Distributed Symmetric Function Computation in Noisy Wireless Sensor Networks with Binary Data

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, each having a recorded bit, the sensor's measurement, which has been set to either "0" or "1". The network has a special node called the fusion center whose goal is to compute a symmetric function of these bits; i.e., a function that depends only on the number of sensors that have a "1." The sensors convey information to the fusion center in a multi-hop fashion to enable the function computation. 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. The main result in this paper is an algorithm whose energy usage is $\Theta\left(n(\log\log n)\left(\sqrt{\frac{\log n}{n}}\right)^\alpha\right)$, and we also show that any algorithm satisfying the performance constraints must necessarily have energy usage $\Omega\left(n\left(\sqrt{\frac{\log n}{n}}\right)^\alpha\right)$. Then, we consider the case where the sensor network observes N events, and each node records one bit per event, thus having N bits to convey. The fusion center now wants to compute N symmetric functions, one for each of the 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 $N=\Omega(\log n)$, and we furthermore show that any other feasible algorithm must have energy usage $\Omega\left(n\left(\sqrt{\frac{\log n}{n}}\right)^\alpha\right)$.**

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 $f(n)\le cg(n)$ for all $n\ge m$.

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

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 one bit; the goal is for the fusion center to compute symmetric functions — those functions determined by the sum of the observed bits. 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 with binary data, which is also called a counting problem in this paper, is as follows: each node is in either state "1" or "0", and the fusion center needs to decide, using information transmitted from the network, the number of sensors in state "1". When nothing is known about the structure of the function to be computed, all bits 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], [10], [11], [9], [8], 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 [8], respectively, where a broadcast network is assumed, in which all nodes can hear all transmissions, and each node has a "1" or a "0." The goal in [2], [8] was to investigate the

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.

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, since both of these are determined if we know how many nodes have a "1."

While the problems considered in [2] and [8] are similar to our problem, a major difference is that in our model, 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 it 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. 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 [6]. 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 [6], 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 [6], where the communication channels are assumed to be reliable, and the processing is to primarily exploit the spatial correlation [6] 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. The local computation in our case is not a simple parity computation as in [2] but as we will see later, the network needs to compute the number of sensors in each local neighborhood (called a *cell*) that have seen a "1." Further, we require that the computation be accurate uniformly over all cells. In addition, we will show that error-correction coding is 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 observation to report, and show that the amount of energy required for counting (i.e., detecting the number of sensors seeing a "1") is $O\left(n(\log\log n)\left(\sqrt{\frac{\log n}{n}}\right)^\alpha\right)$, where n is the number of sensors in the network.
3) We then extend to the case where each sensor has N binary observations, and the symmetric function needs to be computed for each observation. We show that the total transmission energy consumption can be reduced to $O\left(n\left(\max\left\{1, \frac{\log\log n}{N}\right\}\right)\left(\sqrt{\frac{\log n}{n}}\right)^\alpha\right)$ per observation. When $N = \Omega(\log\log n)$, the energy consumption is $\Theta\left(n\left(\sqrt{\frac{\log n}{n}}\right)^\alpha\right)$ per observation, which is a tight bound.
4) If we only want to know roughly (the meaning of "roughly" will be made precise in Section VI) how many sensors have "1." The answer can be obtained with the transmission energy consumption $\Theta\left(n\left(\sqrt{\frac{\log n}{n}}\right)^\alpha\right)$.

The rest of the paper is organized as follows: In Section III, we introduce our sensor network model and give a lower bound on the minimum energy needed. In Section IV, we consider the case where the sensors have only one observation to report. In Section V, we investigate the impact of transmitting N observations. In Section VI, we study the case where only a vague decision is needed. Finally, in Section VII, 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 "1" or "0." 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 is trying to determine γ_c where $\gamma_c = \sum_k b_k$. We say that node A can transmit to node B if the bit error probability of the transmission from A to B is less than or equal to p ($p < 1/2$). The transmission power of each sensor is assumed to be chosen such that the network is connected. The power required to ensure that a node is able to communicate with all neighbors within a radius r is assumed to be proportional to r^α, where $\alpha \geq 2$ is some constant. For simplicity, we will assume that once the transmission power is chosen to ensure connectivity, the channel between any pair of neighbors is a binary symmetric channel with error probability p, i.e., the channel flips the transmitted bit with probability p. 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 number of 1's in the network. Given an algorithm for counting the number of 1's in the network, we define the energy required by the algorithm to be the maximum energy required for the computation over all possible values of the observation bits. 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 $n \to \infty$.

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 majority rule. Then we have following well-known bound [1] on the error probability, where the proof is provided for completeness.

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 majority rule. Then, the probability of decoding error is no greater than

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

Proof: Define m independent binary random variables $\{I_i\}$, where $I_i = 0$ with probability p and $I_i = 1$ with probability $1-p$. Using Chernoff's bound, we have

$$\Pr\left(\sum_{i=1}^{m} I_i < \frac{m}{2}\right) \leq e^{\frac{m}{2}\log(4(1-p)p)} = (4p(1-p))^{\frac{1}{2}m}.$$

■

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.

□

In the sequel, for mathematical concreteness, we assume that each node must explicitly transmit its bit b_k if it is to convey its value, irrespective of the value. However, network scenarios may exist where for example a "0" value can be conveyed implicitly by not transmitting; in such cases our analysis in the sequel can be regarded as applying to the worst-case scenario where each node has observed a "1," and thus must transmit.

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

$$\Omega\left(n\left(\sqrt{\frac{\log n}{n}}\right)^{\alpha}\right) \tag{1}$$

.

Proof: First, connectivity of the network is a necessary condition of correct counting. To guarantee connectivity, it has been shown in [4] that the transmission range of the sensors should be chosen as $\Omega\left(\sqrt{\frac{\log n}{n}}\right)$. Thus, the energy used per sensor transmission is $\Omega\left(\left(\sqrt{\frac{\log n}{n}}\right)^{\alpha}\right)$. There are n sensors in the network, each of which must make at least one transmission; thus, the total transmission energy required is $\Omega\left(n\left(\sqrt{\frac{\log n}{n}}\right)^{\alpha}\right)$. ■

Now, we consider the counting problem in detail. We first define the routing strategy. 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 [7], [13], [12], 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 [12], where $c_1 > 0$, and have following lemma.

Lemma 4 ([12, 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,

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

□

From above lemma, we know that if $c_1 > 8$, 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,

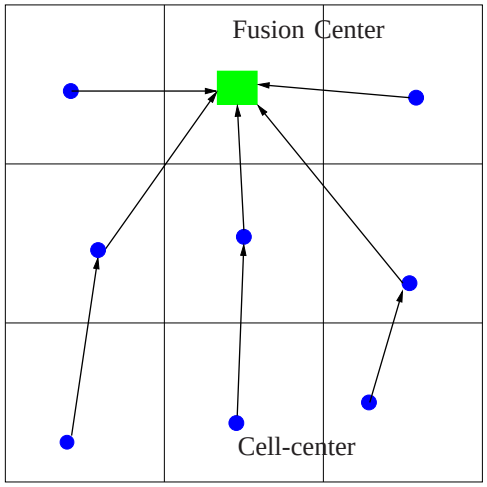

Fig. 1. A Wireless Sensor Network

$C(i)$ 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 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.

Define $\lambda = -\log(4p(1-p))$. Throughout paper, B is chosen as in (2) with

$$c_1 > \max\left\{8, \frac{4}{\lambda}\right\}.$$

IV. An Upper Bound on the Energy Consumption

Here, the conventional coding theorem cannot be used since each sensor has only one bit to transmit. We use the idea in [2] to design an algorithm for which the energy consumed is $\Theta\left(n(\log\log n)\left(\sqrt{\frac{\log n}{n}}\right)^{\alpha}\right)$. 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 bit sent by sensor k. The main idea in [2] is to use the reception diversity to obtain a good estimate of the bit transmitted by sensor k. But it requires additional transmissions among sensors; for example, it takes n more transmissions for n sensors to report the bits 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 bit 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

$$\gamma_i = \sum_{k\in\Delta_i} b_k$$

if the counting is correct.

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

Counting-Algorithm-I:

Adjacent cells may interference with each other, so we adopt the cell scheduling scheme used in [5], [3].

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

(i) The sensors in cell i take turns to transmit their bits. When it is the turn of sensor k, it broadcasts its bit $\lceil \frac{4}{\lambda}(\log\log n)\rceil$ times. Then, all other sensors in the cell will receive $\lceil \frac{4}{\lambda}(\log\log n)\rceil$ bits from sensor k. Sensor j ($j \neq k$) sets α_{jk} to be the majority of the bits received from sensor k, and sets A_j to be

$$A_j = b_j + \sum_{k\in\Delta_i, k\neq j} \alpha_{jk}$$

after all sensors broadcast their bits.

(ii) Select $\left\lfloor \frac{n_i}{\log\log n}\right\rfloor$ sensors in the cell. Each selected sensor j represents A_j using $\lceil \log_2 n_i\rceil$ bits, codes it using a block code with rate R_1 such that $E_r(R_1)/R_1 \geq 1$, 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 sequence $\{\tilde{A}_j\}$.

Cell scheduling for inter-cell transmissions: (1) Let $L = H_{\max}$; (2) Cells with depth L are scheduled according to [5], [3]. If $L \neq 0$, let $L = L-1$ and repeat step (2).

Inter-Cell-Protocol-I:

Define 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

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

where $\tilde{\eta}_j$ is the output of the channel between cell center j and cell center i with input η_j. Since $0 \leq \eta_i \leq n$, note that η_i can be represented using $\lceil \log_2 n\rceil$ 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 $E_r(R_2)/R_2 > 1$.

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

$$\Pr\left(\text{All } \gamma_i \text{ are correct} \;\middle|\; \frac{c_1}{2} \leq \frac{n_i}{\log n} \leq 4c_1 \;\forall i\right) \geq 1 - \frac{1}{c_1 \log n}.$$

Then, in Lemma 6 and Theorem 7, we show that

$$\Pr(\gamma_c \text{ is correct } | \gamma_i \text{ is correct } \forall i) \geq 1 - \frac{4}{c_1 \log n}.$$

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

Lemma 5: Suppose $\frac{c_1}{2}\log n \leq n_i \leq 4c_1 \log n$ for all i. Then, by executing Intra-Cell-Protocol-I, the cell-centers can obtain γ_i with

$$\Pr\left(\gamma_i = \sum_{k\in\Delta_i} b_k \ \forall i \ \middle| \ 4c_1 \geq \frac{n_i}{\log n} \geq \frac{c_1}{2} \ \forall i\right) \geq 1 - \frac{1}{c_1 \log n} \quad (5)$$

and the number of transmissions required in each cell is $\Theta\left((\log n)(\log\log n)\right)$.

Proof: In the following analysis, we assume $\frac{c_1}{2}\log n \leq n_i \leq 4c_1 \log n$ holds for all i. First, the number of transmissions in each cell under Intra-Cell-Protocol-I is

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

Next we investigate the probability that γ_i is correct, i.e., $\gamma_i = \sum_{k\in\Delta_i} b_k$. From Lemma 1, we have

$$\Pr\left(\alpha_{jk} = b_k\right) \geq 1 - (4p(1-p))^{\frac{2\log\log n}{\lambda}}.$$

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

$$\Pr\left(A_j = \sum_{k\in\Delta_i} b_k\right) \geq 1 - n_i(4p(1-p))^{\frac{2\log\log n}{\lambda}} \geq 1 - \frac{4c_1}{\log n}.$$

Consider step (ii) of Intra-Cell-Protocol-I, from Theorem 2,

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

where the last inequality holds because $n_i \geq \frac{c_1}{2}\log n$. Thus,

$$\Pr\left(\tilde{A}_j = \sum_{k\in\Delta_i} b_k\right) \geq 1 - \frac{4c_1+4}{\log n}.$$

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 $\tilde{A}_j = \sum_{k\in\Delta_i} b_k$, and $I_j = 0$ if $\tilde{A}_j \neq \sum_{k\in\Delta_i} b_k$. Since γ_i is the mode of $\{\tilde{A}_j\}$, from Lemma 1, we have

$$\begin{aligned}
\Pr\left(\gamma_i \neq \sum_{k\in\Delta_i} b_k\right) &\leq \Pr\left(\sum_j I_j < \frac{1}{2}n_i\right) \\
&\leq \left(4\left(\frac{4c_1+4}{\log n}\right)\left(1 - \frac{4c_1+4}{\log n}\right)\right)^{\frac{n_i}{2\log\log n}} \\
&\leq e^{-(\log\log n - \log(16c_1+16))\frac{n_i}{2\log\log n}} \\
&\leq e^{-\log n}.
\end{aligned}$$

There are at most $\frac{n}{c_1\log n}$ cells in the network, so

$$\Pr\left(\gamma_i = \sum_{k\in\Delta_i} b_k \ \forall i\right) \geq 1 - \frac{n}{c_1\log n}e^{-\log n} = 1 - \frac{1}{c_1\log n},$$

and the lemma holds. ■

Now, suppose that all γ_i are correct. Since η_i can be represented using $\lceil \log_2 n \rceil$ bits, each cell-center has $\lceil \log_2 n \rceil$ 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:

$$\Pr\left(\gamma_c = \sum_k b_k \middle| \gamma_i = \sum_{k\in\Delta_i} b_k \ \forall i\right) \geq 1 - \frac{4}{c_1\log n}, \quad (6)$$

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

Proof: Suppose all cell-centers have the correct γ_i, then $\gamma_c = \sum_k b_k$ 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,

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

and from the union bound,

$$\begin{aligned}
&\Pr\left(\gamma_c = \sum_k b_k \middle| \gamma_i = \sum_{k\in\Delta_i} b_k \ \forall i\right) \\
= \ &\Pr\left(\text{All } \eta_i\text{'s are correctly received} \middle| \gamma_i = \sum_{k\in\Delta_i} b_k \ \forall i\right) \\
\geq \ &1 - \frac{4n}{c_1\log n}e^{-\log n} \\
= \ &1 - \frac{4}{c_1\log n}.
\end{aligned}$$

■

From Lemma 5 and Lemma 6, we have shown that, under Counting-Algorithm-I, the number of sensors in state "1" can be counted accurately with high probability when the number of sensors is large enough. Based on that, we have following theorem, which provides an upper bound on the energy requirement to solve our counting problem.

Theorem 7: The number of sensors in state "1" can be counted accurately with high probability by total transmission energy consumption

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

and 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 $\frac{7}{c_1\log n}$.

Proof: Recall that

$$c_1 > \max\left\{8, \frac{4}{\lambda}\right\}.$$

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

$$\Pr\left(\gamma_c = \sum_k b_k\right) \geq 1 - \frac{7}{c_1\log n},$$

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

Further, from Lemma 5 and Lemma 6, the number of transmissions under Counting-Algorithm-I is $\Theta(n(\log\log n))$. Since the common transmission range is $\sqrt{\frac{8c_1\log n}{n}}$, the total energy consumption is

$$\Theta\left(n(\log\log n)\left(\sqrt{\frac{\log n}{n}}\right)^{\alpha}\right). \tag{7}$$

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

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. 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. 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 good.

We now discuss whether the techniques in [2] and [8] can be directly used to obtain our energy bounds. In the parity computation problem in [2], even though all nodes can hear each other, the algorithm proceeds by first dividing the network into cells and computes the parity in each cell. A uniform bound is provided on the probability that the computed parity in each cell is correct. However, we cannot use the proof in [2] directly since unlike our problem, in parity computation, one can exploit the fact that if one makes an even number of errors in receiving the bits, the parity of these bits would still be correct. Next consider the algorithm in [2] for the more general problem of knowing all the bits in the network. Since this is a more general problem than ours, one can argue that the algorithm designed in [2] for this purpose can be used within each cell of our sensor network. However, it turns out this algorithm in [2] cannot be used to provide uniform upper bound on the probability of counting error in each cell, which is crucial for our algorithm to work. Finally, the threshold computation in [8] is a special case of our problem and further, the technique in [8] cannot be used to prove a uniform bound on the probability of counting error in each cell.

V. The Impact of Long-Block Observations

In Section IV, we considered the case where each sensor has only one observed bit to transmit. In this section, we will investigate the impact of transmitting N binary observations. 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 bit 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} observed bit of sensor k. The fusion center is interested in determining $\sum_k b_k[h]$ for each fixed h. From Theorem 2, if we have N bits to transmit, there exist block codes with code length

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

and the decoding error probability of each codeword less than

$$4e^{-2\max\{N,\log\log n\}}.$$

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

Counting-Algorithm-II:

Cell schedulings for intra-cell transmissions and inter-cell transmissions are the same as those 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 $\left\lceil \frac{\max\{N,\log\log n\}}{R} \right\rceil$ and suppose that either N or n is large enough such that the decoding error probability of each codeword less than $4e^{-2\max\{N,\log\log n\}}$. The codeword for $\mathbf{b}_k$ is then broadcasted once. Suppose α_{jk} is the output of the binary symmetric channel between sensor k and sensor j with input $\mathbf{b}_k$, sensor j sets

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

after all sensors broadcast their bits.

(ii) Select $\left\lceil \frac{n_i}{\log\log n} \right\rceil$ sensors. Each selected sensor j represents $A_j[h]$ using $\lceil \log_2 n_i \rceil$ bits, encodes it using a block code with rate R_1 such that $E_r(R_1)/R_1 \geq 1$, 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 $\gamma_i[h]$ to be any mode of sequence $\{\tilde{A}_j[h]\}_j$.

Inter-Cell-Protocol-II:

Use $\gamma_i[h]$, $\gamma_c[h]$, $\eta_i[h]$, and $\eta_c[h]$ to denote the values corresponding to the h^{th} bit. Repeat Inter-Cell-Protocol-I for each h.

Theorem 8: Suppose all sensors have N binary observations to report, then the number of "1" in each observation can be counted accurately with high probability by total transmission energy consumption

$$O\left(n\left(\max\left\{1, \frac{\log\log n}{N}\right\}\right)\left(\sqrt{\frac{\log n}{n}}\right)^{\alpha}\right)$$

per observation, and Counting-Algorithm-II is asymptotically correct. Specifically, the probability of computation error at the fusion center is upper bounded by $\frac{5}{c_1 \log n}$.

Proof: Suppose $\frac{c_1}{2}\log n \le n_i \le 4c_1 \log n$ holds for all i. First, consider the number of bits transmitted under Counting-Algorithm-II. There are

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

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

$$\Theta\left(n\left(\max\left\{1, \frac{\log\log n}{N}\right\}\right)\right)$$

per bit. Inter-Cell-Protocol-II is the same as Inter-Cell-Protocol-I, so the number of bits transmitted is $\Theta(n)$ per bit. Thus, under Counting-Algorithm-II, the energy required per observed bit is $\Theta\left(n\left(\max\left\{1, \frac{\log\log n}{N}\right\}\right)\left(\sqrt{\frac{\log n}{n}}\right)^{\alpha}\right)$, which implies that the minimum transmission energy required per bit is

$$O\left(n\left(\max\left\{1, \frac{\log\log n}{N}\right\}\right)\left(\sqrt{\frac{\log n}{n}}\right)^{\alpha}\right)$$

since there may exist other algorithms that consume less energy.

Next, we prove the asymptotic correctness of Counting-Algorithm-II. We will show

$$\Pr\left(\gamma_i = \sum_{k\in\Delta_i} b_k \ \forall i \ \middle|\ 4c_1 \ge \frac{n_i}{\log n} \ge \frac{c_1}{2} \ \forall i\right) \ge 1 - \frac{1}{c_1 \log n},$$

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

First, from Theorem 2, we have

$$\begin{aligned}\Pr\left(\alpha_{jk}[h] = b_k[h]\right) &\ge \Pr\left(\alpha_{jk} = \mathbf{b}_k\right) \\ &\ge 1 - 4e^{-2\max\{N, \log\log n\}} \\ &\ge 1 - \frac{4}{(\log n)^2}.\end{aligned}$$

Since $n_i \le 4c_1 \log n$, from the union bound,

$$\Pr\left(A_j[h] = \sum_{k\in\Delta_i} b_k[h]\right) \ge 1 - n_i \frac{4}{(\log n)^2} \ge 1 - \frac{16c_1}{\log n}.$$

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

$$\Pr\left(\tilde{A}_j[h] = A_j[h]\right) \ge 1 - 4e^{-\frac{E_r(R_1)}{R_1}\log_2 n_i} \ge 1 - 4e^{-\log\log n},$$

and

$$\Pr\left(\tilde{A}_j[h] = \sum_{k\in\Delta_i} b_k[h]\right) \ge 1 - \frac{16c_1 + 4}{\log n}.$$

Define i.i.d. random variables $\{I_j\}$ such that $I_j = 1$ if $\tilde{A}_j[h] = \sum_{k\in\Delta_i} b_k[h]$, and $I_j = 0$ otherwise. From Lemma 1,

$$\begin{aligned}&\Pr\left(\gamma_i[h] \ne \sum_{k\in\Delta_i} b_k[h]\right) \\ \le\ &\Pr\left(\sum_j I_j \le \frac{n_i}{2}\right) \\ \le\ &\left(4\left(1 - \frac{16c_1+4}{\log n}\right)\left(\frac{16c_1+4}{\log n}\right)\right)^{\frac{n_i}{2\log\log n}} \\ \le\ &\left(\frac{64c_1+16}{\log n}\right)^{\frac{n_i}{2\log\log n}} \\ \le\ &e^{-\log n}\end{aligned}$$

for large n.

Thus,

$$\begin{aligned}\Pr\left(\gamma_i[h] = \sum_{k\in\Delta_i} b_k[h] \ \forall i\right) &\ge 1 - \frac{n}{c_1 \log n} e^{-\log n} \\ &\ge 1 - \frac{1}{c_1 \log n},\end{aligned}$$

and the theorem holds. ∎

From the theorem above, when $N = \Omega(\log\log n)$, the transmission energy required per observation 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. The Impact of Vague Counting

In Section IV and Section V, we studied the counting problem where we wanted to know exactly how many sensors have a "1". Since the channels are noisy channels, to obtain the correct answer, we have to use redundant transmissions when each sensor has only one observed bit to report, or transmit N bits as one block when each sensor has more than one observed bit. In this section, we consider a situation where each sensor has only bit to transmit, but the fusion is only interesting in knowing roughly how many sensors have a "1", so only a vague counting is needed. We consider the case where the decision $\overline{\gamma}_c$ of the fusion center is acceptable if for any $\varepsilon > 0$,

$$\lim_{n\to\infty} \Pr\left(\left|\overline{\gamma}_c - \sum_k b_k\right| \le \varepsilon n\right) = 1. \tag{8}$$

Note that, from the above objective, the fusion center is now interested in estimating the fraction of nodes that have a "1," within some resolution ε. In this case, we can use a simple algorithm to obtain an acceptable answer with high probability. The algorithm is as follows:

Counting-Algorithm-III:

Cell schedulings for intra-cell transmissions and inter-cell transmissions are the same as in those Counting-Algorithm-I.
Intra-Cell-Protocol-III (At cell i):

The sensors in cell i take turns to transmit their bits. When it is sensor k's turn, it transmits its bit once to the cell center. After all sensors transmit their bits, the cell center computes the sum of the n_i received bits. This sum is denoted by γ_i.
Inter-Cell-Protocol-III:

The fusion center uses Inter-Cell-Protocol-I to obtain γ_c, and the number of sensors in state "1" is estimated as

$$\bar{\gamma}_c = \left\lceil \frac{\gamma_c - np}{1-2p} \right\rceil. \tag{9}$$

Suppose there are T sensors in state "1," and $\tilde{b}_k$ is the output of the channel between sensor k and the cell-center with input b_k. If $\gamma_c = \sum_k \tilde{b}_k$, then

$$E[\gamma_c] = T(1-p) + (n-T)p = np + T(1-2p).$$

Thus, intuitively $\bar{\gamma}_c$ in (9) is a natural estimate of the total number of 1's in the network.

We first introduce following lemma, which bounds the probability of the summation of n independent (not necessarily identical) binary random variables deviating from the mean.

Lemma 9: Suppose $\{X_k\}$ $(1 \le k \le n)$ are n independent binary random variables. Let $\mu = E[\sum_k X_k]$. Then for any $\varepsilon > 0$,

$$\Pr\left(\left| \sum_k X_k - \mu \right| \ge \varepsilon n \right) \le e^{-2\varepsilon^2 n}.$$

Proof: Lemma 2.1 and Lemma 2.2 of [8]. ■

Based on Lemma 9, we can obtain following theorem.

Theorem 10: For any $\varepsilon > 0$, under Counting-Algorithm-III,

$$\Pr\left(\left| \bar{\gamma}_c - \sum_k b_k \right| \le \varepsilon n \right) \ge 1 - \frac{7}{c_1 \log n},$$

and the transmission energy required is

$$\Theta\left(n \left(\sqrt{\frac{\log n}{n}} \right)^{\alpha} \right).$$

Proof: Suppose $\frac{c_1}{2}\log n \le n_i \le 4c_1 \log n$ holds for all i. First, it is easy to see the number of bits transmitted is $\Theta(n)$, and the energy consumed is $\Theta\left(n \left(\sqrt{\frac{\log n}{n}} \right)^{\alpha} \right)$.

Next, we study the probability that $\bar{\gamma}_c$ satisfies (8). From Lemma 6, under Inter-Cell-Protocol-III, we have

$$\Pr\left(\gamma_c = \sum_i \gamma_i \right) \ge 1 - \frac{4}{c_1 \log n}.$$

Since

$$\Pr\left(\tilde{b}_k = b_k \right) = 1 - p,$$

and

$$E\left[\sum_k \tilde{b}_k \right] = np + \left(\sum_k b_k \right)(1-2p),$$

we have

$$\Pr\left(\left| \frac{\sum_k \tilde{b}_k - np}{(1-2p)} - \sum_k b_k \right| \le \varepsilon n \right)$$
$$= \Pr\left(\left| \sum_k \tilde{b}_k - \mu \right| \le (1-2p)\varepsilon n \right),$$

where $\mu = np + (\sum_k b_k)(1-2p)$. Then, from Lemma 9, we have

$$\Pr\left(\left| \frac{\sum_k \tilde{b}_k - np}{(1-2p)} - \sum_k b_k \right| \le \varepsilon n \right) \ge 1 - 2e^{-(1-2p)^2 \varepsilon^2 n}. \tag{10}$$

Note that $\gamma_c = \sum_k \tilde{b}_k$ if $\gamma_c = \sum_i \gamma_i$. So according to (3), (6) and (10), we have

$$\Pr\left(\left| \bar{\gamma}_c - \sum_k b_k \right| \le \varepsilon n \right) \ge 1 - \frac{7}{c_1 \log n},$$

and the theorem holds. ■

VII. Discussion and Conclusions

In this paper, we investigated counting problems in multi-hop networks with noisy communication channels. First, we considered sensors with single bit measurements, and showed by construction that feasible algorithms exist whose energy consumption is class $O\left(n(\log\log n)\left(\sqrt{\frac{\log n}{n}} \right)^{\alpha} \right)$. Then, we considered the case where the sensors have N bits to report, in which case the transmission energy can be reduced to class $O\left(n\left(\max\left\{ 1, \frac{\log\log n}{N} \right\} \right)\left(\sqrt{\frac{\log n}{n}} \right)^{\alpha} \right)$ per observed bit. Finally, if only vague counting is needed, feasible algorithms are demonstrated from energy class $\Theta\left(n\left(\sqrt{\frac{\log n}{n}} \right)^{\alpha} \right)$.

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 is still needs to be investigated whether $O\left(n(\log\log n)\sqrt{\frac{\log n}{n}}^{\alpha} \right)$ is 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] P. Gupta and P. Kumar. Critical power for asymptotic connectivity in wireless network. In *Stochastic Analysis, Contrtol, Optimization and Applicaitons: a Volume in Honor of W.H.Fleming*, W. McEneaney, G. Yin and Q. Zhang, Eds., 1998

[5] P. Gupta and P. Kumar. The capacity of wireless networks. In *IEEE transactions of Information Theory*, vol. 46, no.2, pp. 388-404, 2000.

[6] 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.

[7] 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.

[8] 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.

[9] S. Rajagopalan and L. J. Schulman. A Coding Theorem for Distributed Computation. In Proc. 26th STOC 790-799, 1994.

[10] L. J. Schulman. Communication on Noisy Channels: A Coding Theorem for Computation. In *Proceeding of 33rd FOCS*, pp. 724-733, 1992.

[11] L. J. Schulman. Deterministic Coding for Interactive Communication. In *Proceeding of the 25th Annual Symposium on Theory of Computing*, pp. 747-756, 1993.

[12] S. Toumpis and A. J. Goldsmith Large wireless network under fading, mobility, and delay constraints. In *Procedings of IEEE INFOCOM*, 2004.

[13] 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.