

DISTRIBUTED GROUP TESTING DETECTION IN SENSOR NETWORKS

Tamara Tošić and Pascal Frossard

Ecole Polytechnique Fédérale de Lausanne (EPFL)
Signal Processing Laboratory (LTS4), Lausanne, 1015-Switzerland
E-mail: {tamara.tosic,pascal.frossard}@epfl.ch.

ABSTRACT

We consider the problem of failure detection in sensor networks and we propose a new distributed detection algorithm based on Group Testing. We examine the presence of defective sensors by employing tests over locally gathered sensor measurements. Tests are represented with binary messages that sensors exchange over dissemination rounds using a gossip algorithm. We propose a novel probabilistic message design that allows the use of a low complexity decoder. Assuming that the maximum number of defective sensors is much smaller than the total number of sensors, we provide a bound on the number of linearly independent messages required for a successful detection of single or multiple defective sensors. Finally, simulations confirm that the proposed method outperforms algorithms based on random walk message gathering in terms of detection accuracy.

Index Terms— Detection, Group Testing, Distributed algorithms, Sensor Network

1. INTRODUCTION

Technology development has caused the emergence of simple and cheap sensors and led to the deployment of sensor networks. A network signal analysis is in general performed on sensor measurements, which may however be inaccurate. It becomes important to detect defective sensors in networks, so that their erroneous measurements do not impact the performance of signal analysis applications [1]. One of the main frameworks that detects defective elements by performing experiments on pools of items is known as Group Testing (GT) [2], [3]. Most of the GT decoding algorithms are based on the idea of detection by elimination performed by a central entity. However, sensor networks have usually a dynamic architecture with loose coordination, which raises increasing demand for collaborative data processing solutions.

In this paper, we design a novel distributed defective sensor detection method based on GT ideas. We assume that the number of defective sensors is much smaller than the total number of sensors in the network. We locally gather sensor measurements and design sets of probabilistic tests within sensor clusters that examine sensors' defectiveness. The test outcome has a binary value, where a nonzero value denotes the participation of a defective sensor in the test. The test functions and their outcomes together form messages that are propagated through the network by a gossip algorithm (rumor mongering) [4] following a pull protocol. When a node receives a novel message from its neighbor, it combines it with its own message. The newly formed message is sent with some probability in the next transmission round. Due to the probabilistic test design and the message dissemination protocol, a simple distance decoder can be used to detect the subset of defective sensors. We provide an upper bound on the number of messages per network clusters that assures

detection with high probability by a distance decoder. Simulations demonstrate the validity of our analysis and show that smaller number of measurements provides satisfactory results in practice. Finally, results also show that our method outperforms methods based on random walk message gathering algorithms in terms of defective sensor detection.

2. GROUP TESTING FRAMEWORK

A sensor network is defined as a connected acyclic graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where the vertices $\mathcal{V} = \{s_i\}_{i=1}^S$ represent the S sensors and the edges \mathcal{E} denote the transmission links between them. Nonzero edge values $e_{i,j} \in \mathcal{E}$ denote an existing communication link between sensors s_i and s_j . We assume that the sensors measure a smooth physical phenomena (for instance, temperature) and the measurements of non-defective neighbor sensors do not differ significantly. Further, we assume that at most $K \ll S$ sensors in the network can be defective.

The sensors perform low-cost boolean algebra operations. In this work, the term *test* denotes an experiment performed on a subset of sensors to detect a set of defective sensors. We mark the participation of sensors in tests with a binary *test matrix* \mathbf{W} of size $B \times S$, where B is the number of tests. The test matrix is obtained by a row-wise gathering of test participation indicators. Let $\mathbf{W}_{:,i}$, and $\mathbf{W}_{:,i}$ denote the i -th test matrix row (i -th test indicator vector) and column (participation of the sensor s_i in tests), respectively. The results of B tests are represented by the *test outcome* vector $\mathbf{g} \in \mathbb{F}_2^B$. Due to the boolean operations performed at sensors, nonzero test outcomes occur when at least one defective item participates in the test.

Group Testing is a centralized detection method that identifies defective sensors in a S -dimensional binary vector \mathbf{f} given the test matrix \mathbf{W} and the test outcome vector \mathbf{g} . Commonly, deterministic methods design the tests in advance and perform the detection based on an elimination method¹. However, when a priori test designs are not feasible, for instance in a large scope sensor network, it is necessary to apply the probabilistic GT approach [5]. The test outcomes are computed as $\mathbf{g} = \mathbf{W} \otimes \mathbf{f}$, where the element-wise boolean matrix multiplication operator \otimes represents the combination of the bitwise OR and the bitwise addition operator. The binary elements of probabilistic test matrix are created as:

$$w_{i,j} = \begin{cases} 1, & \text{with probability } q, \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

For properly selected parameter q , this design creates *disjunct* matrices with high probability [5]. The disjunct property ensures that

¹From the total set of items, non-defective items are eliminated from the zero valued test outcomes.

any test matrix column does not fall in the sub-space formed by any union of K other columns. The test matrix is (K, ϵ) -disjunct if these vectors differ in at least ϵ values, where ϵ is the robustness parameter. The probabilistic GT method exploits the disjunctness test matrix property and employs a distance decoder (Hamming) for detection. The decoder counts the number of different entries in the test matrix columns and the outcome vector \mathbf{g} . ϵ is set to be the value higher than the expected number of nonzero elements in a test matrix column for a given number of tests. Columns that differ in at most ϵ positions correspond to the defective sensor identifiers.

3. DISTRIBUTED GROUP TESTING

While GT is usually performed in centralized settings, we propose a new algorithm for distributed defective sensors detection that is based on novel test design and message dissemination strategy in a new distributed GT framework. The problem is to detect a set \mathcal{K} of up to $K \ll S$ defective sensors in a sensor network with S devices, given $B < S$ linearly independent network test messages.

Each sensor performs successive tasks to form test messages, which are generated and communicated in synchronized rounds. A round $t \in \mathbb{N}$ consists of two phases t_I and t_{II} , as shown in Fig. 1. During the first phase t_I of duration t^- , the sensors obtain messages. They are the estimates about the existence of defective sensors in their vicinity. In the second phase t_{II} , the sensors exchange and combine these messages with neighbors by employing a gossip mechanism.

The message construction in the phase t_I starts with the selection of L sensors, called hereafter *master* nodes. The master nodes and their neighbors cluster the network into disjoint subsets $\mathcal{V}_l \subset \mathcal{V}$, $l \in \{1, \dots, L\}$. They locally collect sensors readings and estimate the presence of defective sensors based on a dissimilarity measure. Sensors in the network participate in a test with probability q . The master node appends a binary value $f(s_i) \in \mathbf{f}$ to each sensor in a neighborhood which participates in the test, where $f(s_i) = 1$ denotes that the sensor s_i is defective. We finally assume that there might be noise in the test messages. Noise is represented by bit-flips in the test matrix. More precisely, noise in tests is modeled with the *activation probability* p . Noise flips the nonzero test matrix elements to zero with probability $1 - p$. To summarize, the *test outcome* computation is given by:

$$g_l = \mathbf{W}_{l,:} \otimes \mathbf{f} = \begin{cases} 1, & \text{sensor}(s) \in \mathcal{K}, \\ 0, & \text{otherwise.} \end{cases} \quad (2)$$

Further, each master node forms a log of the test process that we denote as *message*. The message $(g_l(t^-), \mathbf{W}_{l,:}(t^-))$ contains the test outcome result g_l and the test participation vector $\mathbf{W}_{l,:}$. It is further transmitted from master nodes to their neighbor nodes; this concludes the phase t_I . This phase is illustrated in Fig. 1(a). Further, we describe the dissemination process performed during the phase t_{II} and illustrated in Fig. 1(b). Every sensor s_i requests the message formed in the previous round following a gossip mechanism with a pull protocol from its neighbor s_j . A neighbor s_j chosen uniformly at random responds to this request. The sensor s_i further combines its current message generated in the phase t_I with neighbors' message from the round $t-1$ as follows:

$$\begin{aligned} g_i(t) &\leftarrow g_i(t^-) \oplus g_j(t-1), \\ \mathbf{W}_{i,:}(t) &\leftarrow \mathbf{W}_{i,:}(t^-) \oplus \mathbf{W}_{j,:}(t-1), \end{aligned} \quad (3)$$

where $g_j(t-1)$ denotes the j -th sensor outcome value in the round $t-1$ and $\mathbf{W}_{i,:}(t)$ represents the test indicator vector available at the

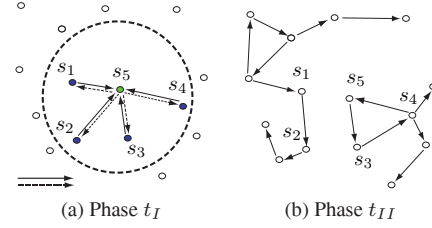


Fig. 1. Illustration of message design and dissemination. (a) Message design steps are illustrated by full and dashed arrows. In the first step, the master sensor collects the measurements from $\{s_1, \dots, s_4\}$ and forms the message $(g_l(t^-), \mathbf{W}_{l,:}(t^-))$. In the second step, it propagates the message to its neighbors. (b) Message dissemination based on a gossip algorithm with pull protocol.

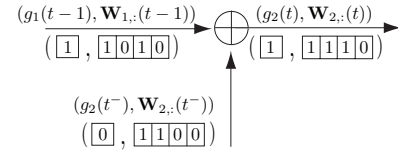


Fig. 2. The message formation at sensor s_2 in round t . We assume that the sensor s_3 is defective ($\mathbf{f} = [001 \dots 0]$) and that s_2 pulls s_1 to send its previous round values (round $t-1$). The outcome value and the test identifier vector are formed by bitwise OR operations.

sensor i in the round t . Since the messages are created probabilistically, the message combinations across the rounds assure that novel information reaches sensors in every round with high probability. A toy example of message dissemination is illustrated in Fig. 2. In this example, the sensor s_2 at the round $t = \tau$ pulls data from the sensor s_1 and constructs a novel message. Messages are represented within brackets and consist of test outcomes and test indicators. A novel message in s_2 is the result of bitwise additions amongst outcomes and test indicators.

The $B \times S$ test matrix is denoted by $\mathbf{W} = [\mathbf{W}_{1,:}; \dots; \mathbf{W}_{B,:}]$. We rewrite the equations (2) and (3) in a matrix form by:

$$\mathbf{g} = \mathbf{W} \otimes \mathbf{f}. \quad (4)$$

This equation resembles the outcome computation in the centralized GT case. However, in distributed GT, \mathbf{W} represents the boolean addition of rows of different disjunct matrix realizations (as in Eq. (1)) over different rounds. Note that for an arbitrary network, the number of network rounds required for collecting at least B linearly independent tests varies and depends on both the network topology and the test participation probability q .

In the multiple defective sensors case, we assume that the number of defective sensors is much smaller than the total number of sensors. Therefore, two messages with nonzero test outcomes are rarely combined together in a new message. However such an event leads in erroneous decoding and the combination of nonzero outcomes at sensors in t_{II} needs to be prevented. We propose to resolve this problem by simple protocol modifications. We buffer the i -th sensor message and forward the original j -th sensor message in the phase t_{II} . Then at the first subsequent round when two messages with zero-valued test outcomes occur at sensor i , the buffered value replaces its message with a zero-valued test. Next, these messages are combined and transmitted.

4. FAILURE DETECTION

4.1. Single defective sensor detection

Below we provide the necessary conditions for detection of a single defective sensor. A defective sensor is detected with high probabil-

ity with a distance decoder if the sensors participate in tests in t_I phases with probability q , (dual to (K, ϵ) -disjunct property, see def. in Sec. 2) and the number of independent messages collected by the sensor is proportional to $\mathcal{O}(K \log(S)/p^3)$, where p stands for the activation probability (see def. in Sec. 3). The formal propositions are given below. The first proposition gives the conditions for centralized detection when clusters' messages are gathered uniformly at random. The second one determines the number of per-cluster messages that ensures accurate detection by a distance decoder with high probability, where the messages are collected distributedly.

Proposition 4.1 *When in total L (K, ϵ_i) -disjunct matrices created as:*

$$w_{i,j} = \begin{cases} 1, & \text{with probability } \alpha_i = q_i, \\ 0, & \text{otherwise,} \end{cases} \quad (5)$$

are linearly combined as in Eq. (3), where $\epsilon = \sum_{i=1}^L \epsilon_i$ and $q = \sum_{i=1}^L q_i$, the resulting test matrix \mathbf{W} permits detection by a distance decoder with high probability as long as it contains in total $B \geq \mathcal{O}(K \log(S)/p^3)$ messages collected from clusters chosen at random.

The complete proof of this proposition is available in [6]. In short, we bound the error causing events by employing a Chernoff bound analysis and show that the probability of decoding failure per cluster is small. Errors occur in cases when the number of column flips in the probabilistic tests matrix generation is higher than ϵ or when its disjunct property is violated. Finally, we show in [6] that, if the number of collected messages is proportional to $\mathcal{O}(K \log(S)/p^3)$, the disjunct property holds for any fixed test matrix with high probability.

We further analyze the proposed distributed algorithm and consider the detection requirements for every sensor in the network. We show that the test messages collected by the sensors during the transmission rounds enable failure detection by the distance decoder with high probability if the number of messages is sufficient, when the decoder operations are performed locally at sensors.

Proposition 4.2 *We assume that L master sensors partition the sensor network in disjunct parts. Test realizations within a cluster form test vectors. Over the rounds, these vectors create in total L (K, ϵ) -disjunct matrices whose elements take values*

$$w_{i,j} = \begin{cases} 1, & \text{with probability } \alpha_i = q_i, \\ 0, & \text{otherwise,} \end{cases} \quad (6)$$

where $q = \sum_{i=1}^L q_i$. If the above assumptions hold and if the number of linearly independent messages received per cluster at every sensor in the network is at least B/L , where $B \geq \mathcal{O}(K \log(S)/p^3)$, the probability that sensors fail to detect the defective sensor by the distance decoder tends to zero as $S \rightarrow \infty$.

For a detailed proof see the corresponding Proposition in [6].

4.2. Multiple defective sensor detection

We assume here that the number of defective sensors is larger than one, but much smaller than the total number of sensors. We propose to modify the communication protocol of our distributed algorithm and to limit the decoder search space to be able to apply the Hamming distance decoder. The communication protocol forbids linear combinations of messages with two positive outcomes, as described in Section 3. In spite of that, the messages generated within clusters that contain more than one defective sensor may still cause error

occurrence. Erroneous messages are generated when a fraction of defective sensors in a cluster participates actively in tests, but this has a negligible probability [6].

We make the assumption that at most one defective sensor exists in any cluster. The decoding proceeds in two main steps. First, the appropriate unions of test matrix columns are created to form the search set space and second, the Hamming distance between the test outcome vector and the vectors of the search set are computed. The minimum Hamming distance indicates the solution of the detection problem.

The outcomes $\mathbf{g} = [\mathbf{g}_0 \ \mathbf{g}_1]^T$ are separated in negative and positive outcome vectors \mathbf{g}_0 and \mathbf{g}_1 , respectively. Subsequently, the rows of the test matrix \mathbf{W} form two sub-matrices \mathbf{W}_0 and \mathbf{W}_1 and Eq. (4) becomes:

$$\begin{bmatrix} \mathbf{g}_0 \\ \mathbf{g}_1 \end{bmatrix} = \begin{bmatrix} \mathbf{W}_0 & 0 \\ 0 & \mathbf{W}_1 \end{bmatrix} \begin{bmatrix} \mathbf{f}_0 \\ \mathbf{f}_1 \end{bmatrix}. \quad (7)$$

Then, we eliminate non-defective sensors from \mathbf{W}_1 using knowledge from \mathbf{W}_0 and obtain \mathbf{W}'_1 . We form sets of unions of up to K columns from the columns of matrix \mathbf{W}'_1 with at least one non-zero value. These columns are classified in sets \mathcal{H} whose size depends on the complete or partial sensor knowledge about cluster affiliation of other sensors in the network. Columns belonging to the same cluster are grouped together in a set \mathcal{H}_i , where $i \in \{1, \dots, L\}$ and L is the number of clusters. The search space \mathcal{U} consists of vectors that are obtained from unions of up to K columns, where each column is picked from a different set \mathcal{H}_i . We choose up to K columns, since the number of defective elements is smaller than or equal to K by the problem definition, while the selection of at most one column from a particular \mathcal{H}_i comes from the above simplifying assumption that at most one defective sensor exists in each cluster.

For the multiple defective sensor case, the transmission protocol ensures that the assumptions behind Proposition 4.2 are verified. We provide below conditions for multiple defective sensors detection with high probability.

Proposition 4.3 *Under the assumption that at most one defective sensor is present in each cluster, that the number of available linearly independent messages at all sensors is at least B/L per cluster, where $B \geq \mathcal{O}(K \log(S)/p^3)$ and that sensors know membership identifiers of all the clusters in the network, the distance decoder detects defective sensors at all sensors in the network with high probability.*

The proof and more details are available in [6].

5. EXPERIMENTAL RESULTS AND DISCUSSION

We evaluate now the performance of the proposed detection algorithm. Extensive simulations are performed on fully connected and irregular networks with $S = 70$ sensors and $L = 5$ master nodes, which are denoted with FG and IG respectively. All results are averaged over 100 simulations per graph realization. The minimum connectivity degree of IG is set to $k \geq 3$ and results are averaged over ten different realizations. The master nodes are chosen either deterministically (DM) or uniformly at random (RM).

Fig. 3 illustrates the influence of the graph connectivity and the value of the test participation probability on defective sensor detection for the proposed method. We compare the performance of FG and IG networks in terms of detection. Results show that for fixed detection probability and fixed network parameters, FG requires less message transmission rounds than IG. Both cases favor those values

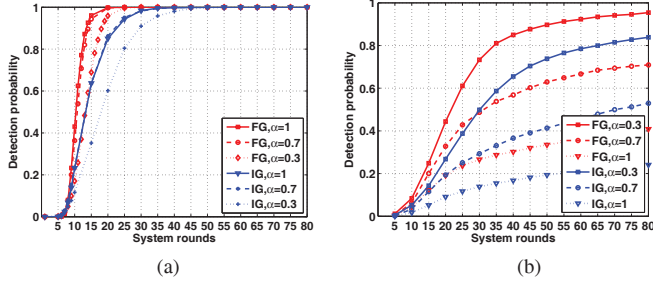


Fig. 3. Defective sensors detection probability of the proposed algorithm. Comparison is given for FG and IG with $k \geq 3$, $(S, L) = (70, 5)$. (a) $K = 1$, RM master nodes selection. (b) $K = 2$, DM master nodes selection.

of the test participation constant $\alpha = qK$ which increase the cardinality of the set of diverse messages. In Fig. 3(a), RM master node selection with high probability creates an innovative message in every round, so higher α values increase detection probability. In Fig. 3(b) DM master node selection partitions the sensors into fixed clusters over the system rounds, so distinct messages are formed for low values of the test participation probability.

The detection performance of the proposed method, denoted as GP, is compared with Random Walk gossip gathering method following a pull protocol dissemination (RWGP) and a Random Walk method (RW). A random walk defines the path of successive random dissemination message exchanges between the neighbor sensors. Both comparison methods initiate measurements collection in L nodes (equal to the number of master nodes in GP). The messages are transmitted by the gossip algorithm which follows a pull protocol. In RWGP the terminal node of a path creates a message based on the dissimilarity measure of the collected measurements. In RW, each node along the path stores the values of all previously “visited” nodes. More details about algorithms are given in [6]. Fig. 4 shows that our method outperforms the comparison detection algorithms in terms of the detection probability for both, FG and IG networks. Finally, we note that RM node selection boosts the detection performance if the values of the parameter α are chosen such that they increase the probability of forming diverse messages in the clusters.

Next, we analyze the theoretical number of messages required for detection with probability close to one and compare it with the corresponding values obtained by simulations in Fig. 3. The number of system rounds required for obtaining a particular number of messages at sensors depends on a network topology. We therefore examine a simple example and analyze FG networks with a single master node. The network parameter selection used in theoretical analysis is described in details in [5], since this specific setup is equivalent to the centralized test design analyzed in the same work. The theoretical results and the network specifications are reported in Table 1 and we see that for small size networks, the number of messages required for detection with high probability is larger than the number of sensors. However, from the FG graphs in Fig. 3 we can see that the required message number (system rounds) is much smaller in practice. With the increase of the number of sensors in the network, the number of messages required in theory (see Proposition 4.2) grows with a smaller rate and it is in general much smaller than the total number of sensors in a medium-size network.

6. CONCLUSIONS

We have studied defective sensor detection problem and proposed a novel distributed algorithm that detects a sparse set of defective sensors. To this aim, we design the network messages probabilistically,

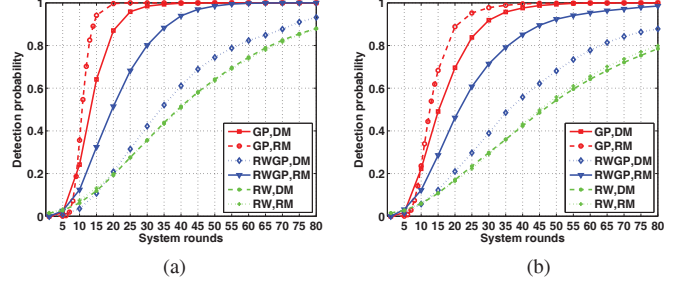


Fig. 4. Detection performance for proposed method (GP) and comparison methods, where $(S, L) = (70, 5)$. (a) fully connected sensor network (FG); (b) irregular sensor network (IG).

$p \in (0.9 - 1)$	S=20		S=70	
	$K = 1$	$K = 2$	$K = 1$	$K = 2$
	130	(115-244)	(174-217)	(125-284)

Table 1. The required theoretical number of messages. Network specifications: $\{S = 20, q \in (0.15 - 0.3), pf_1 = 0.01, pf_2 = 0.01\}$ and $\{S=70, q \in (0.15 - 0.3), pf_1=0.01, pf_2=0.01\}$.

which permits the use of a simple and efficient distance decoder at sensors. The network messages are created locally based on sensor measurements and they are communicated through the network by a gossip algorithm. We have derived the upper bound on the number of linearly independent messages per cluster that ensures detection of defective sensors with high probability. We have experimentally shown that, for fully connected and irregular sensor networks this value is smaller in practice. Finally, we have seen that the proposed method outperforms random walk-based data gathering methods in terms of detection performance on sample networks.

7. ACKNOWLEDGEMENTS

Authors thanks Dr. Nikolaos Thomos for fruitful discussions about the message dissemination algorithm.

8. REFERENCES

- [1] M.Young and R. Boutaba, “Overcoming adversaries in sensor networks,” *Accepted to IEEE Comm. Surveys and Tutorials*, 2011.
- [2] R. Dorfman, “The detection of defective members of large populations,” *Annals of Mathematical Statistics*, vol. 14, pp. 436–440, 1943.
- [3] H. B. Chen, F. K.Hwang, and C. M. Li, “Bounding the number of columns which appear only in positive pools,” *Taiwaneese J. Math.*, vol. 10, no. 4, pp. 927–932, June 2006.
- [4] A. G. Dimakis, S. Kar, J. M.F. Moura, M. G. Rabbat, and A. Scaglione, “Gossip algorithms for distributed signal processing,” *Submitted to IEEE Trans. Inform. Theory*, 2010.
- [5] M. Cheraghchi, A. Hormati, A. Karbasi, and M. Vetterli, “Group testing with probabilistic tests: Theory, design and application,” *Proc. of the 47th Annual Allerton Conf. on Comm., Control, and Computing*, 2010.
- [6] T. Tošić, N. Thomos, and P. Frossard, “Distributed sensor failure detection in sensor networks,” <http://arxiv.org/abs/1109.5636>, 2011.