Distributed Symmetric Function Computation in Noisy Wireless Sensor Networks

Lei Ying, R. Srikant, *Fellow, IEEE*, and Geir E. Dullerud, *Senior Member, IEEE*

***Abstract*—In this correspondence, we consider a wireless sensor network consisting of n sensors, and each sensor has a measurement, which is an integer value belonging to the set $\{0,\ldots,m-1\}$, so that it can be represented by $\lceil \log_2 m \rceil$ bits. The network has a special node called the fusion center whose goal is to compute a symmetric function of these measurements. The problem studied is to minimize the total transmission energy used by the network when computing this function, subject to the constraint that this computation be correct with high probability. We assume the wireless channels are binary symmetric channels with a probability of error p, and that each sensor uses r^α units of energy to transmit each bit, where r is the transmission range of the sensor. For constant m, the main result in this correspondence is an algorithm whose energy usage is $\kappa \lceil \log_2 m \rceil n (\log\log n) \left(8\sqrt{\frac{\log n}{n}}\right)^\alpha$, where $\kappa = \left\lceil \frac{4}{-\log(4p(1-p))} \right\rceil$. Then, we consider the case where the sensor network observes N events. In this case, we demonstrate a network algorithm which has energy usage $\Theta\left(n\left(\sqrt{\frac{\log n}{n}}\right)^\alpha\right)$ per event if the number of events satisfies $N = \Omega(\log\log n)$.**

***Index Terms*—Binary symmetric channel, function computation, reception diversity, sensor network, wireless network.**

Notation: The following notation is used throughout this correspondence. Given a sequence of random variables $X(n)$ indexed by n, and positive function $f(n)$, we will say that

1) $X(n) = O(f(n))$ when there exists a positive constant c such that

$$\lim_{n\to\infty} \Pr(X(n) \le c f(n)) = 1 \text{ holds.}$$

2) $X(n) = \Omega(f(n))$ when there exists a positive constant c such that

$$\lim_{n\to\infty} \Pr(X(n) \ge c f(n)) = 1 \text{ holds.}$$

3) $X(n) = \Theta(f(n))$ when both $X(n) = \Omega(f(n))$ and $X(n) = O(f(n))$ hold.

Note that the above definitions also apply in the obvious way to deterministic functions.

Manuscript received July 8, 2006; revised March 13, 2007. The research was supported by a Vodafone Fellowship, AFOSR URI Grant F49620-01-1-0365, and NSF Grant CNS 05-19535. The material in this correspondence was presented in part at the 4th International Symposium on Modeling and Optimization in Mobile, Ad-Hoc, and Wireless Networks, Boston, MA, April 2006.

L. Ying and R. Srikant are with the Department of Electrical and Computer Engineering and the Coordinated Science Lab., University of Illinois at Urbana-Champaign, Urbana, IL 61801 USA (e-mail: lying@uiuc.edu; @uiuc.edursrikant).

G. E. Dullerud is with the Department of Mechanical and Industrial Engineering and the Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, Urbana, IL 61801 USA (e-mail: dullerud@uiuc.edu).

Communicated by E. Modiano, Associate Editor for Communication Networks.

Color version of Figure 1 in this correspondence is available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIT.2007.909156

I. Introduction

With the wide availability of inexpensive wireless technology and sensing hardware, wireless sensor networks are expected to become commonplace because of their broad range of potential applications. A wireless sensor network consists of sensors that have sensing, computation and wireless communication capabilities. Each sensor monitors the environment surrounding it, collects and processes data, and when appropriate transmits information so as to cooperatively achieve a global detection objective. Here, we consider the common situation where there is a single fusion center, and the network goal is to cooperatively provide information to this fusion center so it can compute some function of the sensor measurements. In this correspondence, we will investigate this problem in multihop networks with noisy communication channels where the measurement of each sensor consists of $\lceil \log_2 m \rceil$ bits; the goal is for the fusion center to compute symmetric functions — those functions determined by the frequency-histogram of the measurements. To achieve this, we would like to design a distributed algorithm while minimizing the total transmission energy consumed by the network.

Specifically, distributed symmetric function computation, which is also called a counting problem in this correspondence, is as follows: the measurement of each node is an integer in $\{0,\ldots,m-1\}$, and the fusion center needs to decide, using information transmitted from the network, the number of sensors having value l, for each $l \in \{0,\ldots,m-1\}$. When nothing is known about the structure of the function to be computed, all measurements must to be transmitted to the fusion center, and this is purely a routing problem when the channels are reliable. When the wireless channels are unreliable, the use of channel coding (see, for example, [2]) makes it possible to convey information in a point-to-point fashion with arbitrarily small amounts of error. However, the use of point-to-point error-correction coding without any in-network processing may result in high energy cost and delay. Our focus in this correspondence is computation of symmetric functions in a noisy wireless sensor network when total energy consumption *is* a major concern.

The algorithms we consider in this correspondence are related to the algorithms for distributed computation over noisy networks, which are studied in [3], [14], [15], [13], [10], and references therein. In both problems, the goal is to compute the value of some function based on the information of the nodes. Our work is closely related to parity computation and threshold detection in noisy radio networks studied in [3] and [10], respectively, where a broadcast network is assumed, in which all nodes can hear all transmissions, and each node has either a "1" or a "0." The goal in [3], [10] was to investigate the minimum number of transmissions required to compute the parity or decide whether the number of nodes in state "1" has exceeded the threshold value. Note that parity and threshold detection are special cases of counting on binary data, since both of these are determined if we know how many nodes have a "1."

While the problems considered in [3] and [10] are similar to our problem, there are two differences. First, in our model, the measurements can take m different values instead of just two values, which is the assumption in [3] and [10]. The second difference is that each node may not be able to hear every other node in the network. This is motivated by the well-known fact that energy usage can be reduced significantly if the transmissions are carried out in a multihop fashion. This is a consequence of the well-known propagation model used to model wireless communication channels, whereby the received energy decreases as $r^{-\alpha}$ with a distance of r, where $\alpha > 2$ is a constant depending upon the environment. The details of this model will be discussed in the next section. Thus, instead of each sensor sending its in-

formation to the fusion center directly, it is more efficient from an energy consumption point of view to send the information through relay nodes. It may be possible to reduce energy consumption even further by using some form of in-network data processing. This may have further benefits; for instance, if all the sensor measurements are to be transmitted from the sensors to the fusion center, then relay nodes closer to the fusion center would transmit more frequently than nodes that are further away from the fusion center. Thus, in-network processing to reduce the number of transmissions could be beneficial for eliminating hot spots.

Fundamentally, application-specific design is the feature that distinguishes multihop wireless networks used for sensing from multihop wireless networks used for communication. In multihop wireless communication networks, the protocols are designed so that they are not application-specific, and therefore the network can support a constantly evolving set of applications. Contrasting this, in multihop sensor networks, the architecture and protocols can be designed for each specific application, exploiting its structure, to reduce the energy usage within the network. This is the motivation for the recent works reported in [4] and [8]. In [4], the authors have designed a block coding scheme to compress the amount of information to be transmitted in a sensor network computing some functions. In [8], the authors investigate the optimal computation time and the minimum energy consumption required to compute the maximum of the sensor measurements. However, the in-network processing that we consider in this correspondence is different from the processing considered in [4] and [8], where the communication channels are assumed to be reliable, and the processing is to primarily exploit the geometric distribution of sensors [8] or the spatio–temporal correlations [4]. In our problem, processing is required not only to reduce the redundancy in the information to be conveyed in the fusion center, but also to introduce some redundancy to combat the effect of the noisy channels in the sensor network. Our results show that the additional redundancy required to combat channel errors does not significantly negate the benefits of in-network computation used to eliminate redundancy in the information, and the combination of in-network computation and channel coding could reduce the number of transmissions required in multihop networks to the same order as the number required in single-hop networks. It is easy to see that the number of transmissions required in a multihop network is lower bounded by the number of transmissions required in a single-hop network with the same number of sensors since in the single-hop network, a transmission can be heard by all nodes in the network.

The main results of the correspondence are as follows:

1) We use the routing protocol in [4] along with ideas from distributed parity computation in noisy networks ([3]) to devise near energy-optimal algorithms for counting in sensor networks. A key difference between our work and the work in [3] is that, in the case of sensor networks, the fusion center does not communicate directly with each of the sensors. Thus, local computation is necessary before conveying some aggregate information in a multihop fashion to the fusion center. Further, we require that the computation be accurate uniformly over all cells; i.e., we require the computation be accurate in all cells with high probability. In addition, error-correction coding, that is more sophisticated than repetition coding, seems to be required in the algorithms to minimize the energy required for counting.
2) Using the above ideas, we first study the case where each sensor has only one measurement to report, and show the amount of energy required for counting is
$$\kappa \lceil \log_2 m \rceil \, n(\log\log n) \left(8\sqrt{\frac{\log n}{n}}\right)^{\alpha}.$$
3) We then extend to the case where each sensor has N measurements, and the symmetric function needs to be computed for each measurement. We show that the total transmission energy consumption can be reduced to
$$O\left(n\left(\max\left\{\lceil \log_2 m \rceil, \frac{\log\log n}{N}\right\} + m\right)\left(\sqrt{\frac{\log n}{n}}\right)^{\alpha}\right)$$
per measurement. When $N = \Omega(\log\log n)$, the energy consumption is
$$\Theta\left(n\left(\sqrt{\frac{\log n}{n}}\right)^{\alpha}\right)$$
per measurement, which is a tight bound.

The rest of the correspondence is organized as follows. In Section II, we introduce our sensor network model and present a straightforward lower bound on the minimum energy needed. In Section III, we consider the case where the sensors have only one measurement to report. In Section IV, we investigate the impact of transmitting N measurements. Finally, in Section V, we conclude the correspondence and point out some future research directions.

II. Model

We consider a random network of n sensors that are uniformly and independently distributed on a unit square. Upon the occurrence of a certain event, sensor k records b_k, where b_k can take a value from $\{0, \ldots, m-1\}$, and thus, can be represented by $\lceil \log_2 m \rceil$ bits. The sensors have the capability to transmit this data over noisy wireless channels, and based on the data transmitted by the sensors in the network, a fusion center tries to evaluate some symmetric function $f(b_1, \ldots, b_n)$, i.e., a function which has the property
$$f(b_1, \ldots, b_n) = f(\sigma(b_1, \ldots, b_n))$$
for any permutation σ. Symmetric functions form a large class of functions, which includes almost all statistical functions like max, min, mean, etc. A key property of a symmetric function is that the function value only depends on the frequency-histogram. So in this correspondence, we will design algorithms to count the number of the sensors that have each value l, for each l, i.e., we compute
$$\sum_{i=1}^{n} 1_{\{b_i = l\}}$$
for each integer $l \in \{0, \ldots, m-1\}$, where $1_{\{b_i=l\}}$ is the indicator function such that $1_{\{b_i=l\}} = 1$ when $b_i = l$, and $1_{\{b_i=l\}} = 0$ otherwise. Since counting and computation are equivalent for symmetric functions, we will interchangeably use the terms counting and computation in this correspondence.

Let S_i denote the location of sensor i and $|S_i - S_j|$ denote the Euclidean distance from sensor i to sensor j. We use the protocol model in [7] with some additional assumptions.

1) All nodes use the same transmission radius r, and the power used to transmit one bit is r^{α}.
2) A transmission from sensor i can be received at sensor j only if $|S_i - S_j| \leq r$ and $|S_k - S_j| \geq (1+\Delta)r$ for each sensor $k \neq i$ which transmits at the same time, where Δ is a protocol-specified guard-factor to prevent interference.
3) A binary modulation scheme is used so that each transmission is either 1 or 0.
4) Even if a transmission is received at the receiver, there is some probability $p < 1/2$ with which the received bit is flipped, i.e., the channel is a binary symmetric channel with error probability p.

Note that this model only holds when the near field effects are negligible, which is assumed in this correspondence. Also note that the received power decreases with distance r. Thus the received power at the nodes which are at a distance r from the transmitter is constant as r decreases (or equivalently, as n increases); for this reason we can model the channels as binary symmetric channels with fixed error probability p. Furthermore, note that in typical sensor networks with small devices, the power levels and PHY layer schemes have to be chosen before deployment, so we assume that the common transmission radius r is chosen *a priori*. We would like to point out that the reason we investigate the sensor network model with assumptions (1)–(4) is that the focus of this correspondence is to propose distributed function computation algorithms without network synchronization and sophisticated signal processing schemes as in [11], [1], [12], which might be used to further reduce energy consumption.

By a computation algorithm, we mean a set of protocols (which may depend on n) to convey the appropriate information from the sensors to the fusion center and a protocol at the fusion center to use the received information to compute the frequency-histogram of the sensor measurements. Given an algorithm for counting, we define the energy required by the algorithm to be the maximum energy required for the computation over all possible values of the measurements. Our goal is to characterize the minimum energy required subject to the constraint that the probability of error in the computation in a random network with n sensors goes to zero as n goes to infinity. In this correspondence, we only consider the transmission energy used for counting, and assume other energy expenditure, due to computation, reception, coordination, etc., is negligible.

Before we investigate the counting problem, we present two well-known results for our convenient reference. First, we study the error probability when using repetition coding. Consider a binary symmetry channel with error probability p where each bit is transmitted M times, and the receiver decodes the data using a majority rule. Then we have following well-known bound [2] on the error probability.

Lemma 1: Suppose one bit of data is transmitted M times over a binary symmetric channel with error probability p, and the receiver decodes the bit using a majority rule. Then, the probability of decoding error is no greater than

$$(4p(1-p))^{\frac{1}{2}M}. \qquad \square$$

We also need the following coding theorem [2] for discrete memoryless channels for our analysis.

Theorem 2 (Gallager's Coding Theorem): For any discrete memoryless channel with capacity C, any positive integer N, and any positive $R < C$, there exist block codes with $M = 2^{NR}$ codewords of length N for which the decoding error probability of each codeword is less than $4e^{-NE_r(R)}$, where $E_r(R)$ is a non-increasing function of R. $\square$

It is obvious that each sensor has to broadcast its value once. Thus, we have the following lemma.

Lemma 3 (A Trivial Lower Bound): The minimum total transmission energy required to count is

$$\lceil \log_2 m \rceil \, n \left(\sqrt{\frac{\log n}{\pi n}} \right)^{\alpha}. \tag{1}$$

Proof: First, connectivity of the network is a necessary condition of correct counting. To guarantee connectivity with probability approaching 1 as the number of nodes goes to infinity, it has been shown in [6] that the transmission range of the sensors should be greater than $\sqrt{\frac{\log n}{\pi n}}$. Thus, the energy used per sensor transmission is $\left(\sqrt{\frac{\log n}{\pi n}} \right)^{\alpha}$. There are n sensors in the network, each of which must transmit $\lceil \log_2 m \rceil$ bits to convey its m–ary value. Thus, the total transmission energy required is at least $\lceil \log_2 m \rceil \, n \left(\sqrt{\frac{\log n}{\pi n}} \right)^{\alpha}$. $\square$

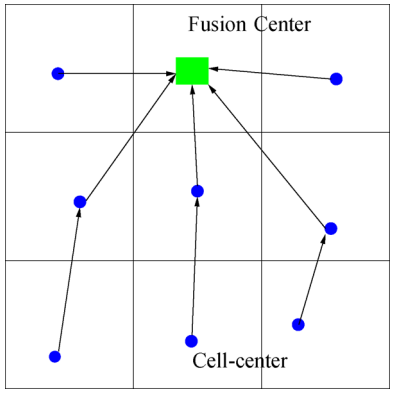

Fig. 1. A wireless sensor network.

Now, we consider the counting problem in detail. We first define the routing strategy, which is the same as the one in [4]. To transmit sensor information to the fusion center, we divide the unit square area into a regular lattice of B cells where $B = D^2$ and D is a positive integer. It is easy to see that

$$E[\text{Number of sensors in each cell}] = \frac{n}{B}.$$

In [9], [17], and [16], it has been shown that the number of sensors in each cell is n/B with high probability when $B = O\left(\frac{n}{\log n}\right)$. Thus, we choose

$$B = \left(\left\lfloor \sqrt{\frac{n}{c_1 \log n}} \right\rfloor \right)^2 \tag{2}$$

according to [16], where $c_1 > 0$, and have following lemma.

Lemma 4 ([16, Lemma 1]): Suppose that the unit square is partitioned into B square cells, where B is chosen as in (2), and further let n_i denote the number of sensors in cell i. Then, for large enough n

$$\Pr\left(\frac{c_1 \log n}{2} \le n_i \le 4 c_1 \log n \ \forall i \right) > 1 - \frac{2n^{\left(1-\frac{c_1}{8}\right)}}{c_1 \log n}. \tag{3}$$

$\square$

From above lemma, we know that if $c_1 \ge 8$, $\max_i n_i = O(\log n)$ and $\min_i n_i = \Omega(\log n)$ both hold.

Then, we adopt the hierarchical architecture of [4]: For each cell, we choose one sensor as the cell-center. Then designating the fusion center as the root, we form a rooted tree like Fig. 1, whose vertices include all the cell-centers, and whose links can only be between cell-centers of adjacent (common edge or corner) cells. Define $P(i)$ to be the parent of cell-center i, $C(i)$ to be the set of the children of cell-center i in the rooted tree, $H_{\max}$ to be the depth of the tree, and $H(i)$ to be the depth of the cell-center i in the tree ($H(\text{fusion center}) = 0$). Further, fix the transmission range

$$r = \sqrt{\frac{8}{B}}, \tag{4}$$

which guarantees a sensor can reach any other sensors within adjacent cells, and thus guarantees the network is connected if there is at least one sensor in each cell.

Now, given the routing strategy, we will next define protocols for intracell and intercell information processing and data aggregation. The protocols will have two distinct parts:

1) Intracell-Protocol: The information within cells is aggregated at the respective cell-centers.

2) Intercell-Protocol: The information aggregated by cell-centers is transmitted, and aggregated further, along the rooted tree to the fusion center.

Throughout the correspondence, B is chosen as in (2) with $c_1 = 8$, giving

$$r = 8\sqrt{\frac{\log n}{n}}.$$

Note that the routing strategy is well-defined under the assumption that the number of nodes in each cell is at least $4\log n$ and no more than $32\log n$, which is typical and happens with probability approaching 1 as n goes to infinity. Thus in the following sections, we will propose distributed algorithms which are asymptotically correct under the same assumption, and we assume that the algorithms report an error if the assumption does not hold. For ease of notation, we also define $\lambda = -\log(4p(1-p))$.

III. An Upper Bound on the Energy Consumption

We use the idea in [3] to design an algorithm for which the energy consumed is

$$\kappa \lceil \log_2 m \rceil \, n \, (\log\log n) \left(8\sqrt{\frac{\log n}{n}}\right)^{\alpha}$$

where $\kappa = \left\lceil \frac{4}{-\log(4p(1-p))} \right\rceil$. In wireless sensor networks, transmissions by a sensor can be heard by any sensor within its transmission range. Suppose there are $\tilde{n}$ sensors in sensor k's transmission range, then there are $\tilde{n}$ independent receptions for each measurement sent by sensor k. The main idea in [3] is to use the reception diversity to obtain a good estimate of the measurement transmitted by sensor k. But it requires additional transmissions among sensors; for example, it takes $\tilde{n}$ more transmissions for $\tilde{n}$ sensors to report the measurement they received from sensor k. We will show how to use in-network processing to reduce the number of transmissions required to exploit the reception diversity.

Now we propose following algorithm, which we call Counting-Algorithm-I. Note that the unit square is divided into square cells, and we assume that each cell contains $\Theta(\log n)$ nodes. In Intracell-Protocol-I, each node first broadcasts its measurement $\Theta(\log\log n)$ times so that the other nodes in the same cell can estimate the measurement correctly with probability $1-\Theta(1/\log n)$. Then every node computes the frequency-histogram of its cell, and reports to the cell-center. Thus every cell-center receives $\Theta(\log n)$ independent estimations, and can compute the frequency-histogram of its cell correctly with probability $1-\Theta(1/n)$. Since the number of cells is $\Theta(n/\log n)$, we further have that, under Intracell-Protocol-I, the probability that all cell-centers obtain the accurate frequency-histograms is $1-\Theta(1/\log n)$. After executing Intracell-Protocol-I, the cell-centers report the frequency-histograms along the rooted-tree under Intercell-Protocol-I. Every cell-center first aggregates its own histogram with the histograms obtained from its children, and then reports to its parent. Under Intercell-Protocol-I, block codes are used to guarantee that the probability that the fusion center obtains the correct counting result is $1-\Theta(1/\log n)$, given that all cell-centers have the correct frequency-histograms of their own cells.

Recall that b_k is the measurement sensor k has. For cell i, define Δ_i as the set of indices of the sensors in cell i, and γ_i as the counting of cell-center i, so γ_i is a vector with length m such that the l^{th} entry is

$$\gamma_i[l] = \sum_{k\in\Delta_i} 1_{\{b_k=l\}}$$

if the counting is correct.

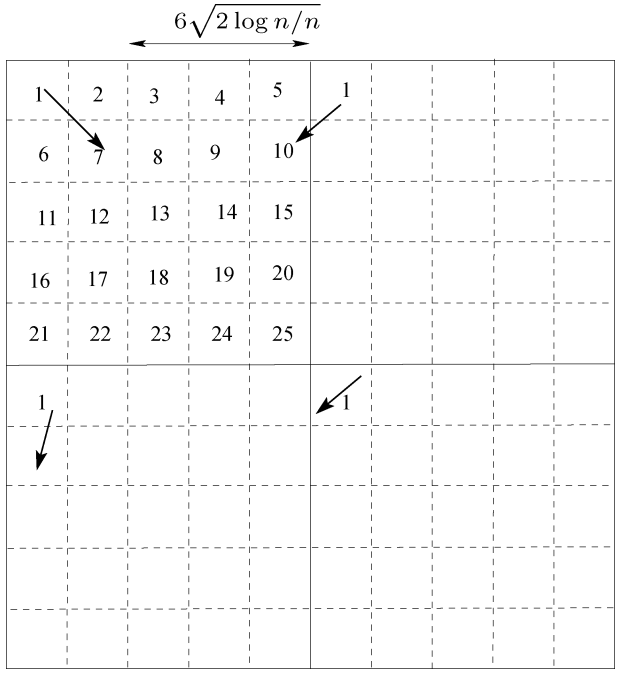

Fig. 2. Cell scheduling ($\Delta = 0.05$)).

A. Counting-Algorithm-I

If the number of nodes in each cell is no less than $4\log n$ and no more than $32\log n$, then the following algorithm is executed.

Simultaneous transmissions may interfere with each other, so we adopt the cell scheduling scheme used in [7], [9], [4]: We group every $C_\Delta \times C_\Delta$ cells into a super-cell, where $C_\Delta = \lceil 2+2\sqrt{2}(1+\Delta)\rceil$, and index the cells from 1 to C_Δ^2. For example, when $\Delta = 0.05$, $C_{0.05} = 5$ and the grouping is as in Fig. 2. We further group every C_Δ^2 time slots into a super time-slot. At time slot i of every super time-slot, the cells with index i are chosen to be active. For example, all cells with index 1 (as in Fig. 2) are active in the first time slot of every super time-slot. In our algorithms, transmissions will occur only within a cell or between neighboring cells. Thus, it is easy to verify that there is only one transmitter within a distance $(1+\Delta)r$ for each receiver, and simultaneous transmissions do not interfere with each other under the cell scheduling. When a cell is active, the nodes in the cell transmit their measurements according to the following intracell-protocol.

1) Intracell-Protocol-I (At Cell i):

1) The sensors in cell i take turns to transmit their $\lceil \log_2 m \rceil$-bit measurement. When it is the turn of sensor k, it broadcasts its measurement $\left\lceil \frac{4}{\lambda}(\log\log n)\right\rceil$ times. Then, all other sensors in the cell will receive $\left\lceil \frac{4}{\lambda}(\log\log n)\right\rceil$ noisy copies from sensor k. Sensor j decodes each bit of b_k using majority rule, and obtains α_{jk}. Then, it sets A_j (a vector with length m) to be

$$A_j[l] = 1_{\{b_j=l\}} + \sum_{k\in\Delta_i, k\neq j} 1_{\{\alpha_{jk}=l\}}$$

after all sensors broadcast their measurements.

2) Select $\left\lceil \frac{n_i}{\log\log n}\right\rceil$ sensors in the cell. Each selected sensor j represents A_j using $m\lceil\log_2 n_i\rceil$ bits ($A_j[h]$ can be represented by $\lceil\log_2 n_i\rceil$ bits), codes it using a block code with rate R_1 such that $mE_r(R_1)/R_1 \geq 1$, and then broadcasts A_j once.

3) Suppose $\tilde{A}_j$ is the output of the binary symmetric channel between the cell-center and sensor j with input A_j. Cell-center i sets γ_i to be any mode of the sequence $\{\tilde{A}_j\}$. For example,

suppose that each sensor measurement can take on one of three values. In this case, the histogram of a cell is a vector of length three. Suppose that a cell center receives four histogram estimates from other nodes in its cell and say the estimates (including the estimate of the cell center) are $(0,3,2),(0,3,2),(0,2,3),(1,2,2)$, and $(0,3,2)$, it sets $\gamma_i = (0,3,2)$ since $(0,3,2)$ is the most frequently occurring vector estimate.

Cell scheduling for intercell transmissions: 1) Let $L = H_{\max}$; 2) cells with depth L are scheduled according to [7], [9], [4]. If $L \neq 0$, let $L = L-1$ and repeat step 2).

2) Intercell-Protocol-I: Define a vector η_i with length m to be the aggregated information of the subtree rooted at cell-center i. When cell-center i is scheduled, cell-center i sets η_i such that

$$\eta_i[l] = \gamma_i[l] + \sum_{j \in C(i)} \tilde{\eta}_j[l]$$

where $\tilde{\eta}_j$ is the output of the channel between cell center j and cell center i with input η_j. Since $0 \leq \eta_i[l] \leq n$ for $0 \leq l \leq m$, η_i can be represented using $m\lceil \log_2 n \rceil$ bits. If i is the fusion center, then $\gamma_c = \eta_i$. Otherwise, it transmits η_i to cell-center $P(i)$ using a block code with rate R_2 such that $mE_r(R_2)/R_2 > 1$.

We now analyze the energy requirement of Counting-Algorithm-I. First, in Lemma 5, we show that under Intracell-Protocol-I

$$\Pr(\text{All } \gamma_i \text{ are correct } \mid 4\log n \leq n_i \leq 32 \log n \ \forall i) \geq 1 - \frac{1}{8\log n}.$$

Then, in Lemma 6, we show that

$$\Pr(\gamma_c \text{ is correct} \mid \gamma_i \text{ is correct } \forall i) \geq 1 - \frac{1}{2\log n}.$$

Finally, Theorem 7 quantifies the energy requirement of Counting-Algorithm-I.

Lemma 5: By executing Intracell-Protocol-I, the cell-centers can obtain γ_i with

$$\Pr\left(\text{All } \gamma_i \text{ are correct} \;\middle|\; 32 \geq \frac{n_i}{\log n} \geq 4 \ \forall i\right) \geq 1 - \frac{1}{8\log n} \quad (5)$$

and the number of transmissions required in cell i is upper bounded by

$$\kappa \lceil \log_2 m \rceil n_i (\log\log n)$$

where $\kappa = \left\lceil \frac{4}{-\log(4p(1-p))} \right\rceil$.

Proof: In the following analysis, we assume $4\log n \leq n_i \leq 32\log n$ holds for all i. First, the number of transmissions in each cell under Intracell-Protocol-I is

$$n_i \lceil \log_2 m \rceil \left\lceil \frac{4}{\lambda}(\log\log n) \right\rceil + \left\lceil \frac{n_i}{\log\log n} \right\rceil \frac{m\lceil \log_2 n_i \rceil}{R_1}.$$

Note that m and R_1 are constants independent of n, and $n_i = \Theta(\log n)$. Thus we have

$$n_i \lceil \log_2 m \rceil \left\lceil \frac{4}{\lambda}(\log\log n) \right\rceil = \Theta((\log n)(\log\log n))$$

and

$$\left\lceil \frac{n_i}{\log\log n} \right\rceil \frac{m\lceil \log_2 n_i \rceil}{R_1} = \Theta(\log n)$$

which implies that for large enough n

$$n_i \lceil \log_2 m \rceil \left\lceil \frac{4}{\lambda}(\log\log n) \right\rceil + \left\lceil \frac{n_i}{\log\log n} \right\rceil \frac{m\lceil \log_2 n_i \rceil}{R_1} < \kappa \lceil \log_2 m \rceil n_i (\log\log n).$$

Next we investigate the probability that γ_i is correct, i.e.,

$$\gamma_i[l] = \sum_{k \in \Delta_i} 1_{\{b_k = l\}}$$

for all l. Recall that sensor j decodes each bit of b_k using majority rule, so from Lemma 1 and the union bound, we have

$$\Pr(\alpha_{jk} = b_k) \geq 1 - \lceil \log_2 m \rceil (4p(1-p))^{\frac{2\log\log n}{\lambda}}.$$

Note that A_j is correct if α_{jk} is correct for all $k \in \Delta_i$. From the union bound, we have

$$\begin{aligned}\Pr\left(A_j[l] = \sum_{k\in\Delta_i} 1_{\{b_k=l\}} \ \forall\, l\right) &\geq 1 - n_i \lceil \log_2 m \rceil (4p(1-p))^{\frac{2\log\log n}{\lambda}} \\ &\geq 1 - \frac{32\lceil \log_2 m \rceil}{\log n}.\end{aligned}$$

Consider step 2) of Intracell-Protocol-I, from Theorem 2

$$\Pr\left(\tilde{A}_j = A_j\right) \geq 1 - 4e^{-\frac{E_r(R_1)}{R_1} m \log_2 n_i} \geq 1 - 4e^{-\log\log n}$$

where the last inequality holds because $n_i \geq 4\log n > \log n$. Thus,

$$\Pr\left(\tilde{A}_j[l] = \sum_{k\in\Delta_i} 1_{\{b_k=l\}} \ \forall\, l\right) \geq 1 - \frac{32\lceil \log_2 m \rceil + 4}{\log n}.$$

Note that $\{\alpha_{jk}\}$ are i.i.d. for all $j \in \Delta_i$, so $\{A_j\}$ are identical and $\{\tilde{A}_j\}$ are i.i.d. Now define i.i.d. random variables $\{I_j\}$ such that $I_j = 1$ if

$$\tilde{A}_j[l] = \sum_{k\in\Delta_i} 1_{\{b_k=l\}}$$

for all l; and $I_j = 0$ otherwise. Since γ_i is the mode of $\{\tilde{A}_j\}$, from Lemma 1, we have

$$\begin{aligned}\Pr(\gamma_i \text{is correct}) &= \Pr\left(\gamma_i[l] = \sum_{k\in\Delta_i} 1_{\{b_k=l\}} \ \forall l\right) \\ &\geq \Pr\left(\sum_j I_j \geq \frac{1}{2} n_i\right) \\ &\geq 1 - \left(4\left(\frac{32\lceil \log_2 m \rceil + 4}{\log n}\right) \times \left(1 - \frac{32\lceil \log_2 m \rceil + 4}{\log n}\right)\right)^{\frac{n_i}{2\log\log n}} \\ &\geq 1 - e^{-(\log\log n - \log(128\lceil \log_2 m \rceil + 16))\frac{n_i}{2\log\log n}} \\ &\geq 1 - e^{-\log n}\end{aligned}$$

where the second inequality follows from Lemma 1. There are at most $\frac{n}{8\log n}$ cells in the network, so

$$\begin{aligned}\Pr\left(\gamma_i[l] = \sum_{k\in\Delta_i} 1_{\{b_k=l\}} \ \forall i, l\right) &\geq 1 - \frac{n}{8\log n} e^{-\log n} \\ &= 1 - \frac{1}{8\log n}\end{aligned}$$

and the lemma holds. □

Now, suppose that all γ_i are correct. Since η_i can be represented using $m\lceil \log_2 n \rceil$ bits, each cell-center has $m\lceil \log_2 n \rceil$ bits to transmit under Intercell-Protocol-I.

Lemma 6: Suppose all cell-centers have the correct γ_i, then under Intercell-Protocol-I, the probability that the fusion center obtains the correct γ_c is bounded as follows:

$$\Pr\left(\gamma_c[l]=\sum_k 1_{\{b_k=l\}}\ \forall l \,\middle|\, \gamma_i[l]=\sum_{k\in\Delta_i} 1_{\{b_k=l\}}\ \forall i,l\right) \geq 1-\frac{1}{2\log n} \quad (6)$$

and the number of transmissions required is $\Theta(n)$.

Proof: Each cell-center transmits $m\lceil\log_2 n\rceil$ bits under Intercell-Protocol-I, and there are $\Theta(n/\log n)$ cell-centers, so that the number of bits transmitted under Intercell-Protocol-I is $\Theta(n)$. Now, suppose all cell-centers have the correct γ_i, then γ_c is also correct if all η_i's are correctly received. From Theorem 2, there exists a block code satisfying the conditions given in Intercell-Protocol-I. Thus, for a given i,

$$\begin{aligned}\Pr(\eta_i \text{ is correctly received}) &\geq 1-4e^{-\frac{E_r(R_2)}{R_2}m\log_2 n}\\ &\geq 1-4e^{-\log n}\end{aligned}$$

and from the union bound

$$\begin{aligned}&\Pr\left(\gamma_c[l]=\sum_k 1_{\{b_k=l\}}\ \forall l \,\middle|\, \gamma_i[l]=\sum_{k\in\Delta_i} 1_{\{b_k=l\}}\ \forall i,l\right)\\ &=\Pr\left(\tilde{\eta}_i=\eta_i\ \forall i \,\middle|\, \gamma_i[l]=\sum_{k\in\Delta_i} 1_{\{b_k=l\}}\ \forall i,l\right)\\ &\geq 1-\frac{4n}{8\log n}e^{-\log n}\\ &=1-\frac{1}{2\log n}.\end{aligned}$$

□

In Lemmas 5 and 6, we have shown that, under Algorithm-I, counting is accurate with high probability when the number of sensors is large enough. Next, we provide an upper bound on the energy requirement to solve our counting problem.

Theorem 7: Any symmetric function can be computed accurately with a total transmission energy consumption no more than

$$\kappa\lceil\log_2 m\rceil\, n(\log\log n)\left(8\sqrt{\frac{\log n}{n}}\right)^{\alpha}$$

where $\kappa=\left\lceil\frac{4}{-\log(4p(1-p))}\right\rceil$. Counting-Algorithm-I is an asymptotically correct algorithm that achieves this energy consumption. Specifically, the probability of computation error at the fusion center is upper bounded by $\frac{7}{8\log n}$.

Proof: From inequalities (3), (5) and (6), we have

$$\Pr\left(\gamma_c[l]=\sum_k 1_{\{b_k=l\}}\ \forall l\right)\geq 1-\frac{7}{8\log n}$$

which converges to one when n goes to infinity. So Counting-Algorithm-I is asymptotically correct.

Further, from Lemma 5 and Lemma 6, it is easy to see the number of transmissions under Counting-Algorithm-I is not more than $\kappa\lceil\log_2 m\rceil\, n(\log\log n)$. Since the common transmission range is $8\sqrt{\frac{\log n}{n}}$, the total energy consumption is

$$\kappa\lceil\log_2 m\rceil\, n(\log\log n)\left(8\sqrt{\frac{\log n}{n}}\right)^{\alpha}. \quad (7)$$

The theorem holds because there may exist other algorithms that consume less energy. □

Remarks:

1) In the Counting-Algorithm-I, block codes are used in the step 2) of the Intracell-Protocol-I and in the Intercell-Protocol-I. Instead, we could use simple repetition coding. Using the repetition coding in the step 2) of the Intracell-Protocol-I will increase the number of transmissions, but will not change the order of magnitude. However, in the Intercell-Protocol, using block codes to transmit the aggregated information η_i is crucial to reduce the number of transmissions. To see this, suppose we transmit each bit M times so that each bit is correctly decoded with probability E_M. We have

$$\frac{n}{8\log n}\times m\lceil\log_2 n\rceil > \frac{mn}{8}$$

bits to transmit, and all of them should to be correctly received. So the probability of obtaining the correct γ_c is upper bounded by

$$(1-E_M)^{\frac{mn}{8}}.$$

To guarantee the error probability to be small, it requires $E_M=O\left(\frac{1}{n}\right)$ and $M=\Omega(\log n)$. So the number of transmissions is

$$\frac{n}{8\log n}\times m\lceil\log_2 n\rceil\times\Omega(\log n)=\Omega(n\log n)$$

which is much larger than $O(n\log\log n)$.

2) A simple lower bound has been obtained in Lemma 3. Comparing it with the upper bound in Theorem 7, we can see that the upper bound differs by a factor of *only* $(\log\log n)$ from the lower bound. But it is still not clear how good our bound is. Consider the case when $m=2$ (binary data), a more general computational problem than ours, i.e., one of knowing all the bits in the network, is considered for a broadcast network in [3]. The number of transmissions required there is also shown to be $O(n(\log\log n))$. This suggests that one may be able to improve our upper bound on the energy usage since counting is easier than detecting all the bits in the network. On the other hand, parity computation which is a simpler problem than counting is also studied in [3], but the number of transmissions needed is again $O(n(\log\log n))$, the same complexity as Counting-Algorithm-I. To the best of our knowledge, this is the best upper bound in the literature for parity computation in broadcast networks if the error is required to go to zero when n increases [5]. Further, our network with its multihop architecture also requires more transmissions for the data from the sensors to reach the fusion center. This suggests that our upper bound on energy usage is quite reasonable.

IV. The Impact of Long-Block Measurements

In Section III, we considered the case where each sensor has only one measurement to transmit. In this section, we will investigate the impact of transmitting N measurements, and each measurement can take the integer value from $\{0,\ldots,m-1\}$, and so can be represented by $\lceil\log_2 m\rceil$ bits. In such a case, we will show that block codes can be used in the intra-cell-protocol, and the number of transmissions can be further reduced. It will be shown that the energy consumption per observation approaches to the lower bound (1) when N increases, and the lower bound is achieved when $N=\Omega(\log\log n)$.

Define $\mathbf{b}_k$ to be a vector with length N, and $b_k[h]$ to be the h^{th} measurement of sensor k. The fusion center is interested in determining $\sum_k 1_{\{b_k[h]=l\}}$ for all l and h. From Theorem 2, if we have $\lceil \log_2 m \rceil N$ bits to transmit, there exist block codes with code length

$$\lceil \max\{N \lceil \log_2 m \rceil, \log\log n\}/R \rceil \tag{8}$$

and the decoding error probability of each codeword less than

$$4e^{-2\max\{N\lceil \log_2 m\rceil, \log\log n\}}$$

where the code length is chosen as in expression (8) to guarantee that the receiver can decode the data with probability at least $1-\Theta(1/\log n)$. In the following algorithm, we use additional block codes in the intracell-protocol to reduce the number of transmissions per measurement.

A. Counting-Algorithm-II

If the number of nodes in each cell is no less than $4\log n$ and no more than $32\log n$, then the following algorithm is executed.

Cell scheduling for intracell transmissions and intercell transmissions is the same as in Counting-Algorithm-I.

1) Intracell-Protocol-II (At Cell i):

1) The sensors in cell i take turns to transmit their bits. If it is sensor k's turn, it encodes $\mathbf{b}_k$ using a block code with code length $\left\lceil \frac{\max\{N\lceil \log_2 m\rceil, \log\log n\}}{R} \right\rceil$ and suppose that either N or n is large enough such that the decoding error probability of each codeword less than $4e^{-2\max\{N\lceil \log_2 m\rceil, \log\log n\}}$. The codeword for $\mathbf{b}_k$ is then broadcast once. Suppose $\boldsymbol{\alpha}_{jk}$ is the output of the binary symmetric channel between sensor k and sensor j with input $\mathbf{b}_k$. After all sensors broadcast their measurements, sensor j obtains an $m \times N$ matrix, where $[l,h]$ element is

$$A_j[l,h] = 1_{\{b_j[h]=l\}} + \sum_{k\in\Delta_i, k\neq j} 1_{\{\alpha_{jk}[h]=l\}}$$

2) Select $\left\lceil \frac{n_i}{\log\log n} \right\rceil$ sensors. Each selected sensor j represents A_j using $mN\lceil \log_2 n_i \rceil$ bits (each entry of A_j can be represented by $\lceil \log_2 n_i \rceil$ bits), encodes it using a block code with rate R_1 such that $mNE_r(R_1)/R_1 \geq 1$, and transmits A_j to the cell center once.

3) Suppose $\tilde{\mathbf{A}}_j$ the output of the binary symmetric channel between the cell-center and sensor j with input $\mathbf{A}_j$. Cell-center i sets γ_i to be any mode of the sequence $\{\tilde{\mathbf{A}}_j\}$.

2) Intercell-Protocol-II: Define a $m \times N$ matrix η_i to be the aggregated information of the subtree rooted at cell-center i. When cell-center i is scheduled, cell-center i sets η_i such that

$$\eta_i[l,h] = \gamma_i[l,h] + \sum_{j\in C(i)} \tilde{\eta}_j[l,h]$$

where $\tilde{\eta}_j$ is the output of the channel between cell center j and cell center i with input η_j. Since $0 \leq \eta_i[l,h] \leq n$ for $0 \leq l \leq m$, η_i can be represented using $mN\lceil \log_2 n \rceil$ bits. If i is the fusion center, then $\gamma_c = \eta_i$. Otherwise, it transmits η_i to cell-center $P(i)$ using a block code with rate R_2 such that $mNE_r(R_2)/R_2 > 1$.

Theorem 8: Suppose all sensors have N measurements to report, then the frequency-histogram can be computed accurately with high probability and with a total transmission energy consumption

$$O\left(n\left(\max\left\{\lceil \log_2 m\rceil, \frac{\log\log n}{N}\right\}+m\right)\left(\sqrt{\frac{\log n}{n}}\right)^{\alpha}\right)$$

per measurement, and Counting-Algorithm-II is asymptotically correct. Specifically, the probability of computation error at the fusion center is upper bounded by $\frac{5}{8\log n}$.

Proof: Suppose $4\log n \leq n_i \leq 32\log n$ holds for all i. First, consider the number of bits transmitted under Counting-Algorithm-II. There are

$$\left\lceil \frac{\max\{N\lceil \log_2 m\rceil, \log\log n\}}{R} \right\rceil n_i + \frac{mN\lceil \log_2 n_i\rceil}{R_1}\left\lceil \frac{n_i}{\log\log n}\right\rceil$$

bits transmitted in each cell under Intracell-Protocol-II. Thus, the total number of bits transmitted in the network under Intracell-Protocol-II is

$$\Theta\left(n\left(\max\left\{\lceil \log_2 m\rceil, \frac{\log\log n}{N}\right\}+m\right)\right)$$

per measurement. The number of bits transmitted under Intercell-Protocol-II is $\Theta(mn)$ per measurement. Thus, under Counting-Algorithm-II, the energy required per measurement is

$$\Theta\left(n\left(\max\left\{\lceil \log_2 m\rceil, \frac{\log\log n}{N}\right\}+m\right)\left(\sqrt{\frac{\log n}{n}}\right)^{\alpha}\right)$$

which implies that the minimum transmission energy required per observation is

$$O\left(n\left(\max\left\{\lceil \log_2 m\rceil, \frac{\log\log n}{N}\right\}+m\right)\left(\sqrt{\frac{\log n}{n}}\right)^{\alpha}\right)$$

since there may exist other algorithms that consume less energy.

Next, we study the probability of correct counting under Counting-Algorithm-II. We will show

$$\Pr\left(\gamma_i[l,h] = \sum_{k\in\Delta_i} 1_{\{b_k[h]=l\}}\ \forall i,l,h \;\middle|\; 32 \geq \frac{n_i}{\log n} \geq 4\ \forall i\right) \geq 1 - \frac{1}{8\log n}$$

and the remainder of the proof follows from Lemma 6 and Theorem 7.

First, from Theorem 2, we have

$$\Pr(\alpha_{jk} = \mathbf{b}_k) \geq 1 - 4e^{-2\max\{N\lceil \log_2 m\rceil, \log\log n\}} \geq 1 - \frac{4}{(\log n)^2}.$$

Since $n_i \leq 32\log n$, from the union bound,

$$\Pr\left(A_j[l,h] = \sum_{k\in\Delta_i} 1_{\{b_k[h]=l\}}\ \forall\, l,h\right) \geq 1 - n_i\frac{4}{(\log n)^2} \geq 1 - \frac{128}{\log n}.$$

Now suppose $\tilde{A}_j$ is the output of the channel between the cell center and sensor j with input A_j, so from Theorem 2

$$\Pr\left(\tilde{A}_j = A_j\right) \geq 1 - 4e^{-\frac{E_r(R_1)}{R_1}mN\log_2 n_i} \geq 1 - 4e^{-\log\log n}$$

and

$$\Pr\left(\tilde{A}_j \text{is correct}\right) = \Pr\left(\tilde{A}_j[l,h] = \sum_{k\in\Delta_i} 1_{\{b_k[h]=l\}}\ \forall\, l,h\right) \geq 1 - \frac{132}{\log n}.$$

Define independent and identically distributed (i.i.d.) random variables $\{I_j\}$ such that $I_j = 1$ if $\tilde{A}_j$ is correct and $I_j = 0$ otherwise. From Lemma 1,

$$\begin{aligned}\Pr(\gamma_i \text{ is correct}) &= \Pr\left(\gamma_i[l,h] = \sum_{k\in\Delta_i} 1_{\{b_k[h]=l\}}\ \forall\, l,h\right)\\ &\geq 1 - \Pr\left(\sum_j I_j \leq \frac{n_i}{2}\right)\\ &\geq 1 - \left(4\left(1-\frac{132}{\log n}\right)\left(\frac{132}{\log n}\right)\right)^{\frac{n_i}{2\log\log n}}\\ &\geq 1 - \left(\frac{528}{\log n}\right)^{\frac{n_i}{2\log\log n}}\\ &\geq 1 - e^{-\log n}\end{aligned}$$

for large n.

Thus

$$\begin{aligned}\Pr(\gamma_i \text{ is correct } \forall i) &\geq 1 - \frac{n}{8\log n}e^{-\log n}\\ &\geq 1 - \frac{1}{8\log n}\end{aligned}$$

and the theorem holds. □

From the theorem above, when $N = \Omega(\log\log n)$, the transmission energy required per measurement is $O\left(n\left(\sqrt{\frac{\log n}{n}}\right)^{\alpha}\right)$. Then, from the lower bound (1), we can conclude that when $N = \Omega(\log\log n)$, the transmission energy required is $\Theta\left(n\left(\sqrt{\frac{\log n}{n}}\right)^{\alpha}\right)$, which is tight.

V. Discussion and Conclusions

In this correspondence, we investigated counting problems in multihop networks with noisy communication channels. First, we considered sensors, each with a single measurement, and showed by construction that feasible algorithms exist whose energy consumption is class

$$O\left(\lceil\log_2 m\rceil\, n(\log\log n)\left(\sqrt{\frac{\log n}{n}}\right)^{\alpha}\right)$$

. Then, we considered the case where the sensors have N bits to report, in which case the transmission energy can be reduced to class

$$O\left(n\left(\max\left\{\lceil\log_2 m\rceil, \frac{\log\log n}{N}\right\} + m\right)\left(\sqrt{\frac{\log n}{n}}\right)^{\alpha}\right)$$

per measurement.

There are several directions for future work. First, while the ratio of the upper bound to the lower bound is only of the order of $\log\log n$, it still needs to be investigated whether

$$O\left(\lceil\log_2 m\rceil\, n(\log\log n)\sqrt{\frac{\log n}{n}}^{\alpha}\right)$$

is the best upper bound. Second, we have shown that the lower bound can be achieved if each sensor has $N = \Omega(\log\log n)$ bits, it is still an open problem whether $\log\log n$ is the minimum number of observed bits needed to achieve the lower bound.

References

[1] H. El Gamal, "On the scaling laws of dense wireless sensor networks," *IEEE Trans. Inf. Theory*, vol. 51, pp. 1229–1234, Mar. 2005.

[2] R. G. Gallager, *Information Theory and Reliable Communication*. New York, NY: Wiley, 1968.

[3] R. G. Gallager, "Finding parity in a simple broadcast network," *IEEE Trans. Inf. Theory*, vol. 34, pp. 176–180, 1988.

[4] A. Giridhar and P. R. Kumar, "Computing and communicating functions over sensor networks," *IEEE J. Sel. Areas Commun.*, vol. 23, no. 4, pp. 455–764, Apr. 2005.

[5] N. Goyal, G. Kindler, and M. Saks, "Lower bounds for the noisy broadcast problem," in *Proc. Ann. IEEE Symp. Found. Comput. Sci.*, Oct. 2005, pp. 40–49.

[6] P. Gupta and P. Kumar, "Critical power for asymptotic connectivity," in *Proc. Conf. Dec. Contr.*, 1998, pp. 1106–1110.

[7] P. Gupta and P. Kumar, "The capacity of wireless networks," *IEEE Trans. Inf. Theory*, vol. 46, pp. 388–404, 2000.

[8] N. Khude, A. Kumar, and A. Karnik, "Time and energy complexity of distributed computation in wireless sensor networks," in *Proc. IEEE INFOCOM*, 2005, pp. 2625–2637.

[9] S. R. Kulkarni and P. Viswanath, "A deterministic approach to throughput scaling in wireless networks," *IEEE Trans. Inf. Theory*, vol. 50, pp. 1041–1049, 2004.

[10] E. Kushilevitz and Y. Mansour, "Computation in noisy radio networks," in *Proc. 9th Ann. ACM-SIAM Symp. Discr. Algor.*, 1998, pp. 236–243.

[11] K. Liu and A. Sayeed, "Optimal distributed detection strategies for wireless sensor networks," in *Proc. 42nd Ann. Allerton Conf. Commun., Contr. Comput.*, Monticello, IL, Oct. 2004.

[12] G. Mergen and L. Tong, "Type based estimation over multiaccess channels," *IEEE Trans. Signal Processing*, vol. 54, no. 2, pp. 613–626, Feb. 2006.

[13] S. Rajagopalan and L. J. Schulman, "A coding theorem for distributed computation," in *Proc. 26th Ann. ACM Symp. Theory Comput.*, 1994, pp. 790–799.

[14] L. J. Schulman, "Communication on noisy channels: A coding theorem for computation," in *Proc. Ann. IEEE Symp. Found. Comput. Sci.*, Pittsburgh, PA, Oct. 1992, pp. 724–733.

[15] L. J. Schulman, "Deterministic coding for interactive communication," in *Proc. 25th Ann. ACM Symp. Theory Comput.*, 1993, pp. 747–756.

[16] S. Toumpis and A. J. Goldsmith, "Large wireless networks under fading, mobility, and delay constraints," in *Proc. IEEE INFOCOM*, 2004, vol. 1, pp. 619–627.

[17] F. Xue and P. R. Kumar, "The number of neighbors needed for connectivity of wireless networks," *Wireless Netw.*, vol. 10, no. 2, pp. 169–181, Mar. 2004.