

# Compressed Sensing with Probabilistic Measurements: A Group Testing Solution

Mahdi Cheraghchi, Ali Hormati, Amin Karbasi, Martin Vetterli  
 EPFL, School of Computer and Communication Sciences  
 1015 Lausanne, Switzerland  
 {mahdi.cheraghchi, ali.hormati, amin.karbasi, martin.vetterli}@epfl.ch

**Abstract**—Detection of defective members of large populations has been widely studied in the statistics community under the name “group testing”, a problem which dates back to World War II when it was suggested for syphilis screening. There, the main interest is to identify a small number of infected people among a large population using *collective samples*. In viral epidemics, one way to acquire collective samples is by sending agents inside the population. While in classical group testing, it is assumed that the sampling procedure is fully known to the reconstruction algorithm, in this work we assume that the decoder possesses only *partial* knowledge about the sampling process. This assumption is justified by observing the fact that in a viral sickness, there is a chance that an agent remains healthy despite having contact with an infected person. Therefore, the reconstruction method has to cope with two different types of uncertainty; namely, identification of the infected population and the partially unknown sampling procedure.

In this work, by using a natural probabilistic model for “viral infections”, we design non-adaptive sampling procedures that allow successful identification of the infected population with overwhelming probability  $1 - o(1)$ . We propose both probabilistic and explicit design procedures that require a “small” number of agents to single out the infected individuals. More precisely, for a contamination probability  $p$ , the number of agents required by the probabilistic and explicit designs for identification of up to  $k$  infected members is bounded by  $m = O(k^2(\log n)/p^2)$  and  $m = O(k^2(\log^2 n)/p^2)$ , respectively. In both cases, a simple decoder is able to successfully identify the infected population in time  $O(mn)$ .

## I. INTRODUCTION

Suppose that we have a large population in which only a small number of people are infected by a certain viral disease (e.g., one may think of a flu epidemic), and that we wish to identify the infected ones. By testing each member of the population individually, we can expect the cost of the testing procedure to be large. If we could instead pool a number of samples together and then test the pool collectively, the number of tests required might be reduced. This is the main conceptual idea behind the classical *group testing* problem which was introduced by Dorfman [1] and later found applications in variety of areas. A few examples of such applications include testing for defective items (e.g., defective light bulbs or resistors) as a part of industrial quality assurance [2], DNA sequencing [3] and DNA library screening in molecular biology (see, e.g., [4],

This work was supported by Swiss NSF grants 200020-115983/1, 5005-67322 and 200020-103729.

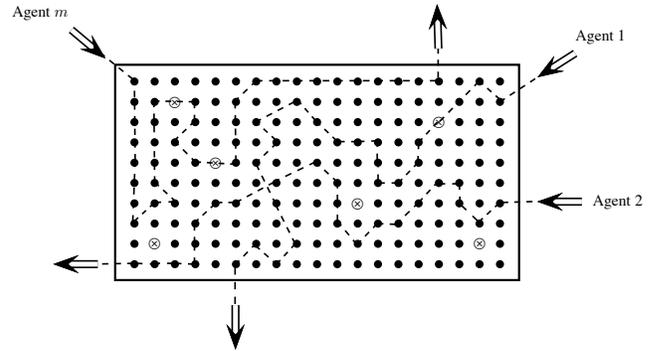


Fig. 1. Collective sampling using agents.  $\otimes$  symbols represent infected people among healthy people indicated by  $\bullet$  symbols. The dashed lines show the individuals contacted by the agents.

[5], [6], [7], [8] and the references therein), multiaccess communication [9], data compression [10], pattern matching [11], streaming algorithms [12], software testing [13], and compressed sensing [14]. See the books by Du and Hwang [15], [16] for a detailed account of the major developments in this area.

One way to acquire collective samples is by sending agents inside the population whose task is to contact people (see Fig. 1). The agents can also be chosen as ATM machines, cashiers in supermarkets, among other possibilities. Once an agent has made contact with an “infected” person, there is a *chance* that he gets infected, too. By the end of the testing procedure, all agents are gathered and tested for the disease. Here, we assume that each agent has a *log file* by which one can figure out with whom he has made contact. One way to implement the log in practice is to use identifiable devices (for instance, cell phones) that can exchange unique identifiers when in range. This way, one can for instance ask an agent to randomly meet a certain number of people in the population and at the end learn which individuals have been met from the data gathered by the device that is carried by the agent. Note that, even if an agent contacts an infected person, he will not get infected with certainty. Hence, it may well happen that an agent’s result is negative (meaning that he is not infected) despite a contact with some infected person. We will assume that when an agent gets infected, the resulting infection will not be contagious, i.e., an agent never infects

other people. Our ultimate goal is to identify the infected persons with the use of a simple recovery algorithm, based on the test results<sup>1</sup>. We remark that this model is applicable in certain scenarios different from what we described as well. For instance, in classical group testing, “dilution” of a sample might make some of the items present in a pool ineffective. The effect of dilution can be captured by the notion of contamination in our model.

It is important to notice the difference between this setup and the classical group testing where each contact with an infected person will infect the agent with certainty. In other words, in the classical group testing the decoder fully knows the sampling procedure, whereas in our setup, it has only uncertain knowledge. Hence, in this scenario the decoder has to cope simultaneously with two sources of uncertainty, the unknown group of infected people and the partially unknown (or stochastic) sampling procedure.

The collective sampling can be done in adaptive or non-adaptive fashions. In the former, samplings are carried out one at a time, possibly depending the outcomes of the previous agents. However, in the latter, the sampling strategy is specified and fixed before seeing the the test outcome for any of the agents. In this paper we only focus on non-adaptive sampling methods, which is more favorable for applications.

The idea behind our setup is mathematically related to compressed sensing [17], [18]. Nevertheless, they differ in a significant way: In compressed sensing, the samples are gathered as linear observations of a sparse real signal and typically tools such as linear programming methods is applied for the reconstruction. To do so, it is assumed that the decoder knows the measurement matrix a priori. However, this is not the case in our setup. In other words, using the language of compressed sensing, in our scenario the measurement matrix might be “noisy” and is not precisely known to the decoder. As it turns out, by using a sufficient number of agents this issue can be resolved.

## II. PROBLEM SETTING AND SUMMARY OF THE RESULTS

To model the problem, we enumerate the individuals from 1 to  $n$  and the agents from 1 to  $m$ . Let the non-zero entries of  $\mathbf{x} := (x_1, x_2, \dots, x_n) \in \mathbb{F}_2^n$  indicate the infected individuals within the population. Moreover, we assume that  $\mathbf{x}$  is a  $k$ -sparse vector, i.e., it has at most  $k$  nonzero entries (corresponding to the infected population). We refer to the *support set* of  $\mathbf{x}$  as the the set which contains positions of the nonzero entries.

As typical in the literature of group testing and compressed sensing, to model the non-adaptive samplings done by the agents, we introduce an  $m \times n$  boolean *contact* matrix  $M^c$  where we set  $M_{ij}^c$  to one if and only if the  $i$ th agent contacts the  $j$ th person. As we see, the matrix  $M^c$  only shows which agents contact which persons. In particular it does not indicate whether the agents eventually get affected by the

contact. Let us assume that at each contact with a sick person an agent gets infected independently with probability  $p$  (a fixed parameter that we call the *contamination probability*). Therefore, the real *sampling* matrix  $M^s$  can be thought of as a variation of  $M^c$  in the following way:

- Each non-zero entry of  $M^c$  is flipped to 0 independently with probability  $1 - p$ ;
- The resulting matrix  $M^s$  is used just as in classical group testing to produce the *outcome* vector  $\mathbf{y} \in \mathbb{F}_2^m$ ,

$$\mathbf{y} = M^s \cdot \mathbf{x}, \quad (1)$$

where the arithmetic is boolean (i.e., multiplication with the logical AND and addition with the logical OR).

The contact matrix  $M^c$ , the outcome vector  $\mathbf{y}$ , the number of non-zero entries  $k$ , and the contamination probability  $p$  are known to the decoder, whereas the sampling matrix  $M^s$  (under which the collective samples are taken) and the input vector  $\mathbf{x}$  are unknown. The task of the decoder is to identify the  $k$  non-zero entries of  $\mathbf{x}$  based on the known parameters.

*Example 1:* As a toy example, consider a population with 6 members where only two of them (persons 3 and 4) are infected. We send three agents to the population, where the first one contacts persons 1, 3, 5, the second one contacts persons 2, 4, 6, and the third one contacts persons 2, 3, 5, 6. Therefore, the contact matrix and the input vector have the following form

$$\begin{aligned} \mathbf{x} &= (0 \ 0 \ 1 \ 1 \ 0 \ 0)^\top, \\ \text{supp}(\mathbf{x}) &= \{3, 4\}, \\ M^c &= \begin{pmatrix} 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 1 & 1 \end{pmatrix}. \end{aligned}$$

Let us assume that only the second agent gets infected. This means that the outcome vector is

$$\mathbf{y} = (0 \ 1 \ 0)^\top.$$

As we can observe, there are many possibilities for the sampling matrix, all of the following form:

$$M^s = \begin{pmatrix} ? & 0 & ? & 0 & ? & 0 \\ 0 & ? & 0 & ? & 0 & ? \\ 0 & ? & ? & 0 & ? & ? \end{pmatrix},$$

where the question marks are 0 with probability  $1 - p$  and 1 with probability  $p$ . It is the decoder’s task to figure out which combinations make sense based on the outcome vector. For example, the following matrices and input vectors

<sup>1</sup>In this work we focus on the exact reconstruction of the set of infected individuals in the worst case (i.e., regardless of the choice of this set).

fit perfectly with  $\mathbf{y}$ :

$$\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \end{pmatrix},$$

$$\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 1 \end{pmatrix}.$$

More formally, the goal of our scenario is two-fold:

- 1) Designing the contact matrix  $M^c$  so that it allows unique reconstruction of *any* sparse input  $\mathbf{x}$  from outcome  $\mathbf{y}$  with overwhelming probability  $(1 - o(1))$  over the randomness of the sampling matrix  $M^s$ .
- 2) Proposing a recovery algorithm with low computational complexity.

In this work, we present a probabilistic and a deterministic approach for designing contact matrices suitable for our problem setting along with a simple decoding algorithm for reconstruction. Our approach is to first introduce a rather different setting for the problem that involves no randomness in the way the infection spreads out. Namely, in the new setting an adversary can arbitrarily decide whether a certain contact with an infected individual results in a contamination or not, and the only restriction on the adversary is on the total amount of contaminations being made. In this regard, the relationship between the adversarial variation of the problem and the original (stochastic) problem can be thought of akin to the one between the combinatorial problem of designing block codes with large minimum distances as opposed to designing codes for stochastic communication channels. The reason for introducing the adversarial problem is its combinatorial nature that allows us to use standard tools and techniques already developed in combinatorial group testing. Fortunately it turns out that solving the adversarial variation is sufficient for the original (stochastic) problem. We discuss this relationship and an efficient reconstruction algorithm in Section III.

Our next task is to design contact matrices suitable for the adversarial (and thus, stochastic) problem. We extend two standard techniques from group testing to our setting. Namely, we give a probabilistic and an explicit construction of the contact matrix in Sections IV and V, respectively. The probabilistic construction requires each agent to independently contact any individual with a certain well-chosen probability and ensures that the resulting data gathered at the end of the experiment can be used for correct identification of the infected population with overwhelming probability, provided that the number of agents is sufficiently large. Namely, for contamination probability  $p$ , we require  $O(k^2(\log n)/p^2)$  agents, where  $k$  is the estimate on the size of the infected population. The explicit construction, on the other hand,

precisely determines which agent should contact which individual, and guarantees correct identification with certainty in the adversarial setting and with overwhelming probability (over the randomness of the contaminations) in the stochastic setting. This construction requires  $O(k^2(\log^2 n)/p^2)$  agents which is inferior than what achieved by the probabilistic construction by a factor  $O(\log n)$ .

We point out that, very recently, Atia and Saligrama [19] developed an information theoretic perspective applicable to a variety of group testing problems, including a ‘‘dilution model’’ which is closely related to what we consider in this work. Contrary to our combinatorial approach, they use information theoretic techniques to obtain bounds on the number of required measurements. Their bounds are with respect to random constructions and typical set decoding as the reconstruction method. Specifically, in our terminology with contamination probability  $p$ , they obtain an information theoretic upper bound of  $O(k^2 \log n/p^2)$  on the number of measurements, which is comparable to what we obtain in our probabilistic construction.

*Remark:* As is customary in the standard group testing literature, we think of the sparsity  $k$  as a parameter that is noticeably smaller than the population size  $n$ ; for example, one may take  $k = O(n^{1/3})$ . Indeed, if  $k$  becomes comparable to  $n$ , there would be little point in using a group testing scheme and in practice, for large  $k$  it is generally more favorable to perform trivial tests on the individuals. Nevertheless it is easy to observe that our probabilistic scheme can in general achieve  $m = O(k^2 \log(n/k)/p^2)$ , but we ignore such refinements for the sake of clarity.

### III. ADVERSARIAL SETTING

The problem described in Section II has a stochastic nature, in that the sampling matrix is obtained from the contact matrix through a random process. In this section we introduce an adversarial variation of the problem that we find more convenient to work with.

In the adversarial variation of the problem, the sampling matrix is obtained from the contact matrix by flipping up to  $e$  arbitrary entries to 0 on the support (i.e., the set of nonzero entries) of each column of  $M^c$ , for some *error parameter*  $e$ . The goal is to be able to exactly identify the sparse vector despite the perturbation of the contact matrix and regardless of the choice of the altered entries. Note that the classical group testing problem corresponds to the special case  $e = 0$ . Thus the only difference between the adversarial problem and the stochastic one is that in the former problem the flipped entries of the contact matrix are chosen arbitrarily (as long as there are not too many flips) while in the latter they are chosen according to a specific random process.

It turns out that the combinatorial tool required for solving the adversarial problem is precisely the notion of *disjunct* matrices that is well studied in the group testing literature. The formal definition is as follows.

*Definition 2:* A boolean matrix  $M$  with  $n$  columns  $M_1, \dots, M_n$  is called  $(k, e)$ -disjunct if, for every subset

$S \subseteq [n]$  of the columns with  $|S| \leq k$ , and every  $i \notin S$ , we have

$$\left| \text{supp}(M_i) \setminus \left( \bigcup_{j \in S} \text{supp}(M_j) \right) \right| > e,$$

where  $\text{supp}(M_i)$  denotes the support of the column  $M_i$ .

The following proposition shows a one-to-one correspondence between contact matrices suitable for the adversarial problem and disjoint matrices:

*Proposition 3:* Let  $M$  be a  $(k, e)$ -disjunct matrix. Then taking  $M$  as the contact matrix solves the adversarial problem for  $k$ -sparse vectors with error parameter  $e$ . Conversely, any matrix that solves the adversarial problem must be  $(k - 1, e)$ -disjunct.

*Proof:* Let  $M$  be a  $(k, e)$ -disjunct matrix and consider  $k$ -sparse vectors  $x, x'$  supported on different subsets  $S, S' \subseteq [n]$ . Take an element  $i \in S'$  which is not in  $S$ . By Definition 2, we know that the column  $M_i$  has more than  $e$  entries on its support that are not present in the support of any  $M_j, j \in S$ . Therefore, even after  $e$  bit flips in  $M_i$ , at least one entry in its support remains that is not present in the measurement outcome of  $x'$ , and this makes  $x$  and  $x'$  distinguishable.

For the reverse direction, suppose that  $M$  is not  $(k - 1, e)$ -disjunct and take any  $i \in [n]$  and  $S \subseteq [n]$  with  $|S| \leq k - 1, i \notin S$  which demonstrate a counterexample for  $M$  being  $(k - 1, e)$ -disjunct. Consider  $k$ -sparse vectors  $x$  and  $x'$  supported on  $S$  and  $S \cup \{i\}$ , respectively. An adversary can flip up to  $e$  bits on the support of  $M_i$  from 1 to 0, leave the rest of  $M$  unchanged, and ensure that the measurement outcomes for  $x$  and  $x'$  coincide. Thus  $M$  is not suitable for the adversarial problem. ■

Of course, posing the adversarial problem is only interesting if it helps in solving the original stochastic problem from which it originates. Below we show that this is indeed the case; and in fact the task of solving the stochastic problem reduces to that of the adversarial problem; and thus after this point it suffices to focus on the adversarial problem.

*Proposition 4:* Suppose that  $M$  is an  $m \times n$  contact matrix that solves the adversarial problem for  $k$ -sparse vectors with some error parameter  $e$ . Moreover, suppose that the weight of each column of  $M$  is between  $(1 - \delta)qm$  and  $qm$ , for a parameter  $q \in (0, 1)$  and a constant  $\delta \in (0, 1)$ , and that  $e = (1 - p)(1 + \delta)qm$ , for a constant  $p \in (0, 1)$ . Then  $M$  can be used for the stochastic problem with contamination probability  $p$ , and achieves error probability at most  $n2^{-\Omega(qm)}$ , where probability is taken over the randomness of sampling (and the constant behind  $\Omega(\cdot)$  depends on  $p$  and  $\delta$ ).

*Proof:* Take any column  $M_i$  of  $M$ , and let  $w_i$  be its weight. After the bit flips, we expect the weight of the column to reduce to  $pw_i$ . Moreover, by Chernoff bounds, the probability that (for ‘‘small’’  $\delta$ ) the amount of bit flips exceeds  $(1 - p)w_i(1 + \delta)$  is at most

$$\begin{aligned} \exp(-\delta^2(1 - p)w_i/4) &\leq \\ \exp(-\delta^2(1 - \delta)(1 - p)qm/4) &= 2^{-\Omega(qm)}. \end{aligned}$$

Thus, by a union bound, the probability that the amount of bit flips at some column is not tolerable by  $M$  is at most  $n2^{-\Omega(qm)}$ . ■

*Remark:* Note that, as we mentioned earlier, the adversarial problem is stronger than classical group testing, and thus, any lower bound on the number of measurements required for classical group testing applies to our problem as well. It is known that any measurement matrix that avoids confusion in standard group testing requires at least  $\Omega(k^2 \log_k n)$  measurements [20], [21], [22]. Thus we must necessarily have  $m = \Omega(k^2 \log_k n)$  as well, and this upper bounds the error probability given by Proposition 4 by at most  $n^{1 - \Omega(qk^2 / \log k)} = o(1)$ .

#### A. Decoding

Suppose that the contact matrix  $M^c$  is  $(k, e)$ -disjunct. Therefore, by Proposition 3 it can combinatorially distinguish between  $k$ -sparse vectors in the adversarial setting with error parameter  $e$ . In this work we consider a very simple decoder that works as follows.

**Distance decoder:** For any column  $c_i$  of the contact matrix  $M^c$ , the decoder verifies the following:

$$|\text{supp}(c_i) \setminus \text{supp}(y)| \leq e, \quad (2)$$

where  $y$  is the vector consisting of the measurement outcomes. The coordinate  $x_i$  is decided to be nonzero if and only if the inequality holds.

*Lemma 5:* The distance decoder correctly identifies the correct support of any  $k$ -sparse vector (with the above disjointness assumption on  $M$ ).

*Proof:* Let  $x$  be a  $k$ -sparse vector and  $S := \text{supp}(x)$ ,  $|S| \leq k$ , and  $M_S^c$  denote the corresponding set of columns in the sampling matrix. Obviously all the columns in  $M_S^c$  satisfy (2) (as no column is perturbed in more than  $e$  positions) and thus the reconstruction includes the support of  $x$  (this is true regardless of the disjointness property of  $M$ ). Now let the vector  $\hat{y}$  be the bitwise OR of the columns in  $M_S^c$  so that  $\text{supp}(y) \subseteq \text{supp}(\hat{y})$ , and assume that there is a column  $c$  of  $M^c$  outside  $S$  that satisfies (2). Thus we will have  $|\text{supp}(c) \setminus \text{supp}(\hat{y})| \leq e$ , and this violates the assumption that  $M^c$  is  $(k, e)$ -disjunct. Therefore, the distance decoder outputs the exact support of  $x$ . ■

## IV. PROBABILISTIC DESIGN

In light of Propositions 3 and 4, we know that in order to solve the stochastic problem with contamination probability  $p$  and sparsity  $k$ , it is sufficient to construct a  $(k, e)$ -disjunct matrix for an appropriate choice of  $e$ . In this section, we consider a probabilistic construction for  $M^c$ , where each entry of  $M^c$  is set to 1 independently with probability  $q := \alpha/k$ , for a parameter  $\alpha$  to be determined later, and 0 with probability  $1 - q$ . We will use standard arguments to show that, if the number of measurements  $m$  is sufficiently large, then the resulting matrix  $M^c$  is suitable with all but a vanishing probability.

Let  $\delta > 0$  be an arbitrary (and small) constant. Using Chernoff bounds, we see that if  $m \gg \log n$  (which will be the case), with probability  $1 - o(1)$  no column of  $M^c$  will have weight greater than  $q(1 + \delta)m$  or less than  $q(1 - \delta^2)m$ . Thus in order to be able to apply Proposition 4, it suffices to set  $e := (1 - p)(1 + 3\delta)qm$  as this value is larger than the error parameter  $(1 - p)(1 + \delta)^2qm$  required by the proposition.

*Lemma 6:* For the above choices of the parameters  $q$  and  $e$ , the probabilistic construction obtains a  $(k, e)$ -disjunct matrix with probability  $1 - o(1)$  using  $m = O(k^2(\log n)/p^2)$  measurements.

*Proof:* Consider any set  $S$  of  $k$  columns of  $M^c$ , and any column outside these, say the  $i$ th column where  $i \notin S$ . First we upper bound the probability of a *failure* for this choice of  $S$  and  $i$ , i.e., the probability that the number of the positions at the  $i$ th column corresponding to which all the columns in  $S$  have zeros is at most  $e$ . Clearly if this event happens the  $(k, e)$ -disjunct property is violated. On the other hand, if for no choice of  $S$  and  $i$  a failure happens the matrix is indeed  $(k, e)$ -disjunct.

Now we compute the failure probability  $p_f$  for a fixed  $S$  and  $i$ . A row is *good* if at that row the  $i$ th column has a 1 but all the columns in  $S$  have zeros. For a particular row, the probability that the row is good is  $q(1 - q)^k$ . Then failure corresponds to the event that the number of good rows is at most  $e$ . The distribution on the number of good rows is binomial with mean  $\mu = q(1 - q)^k m$ . By a Chernoff bound, the failure probability is at most

$$\begin{aligned} p_f &\leq \exp(-(\mu - e)^2/(2\mu)) \\ &= \exp(-mq((1 - q)^k - \\ &\quad (1 - p)(1 + 3\delta))^2/(2(1 - q)^k)) \\ &\leq \exp(-mq(1/3^\alpha - (1 - p)(1 + 3\delta))^2/2^{1-\alpha}) \end{aligned}$$

where the last inequality is due to the fact that  $(1 - q)^k = (1 - \alpha/k)^k$  is always between  $1/3^\alpha$  and  $1/2^\alpha$ . Let  $\gamma := (1/3^\alpha - (1 - p)(1 + 3\delta))^2/2^{1-\alpha}$ . Note that by choosing the parameters  $\alpha$  and  $\delta$  as sufficiently small constants,  $\gamma$  can be made arbitrarily close to  $p^2/2$ .

Now if we apply a union bound over all possible choices of  $S$  and  $i$ , the probability of coming up with a bad choice of  $M^c$  would be at most  $n \binom{n}{k} \exp(-mq\gamma)$ . This probability vanishes so long as  $m > k^2 \log(n/k)/(\alpha\gamma) = O(k^2(\log n)/p^2)$ . ■

Along with Propositions 3 and 4, the result above immediately gives the following:

*Theorem 7:* The probabilistic design for construction of an  $m \times n$  contact matrix  $M^c$  achieves  $m = O(k^2(\log n)/p^2)$  measurements and error probability at most  $n^{-\Omega(k/\log k)} = o(1)$  for the stochastic problem using distance decoder as the reconstruction method.

The probabilistic construction results in a rather sparse matrix, namely, one with density  $O(1/k)$  that decays with the sparsity parameter  $k$ . Below we show that sparsity is necessary condition for the construction to work:

*Lemma 8:* Let  $M$  be an  $m \times n$  boolean random matrix, where  $m = O(k^2 \log n)$  for an integer  $k > 0$ , which is

constructed by setting each entry independently to 1 with probability  $q$ . Then either  $q = O(\log k/k)$  or otherwise the probability that  $M$  is  $(k, e)$ -disjunct (for any  $e \geq 0$ ) approaches to zero as  $n$  grows.

*Proof:* Suppose that  $M$  is an  $m \times n$  matrix that is  $(k, e)$ -disjunct. Observe that, for any integer  $t \in (0, k)$ , if we remove any  $t$  columns of  $M$  and all the rows on the support of those columns, the matrix must remain  $(k - t, e)$ -disjunct. This is because any counterexample for the modified matrix being  $(k - t, e)$ -disjunct can be extended to a counterexample for  $M$  being  $(k, e)$ -disjunct by adding the removed columns to its support.

Now consider any  $t$  columns of  $M$ , and denote by  $m_0$  the number of rows of  $M$  at which the entries corresponding to the chosen columns are all zeros. The expected value of  $m_0$  is  $(1 - q)^t m$ . Moreover, for every  $\delta > 0$  we have

$$\Pr[m_0 > (1 + \delta)(1 - q)^t m] \leq \exp(-\delta^2(1 - q)^t m/4) \quad (3)$$

by a Chernoff bound.

Let  $t_0$  be the largest integer for which  $(1 + \delta)(1 - q)^{t_0} m \geq \log n$ . If  $t_0 < k - 1$ , we let  $t := 1 + t_0$  above, and this makes the right hand side of (3) upper bounded by  $o(1)$ . So with probability  $1 - o(1)$ , the chosen  $t$  columns of  $M$  will keep  $m_0$  at most  $(1 + \delta)(1 - q)^t m$ , and removing those columns and  $m_0$  rows on their union leaves the matrix  $(k - t_0 - 1, e)$ -disjunct, which obviously requires at least  $\log n$  rows (as even a  $(1, 0)$ -disjunct matrix needs so many rows). Therefore, we must have

$$(1 + \delta)(1 - q)^t m \geq \log n$$

or otherwise (with overwhelming probability)  $M$  will not be  $(k, e)$ -disjunct. But the latter inequality is not satisfied by the assumption on  $t_0$ . So if  $t_0 < k - 1$ , little chance remains for  $M$  to be  $(k, e)$ -disjunct. Now consider the case  $t_0 \geq k - 1$ . By a similar argument as above, we must have

$$(1 + \delta)(1 - q)^k m \geq \log n$$

or otherwise the matrix will not be  $(k, e)$ -disjunct with overwhelming probability. The above inequality implies that we must have

$$q \leq \frac{\log(m(1 + \delta)/\log n)}{k},$$

which, for  $m = O(k^2 \log n)$  gives  $q = O(\log k/k)$ . ■

## V. EXPLICIT DESIGN

In the previous section we showed how a random construction of the contact matrix achieves the desired properties for the adversarial (and thus, stochastic) model that we consider in this work. However, in principle an unfortunate choice of the contact matrix might fail to be of use (for example, it is possible though very unlikely that the contact matrix turns out to be all zeros) and thus it is of interest to have an explicit and deterministic construction of the contact matrix that is guaranteed to work.

In this section, we demonstrate how a classical construction of superimposed codes due to Kautz and Singleton [23]

can be extended to our setting by a careful choice of the parameters. This is given by the following theorem.

*Theorem 9:* There is an explicit construction for an  $m \times n$  contact matrix  $M^c$  that is guaranteed to be suitable for the stochastic problem with contamination probability  $p$  and sparsity parameter  $k$ , and achieves  $m = O(k^2(\log^2 n)/p^2)$ .

*Proof:* Let  $m$  be an even power of a prime, and  $n' := \sqrt{m}$ . Consider a Reed-Solomon code of length  $n'$  and dimension  $k'$  over an alphabet of size  $n'$ . The contact matrix  $M^c$  is designed to have  $n'^{k'}$  columns, one for each codeword. Consider a mapping  $\varphi: \mathbb{F}_{n'} \rightarrow \mathbb{F}_2^{n'}$  that maps each element of  $\mathbb{F}_{n'}$  to a unique canonical basis vector of length  $n'$ ; e.g.,  $0 \mapsto (1, 0, 0, \dots, 0)^\top$ ,  $1 \mapsto (0, 1, 0, \dots, 0)^\top$ , etc. The column corresponding to a codeword  $c$  is set to the binary vector of length  $m$  that is obtained by replacing each entry  $c_i$  of  $c$  by  $\varphi(c_i)$ , blowing up the length of  $c$  from  $n'$  to  $n'^2$ .

Note that the number of columns of  $M^c$  is  $n := n'^{k'} = m^{k'/2}$ , and each column has weight exactly  $n' = m/n'$ . Moreover, the support of any two distinct columns intersect at less than  $k'$  entries, because of the fact that the underlying Reed-Solomon code is an MDS code and has minimum distance  $n' - k' + 1$ . Thus in order to ensure that  $M^c$  is  $(k, e)$ -disjunct, it suffices to have  $n' - kk' > e$  (so that no set of  $k$  columns of  $M^c$  can cover too many entries of any column outside the set), or equivalently,

$$\sqrt{m} - 2k(\log n / \log m) > e. \quad (4)$$

By Proposition 4, we need to set  $e := (1 - p)(1 + \delta)m/n'$  for an arbitrary constant  $\delta > 0$ . Thus in order to satisfy (4), it suffices to have  $\sqrt{m}(1 - (1 - p)(1 + \delta)) > 2k \log n$ , which gives  $m > 4k^2 \log^2 n / (1 - (1 - p)(1 + \delta))^2$ . As  $\delta$  can be chosen arbitrarily small, the denominator can be made arbitrarily close to  $p^2$  and thus we conclude that this construction achieves  $m = O(k^2 \log^2 n / p^2)$  measurements, which is essentially larger than the amount achieved by the probabilistic construction by a factor  $O(\log n)$ . ■

Observe that, unlike the probabilistic construction of the previous section, the explicit construction above guarantees a correct reconstruction in the adversarial setting (where up to a  $1 - p$  fraction of the entries on the support of each column of the contact matrix might be flipped to zero). Moreover, in the original stochastic setting with contamination probability  $p$ , a single matrix given by the explicit construction guarantees correct reconstruction with overwhelming probability, where the probability is only over the randomness of the testing procedure. This is in contrast with the probabilistic construction where the failure probability is small, but originates from two sources; namely, unfortunate outcome of the testing

procedure as well as unfortunate choice of the contact matrix  $M^c$ .

## REFERENCES

- [1] R. Dorfman, "The detection of defective members of large populations," *Annals of Mathematical Statistics*, vol. 14, pp. 436–440, 1943.
- [2] M. Sobel and P. Groll, "Group-testing to eliminate efficiently all defectives in a binomial sample," *Bell Systems Technical Journal*, vol. 38, pp. 1179–1252, 1959.
- [3] P. Pevzner and R. Lipshutz, "Towards DNA sequencing chips," in *Proceedings of MFCS*, ser. Lecture Notes in Computer Science, vol. 841, 1994, pp. 143–158.
- [4] H. Ngo and D. Du, "A survey on combinatorial group testing algorithms with applications to DNA library screening," *DIMACS Series on Discrete Math. and Theoretical Computer Science*, vol. 55, pp. 171–182, 2000.
- [5] A. Schliep, D. Torney, and S. Rahmann, "Group testing with DNA chips: Generating designs and decoding experiments," in *Proc. Comp. Syst. Bioinformatics*, 2003.
- [6] A. Macula, "Probabilistic nonadaptive group testing in the presence of errors and DNA library screening," *Annals of Combinatorics*, vol. 3, no. 1, pp. 61–69, 1999.
- [7] —, "Probabilistic nonadaptive and two-stage group testing with relatively small pools and DNA library screening," *J. Comb. Optim.*, vol. 2, pp. 385–397, 1999.
- [8] Y. Cheng and D.-Z. Du, "New constructions of one- and two-stage pooling designs," *Journal of Computational Biology*, vol. 15, no. 2, pp. 195–205, 2008.
- [9] J. Wolf, "Born-again group testing: multiaccess communications," *IEEE Transactions on Information Theory*, vol. 31, pp. 185–191, 1985.
- [10] E. Hong and R. Ladner, "Group testing for image compression," in *Data Compression Conference*, 2000, pp. 3–12.
- [11] R. Clifford, K. Efremenko, E. Porat, and A. Rothschild, " $k$ -mismatch with don't cares," in *Proceedings of ESA*, ser. LNCS, vol. 4698, 2007, pp. 151–162.
- [12] G. Cormode and S. Muthukrishnan, "What's hot and what's not: tracking most frequent items dynamically," *ACM Trans. on Database Syst.*, vol. 30, no. 1, pp. 249–278, 2005.
- [13] A. Blass and Y. Gurevich, "Pairwise testing," *Bulletin of the EATCS*, vol. 78, pp. 100–132, 2002.
- [14] G. Cormode and S. Muthukrishnan, "Combinatorial algorithms for compressed sensing," in *Proceedings of Information Sciences and Systems*, 2006, pp. 198–201.
- [15] D.-Z. Du and F. Hwang, *Combinatorial Group Testing and its Applications*, 2nd ed. World Scientific, 2000.
- [16] —, *Pooling Designs and Nonadaptive Group Testing*. World Scientific, 2006.
- [17] E. Candes, J. Romberg, and T. Tao, "Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inform. Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006.
- [18] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inform. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.
- [19] G. Atia and V. Saligrama, "Boolean compressed sensing and noisy group testing," 2009, preprint (arXiv: cs.IT/0907.1061).
- [20] A. D'yachkov and V. Rykov, "Bounds of the length of disjunct codes," *Problems of Control and Information Theory*, vol. 11, pp. 7–13, 1982.
- [21] Ruzinkó, "On the upper bound of the size of the  $r$ -cover-free families," *J. Comb. Theory., Series A*, vol. 66, pp. 302–310, 1994.
- [22] Z. Füredi, "On  $r$ -cover-free families," *J. Comb. Theory, Series A*, vol. 73, pp. 172–173, 1996.
- [23] W. Kautz and R. Singleton, "Nonrandom binary superimposed codes," *IEEE Transactions on Information Theory*, vol. 10, pp. 363–377, 1964.