On Multiple Access for Distributed Dependent Sources: A Content-Based Group Testing Approach

Yao-Win Hong
School of ECE, Cornell University
E-mail: yh84@cornell.edu

Anna Scaglione
School of ECE, Cornell University
E-mail: anna@ece.cornell.edu

Abstract — **In this paper we consider the multiple access problem with distributed dependent sources. We derive the optimal designs for the case of N correlated binary sources whose data are modelled as a two-state Markov chain. The solution can be classified as a group testing technique where data values at the sensors are determined through the successive refinements of the tests over smaller groups. The tests form, progressively, an accurate map of the sensor data at the central receiver. We derive the conditions on the parameters of the data model for which the group testing approach is superior to time sharing. In contrast to standard multiple access techniques, this is the first method proposed for data retrieval from distributed dependent sources which is content-based rather than user-based.**[1]

I. Introduction

The goal of our work is to design multiple access communication strategies to transmit the information from a set of distributed dependent sources to a central receiver through a multiple access channel. In sensor networks, sensor nodes are often deployed in large scale to observe physical events or measurements from the environment. The detected events or quantized measurements at the sensors naturally classify them into groups of the same state. For example, in the binary detection problem, the sensors that have detected a certain event will be grouped into one class, while the other nodes will be grouped into another class. In the case of quantized measurements, all sensors observing data within the same quantization level also constitute a certain class. By reliably identifying the class for which each sensor resides, the central node is able to reconstruct the entire sensor field or to accurately locate the occurrence of an event.

The key idea of this work is to utilize the concept of group testing to efficiently acquire the states of the sensors instead of polling the sensors one-by-one. This is a revolution in multiple access since the method we propose is not aimed at retrieving the information within a certain node but rather to locate the sensors containing a certain information. This leads to the so called *content-based multiple access technique*. For a set of distributed sources that have low aggregate entropy, we show that this technique can be more bandwidth efficient than its standard multiple access alternatives.

[1]This work is supported in part by the National Science Foundation under grant CCR-0227676 and ONR Contract N00014-00-1-0564.

The concept of *group testing* was first introduced by Dorfman [1] in World War II to efficiently identify all syphilitic men called up for induction into the armed forces. The method significantly reduces the number of blood tests necessary by pooling a number of blood samples together, and testing the pooled samples instead of testing them individually. This method efficiently solves the problem of classifying the states of all individuals in a large population. The work by Sobel and Groll [2] further extended this method to many industrial applications.

Group testing has also been proposed as an efficient solution for random access scheduling [3, 4]. In this case, the probability that a particular sensor has a message to transmit is independent from sensor to sensor, therefore, classical group testing strategies can be applied directly since they are mostly based on the same assumption. In sensor networks, an analogous model could arise when unexpected independent events trigger alarms in isolated sensors. However, in general, the observations made at the sensors are often both spatially and temporally correlated, which violates the assumption in classical problems. *To the best of our knowledge, we are the first to propose its application to directly retrieve information from a distributed sensor network, and in this lies our major contribution.* To demonstrate that this strategy has wide applicability in the context of distributed sources with low aggregate entropy, we explicitly derive the optimum group testing strategy for a special case of correlated sources, which are modelled as a two-state Markov chain.

Group testing can in fact be applied to transmit efficiently a much wider class of information data from a set of distributed sources that have low aggregate entropy. Given a certain group testing strategy, the answers to the sequence of tests imposed upon the set of sensors constitute a code [4] that maps to the corresponding information at the sensors. If group testing is performed optimally, the average code length resulting from the tests should be shorter or at most equal to testing each node individually. In the case where no error occurs in the tests, group testing may be seen as a special case of the zero-error data compression problem. Specifically, group testing has been applied in the context of image compression [5, 6]. However, in contrast to these applications, the information in sensor networks is distributed among sensors instead of being known at the central processor, therefore, the strategies used in the image compression literature cannot be applied directly. The importance of our work is that we simultaneously schedule the transmission of each sensor while compressing the sensor information along

0-7803-8720-1/04/$20.00 ©2004 IEEE

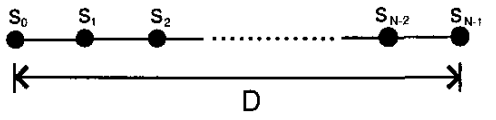

Figure 1: The spatial distribution of the sensors.

the spatial domain [7]. Therefore, *our method solves jointly the source coding and the multiple access channel coding problems without separation.*Compared to other works that proposed the Slepian-Wolf distributed source coding solution [8, 9, 10] to reduce the aggregate rate of sensor networks, this method does not require the joint statistics of the data at the nodes. Furthermore, it automatically achieves the cooperative compression effect without requiring the sequential application of compression algorithms over increasingly large sets of data [11].

II. Problem Setup: Correlated Information

Consider $N = 2^M$ sensors uniformly deployed on the real line within the interval $[0, D]$, as shown in Fig. 1. Denote the network of sensors by $\mathcal{S} = \{s_0, s_1, \cdots, s_{N-1}\}$. The set of sensors observe a sequence of spatially correlated sensor information $\mathbf{X} = \{X_0, X_1, \cdots, X_{N-1}\}$, where X_i is the observation made at node s_i and that $X_i \in \{0, 1\}$ for all i, *i.e.* binary sources. For example, the binary bits at each sensor may represent the local decisions of a binary hypothesis testing, or the information of the quantized samples that the sensors observe, both of these cases are correlated over the spatial area. Our goal is to allow a central node to efficiently acquire the information contained at each local sensor using the minimum number of channel access. More specifically, we consider a simplified model for the channel and make the following assumptions: 1) the medium is broadcast; 2) the answers to each test are reliable, *i.e.* zero errors.

We envision that the optimum physical-layer strategy supporting this transmission will be highly dependent on the power constraints, propagation model and the topology of the sensor field. The transmission may be achieved with or without user cooperation. Clearly, it is possible to benefit from asking more articulate questions and receiving a greater number of aggregate bits per test. However, we choose to leave the study of deep physical layer aspects for future work. Particularly, it will be important to establish what is the minimum physical cost in terms of transmit power that this channel model implies, what is the test channel cut-off rate and how to combine error protection with each test effectively. For the rest of the paper we associate the cost of transmission solely to the number of tests necessary to reconstruct the sensor field, assuming that each test is reliable. In this sense, what we are measuring in this paper is primarily the efficiency of transmission of our distributed compression technique. The next natural step is to combine each group testing with the optimum channel coding strategy.

In classical group testing problems, e.g. the blood testing problem, each item in the population of size N can be either defective or non-defective. It is common to assume that each node is an independent Bernoulli trial with the identical probability p of being defective. When $p \ll 1$, the event that a particular item is defective is very unlikely, therefore, testing each node individually, which is analogous to the TDMA in multiple access applications, is inefficient. Suppose we are able to apply tests on groups of items with size $n > 1$ such that a "positive" result is given if one or more items are defective, and a "negative" result if and only if the items are all non-defective. When the "positive" result is obtained, it means that the group of n items contains at least one defective item, therefore, we must split the group into smaller subgroups to identify, specifically, which item or items are defective. It has been shown that group testing reduces significantly the average number of tests necessary when p is less than the cutoff point $p^* = \frac{1}{2}(3 - \sqrt{5})$ [12].

In relating our problem to classical group testing, we correspond sensor s_i to be defective if $x_i = 1$, while otherwise non-defective. To obtain the information $\mathbf{X}$, the central node must impose a sequence of tests $T_0, T_1, \cdots, T_{L-1}$ to the subsets $U_0, U_1, \cdots, U_{L-1}$, where $U_l \subset \mathcal{S}$ for all l and L is random variable representing the number of tests necessary to reconstruct the sensor field. Within each test T_l, the distributed sensors in the subset U_l reply with only a simple pulse transmission. For example, when the central node sends the test "Do you have bit 1?" to the sensors in U_l, a sensor $s_i \in U_l$ will emit a pulse indicating that it is "defective" if $x_i = 1$. Assuming that the central node utilizes only an energy detector, it is not able to distinguish between the number of defective items within the group, given that the number is greater than 1. Therefore, if the contents of the sensors in U_l were not completely resolved after T_l, the group of sensors U_{l+1} chosen for the next test must contain a subset of U_l in order to identify exactly the sensors for which $x_i = 1$.

Although the sensor problem considered seems similar to the classical group testing problem, the efficiency can be improved over the traditional scheme by utilizing two major features:

(I.1) the bit information is correlated over sensors;

(I.2) each sensor has knowledge of its own observation.

In considering **(I.1)**, one must adjust the grouping strategy accordingly with respect to the result of previous tests. For example, in the case where sensors are highly correlated, a set of m sensors tested to be non-defective may imply that the whole network of N sensors are non-defective with probability close to 1. In applying **(I.2)**, the central node is allowed to impose different questions upon the set of sensors. In other words, instead of asking each time the question "Do you have bit 1?", a central node may also ask the question "Do you have bit 0?" when it is more efficient to do so. Since the sensor has knowledge of its own observation, it is able to answer the latter question using the same pulse transmission. This is generalizing the reversing technique [3] applied by Berger *et al* in their reservation-based multiple access protocol.

A. Sensor Model: Two State Markov Chain

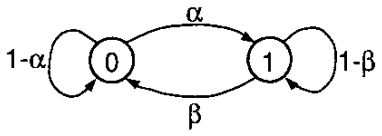

Figure 2: The two state Markov chain.

Consider the case where the sequence of binary observations $\mathbf{X} = (X_0, X_1, \cdots, X_{N-1})$ are modelled using a two state Markov chain, as shown in Fig. 2, where the two states 0 and 1 represent the realization of the observations at each node. Let α and β be the transition probability from state 0 to state 1 and state 1 to state 0, respectively. Assume that the initial probabilities are the steady state distributions such that

$$\Pr\{X_i = 1\} = \frac{\alpha}{\alpha + \beta} = p \tag{1}$$

and, similarly, $\Pr\{X_i = 0\} = \frac{\beta}{\alpha+\beta} = 1 - p = q$. Therefore, $\sigma^2_{X_i} = \mathbf{E}[X_i - E(X_i)]^2 = p(1-p)$ and $\mathrm{Cov}(X_i, X_{i+1}) = p(1-p)[1-(\alpha+\beta)]$. Thus, the correlation coefficient

$$\rho = \frac{\mathrm{Cov}(X_i, X_{i+1})}{\sigma_{X_i}\sigma_{X_{i+1}}} = 1 - (\alpha + \beta).$$

We note that (p, ρ) uniquely specifies the pair of transition probabilities (α, β). In this work, we restrict our problem to the case where $\rho \in [0, 1]$ such that group of nodes adjacent to each other would have the highest correlation [c.f. Section III].

B. Information Lower Bound

During each access of the channel, the central node asks a question which is responded by the binary answer — "yes" or "no". With the sequence of answers obtained from the tests, the central node is able to obtain a lossless reconstruction of the sensor field. Therefore, the total number of access can be trivially lower-bounded by the entropy of the sensor data $\mathbf{X}$. The entropy of $\mathbf{X}$ is given as follows:

$$H(\mathbf{X}) = h(p) + (N-1)H(X_1|X_0) \tag{2}$$

where $H(X_1|X_0) = p \cdot h[(1-p)(1-\rho)] + (1-p) \cdot h[p(1-\rho)]$ and $h(p) = -p\log p - (1-p)\log(1-p)$.

In the following section, we derive the optimal grouping strategy by exploiting the knowledge of the correlation.

III. Optimal Strategy

In considering **(I.2)**, we fix the strategy such that the central node asks both questions each time it imposes a test upon a certain group. We note that the true optimal strategy would be to ask one question each time, but to choose the best question depending on the sensor model and the results of the previous tests. However, to be consistent with the strategies discussed throughout this paper, we derive, in this section, the optimal grouping given that both questions are asked during each test.

In this case, the central station accesses the channel twice during each test: it first asks if any sensor within the group has the bit 1, then asks whether any sensor has the bit 0. Each sensor within the tested group replies through a noiseless OR channel, thus, for group U_l, the feedback obtained from the test is

$$(Z_{U_l}, \bar{Z}_{U_l}) = \left(\vee_{s_i \in U_l} X_i, \vee_{s_i \in U_l} \bar{X}_i\right). \tag{3}$$

The outcome of the test T_l may be one of the following pairs:
(r.1) $(Z_{U_l}, \bar{Z}_{U_l}) = (0, 1)$;
(r.2) $(Z_{U_l}, \bar{Z}_{U_l}) = (1, 0)$;
(r.3) $(Z_{U_l}, \bar{Z}_{U_l}) = (1, 1)$.
If the result of test T_l is (r.1), then the central station knows that $X_i = 0$ for all $s_i \in U_l$; and *vice versa* for (r.2). However, when the test results in (r.3), the content of the group U_l is not resolved since both 1 and 0 are contained in the group, we refer to this state as the *erasure*. When (r.3) occurs, further testing should be performed on a subgroup of U_l in order to resolve completely the contents within. This is equivalent to the ternary 0, 1, e feedback proposed in [3].

Given any particular group in the group testing strategy, it is most desirable to have each test result in either (r.1) or (r.2) since no further testing is necessary in these cases and the total number of tests necessary to resolve the whole field may be reduced. In modelling the sensors with a two-state Markov Chain and for any given group size n, choosing sensors that are adjacent to each other has the highest probability of obtaining (r.1) or (r.2) since these group of sensors have the highest correlation among each other. Therefore, we restrict our choice of the groups only to adjacent sensors (see Fig. 1). Furthermore, a sensor with a lower subscript is always chosen over the sensor with higher subscript if the content of that sensor has not yet been determined. For example, if $U_l = \{s_{k_l}, s_{k_l+1}, \cdots, s_{k_l+|U_l|-1}\}$ for some integer k_l, then the bits $x_0, \cdots, x_{k_l-1}$ are known through the tests $T_0, \cdots, T_{k_l-1}$.

Let L be the random variable that denotes the minimum number of tests necessary to reconstruct the sensor field. To derive the expectation of L, *i.e.* $\mathbf{E}\{L\}$, it is necessary to introduce the following notations:

$G_a(n)$: the expected number of tests to resolve the values of $U_l = \{x_{k_l}, \cdots, x_{k_l+n-1}\}$ given $x_{k_l-1} = a$, when the central node has no knowledge U_l.

$F_a(m)$: the expected number of tests to resolve the values of $U_l = \{x_{k_l}, \cdots, x_{k_l+m-1}\}$ given $x_{k_l-1} = a$ when the outcome of the test $(Z_{U_l}, \bar{Z}_{U_l}) = (1, 1)$.

Therefore, given that there are N nodes in the network, the expected number of tests necessary for the central node to completely reconstruct the sensor field is

$$\mathbf{E}\{L\} = (1-p) \cdot G_0(N) + p \cdot G_1(N). \tag{4}$$

Since $a \in \{0, 1\}$ is the value of the bit x_{k_l-1} preceding the tested group, we denote by $\Pr_a(E)$ the probability of the event

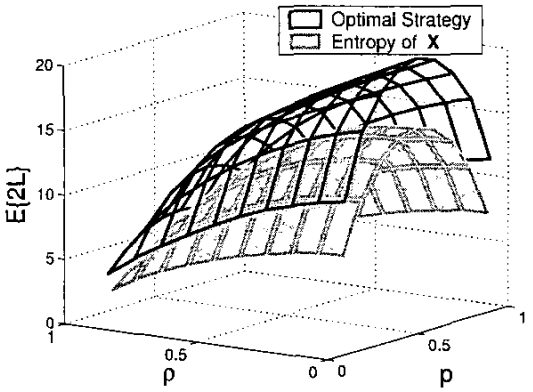

Figure 3: For $N = 16$, we show the expected number of access $E\{2L\}$ for each (p, ρ) pair.

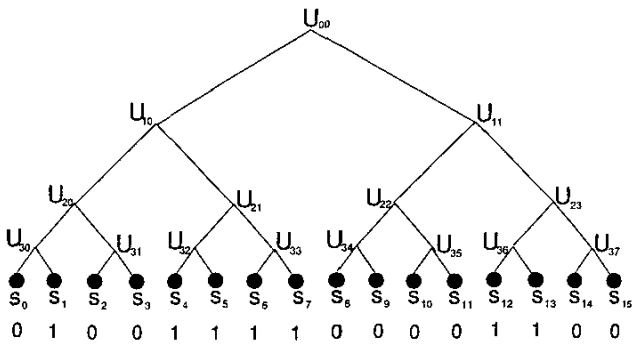

Figure 4: Example of the realization of a sensor field with the binary sequence 0100111100001100.

E conditioned on the preceding bit equal to a. The values of $G_a(n)$ and $F_a(m)$ can be written as the following recursion:

$$G_a(n) = 1 + \min_{1\le x\le n} \Pr_a(A_x)G_0(n-x) + \Pr_a(B_x)G_1(n-x) + \Pr_a(C_x)\left[F_a(x) + \sum_{i=2}^{x} \Pr_a(D_i^0|C_x)G_1(n-i) + \sum_{i=2}^{x} \Pr_a(D_i^1|C_x)G_0(n-i)\right] \quad (5)$$

$$F_a(m) = 1 + \min_{1\le x\le m-1} \Pr_a(A_x|C_m)F_0(m-x) + \Pr_a(B_x|C_m)F_1(m-x) + \Pr_a(C_x|C_m)F_a(x) \quad (6)$$

where the boundary conditions are $G_a(0) = 0$ and $F_a(1) = 0$; and the events are defined as follows:

$$\begin{aligned} A_x &= \{\text{All } x \text{ items have bit 0}\} \\ B_x &= \{\text{All } x \text{ items have bit 1}\} \\ C_x &= \{\text{Not all } x \text{ items have the same bit}\} \\ D_x^0 &= \{\text{First } x-1 \text{ items are 0 and the } x\text{th item is 1}\} \\ D_x^1 &= \{\text{First } x-1 \text{ items are 1 and the } x\text{th item is 0}\} \end{aligned}$$

The probabilities of these events are derived in the appendix.

In Fig. 3, we show, for each value of (p, ρ), the expected number of channel access $E\{2L\}$ required for the central node to reconstruct the sensor field using the optimal strategy. We compare the performance to the information lower bound of $\mathbf{X}$ which is also symmetric with respect to p since both question were asked during each test. We also observe that $E\{2L\}$ decreases as the correlation ρ increases, which is expected since, with high correlation, the sensor bits is likely to be the same, therefore, the values of a large group can be resolved with a single test. The gap between the optimal strategy and the information lower bound comes from the fact that we choose to ask both questions at each test even when it is not necessary. We could further optimize the strategy by taking into consideration the choice of the question along with the optimal group. However, this will not be treated in this paper.

In the optimal strategy, the number of nodes taken in each group is optimized depending on the outcome of the previous tests. This dynamic strategy may be computationally inefficient when the number of nodes are large. In the next section, we propose a suboptimal tree algorithm that is easily implementable and nonparametric to the probability distribution and the number of nodes in the network.

IV. Suboptimal Strategy

Consider a network of 16 nodes, as shown in Fig. 4. Let each vertex U_{ij} denote a group consisting of sensors within its subtree. In the suboptimal scheme, the sequence of tests starts from the subgroups of U_{00}, *i.e.* U_{10} and U_{11}, and continues the tests upon each subgroup in the order of their size. If the result of the test on U_{ij} is either (r.1) or (r.2), then the values of the sensors within U_{ij} are all resolved and the central station moves on to test the group $U_{i,j+1}$. However, if the test results in an erasure, *i.e.* (r.3), the central station does a binary splitting among the nodes within the subtree, and chooses the group $U_{i+1,2j}$ to be tested next. For the example shown in Fig. 4, the sequence of tests are done in the order of the following groups: $U_{10} \to U_{20} \to U_{30} \to U_{40} \to U_{31} \to U_{21} \to U_{11} \to U_{22} \to U_{23} \to U_{36} \to U_{37}$. After the test of U_{40}, the test on U_{41} is not necessary since we already know that U_{30} is in conflict.

The importance of the binary tree splitting algorithm is that the algorithm is nonparametric and it is applicable to all cases independent of the probability model. The performance of this scheme serves as an upper bound to the minimum number of tests that can be achieved through group testing. This strategy can be applied to a number of nodes equal to the power of 2. In the following, we use Capetanakis' approach [13] to calculate the expected number of tests necessary to completely reconstruct the sensor field. Assuming that the entire group is immediately split into two groups, we need a minimum of 2 tests

to resolve the field. Then, the number of tests can be expressed as follows:

$$L = 2 + 2\sum_{i=1}^{M-2}\sum_{j=0}^{2^i-1} n_{ij} + \sum_{j=0}^{2^{M-1}-1} n_{ij} \quad (7)$$

where

$$n_{ij} = \begin{cases} 1 & \text{if the subgroups of } U_{ij} \text{ are tested} \\ 0 & \text{otherwise.} \end{cases} \quad (8)$$

The factor of 2 in the second term of (7) represents the binary splitting of each group U_{ij}, therefore, two subgroups of U_{ij} is tested if $n_{ij} = 1$. Since we require the central node to ask both questions during each test, it is necessary to test subgroups of U_{ij} if and only if the set of sensors in U_{ij} do not contain the exact same bit of information. However, for $i = M-1$, each subgroup of $U_{M-1,j}$ consists of only one node. Therefore, having resolved the content of the first node gives you knowledge of the second node provided that $n_{ij} = 1$, since it is known that the two sensors do not contain the same bit of information. The probability of the event that $n_{ij} = 1$ is

$$\Pr(n_{ij} = 1) = \psi(M-i, \alpha, \beta) \quad (9)$$

where

$$\psi(M-i,\alpha,\beta) = 1 - \frac{\alpha}{\alpha+\beta}(1-\beta)^{2^{M-i}-1} - \frac{\beta}{\alpha+\beta}(1-\alpha)^{2^{M-i}-1}.$$

Then the expected number of tests can be calculated as

$$\mathbf{E}\{L\} = 2 + \sum_{i=0}^{M-2} 2^{i+1}\psi(M-i,\alpha,\beta) + 2^{M-1}\psi(1,\alpha,\beta). \quad (10)$$

If $\alpha, \beta \ll 1$, *i.e.* $\rho \approx 1$, we can approximate (9) as

$$\begin{aligned}\Pr(n_{ij}=1) &\cong 1 - \frac{\alpha[1-\beta(2^{M-i}-1)]}{\alpha+\beta} - \frac{\beta[1-\alpha(2^{M-i}-1)]}{\alpha+\beta} \\ &= \frac{2\alpha\beta}{\alpha+\beta}\cdot(2^{M-i}-1). \end{aligned} \quad (11)$$

Then, we can approximate the expected number of tests as

$$\begin{aligned}\mathbf{E}\{L\} &\cong 2 + \sum_{i=1}^{M-2} 2^{i+1}\cdot\frac{2\alpha\beta}{\alpha+\beta}\cdot(2^{M-i}-1) + 2^{M-1}\cdot\frac{2\alpha\beta}{\alpha+\beta} \\ &= 2 + [4N\cdot\log_2 N - 9N + 8]\cdot p(1-p)\cdot(1-\rho). \end{aligned} \quad (12)$$

In the TDMA scheme where each sensor is tested individually, the expected number of tests is equal to the total number of sensors N. Compared to the tree algorithm proposed above, the TDMA system is advantageous only when

$$2\{2 + [4N\cdot\log_2 N - 9N + 8]\cdot p(1-p)\cdot(1-\rho)\} > N, \quad (13)$$

i.e. when

$$p(1-p)(1-\rho) > \frac{N/2 - 2}{4N\cdot\log_2 N - 9N + 8}. \quad (14)$$

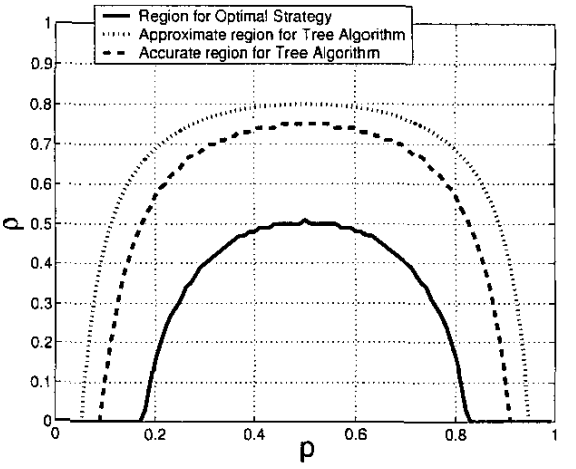

Figure 5: Let $N = 16$. TDMA is optimal in the region under the performance curves shown for the approximation and accurate derivation of the Tree Algorithm and the Optimal Strategy.

The factor of 2 on the LHS of (13) represents the two channel accesses that is needed in each testing while TDMA requires only one channel slot for each sensor. As shown in Fig. 5, the region for which TDMA is optimal compared to both the tree algorithm and the optimal strategy (c.f. Section III) is the area under the concave curve. Specifically, for a fixed number of sensors N, the tree algorithm performs better than TDMA when the correlation is high. This is reasonable since it is inefficient to ask each node individually if they are likely to contain the same bit. The region in (14) is symmetric with respect to p, as opposed to that in [12], since we ask both questions during each testing.

Although the binary splitting algorithm provides a scheme that is nonparametric to the distribution or the number of nodes in the network, it may be more desirable for one to split the root of the tree into 2^K branches, for $K > 1$, instead of just two branches, when further knowledge of the probability distribution of the sensor bits can be utilized. The optimization can be done by following the same approach as that in [13]. Intuitively, choosing a larger group of sensors to be tested is advantageous when the correlation among nodes are high, *i.e.* when the group of sensors tested has a higher probability of containing the same bit of information. However, when the correlation is low, one should choose a smaller group since a larger group would more likely result in a conflict. We claim that the optimal splitting of the root node varies monotonically with respect to the correlation coefficient ρ. This is different from that considered in [13] where the size of the tested group depends only on the probability p since all the sensors are modelled as *i.i.d.* Bernoulli.

Conjecture 1 *The number of branches 2^K for the optimal splitting of the root node decreases monotonically with the correlation coefficient ρ, for $0 \le \rho \le 1$.*

Given that the conjecture is true, we can determine the optimal K (denoted by K^*) as a function of ρ. Since the rela-

tion between ρ and K is monotone, there must exist an interval $(\hat{\rho}_{K+1,M}, \hat{\rho}_{K,M}]$ of ρ for each K such that $K^* = K$ for all $\rho \in (\hat{\rho}_{K+1,M}, \hat{\rho}_{K,M}]$. To simplify our analysis, we assume that each subgroup of the groups resulting in erasure are tested even when each subgroup contains only one node. This makes the algorithm slightly inefficient by not utilizing the advantage of the case of $i = M - 1$ as described previously. However, in this case, splitting the root node into 2^K branches is equivalent to starting the test from the the layer $i = K - 1$, which is equivalent to having 2^{K-1} binary trees with $2^{(M-K+1)}$ sensors in each tree. Therefore, it follows similar to that in [13] that

$$\mathbf{E}\{L|p, \rho, M, K\} = 2^{K-1}\mathbf{E}\{L|p, \rho, M - K + 1, 1\} \quad (15)$$

where the conditioning on K on the left-hand side indicates the logarithm of the number of branches the root node is split into, and M is the logarithm of the total number of sensors in the network. Similarly on the right-hand side. Furthermore, from Conjecture 1 and the continuity of the minimized expectation due to optimal splitting with respect to ρ, we solve for $\hat{\rho}_{K,M}$ by

$$\mathbf{E}\{L|p, \rho, M, K\} = \mathbf{E}\{L|p, \rho, M, K - 1\}. \quad (16)$$

From (15) and (16), we obtain the following condition for which K is the optimal splitting:

$$\psi(M - K + 1, \alpha, \beta) = \frac{1}{2}. \quad (17)$$

By substituting p, q and ρ into the previous equation, we obtain:

$$p[1 - q(1-\rho)]^{2^{M-K+1}-1} + q[1 - p(1-\rho)]^{2^{M-K+1}-1} = \frac{1}{2}. \quad (18)$$

For $p = 0.5$ and a given value of the correlation coefficient ρ, it is optimal to split the root node into 2^K branches where

$$K = \left\lfloor M + 1 - \log\left(1 + \frac{1}{1 - \log(1+\rho)}\right)\right\rfloor. \quad (19)$$

Furthermore, by approximating (17) the same way as in (11), we can obtain for $\alpha, \beta \ll 1$ that

$$K \approx \left\lfloor M + 1 - \log\left(1 + \frac{1}{4p(1-p)(1-\rho)}\right)\right\rfloor.$$

From (19), we observe that it is optimal to test each individual sensor one at a time for $(p, \rho) = (0.5, 0)$ since $K = M$ in this case. It is most likely that the sensors within a group will contain different symbols. This also achieves the information lower bound since $H(\mathbf{X}) = N \cdot H(X_1) = N$.

A. The probabilities used in (5) and (6)

Note that $\Pr_a(C_1) = 0$, then

$$\begin{aligned}
\Pr_0(A_x) &= (1-\alpha)^x; \\
\Pr_0(B_x) &= \alpha(1-\beta)^{x-1}; \\
\Pr_0(C_x) &= 1 - (1-\alpha)^x - \alpha(1-\beta)^{x-1}; \\
\Pr_0(D_i^0|C_x) &= \frac{\alpha(1-\alpha)^{i-1}}{1-(1-\alpha)^x - \alpha(1-\beta)^{x-1}}; \\
\Pr_0(D_i^1|C_x) &= \frac{\alpha\beta(1-\beta)^{i-2}}{1-(1-\alpha)^x - \alpha(1-\beta)^{x-1}}; \\
\Pr_0(A_x|C_m) &= \frac{(1-\alpha)^x \cdot [1-(1-\alpha)^{m-x}]}{1-(1-\alpha)^m - \alpha(1-\beta)^{m-1}}; \\
\Pr_0(B_x|C_m) &= \frac{\alpha(1-\beta)^{x-1} \cdot [1-(1-\beta)^{m-x}]}{1-(1-\alpha)^m - \alpha(1-\beta)^{m-1}}; \\
\Pr_0(C_x|C_m) &= \frac{1-(1-\alpha)^x - \alpha(1-\beta)^{x-1}}{1-(1-\alpha)^m - \alpha(1-\beta)^{m-1}}.
\end{aligned}$$

for $i \geq 2$. Similarly,

$$\begin{aligned}
\Pr_1(A_x) &= \beta(1-\alpha)^{x-1}; \\
\Pr_1(B_x) &= (1-\beta)^x; \\
\Pr_1(C_x) &= 1 - \beta(1-\alpha)^{x-1} - (1-\beta)^x; \\
\Pr_1(D_i^0|C_x) &= \frac{\alpha\beta(1-\alpha)^{i-2}}{1-\beta(1-\alpha)^{x-1} - (1-\beta)^x}; \\
\Pr_1(D_i^1|C_x) &= \frac{\beta(1-\beta)^{i-1}}{1-\beta(1-\alpha)^{x-1} - (1-\beta)^x}; \\
\Pr_1(A_x|C_m) &= \frac{\beta(1-\alpha)^{x-1} \cdot [1-(1-\alpha)^{m-x}]}{1-\beta(1-\alpha)^{m-1} - (1-\beta)^m}; \\
\Pr_1(B_x|C_m) &= \frac{(1-\beta)^x \cdot [1-(1-\beta)^{m-x}]}{1-\beta(1-\alpha)^{m-1} - (1-\beta)^m}; \\
\Pr_1(C_x|C_m) &= \frac{1-\beta(1-\alpha)^{x-1} - (1-\beta)^x}{1-\beta(1-\alpha)^{m-1} - (1-\beta)^m}.
\end{aligned}$$

References

[1] Robert Dorfman. The detection of defective members of large population. *The Annals of Mathematical Statistics*, 14(4):436–440, December 1943.

[2] Milton Sobel and Phyllis A. Groll. Group testing to eliminate efficiently all defectives in a binomial sample. *The Bell System Technical Journal*, 38:1179–1253, September 1959.

[3] Toby Berger, Nader Mehravari, Don Towsley, and Jack Wolf. Random multiple-access communication and group testing. *IEEE Trans. Commun.*, 32(7):769–779, July 1984.

[4] Jack K. Wolf. Born again group testing: Multiaccess communications. *IEEE Trans. Inform. Theory*, 31(2):185–191, March 1985.

[5] J.M. Shapiro. Embedded image coding using zerotrees of wavelet coefficients. *IEEE Trans. Signal Processing*, 41(12):3445–3462, December 1993.

[6] E.S. Hong and R.E. Ladner. Group testing for image compression. *IEEE Trans. Image Processing*, 11(8):901–911, August 2002.

[7] Prakash Ishwar, Animesh Kumar, and Kannan Ramchandran. Distributed sampling in dense sensor networks: a "bit-conservation" principle. In *Proceedings of International Workshop on Information Processing in Sensor Networks*, Palo Alto, CA, April 2003. Springer-Verlag Heidelberg.

[8] Qian Zhao and M. Effros. Lossless and near-lossless source coding for multiple access networks. *IEEE Trans. Inform. Theory*, 49(1):112–128, January 2003.

[9] S. S. Pradhan and K. Ramchandran. Distributed source coding using syndromes (discus): Design and construction. In *Proc. IEEE Data Compression Conf.*, 1999.

[10] S. D. Servetto. Lattice quantization with side information. In *Proceedings of Data Compression Conference (DCC)*, pages 510–519, March 2000.

[11] A. Scaglione and S. Servetto. On the interdependence of routing and data compression in multi-hop sensor networks. In *Proceedings of the 8th annual international conference on Mobile computing and networking*, Atlanta, GA, September 2002.

[12] Peter Ungar. The cutoff point for group testing. *Communications on Pure and Applied Mathematics*, 13:49–54, 1960.

[13] J. Capetanakis. Generalized TDMA: The Multi-Accessing Tree Protocol. *IEEE Trans. Commun.*, 27(10):1476–1484, October 1979.