on the set of *all selected nodes* for a more holistic view of the correlation among selected nodes. We show in Section IV-B the performance benefits of our proposed algorithm compared to the algorithms in [5], [8].

IV. Performance Evaluation

In this section, we evaluate the performance of the proposed algorithms by comparing against related work through the following simulations.

A. Simulation Setup

We randomly distribute N sensors over a circular area of radius R with the BS placed at the center. The sensor data $\mathbf{x}$ is generated from a Gaussian source field with zero mean and normalized standard deviation. Different correlation levels (and hence different compression ratios η) among sensors are considered through controlling κ in (2). To quantize the data, an 8-bit uniform quantizer (quantization step Δ is 1 / 256) is applied. We use normalized mean squared error (MSE) of reconstruction data as the performance metric δ, which is the average of $\frac{\|\mathbf{x}-\hat{\mathbf{x}}\|^2}{\|\mathbf{x}\|^2}$ over 100 trials. The sample covariance matrix used for calculating the PCA sparsification basis can be obtained by gathering data from all sensors for a period of time. We assume that BS is given information about the covariance matrix and hyperparameters $\boldsymbol{\Lambda}$ and α_0 for Bayesian estimation.

Authorized licensed use limited to: Carleton University. Downloaded on November 05,2020 at 09:30:02 UTC from IEEE Xplore. Restrictions apply.

In each data gathering period, the maximum available resource that can be allocated is T time slots and we assume each data takes one time slot for transmission. The maximum transmission power of each sensor is set to 0 dBm based on the 3GPP simulation parameters for low-cost MTC UEs. We follow the propagation model implemented in the 6TiSCH simulator [13] designed for ultra-reliable, ultra low-power wireless mesh networks. The path loss model calculates the initial RSSI value of each link based on the distance between the pair of nodes and the frequency band as follows (d in km and f in GHz):

$$P_{rx} = P_{tx} - (20\log_{10}(d) + 20\log_{10}(f) + 92.45) - 40 \cdot rand(0,1). \quad (17)$$

To model link loss, the RSSI values are then converted to the Packet Delivery Ratio (PDR) values by a conversion table based on real-world deployments as used in [13]. The PDR value is the ratio of packets successfully received to the total sent, and the reciprocal of the PDR value is the ETX (expected transmission count) metric, which is the number of expected transmissions of a packet necessary for it to be received without error at its destination.

B. Comparison of Algorithms

To compare the performance of the proposed algorithms, we consider the algorithms proposed in related work [5], [6], [8]. The **"ECB-DNS"** algorithm proposed in [5] as well as the **"STCS"** algorithm proposed in [8] have been described in Section III. The **"GA-LS"** (Genie-Aided Least-Square estimation) algorithm proposed in [6] considers packet loss in the problem formulation similar to ours. The difference is that in **"GA-LS"** each selected sensor is still allocated only one unit of resource block (irrespective of the packet loss probability) and the objective is to maximize the expected data quality by taking probabilistic data loss into consideration. As a result, **"GA-LS"** tends to select these nodes *with better link quality but possibly not important data.* In contrast, our proposed **"Greedy"** algorithm intends to allocate additional resource for important nodes to increase the probability of successful data delivery.

In Fig. 1, we compare the reconstruction error δ (y-axis) of different algorithms versus different amounts of available resource T (x-axis). The number of nodes N is 50 and the transmission power is 0dBm such that the average PDR for all nodes is slightly over 60%. In the figure, **"CE"** is the node selection algorithm derived from [9] based on the CE algorithm and "**Greedy**" is the proposed greedy algorithm in this paper. We can find that the meta-heuristic **"CE"** algorithm outperforms all other methods. It is worth noting that the proposed **"Greedy"** algorithm performs better than the other 3 algorithms proposed in related work. The result implies that *allocating more resource blocks for important nodes to increase the probability of successful data delivery indeed works*, in contrast to the **"GA-LS"** algorithm that selects nodes with good link quality and allocates only one resource block to each selected node. The **"ECB-DNS"** algorithm performs the worst since the way it selects informative sensors are flawed as described in Section III. It is also noted that when the radio resource becomes less, the performance of the **"Greedy"** algorithm becomes closer to the **"CE"** algorithm.

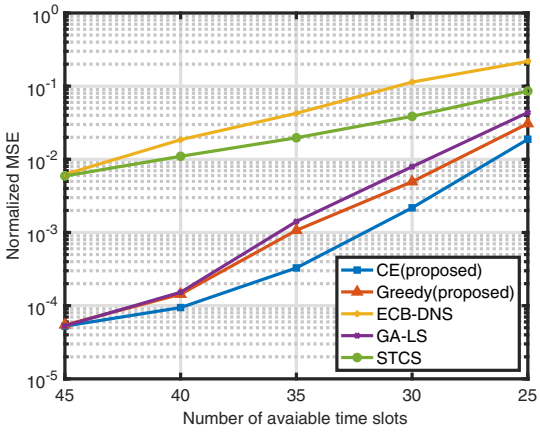

Fig. 1. Reconstruction error versus available resource when $N = 50$, $R = 35$, $\eta = 0.6$, and transmission power = 0 dBm.

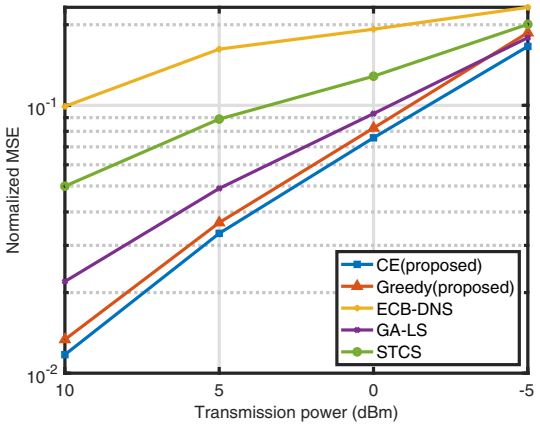

Fig. 2. Reconstruction error versus transmission power when $N = 50$, $R = 35$, $\eta = 0.6$, and available time slots $T = 30$.

To investigate the impact of PDR on the performance of the algorithms, we change the transmission power of each sensor node. Based on the propagation model and the RSSI-to-PDR table [13], the average PDR decreases from 90% to 50% when we change the transmission power from 10 dBm to -5 dBm. In Fig. 2, the horizontal axis represents the transmission power of each sensor node and the vertical axis reflects the reconstruction error. We can find that the **"CE"** algorithm outperforms as before and the **"Greedy"** algorithm follows closely. In addition, our proposed algorithm can consistently perform better than those proposed in related work when the values of transmission powers change.

We further use Fig. 3 to show that the nodes selected by the **"Greedy"** algorithm are indeed more "informative" than **"ECB-DNS"** and **"STCS"** that are also based on conditional

Authorized licensed use limited to: Carleton University. Downloaded on November 05,2020 at 09:30:02 UTC from IEEE Xplore. Restrictions apply.

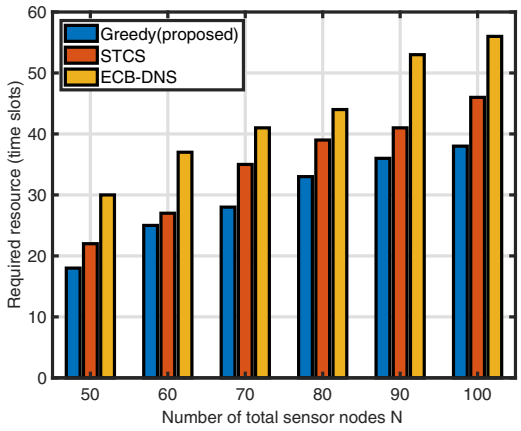

Fig. 3. Required resource versus different number of total nodes N when compression ratio $\eta = 0.6$.

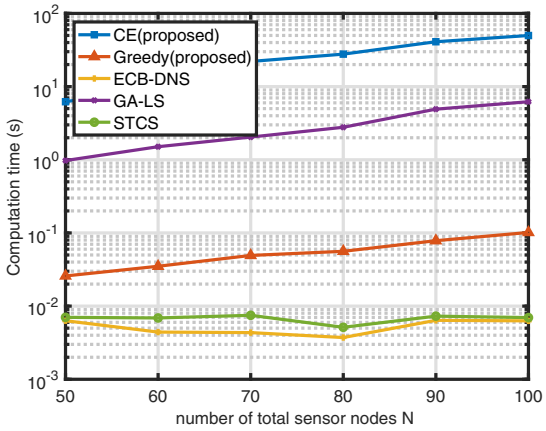

Fig. 4. Computation time versus total nodes (network scale) N.

variance. The bars in the figure shows the required resource of different algorithms when the target reconstruction error is set to be less than 0.1 ($\delta \leq 0.1$). Different network sizes are shown while fixing the compression ratio (correlation factor) $\eta = 0.6$. Clearly, the required resource (time slot) of the proposed **"Greedy"** algorithm is less than the other two algorithms. This means that *more nodes can enter the sleeping mode and hence the lifetime of network can be prolonged.* Besides, less resource means that the **"Greedy"** algorithm *requires lower latency for gathering the same quality of data*, which is favorable for latency-sensitive applications.

Note that while the **"CE"** algorithm performs the best in terms of the reconstruction error, we show in Fig. 4 the computation time complexity of different algorithms. We can find that although the reconstruction error of the **"Greedy"** algorithm is not as good as the **"CE"** algorithm, *it strikes a rather good balance in terms of performance and complexity.* The computation complexity of the **"Greedy"** algorithm is indeed higher than that of **"ECB-DNS"** and **"STCS"** owing to the way conditional variance is calculated. However, the performance benefits as evident from Fig. 1, Fig. 2, and Fig. 3 justify the slight increase in the computation complexity.

V. Conclusions

We have considered in this paper the problem of correlated data gathering from a set of sensors that communicate directly to the base station (BS) using machine-type communications with lossy links. Since the BS has limited radio resources to support all sensors, we investigate the optimization problem of node selection for optimizing the quality of data reconstructed from the selected sensors. We have investigated two algorithms to solve the problem and we have compared the performance of the proposed algorithms against related work. Evaluation results have shown that the proposed solution for node selection has performance benefits in terms of radio resource usage and energy conservation.

Acknowledgment

This work was supported in part by funds from the Ministry of Science and Technology and National Taiwan University under Grants MOST-108-2221-E-002-037, MOST-108-2218-E-002-060, and MOST-106-2923-E-002-015-MY3.

References

[1] S. Joshi and S. Boyd, "Sensor selection via convex optimization," *IEEE Transactions on Signal Processing*, vol. 57, no. 2, pp. 451–462, 2009.

[2] L. Liu, X. Zhang, and H. Ma, "Optimal node selection for target localization in wireless camera sensor networks," *IEEE Transactions on Vehicular Technology*, vol. 59, pp. 211–244, Sep. 2009.

[3] B. Cheng, Z. Xu, C. Chen, and X. Guan, "Spatial correlated data collection in wireless sensor networks with multiple sinks," in *Proceedings of IEEE International Workshop on Wireless Sensor, Actuator and Robot Networks (INFOCOM Workshop)*, 2011, pp. 578–583.

[4] A. Alagha, S. Singh, R. Mizouni, A. Ouali, and H. Otrok, "Data-driven dynamic active node selection for event localization in IoT applications - A case study of radiation localization," *IEEE Access*, vol. 7, pp. 16 168 – 16 183, Jul. 2019.

[5] M. Hooshmand, M. Rossi, D. Zordan, and M. Zorzi, "Covariogram-based compressive sensing for environmental wireless sensor networks," *IEEE Sensors Journal*, vol. 16, no. 6, pp. 1716–1729, 2015.

[6] W. Chen and I. J. Wassell, "Optimized node selection for compressive sleeping wireless sensor networks," *IEEE Transactions on Vehicular Technology*, vol. 65, p. 827–836, Feb. 2015.

[7] S. Ji, Y. Xue, and L. Carin, "Bayesian compressive sensing," *IEEE Transactions on Signal Processing*, vol. 56, pp. 2346–2356, May. 2008.

[8] M. Kortas, V. Meghdadi, A. Bouallegue, T. Ezzeddine, O. Habachi, and J. Cances, "Routing aware space-time compressive sensing for wireless sensor networks," in *Proceedings of IEEE PIMRC*, Oct 2017, pp. 1–6.

[9] H.-Y. Hsieh, C.-H. Chang, and W.-C. Liao, "Not every bit counts: Data-centric resource allocation for correlated data gathering in machine-to-machine wireless networks," *ACM Transactions on Sensor Networks*, vol. 11, no. 2, pp. 38:1–38:33, Mar. 2015.

[10] G. Quer, R. Masiero, G. Pillonetto, M. Ross, and M. Zorzi, "Sensing, compression, and recovery for WSNs: Sparse signal modeling and monitoring framework," *IEEE Transactions on Wireless Communications*, vol. 11, no. 10, p. 3447–3461, Oct. 2012.

[11] E. C and J. Romberg, "ℓ1-magic: Recovery of sparse signals via convex programming," 2005.

[12] M. E. Tipping, "Sparse bayesian learning and the relevance vector machine," *Journal of Machine Learning Research*, vol. 1, pp. 211–244, 2001.

[13] E. Municio, G. Daneels, M. Vucinic, S. Latre, J. Famaey, Y. Tanaka, K. Brun, K. Muraoka, X. Vilajosana, and T. Watteyne, "Simulating 6TiSCH networks," *Transactions on Emerging Telecommunications Technologies*, vol. 30, no. 3, p. e3494, Sep. 2019.

Authorized licensed use limited to: Carleton University. Downloaded on November 05,2020 at 09:30:02 UTC from IEEE Xplore. Restrictions apply.