Content-Based Multiple Access: Combining Source and Multiple Access Coding for Sensor Networks

Yao-Win Hong and Anna Scaglione
432 Rhodes Hall, Cornell University
Ithaca, New York, USA
Telephone: (607) 255-1246; Fax: (607) 255-9072
Email: yh84@cornell.edu, anna@ece.cornell.edu

Abstract—In this work, we explore the concept of group testing to efficiently acquire data from a distributed sensor field and to reconstruct the sensor field at a central station. We show that group testing techniques are not only an efficient tool to schedule multiple access transmissions, they are also transmission techniques that allow the central node to rapidly discriminate the information from the sensor field when a large number of sources generates data with low aggregate entropy. Our method enables the sensors to reconstruct a map of the entire sensor field with bandwidth requirements that depend on the precision of the reconstructed field and, thus, do not grow linearly with the increased number of nodes when the network density increases.

I. INTRODUCTION

In sensor network communications, the crucial problem is in gathering rapidly the information that the network produces. In several sensor applications, the data collected at each individual sensor rarely contain very significant information. This is why coverage is so important; the sensors are there to capture rare events. However, gathering the information from each sensor by scheduling their individual transmissions to each other or to a central fusion center may become overwhelmingly complex as the size of the network increases. In fact, it has been shown that this procedure does not scale [1] in the application of wireless networks. We contend that this procedure is not capitalizing on the main feature of sensor network applications — the individual sensors have very little to communicate most of the time and when there is something to communicate, the sensor data are often dependent. 

Consider a wireless sensor network $\mathcal{V}$ which contains the set of sensors $\{v_1, v_2, \ldots, v_N\}$. Each sensor wants to transmit a message $c$ out of the common codebook $\mathcal{C}$ which contains the information about its own observation. In most communication systems, each sensor $v \in \mathcal{V}$ is assigned a distinct channel where its message $c$ is deposited. In this case, the channel assignment is driven by the identity of the users instead of by the message contained in the sensor field. In this paper, we propose an efficient multiple access technique that is driven by the sensor data at each node. The main idea is to assign a channel to each message $c \in \mathcal{C}$ instead of to each user $v \in \mathcal{V}$ and use this channel to test whether or not a chosen group of sensors contain the message $c$. The major contribution of our work is to show that the optimal multiple access techniques should be driven by the information of the sensor field instead of by the distributed users.

In the channel assigned for message $c$, a certain group of sensors are asked whether or not the contain the message $c$, these sensors then respond with a positive answer if it contains the message $c$; otherwise, it remains silent. In fact, the type of test we consider well fits a very simple wireless access model where we assume that an incoherent energy detector can be sensitive enough to detect reliably the presence of an additive contribution to the received signal in the presence of receiver noise, even if only one member of the group sends a signal (a positive answer). In other words, we assume that the capacity of the channel from any group of users we test to the destination, with no knowledge of the channel state at either sides, can always carry at least 1 bit of aggregate information reliably. How to build a reliable test is beyond the scope of this work and will be subject of future investigation. Also, more powerful schemes could be implemented if it was possible to discriminate how many users provide a positive answer, but this will not be considered in this case.

A. Background

Group testing was first introduced by Dorfman [2] in World War II to efficiently test all syphilitic men that were called up for induction into the US army. This method reduces significantly the number of blood tests necessary to identify all men with the syphilitic disease by pooling the blood samples of several men into one test instead of testing them individually. In the work by Berger [3] and Wolf [4], group testing was first exploited as an effective method to schedule the transmission of a group of users in a multiple access channel. However, this multiple access strategy is driven by the users instead of by the information contained in the users. Therefore, it would be inefficient for sensor networks that have a large number of users while the observation field consists of low aggregate entropy.

The application of group testing in the wireless channel access problem [3], [4] has been used to resolve efficiently the channel contention among network nodes. The sequence of tests in [3] and [4] are aimed at retrieving the information of the random event that a node has a packet to transmit. In
this context, it has been shown that group testing allows the network to rapidly identify the nodes that intend to transmit and enable the transmission of these nodes respectively instead of letting them collide at random to acquire the channel [5].

Interestingly, group testing has also been proposed as an image compression tool [6] such as that of the popular zero-tree encoding [7]. The idea behind these compression schemes is that, after transforming correlated data into a domain where most of the energy is concentrated in few coefficients, transmitting the map of the coefficients that are above a certain quantization threshold is more efficient than compressing the image bits directly. It is a fortunate coincidence that both the multiple access and the image compression problem can benefit from the application of group testing. In the following, we show how advantageous it may be to exploit this relation in sensor network applications.

B. Our Contribution

To the best of our knowledge, our work utilizes for the first time the idea of group testing as a data centric physical layer method to transmit information in a multiple access channel. Our idea is that, by formulating the right tests to the right groups, the network can gather the answers of several sensors directly and efficiently in a compressed form. The tests for each group is mapped one-to-one onto an orthogonal channel (a time slot, a frequency band or a spreading code) supporting a reliable aggregate rate of at least one bit towards the destination that can be either a fusion center or the nodes themselves. Hence, assuming that the protocol for choosing the sequence of tests and the groups to test is known a priori, our scheme solves simultaneously the distributed compression problem as well as the distributed communication problem. The number of tests \( L \) is the minimum number of one bit channels, i.e. the total number of bits \( \times \text{sec} \times \text{Hz} \) needed to gather all the network data.

We consider the case of a random correlated field and provide an achievable upper-bound and a lower bound for the average number of tests needed. The upper-bound is obtained by a novel non-parametric group testing procedure (the tree algorithm) for which we are able to obtain bounds for the average length of the test. Because compression and channel coding are done simultaneously we conjecture that optimal group testing may lead to the most efficient joint source and multiple access coding protocol in networks where the nodes’ data have low aggregate entropy.

II. SYSTEM MODEL

We consider a network of sensors \( V = \{v_1, v_2, \ldots, v_N\} \) deployed on a unit area where each sensor \( i \) observes a sample \( S(v_i) \) from the bi-dimensional random field such that \( S = [S(v_1), S(v_2), \ldots, S(v_N)]^T \) is the vector of samples observed by the sensors. The goal of the system is to acquire efficiently an estimate of the entire sensor field \( \hat{S} \) from the distributed sensors, for a certain level of the total distortion \( d(\hat{S}, S) = D \).

An obvious lower bound on the total number of tests necessary to acquire the sensor field is provided, in this case, by the aggregate rate-distortion function, i.e.

\[
L_{\text{min}} \geq R(D) = \min_{p(S|\hat{S})} I(S; \hat{S})
\]  

(1)

Alternatively, we can assume that the sensors quantize their samples \( S(v) \) with a high resolution \( \Delta \ll 1 \) and that the lower bound on the minimum number of tests required is the joint entropy of the quantized field:

\[
L_{\text{min}} \geq H^\Delta(D).
\]  

(2)

For group testing problems with a Bernoulli i.i.d. random field the fact that such lower bounds are generally achievable cannot be proven. But it is interesting to note that, for both lower bounds, considering as distortion measure the mean square error, we have shown that both \( R(D) \) and \( H^\Delta(D) \) are \( O(\log(N/D)) \) [8] when the correlation of the sensor fields increases smoothly with the distance. If the lower bound is attainable, then it is necessary to accept a constant per node distortion, i.e. \( D/N = \text{constant} \), in order to require only a finite aggregate channel capacity \( L \) as \( N \to \infty \). However, we believe that this process is still unjustifiably skewed towards the sensor’s identity or location rather than the data themselves. In fact, it might be useless to determine, beyond a certain point, who is exactly the sensor that wants to transmit the message \( c \) within a certain group. Hence, spatial resolution in addition to resolution on the sensor value, is a reasonable quantity to consider finite, even when the density of the sensors grows to infinity.

Our goal in the remaining part of this paper is to identify a suboptimum strategy that, in a vast class of problems, leads to a scalable procedure to acquire the sensor field within a certain resolution on each sample \( d(\hat{S}(v), S(v)) \leq D/N \). This method provides us with an achievable upper-bound.

III. AN ACHIEVABLE RATE

Let the codebook \( C \) contain the levels of a uniform quantizer with the range \([-W, W] \). We assume that the levels are separated by \( \Delta = W/2^b \) where \( b \) is the number of bits necessary to represent the local samples. Given that sensor \( v \) observes the sample \( S(v) \), the message generated by \( v \) is

\[
X(v) = \arg \min_{c \in C} |S(v) - Q(c)|
\]  

(3)

where \( Q(c) \) is the quantization level corresponding to the message \( c \). Therefore, for \( Q(c) \) being the inner quantization level, the probability of \( X(v) = c \) is

\[
p(\text{X}(v) = c) = P(|S(v) - Q(c)| < \Delta),
\]  

(4)

We assume that the sensing goal is met if we receive the messages \( X(v) \) for all \( v \in V \), i.e. obtaining the estimates \( \hat{S}(v) = X(v) \) is sufficient to achieve \( d(\hat{S}(\theta), S(\theta)) \leq D/N \) since \( \Delta \) is chosen such that

\[
d(\hat{S}, S) = d(Q(X), S) \leq D.
\]  

(5)
During each channel access, a group of nodes \( G \subset V \) are given a certain test \( T(c) \): “Do you have the message \( c \) to transmit?” and the sensors within the group responds with a boolean answer

\[
B[v; T(c)] = \begin{cases} 
1 & \text{if } X(v) = c \\
0 & \text{otherwise.} 
\end{cases}
\]

for all \( v \in G \). By using the incoherent energy detector as the wireless access model, the fusion center is able to obtain the logic OR of the boolean variables, i.e.

\[
B[G; T(c)] = \begin{cases} 
1 & \text{if } \sum_{v \in G} B[v; T(c)] > 0 \\
0 & \text{otherwise.} 
\end{cases}
\]

If we impose the set of tests \( T = \{T(c) : c \in C\} \) consecutively on each chosen group \( G \), we know that \( X(v) = c \) for all \( v \in G \) if and only if \( B[G; T(c')] = 0 \) for all \( c' \neq c \). Therefore, the information obtained by testing a certain group can be written as the binary vector \( \mathbf{B}[G] = [B[G; T(c_0)], B[G; T(c_1)], \ldots, B[G; T(c_{2^B-1})]]^T \). If \( K \) uses of the tests \( T \) are necessary to resolve the contents of the sensors in \( G \), then the sensors transmit a total of \( K2^B \) bits to the fusion center while a time sharing strategy among individual sensors would require \( B \cdot |G| \) bits. Although it is not necessary to use the whole set of tests \( T \) on \( G \) since the information obtained from the tests on the superset \( F \supset G \) may already have indicated some events impossible, this improvement is neglected for the simplicity of the analysis.

Let \( G \) be the set of sensors tested at each use of \( T \) and let \( \overline{G} \subset G \) be the set of sensors that has not yet conveyed its message to the fusion center. Then, we can express the procedure to retrieve the vector of messages \( X \) as follows:

1) \( k \leftarrow 0; \)
2) \( G \leftarrow G_k; \overline{G} \leftarrow \overline{G}_k; \)
3) IF \( G \neq \emptyset \) and \( \overline{G} \neq \overline{G}_k \), then DO
   \[ \mathbf{B}_k \leftarrow [B[G; T(c_0)], B[G; T(c_1)], \ldots, B[G; T(c_{2^B-1})]]^T; \]
4) \( k \leftarrow k + 1; \)
5) IF \( k < K \) (the total number of sets), GOTO 2.

Clearly, the most important parameter of this algorithm is the choice of the groups that are tested. A sequential partition of groups is considered as non-ambiguous, if it allows to identify the messages \( X(v) \) at each node \( v \in V \) with no ambiguity. A trivial example of a non-ambiguous partition is \( G_k = v_k, \forall v_k \in V \). This selection corresponds to time sharing among individual sensors, and it is a particularly bad choice when the sensor information is highly correlated, since it requires the total number of tests \( L = O(N) \). In group testing, advantages are obtained by initially partitioning the network into larger groups and then refining the tests progressively into smaller subgroups when necessary.

To derive an upper bound on the optimal \( L \), we calculate the expected value of \( L \) for the binary tree splitting scenario which is a special case of the non-ambiguous partitioning. We assume that the partitions separate the nodes so that they belong to two sets that cover an equal connected geographical area. In the case of a one dimensional network where the nodes \( v_1 \) to \( v_N \) are labelled from left to right as shown in Fig. 1, we define the groups \( G_{ij} = \{v_{i2^M-1}, \ldots, v_{(j+1)2^M-1}\} \) where \( i = 0, \ldots, M, j = 0, \ldots, 2^l - 1 \) and \( M = \log_2 N \). An example is shown in Fig. 2 for a network of 16 nodes. Each of these groups is tested in the order of the size from large to small. However, the procedure allows us to test the group \( G \) only when there exists a \( v \in G \) whose content is not yet resolved by the previous tests.

Following the approach in [9], the number of tests necessary to acquire the sequence of quantized samples \( X \) for each realization of the sensor field can be expressed as

\[
L = |C| \cdot \left( 1 + 2 \sum_{i=0}^{B-1} \sum_{j=0}^{2^i-1} n_{ij} \right)
\]

where

\[
n_{ij} = \begin{cases} 
1 & \text{if the group } G_{ij} \text{ is split} \\
0 & \text{otherwise.} 
\end{cases}
\]

The multiplicative factor \( |C| \) is due to fact that all tests \(^1\) within \( T \) are imposed upon all groups.

The event \( \{n_{ij} = 1\} \) is the event that group \( G_{ij} \) is split into two subgroups. Since we impose the whole set of tests \( T \) upon each group, the group \( G_{ij} \) will be split if and only if the sensors within the group have more than one message to transmit. Let the messages \( c_0, c_1, \ldots, c_{2^B-1} \) be ordered such that \( Q(c_i) < Q(c_j) \) for all \( i < j \) and, thus, \( c_0 \) and \( c_{2^B-1} \) are the outer quantization levels. The random vector of observations made by the group \( G_{ij} \) are denoted by \( S_{ij} \). Then,

\(^1\)This is not necessary since some tests can be eliminated due to the result of previous tests.
the probability of the event \( \{n_{ij} = 1\} \) is

\[
\Pr\{n_{ij} = 1\} = 1 - \left( \sum_{k=1}^{q_i^2-2} \Pr\{(Q(c_k) - \Delta) \leq S_{ij} \leq (Q(c_k) + \Delta)\} \right.
\]

\[
- \Pr\{S_{ij} < (Q(c_k) + \Delta)\} - \Pr\{S_{ij} > (Q(c_{2q_k-1}) - \Delta)\} \right) .
\]

If we assume that the distribution is homogeneous over the sensor field, then the probability in (9) is dependent only on the index \(i\). Therefore, we can define \( \Pr(n_{ij} = 1) \triangleq \psi(i, B) \) and the expected number of tests can be written as

\[
E(L) = |C| \cdot \left( 1 + \sum_{i=0}^{B-1} 2^{i+1} \psi(i, B) \right) .
\]

In the following, we show the numerical simulation of the case of spatially correlated gaussian samples.

**IV. Example: Correlated Gaussian Samples**

Consider a one-dimensional network \( \mathcal{V} \) with nodes uniformly deployed within the interval \([0, 1]\), as shown in Fig. 1, and let \( x_k \) be the spatial location of sensor \( v_k \). For a network of \(|\mathcal{V}| = N\) sensors, assume that the correlation function of the samples is spatially homogeneous such that \( R_S(x_k, x_i) = R_S(x_k - x_i) = R_S(\frac{x_k - x_i}{\sigma}) \) and that it is smooth over the interval \([-1, 1]\). When the number of nodes increases, the distance between neighboring nodes will decrease as \(1/N\).

Therefore, the correlation function can be written as

\[
R_S = \begin{bmatrix}
R_S(0) & R_S(\varepsilon) & \ldots & R_S((N-1)\varepsilon) \\
R_S(\varepsilon) & R_S(0) & \ldots & R_S((N-2)\varepsilon) \\
\vdots & \vdots & \ddots & \vdots \\
R_S((N-1)\varepsilon) & R_S((N-2)\varepsilon) & \ldots & R_S(0) \\
\end{bmatrix}
\]

where \(\varepsilon = 1/N\).

Take the sensors within a group \( g_{ij} \) from the tree algorithm and let \( S_{ij} \) be the observations of this group of sensors. When the density of the sensors \( N \) is sufficiently large or when the group size \( N_{ij} = |g_{ij}| \) is sufficiently small, we can approximate the correlation matrix \( R_{S_{ij}} \) by the following:

\[
R_{S_{ij}} \approx R_S(0)1 \cdot 1^T + \frac{\varepsilon^2}{2} \bar{R}_S(0)J_1 + \frac{(2\varepsilon)^2}{2} \bar{R}_S(0)J_2 + \ldots + \frac{((N_{ij} - 1)\varepsilon)^2}{2} \bar{R}_S(0)J_{N_{ij}-1}
\]

\[
\approx R_S(0)1 \cdot 1^T
\]

where \( J_i \) is the matrix with all 1’s on the \( i \)-th and \((-i)\)-th off-diagonal and zero everywhere else. In this case, it is easy to show that \( R_{S_{ij}} \) has only one nonzero eigenvalue \( \lambda_1 = R_S(0) \) corresponding to the eigenvector \( u_1 \approx \sqrt{2}1 \). Then, one can transform the samples \( S_{ij} \) into a set of independent random variables \( Y_i(c_k) \) such that \( Y_1(c_k) \sim N(0, N_{ij}R_S(0)) \) and \( Y_2(c_k) = 0 \) a.e. for \( j \neq 1 \). Hence, under the condition that \((\varepsilon N_{ij})^2|\bar{R}_S(0)| \ll |R_S(0)|\), the probability in (9) becomes

\[
\Pr(n_{ij} = 1) \approx 1 - \Pr\left\{ -\infty < \frac{Y_1(c_k)}{\sqrt{N_{ij}}} < \infty \right\} = 0
\]

The condition \((\varepsilon N_{ij})^2|\bar{R}_S(0)| \ll |R_S(0)|\) implies that the test is terminated with probability close to 1 for a group in the \( i \)-th level of the tree. Hence, the efficiency of the tree algorithm depends on the second derivative of the correlation function.

Consider the case where the Gaussian samples are uniformly quantized into \( q \) levels within the range \([-\sigma^2, \sigma^2]\). The vector of samples \( S \) has mean 0 and the covariance matrix

\[
R_S(\xi, \nu) = \text{sinc}(\pi \xi)\text{sinc}(\pi \nu).
\]

In Fig. 3, we show the average number of tests necessary to obtain the samples of the quantized field. With the distortion of the chosen quantizer, we calculate the rate-distortion lower-bound for this example. We observe that the average number of tests for the tree algorithm does not increase linearly with \( N \). In fact, the growth rate increases similar to that of the rate-distortion bound, although a degradation in performance is observed due to its sub-optimal structure.

**REFERENCES**


