

Distributed Symmetric Function Computation in Noisy Wireless Sensor Networks

Lei Ying, R. Srikant, and Geir E. Dullerud
University of Illinois at Urbana-Champaign
{lying,rsrikant,dullerud}@uiuc.edu

Abstract

We consider a wireless sensor network consisting of n sensors, and each sensor has a measurement, which is an integer value belonging to $\{0, \dots, m-1\}$, 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 is 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 paper 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 $N = \Omega(\log \log n)$.

I. NOTATIONS

The following notations are used throughout this paper, given non-negative functions $f(n)$ and $g(n)$:

- (i) $f(n) = O(g(n))$ means there exist positive constants c and m such $f(n) \leq cg(n)$ for all $n \geq m$.
- (ii) $f(n) = \Omega(g(n))$ means there exist positive constants c and m such that $f(n) \geq cg(n)$ for all $n \geq m$.

Namely, $g(n) = O(f(n))$.

- (iii) $f(n) = \Theta(g(n))$ means that both $f(n) = \Omega(g(n))$ and $f(n) = O(g(n))$ hold.

The research was supported by a Vodafone Fellowship and an AFOSR URI Grant F49620-01-1-0365

The first two authors are with the Department of Electrical and Computer Engineering and the Coordinated Science Lab and the third author is with the Department of Mechanical and Industrial Engineering and the Coordinated Science Lab.

II. 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 paper we will investigate this problem in multi-hop 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 paper, is as follows: the measurement of each node is an integer belonging to $\{0, \dots, m-1\}$, and the fusion center needs to decide, using information transmitted from the network, the number of sensors having value l , for all $l \in \{0, \dots, m-1\}$. 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, [1]) 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 paper 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 paper are related to the algorithms for distributed computation over noisy networks, which are studied in [2], [11], [12], [10], [9], and references within. 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 [2] and [9], 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 [2], [9] 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 [2] and [9] 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 [2] and [9]. The second difference is that each node may not be able to hear every other node in the network. The reason for this is that energy consumption can be an important consideration in wireless networks and it is well-known that energy usage can be reduced significantly if the transmissions are carried out in a multi-hop fashion. This is a consequence of the well-known propagation model used to model wireless communication channels, whereby the energy required to transmit over a distance of r is proportional to r^α , where $\alpha \geq 2$ 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 information 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 be depleted of their energy faster 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, this is the distinction between multi-hop wireless networks used for communication and multi-hop wireless networks used for sensing. In multi-hop 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 multi-hop 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 [3] and [7]. In [3], 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 [7], 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 paper is different from the processing considered in [3] and [7], where the communication channels are assumed to be reliable, and the processing is to primarily exploit the spatial correlation [7] or the spatio-temporal correlations [3]. 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 multi-hop networks to the same order as the number required in single-hop networks.

The main results of the paper are as follows:

- 1) We use the routing protocol in [3] along with ideas from distributed parity computation in noisy networks ([2]) to devise near energy-optimal algorithms for counting in sensor networks. A key difference between our work and the work in [2] 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 multi-hop fashion to the fusion center. Further, we require that the computation be accurate uniformly over all cells. In addition, sophisticated error-correction coding, not just 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 $\kappa \lceil \log_2 m \rceil n (\log \log n) \left(8 \sqrt{\frac{\log n}{n}} \right)^\alpha$, where n is the number of sensors in the network.
- 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 paper is organized as follows: In Section III, we introduce our sensor network model and present a straightforward lower bound on the minimum energy needed. In Section IV, we consider the case where the sensors have only one measurement to report. In Section V, we investigate the impact of transmitting N measurements. Finally, in Section VI, we conclude the paper and point out some future research directions.

III. MODEL

We consider a 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 be taken value from $\{0, \dots, m-1\}$, so 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 compute some symmetric function $f(b_1, \dots, b_n)$, i.e., a function which has the property

$$f(b_1, \dots, b_n) = f(\sigma(b_1, \dots, b_n)),$$

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 paper, we will design algorithms to count the number of the sensors having the same value l , for each l , i.e., to compute

$$\sum_{i=1}^n 1_{\{b_i=l\}}$$

for each integer $l \in \{0, \dots, m-1\}$, 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 paper.

Let S_i denote the location of sensor i and $|S_i - S_j|$ denote the distance from sensor i to sensor j . We use the protocol model in [6] with some additional assumptions.

- (1) A transmission from sensor i can be received at sensor j only if

$$|S_k - S_j| \geq (1 + \Delta)|S_i - S_j|$$

for each sensor $k \neq i$ which transmits at the same time, where Δ is a protocol-specified guard-zone to prevent interference.

- (2) A binary modulation scheme is used so that each transmission is either 1 or 0.
- (3) 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 .
- (4) The power required to transmit to a distance r is r^α .

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 goes to zero as $n \rightarrow \infty$.

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 [1] 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

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

□

We also need the following coding theorem [1] 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 .

□

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

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

Proof: First, connectivity of the network is a necessary condition of correct counting. To guarantee connectivity, it has been shown in [5] 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 make at least one transmission; thus, the total transmission energy required is $\lceil \log_2 m \rceil n \left(\sqrt{\frac{\log n}{\pi n}} \right)^\alpha$. ■

Now, we consider the counting problem in detail. We first define the routing strategy, which is the same as the one in [3]. 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 [8], [14], [13], 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

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

according to [13], where $c_1 > 0$, and have following lemma.

Lemma 4 ([13, 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 ,

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

□

From above lemma, we know that if $c_1 \geq 8$, all cells have at least $\frac{c_1 \log n}{2}$ sensors, and at most $4c_1 \log n$ sensors, which guarantees $n_i = \Theta(\log n)$ for all i with high probability.

Then, we adopt the hierarchical architecture of [3]: 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 Figure 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-

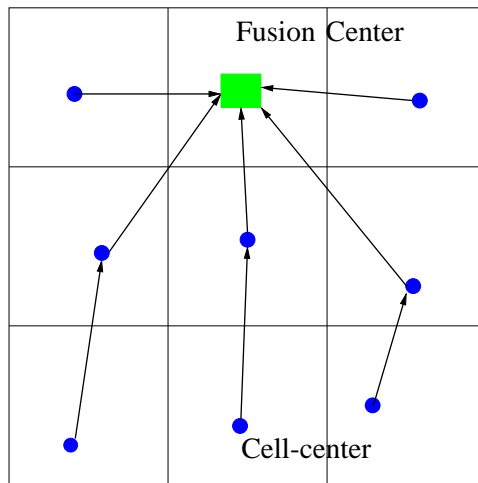


Fig. 1. A Wireless Sensor Network

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

$$r = \sqrt{\frac{8}{B}}, \quad (4)$$

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 intra-cell and inter-cell information processing and data aggregation. The protocols will have two distinct parts:

- (1) Intra-Cell-Protocol: The information within cells is aggregated at the respective cell-centers.
- (2) Inter-Cell-Protocol: The information aggregated by cell-centers is transmitted, and aggregated further, along the rooted tree to the fusion center.

Throughout paper, B is chosen as in (2) with $c_1 = 8$, so

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

We also define $\lambda = -\log(4p(1-p))$.

IV. AN UPPER BOUND ON THE ENERGY CONSUMPTION

We use the idea in [2] 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 n sensors in sensor k 's transmission range, then there are n independent receptions for each measurement sent by sensor k . The main idea in [2] 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 n more transmissions for 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.

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.

Now we propose following algorithm, which we call Counting-Algorithm-I.

Counting-Algorithm-I:

Adjacent cells may interfere with each other, so we adopt the cell scheduling scheme used in [6], [3].

Intra-Cell-Protocol-I (At cell i):

- (i) 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 $\lceil \frac{4}{\lambda} (\log \log n) \rceil$ 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.

- (ii) Select $\lceil \frac{n_i}{\log \log n} \rceil$ 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.
- (iii) 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 $\{\tilde{A}_j\}$.

Cell scheduling for inter-cell transmissions: (1) Let $L = H_{\max}$; (2) Cells with depth L are scheduled according to [6], [3]. If $L \neq 0$, let $L = L - 1$ and repeat step (2).

Inter-Cell-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 Intra-Cell-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 and Theorem 7, 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: Suppose $4 \log n \leq n_i \leq 32 \log n$ for all i . Then, by executing Intra-Cell-Protocol-I, the cell-centers can obtain γ_i with

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

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 Intra-Cell-Protocol-I is

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

It is easy to see 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 (ii) of Intra-Cell-Protocol-I, from Theorem 2,

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

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

$$\begin{aligned} \Pr(\gamma_i \text{ is correct}) &= \Pr\left(\gamma_i[l] = \sum_{k \in \Delta_i} 1_{\{b_k=l\}} \quad \forall l\right) \geq \Pr\left(\sum_j I_j \geq \frac{1}{2}n_i \quad \forall l\right) \\ &\geq 1 - \left(4 \left(\frac{32 \lceil \log_2 m \rceil + 4}{\log n}\right) \left(1 - \frac{32 \lceil \log_2 m \rceil + 4}{\log n}\right)\right)^{\frac{n_i}{2 \log \log n}} \\ &\geq 1 - e^{-(\log \log n - \log(64 \lceil \log_2 m \rceil + 16)) \frac{n_i}{2 \log \log n}} \geq 1 - e^{-\log n}. \end{aligned}$$

There are at most $\frac{n}{8 \log n}$ cells in the network, so

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

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 Inter-Cell-Protocol-I.

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

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

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

Proof: First, it is easy to see that the number of bits transmitted 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 step (i) of Inter-Cell-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,

$$\begin{aligned} &\Pr\left(\gamma_c[l] = \sum_k 1_{\{b_k=l\}} \quad \forall l \mid \gamma_i[l] = \sum_{k \in \Delta_i} 1_{\{b_k=l\}} \quad \forall i, l\right) \\ &= \Pr\left(\text{All } \eta_i \text{'s are correctly received} \mid \gamma_i[l] = \sum_{k \in \Delta_i} 1_{\{b_k=l\}} \quad \forall i, l\right) \\ &\geq 1 - \frac{4n}{8 \log n} e^{-\log n} \\ &= 1 - \frac{1}{2 \log n}. \end{aligned}$$

■

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 high probability and 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

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

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. ■

Remark:

(i) In the Counting-Algorithm-I, block codes are used in the step (ii) of the Intra-Cell-Protocol-I and in the Inter-Cell-Protocol-I. Instead, we could use simple repetition coding. Using the repetition coding in the step (ii) of the Intra-Cell-Protocol-I will increase the number of transmissions, but will not change the order of magnitude. However, in the Inter-Cell-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 $O(n \log \log n)$.

(ii) 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 [2]. 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 [2], 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 [4]. Further, our network with its multi-hop 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.

V. THE IMPACT OF LONG-BLOCK MEASUREMENTS

In Section IV, 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 $\{0, \dots, m-1\}$, 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 observed 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

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

and the decoding error probability of each codeword less than

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

In the following algorithm, we use additional block codes in the intra-cell-protocol to reduce the number of transmissions per measurement.

Counting-Algorithm-II:

Cell scheduling for intra-cell transmissions and inter-cell transmissions is the same as in Counting-Algorithm-I.

Intra-Cell-Protocol-II (At cell i):

- (i) 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 α_{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

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

- (ii) 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.
- (iii) 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 $\{\tilde{\mathbf{A}}_j\}$.

Inter-Cell-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 Intra-Cell-Protocol-II. Thus, the total number of bits transmitted in the network under Intra-Cell-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 Inter-Cell-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

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

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

First, from Theorem 2, we have

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

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

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

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,

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

and

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

Define 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.

VI. DISCUSSION AND CONCLUSIONS

In this paper, we investigated counting problems in multi-hop 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] R. G. Gallager. Information Theory and Reliable Communication. John Wiley & Sons, New York, 1968.
- [2] R. G. Gallager. Finding parity in a simple broadcast network. In *IEEE Transactions on Information Theory*, vol. 34, pp 176-180, 1988.
- [3] A. Giridhar and P. R. Kumar. Computing and communicating functions over sensor networks. In *IEEE Journal on Selected Areas in Communications*, pp. 755–764, vol. 23, no. 4, April 2005.
- [4] N. Goyal, G. Kindler and M. Saks. Lower Bounds for The Noisy Broadcast Problem. Preprint.
- [5] P. Gupta and P. Kumar. Critical power for asymptotic connectivity in wireless network. In *Stochastic Analysis, Control, Optimization and Applications: a Volume in Honor of W.H.Fleming*, W. McEneaney, G. Yin and Q. Zhang, Eds., 1998
- [6] P. Gupta and P. Kumar. The capacity of wireless networks. In *IEEE transactions of Information Theory*, vol. 46, no.2, pp. 388-404, 2000.
- [7] N. Khude, A. Kumar and A. Karnik. Time and Energy Complexity of Distributed Computation in Wireless Sensor Networks. In *Proceedings of the IEEE Infocom*, 2005.
- [8] S. R. Kulkarni and P. Viswanath. A Deterministic Approach to Throughput Scaling in Wireless Networks. In *IEEE Trans. on Information Theory*, Vol. 50, No.6, pp. 1041-1049, June 2004.
- [9] E. Kushilevitz and Y. Mansour. Computation in Noisy Radio Networks In *Proceedings of the ninth annual ACM-SIAM symposium on Discrete algorithms*, pp. 236-243, 1998.
- [10] S. Rajagopalan and L. J. Schulman. A Coding Theorem for Distributed Computation. In Proc. 26th STOC 790-799, 1994.
- [11] L. J. Schulman. Communication on Noisy Channels: A Coding Theorem for Computation. In *Proceeding of 33rd FOCS*, pp. 724-733, 1992.
- [12] L. J. Schulman. Deterministic Coding for Interactive Communication. In *Proceeding of the 25th Annual Symposium on Theory of Computing*, pp. 747-756, 1993.
- [13] S. Toumpis and A. J. Goldsmith Large wireless network under fading, mobility, and delay constraints. In *Proceedings of IEEE INFOCOM*, 2004.
- [14] F. Xue and P. R. Kumar. The number of neighbors needed for connectivity of wireless networks. *Wireless Networks*, pp. 169–181, vol.10, no. 2, March 2004.