Improved Bounds For Efficiently Decodable Probabilistic Group Testing With Unreliable Items

Sarthak Jain, Martina Cardone, Soheil Mohajer
University of Minnesota, Minneapolis, MN 55455, USA, Email: {jain0122, mcardone, soheil}@umn.edu

2023 IEEE Information Theory Workshop (ITW) | 979-8-3503-0149-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/ITW55543.2023.10161616

***Abstract*—This work uses non-adaptive probabilistic group testing to find a set of L defective items out of n items. In contrast to traditional group testing, in the considered setup each item can hide itself (or become inactive) during any given test with probability $1-\alpha$ and is active with probability α. The authors of [Cheraghchi et al.] proposed an efficiently decodable probabilistic group testing scheme which requires $O\left(\frac{L\log(n)}{\alpha^3}\right)$ tests for the per-instance scenario (where the group testing matrix works for any arbitrary, but fixed, set of L defective items) and $O\left(\frac{L^2\log(n/L)}{\alpha^3}\right)$ tests for the universal scenario (where the same group testing matrix works for all possible defective sets of L items). The contribution of this work is two-fold: (i) with a slight modification in the construction of the group testing matrix proposed by [Cheraghchi et al.], the corresponding bounds on the number of sufficient tests are improved to $O\left(\frac{L\log(n)}{\alpha^2}\right)$ and $O\left(\frac{L^2\log(n/L)}{\alpha^2}\right)$ for the per-instance and universal scenarios respectively, while still using their efficient decoding method; and (ii) it is shown that the same bounds also hold for the fixed pool-size probabilistic group testing scenario, where in every test a fixed number of items are included for testing.**

I. Introduction

Group testing, introduced back in 1943 by Dorfman [1], has numerous applications ranging from medicine [2] to computer science [3]. It is a well studied methodology used to identify L defective items among n items whenever $L << n$, by incorporating efficient strategies of testing *groups* of items at a time, instead of testing the items one by one. Two principal types of group testing methodologies, namely *probabilistic* and *combinatorial* group testing are prevalent in the literature. In combinatorial group testing, the goal is to identify the set of L defective items among n items for any size n with a zero probability of error [4]. In probabilistic group testing, the probability of error goes to 0 as $n \to \infty$ [5]–[8]. The results of a group test can be noisy in mainly two ways: (i) *unreliable items*, i.e., among the items that are selected in any given test, some items can choose to hide themselves (or become inactive) with a certain probability [9]–[12] and hence, they do not contribute in that specific test; or (ii) *noisy/unreliable tests*, i.e., the result can itself be noisy, that is, the test results can themselves get flipped [5], [13], [14].

In this work, we focus on the first case of unreliable items, where in any given test, each item selected for that test is active with probability α and inactive with probability $1-\alpha$. This effect is often referred to as the dilution effect in the literature [9], [10], because of its relevance in biological experiments and viral epidemics [15]. Note that in this work we focus on noise-level-independent test design, where the test matrix is constructed independently of α. In [12], the authors derived an achievable bound on the number of tests required for identifying the L defective items. Their scheme is based on a noise-level-independent Bernoulli test design and requires $O\left(\frac{L\log(n)}{\alpha^2}\right)$ tests for the per-instance scenario (where the designed group testing matrix works for any arbitrary, but fixed, set of L defective items) and $O\left(\frac{L^2\log(n/L)}{\alpha^2}\right)$ tests for the universal scenario (where the designed group testing matrix works for all possible defective sets of L items). However, their scheme uses maximum likelihood decoding and checks all the $\binom{n}{L}$ sets to see which set is most likely to be defective, making the decoding computationally expensive. The authors of [10] proposed an efficiently decodable group testing scheme based on a distance decoder. However, this scheme requires more tests, namely, $O\left(\frac{L\log(n)}{\alpha^3}\right)$ for the per-instance scenario and $O\left(\frac{L^2\log(n/L)}{\alpha^3}\right)$ for the universal scenario.

In this work, we still use the efficient distance decoder of [10] (but with more fine-tuned parameters), and improve the bound on the number of tests to $O\left(\frac{L\log(n)}{\alpha^2}\right)$ for the per-instance scenario and to $O\left(\frac{L^2\log(n/L)}{\alpha^2}\right)$ for the universal scenario, by making a few modifications in the group testing scheme of [10]. Thus, our constructions are both computationally efficient and achieve the same bounds as the maximum likelihood-based group testing scheme of [12]. Note that our test matrix construction is noise-level-independent, which makes our scheme more robust to errors in the estimate of α. With a noise-level-dependent test construction, the achievable bounds can be even further improved [16], [17]. Furthermore, in both [10] and [12], there is no restriction on the pool-size. For the noiseless case, group testing with pool-size constraints has been well studied [18]–[20]. In this work, we study a specific case of pool-size-constrained group testing for the dilution model. In particular, we extend our achievability result to the fixed pool-size case, and we show that our achievable bounds also hold when we want the same number of items to be selected for testing in each test. The fixed pool size case is practically relevant, for instance, in distributed computing settings where the server, in order to retrieve the result of a computation, might need to aggregate the results received from a fixed size of worker nodes [3], [21].

This research was supported in part by the U.S. National Science Foundation under Grant CCF-1907785. The authors would also like to thank Dr. M. Cheraghchi for discussing the improved bound, and his encouragement to submit this work.

Paper Organization. In Section II, we present the group testing framework under consideration. In Section III, we present the achievable bounds for probabilistic group testing with unreliable items where there is no restriction on the pool-size. Finally, in Section IV, we show that the same bounds also hold true for the fixed pool-size case.

Notation. We use uppercase calligraphic letters to represent sets (example: $\mathcal{L}$), capital serif letters to represent matrices (example: M), lowercase boldface letters to represent vectors (example: $\boldsymbol{x}$), and uppercase boldface letters to denote random vectors (example: $\boldsymbol{X}$). For a matrix M, we use $\mathsf{M}_{i,:}$ and $\mathsf{M}_{:,j}$ to represent its ith row and jth column, respectively. Moreover, for sets $\mathcal{A}$ and $\mathcal{B}$, $\mathsf{M}_{\mathcal{A},\mathcal{B}}$ is the submatrix of M where only the rows in $\mathcal{A}$ and the columns in $\mathcal{B}$ are retained. For a vector $\boldsymbol{z}$, we define $\mathsf{supp}(\boldsymbol{z}) \triangleq \{i : \boldsymbol{z}_i \neq 0\}$ and $\|\boldsymbol{z}\|_0 = |\mathsf{supp}(\boldsymbol{z})|$. For a positive integer n, we define $[n] \triangleq \{1, 2, \ldots, n\}$.

II. System Model

We consider n items labeled by integers in $[n]$. A subset $\mathcal{L} \subseteq [n]$ with $|\mathcal{L}| = L$ of the items are defective, while the remaining $[n] \setminus \mathcal{L}$ items are non-defective. Our goal is to identify the set $\mathcal{L}$. To this end, we can perform tests on individual, or groups of, items and tell whether the tested group includes any defective item or not. However, the defective items are not reliable and may be active (and act as a defective item) or inactive (and act as a non-defective item) in each test. Hence, our tests are noisy, and their results will be negative, if all the defective items in the selected group are inactive. Equivalently, a test result will be positive if and only if there is at least one active defective item in the tested group $\mathcal{G}$. The probability of a defective item being active in a test is α, independent of the other defective items and other tests.

Let M be the total number of tests. We represent these tests using a *contact* matrix $\mathsf{M}^{(c)} \in \{0,1\}^{M \times n}$, where $\mathsf{M}^{(c)}_{i,j} = 1$ if and only if the jth item is included in the ith test. Moreover, we use a vector $\boldsymbol{x} \in \{0,1\}^{n \times 1}$, to indicate whether or not each item is defective. More precisely, $\boldsymbol{x}_j = 1$ if and only if the jth item is defective. In an ideal setting (where all the defective items are always active, i.e., $\alpha = 1$), the result of the tests can be represented by a vector $\boldsymbol{y}^{(c)} \in \{0,1\}^M$ as

$$\boldsymbol{y}^{(c)} = \mathsf{M}^{(c)} \odot \boldsymbol{x}, \tag{1}$$

where the multiplication and addition are logical `and` and `or`, respectively. More precisely, we have $\boldsymbol{y}^{(c)}_i = \bigvee_{j=1}^{n} (\mathsf{M}^{(c)}_{i,j} \wedge \boldsymbol{x}_j)$.

To account for the unreliable behavior of the defective items, we adopt the notation used in [10], and define a *sampling* matrix $\mathsf{M}^{(s)}$ which is obtained from the contact matrix $\mathsf{M}^{(c)}$ as follows. Each non-zero entry of $\mathsf{M}^{(c)}$ is flipped with probability $1-\alpha$. In other words, each entry $\mathsf{M}^{(c)}_{i,j}$ with $i \in [M]$ and $j \in [n]$ is passed through a Z-channel, and we have

$$\mathsf{M}^{(s)}_{i,j} = \begin{cases} 0 & \text{if } \mathsf{M}^{(c)}_{i,j} = 0, \\ 1 \text{ w.p. } \alpha & \text{if } \mathsf{M}^{(c)}_{i,j} = 1, \\ 0 \text{ w.p. } 1-\alpha & \text{if } \mathsf{M}^{(c)}_{i,j} = 1. \end{cases} \tag{2}$$

Note that if $\alpha = 1$, the problem reduces to the classical group testing problem [1]. So, we focus on a range of α that is bounded away from 1, i.e., upper bounded by any constant less than 1, and without loss of generality, we assume $\alpha \leq \frac{1}{2}$. The result of the actual tests (in the presence of unreliable defective items) can be represented as

$$\boldsymbol{y} = \mathsf{M}^{(s)} \odot \boldsymbol{x}. \tag{3}$$

Assume that item j is selected to be included in test i. Then, we have $\mathsf{M}^{(c)}_{i,j} = 1$. Now, assume that item j is defective ($\boldsymbol{x}_j = 1$), but inactive in test i. Since $\mathsf{M}^{(c)}_{i,j} = 1$, we have $\boldsymbol{y}^{(c)}_i = 1$. However, the inactive behavior of item j in test i is captured by flipping $\mathsf{M}^{(c)}_{i,j}$ and setting $\mathsf{M}^{(s)}_{i,j} = 0$. In this case, the inactive defective item j does not lead to $\boldsymbol{y}_i = 1$.

An example of a contact matrix and a sampling matrix for $n = 5$ items and $M = 3$ tests is given as

$$\mathsf{M}^{(c)} = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \end{bmatrix}, \ \mathsf{M}^{(s)} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}. \tag{4}$$

In this example, the third test consists of testing items $\{2, 4\}$. However, $\mathsf{M}^{(c)}_{3,2} = 1$ and $\mathsf{M}^{(s)}_{3,2} = 0$ imply that item 2 was selected in test 3, but was inactive. Note that we do *not* have access to $\mathsf{M}^{(s)}$, and $\mathsf{M}^{(s)}$ is only used to model the random behavior of the defective items.

Remark 1. *Note that the tests are always governed by the contact matrix $\mathsf{M}^{(c)}$, but due to the unreliable items, the test results are given by $\boldsymbol{y}$ in* (3) *(rather than $\boldsymbol{y}^{(c)}$ in* (1)*).*

The goal of our work is to design $\mathsf{M}^{(c)}$ with as few rows as possible, such that, using $\mathsf{M}^{(c)}$ and $\boldsymbol{y}$, we can identify, with an arbitrarily small error probability (as $n \to \infty$), the set $\mathcal{L}$ of defective items. In particular, we are interested in characterizing how the number of tests M should scale with respect to the underlying parameters, n, L and α, to achieve a vanishing error probability. As discussed in [10], there are usually two scenarios of interest: (i) the per-instance scenario; and (ii) the universal scenario. These are defined as follows,

1) **Per-instance scenario:** When the contact matrix $\mathsf{M}^{(c)}$ works for any arbitrary, yet fixed, subset $\mathcal{L} \subseteq [n]$ of size L, but the same $\mathsf{M}^{(c)}$ may not work for all other subsets $\mathcal{L}' \neq \mathcal{L}$ of size L.
2) **Universal scenario:** When the same contact matrix $\mathsf{M}^{(c)}$ works for all possible subsets $\mathcal{L} \subseteq [n]$ of size L.

In this paper, we assume $L = O(n^\zeta)$ for $0 \leq \zeta < 1$ (i.e., sparse regime), because if $L = O(n)$ (i.e., linear regime), then it has been shown that the number of tests required is $\Omega(n)$, which is achievable by testing the items individually [22], [5].

III. Probabilistic Group Testing: Improving the achievable bound

In [10], the authors proposed a probabilistic method to construct the contact matrix $\mathsf{M}^{(c)}$ with $O\left(\frac{L \log(n)}{\alpha^3}\right)$ and $O\left(\frac{L^2 \log(n/L)}{\alpha^3}\right)$ rows for the per-instance and universal scenarios, respectively, to identify the L defective items, with vanishing error probability (as $n \to \infty$). In this section, we

improve their bounds by a factor of α by slightly modifying their proof and show that $O\left(\frac{L\log(n)}{\alpha^2}\right)$ and $O\left(\frac{L^2\log(n/L)}{\alpha^2}\right)$ tests suffice for the per-instance and universal cases, respectively.

Before presenting the construction of $\mathsf{M}^{(c)}$, we describe the decoding rule that is used to identify the defective items from the test results, or equivalently, to find $\hat{\boldsymbol{x}}$, an estimate of $\boldsymbol{x}$, from $\mathsf{M}^{(c)}$ and $\boldsymbol{y}$.

The e-Distance Decoder: Given a contact matrix $\mathsf{M}^{(c)}$ and test results $\boldsymbol{y}$, an item j will be marked as defective (i.e., $\hat{\boldsymbol{x}}_j = 1$) if and only if the jth column $\mathsf{M}^{(c)}_{:,j}$ satisfies

$$\left|\mathsf{supp}\left(\mathsf{M}^{(c)}_{:,j}\right) \setminus \mathsf{supp}(\boldsymbol{y})\right| \leq e. \tag{5}$$

Lemma 1 below presents sufficient conditions for the matrices $\mathsf{M}^{(c)}$ and $\mathsf{M}^{(s)}$ to guarantee a correct reconstruction of $\boldsymbol{x}$.

Lemma 1. *If for some arbitrary parameter e, the contact matrix $\mathsf{M}^{(c)}$ and the sampling matrix $\mathsf{M}^{(s)}$ are such that:*

(1) In each column $j \in \mathcal{L}$, there are no more than e flips from $\mathsf{M}^{(c)}$ to $\mathsf{M}^{(s)}$;

(2) For every column $j \in [n]\setminus\mathcal{L}$, there exist at least $e+1$ rows $\mathcal{X} = \{i_1, \ldots, i_{e+1}\}$ such that $\mathsf{M}^{(c)}_{\mathcal{X},j} = 1$ and $\mathsf{M}^{(s)}_{\mathcal{X},\mathcal{L}} = 0$,

then the e-distance decoder correctly identifies all the defective items in $\mathcal{L}$.

Proof of Lemma 1. We prove that the condition in (5) holds for every $j \in \mathcal{L}$, and is violated for every $j \in [n] \setminus \mathcal{L}$. First, consider some $j \in \mathcal{L}$ (i.e., $\boldsymbol{x}_j = 1$). Since $\boldsymbol{y}_i = 0$ occurs only if $\mathsf{M}^{(s)}_{i,\mathcal{L}} = 0$, and in particular, $\mathsf{M}^{(s)}_{i,j} = 0$, we can write

$$\begin{aligned}\left|\mathsf{supp}\big(\mathsf{M}^{(c)}_{:,j}\big) \setminus \mathsf{supp}(\boldsymbol{y})\right| &= \left|\{i : \mathsf{M}^{(c)}_{i,j} = 1, \boldsymbol{y}_i = 0\}\right| \\ &\overset{(a)}{=} \left|\{i : \mathsf{M}^{(c)}_{i,j} = 1, \boldsymbol{y}_i = 0, \mathsf{M}^{(s)}_{i,j} = 0\}\right| \\ &\overset{(b)}{\leq} \left|\{i : \mathsf{M}^{(c)}_{i,j} = 1, \mathsf{M}^{(s)}_{i,j} = 0\}\right| \overset{(c)}{\leq} e.\end{aligned}$$

where (a) holds since for $j \in \mathcal{L}$ we have that $\boldsymbol{y}_i = 0$ only if $\mathsf{M}^{(s)}_{i,j} = 0$, (b) holds since the set in the left-hand side is a subset of the one in the right-hand side, and finally (c) follows from Condition (1) in Lemma 1.

Next, consider some $j \in [n] \setminus \mathcal{L}$. The fact that $\mathsf{M}^{(s)}_{\mathcal{X},\mathcal{L}} = 0$ implies that $\boldsymbol{y}_{\mathcal{X}} = 0$. Then, we have that

$$\left|\mathsf{supp}\big(\mathsf{M}^{(c)}_{:,j}\big) \setminus \mathsf{supp}(\boldsymbol{y})\right| = \left|\{i : \mathsf{M}^{(c)}_{i,j} = 1, \boldsymbol{y}_i = 0\}\right| \geq |\mathcal{X}| = e+1,$$

and thus, the condition in (5) is violated. This completes the proof of Lemma 1. □

Remark 2. *Lemma 1 follows from Proposition 3 in [10], with a minor but important modification. The second condition in [10, Proposition 3] requires the existence of a set $\mathcal{X}$ of rows with $|\mathcal{X}| = e + 1$ such that $\mathsf{M}^{(c)}_{\mathcal{X},j} = 1$, $j \in [n] \setminus \mathcal{L}$, and $\mathsf{M}^{(c)}_{\mathcal{X},\mathcal{L}} = 0$, instead of $\mathsf{M}^{(s)}_{\mathcal{X},\mathcal{L}} = 0$. Note that the test result vector is given by $\boldsymbol{y} = \mathsf{M}^{(s)} \odot \boldsymbol{x}$, and it is $\mathsf{M}^{(s)}$ that determines whether the inequality in* (5) *holds or not. Moreover, the Z-channel from $\mathsf{M}^{(c)}$ to $\mathsf{M}^{(s)}$ allows for $\mathsf{M}^{(c)}_{x,\ell} = 1$ but $\mathsf{M}^{(s)}_{x,\ell} = 0$ for some $x \in \mathcal{X}$ and $\ell \in \mathcal{L}$. Therefore, the condition of [10] is more restrictive that ours. By relaxing this condition, we improve the achievable bound by a factor of α.*

A. Construction of the Contact Matrix $\mathsf{M}^{(c)}$

We consider the following construction for the contact and sampling matrices. The constructions and the error probability analysis follow similar lines as [10], with a few modifications, which are highlighted in the following.

Probabilistic construction for $\mathsf{M}^{(c)}$ and $\mathsf{M}^{(s)}$: *We generate $\mathsf{M}^{(c)}$ randomly, where each entry is drawn from a Bernoulli distribution with parameter $q = \frac{\theta}{L}$ (where the parameter θ will be determined later), independent of all other entries. We also generate the sampling matrix $\mathsf{M}^{(s)}$ from $\mathsf{M}^{(c)}$ by passing each entry through the Z-channel described in* (2)*.*

The following theorem shows that the construction above can decode the vector $\boldsymbol{x}$ with an overwhelming probability.

Theorem 1 (The per-instance scenario)**.** *For every arbitrary, but a priori fixed, L-sparse vector $\boldsymbol{x}$, the contact and sampling matrices $\mathsf{M}^{(c)}$ and $\mathsf{M}^{(s)}$ with $M = O(L\log(n)/\alpha^2)$ rows generated by the probabilistic construction above, can decode the vector $\boldsymbol{x}$ using the distance decoder with a proper parameter, and the probability of error $P_e = \mathbb{P}[\hat{\boldsymbol{X}} \neq \boldsymbol{x} | \boldsymbol{X} = \boldsymbol{x}]$ goes to 0 as $n \to \infty$.*

As a consequence of Theorem 1, the following corollary presents a similar result for the universal scenario.

Corollary 1 (The universal scenario)**.** *Using the contact and sampling matrices $\mathsf{M}^{(c)}$ and $\mathsf{M}^{(s)}$ generated by the probabilistic construction above with $M = O(L^2\log(n/L)/\alpha^2)$ rows, and the distance decoder with a proper parameter, we get $P_e = \mathbb{P}[\hat{\boldsymbol{X}} \neq \boldsymbol{x} | \boldsymbol{X} = \boldsymbol{x}] \to 0$ as $n \to \infty$, for every possible L-sparse vector $\boldsymbol{x}$.*

Before providing the proof of Theorem 1, we present two remarks regarding our scheme's robustness and related works.

Remark 3. *Our proposed scheme is robust to over-estimating L. In other words, if our estimate $\hat{L}$ is such that $\hat{L} \geq L$, then the number of tests in Theorem 1 and Corollary 1 still hold by replacing L with $\hat{L}$, i.e., by using the exact same scheme (but with the parameters now designed according to $\hat{L}$), we still achieve vanishing error probabilities.*

Remark 4. *In our design, the construction of $\mathsf{M}^{(c)}$ is noise-level-independent (that is, independent of α). With a noise-level-dependent design, the bound in Theorem 1 can be further improved. In particular, by choosing $q = \frac{\log(2)}{L\alpha}$, the authors in [16] showed an achievable bound of $M = O(L\log(n)/\alpha)$ tests for the per-instance scenario. However, their scheme would not work when $\alpha \leq \frac{\log(2)}{L}$ because $q \geq 1$ in this regime. The authors in [17] used a heavy machinery to provide achievable bounds for a much general noise model, of which the dilution model is a special case. For the dilution model, their bound is tighter than $M = O(L\log(n)/\alpha^2)$ for all $\alpha \in (0, 1)$. However their scheme has a few shortcomings compared to our work. In particular: (i) their test matrix is noise-level-dependent, (ii) errors and bounds are sensitive to small perturbations in the value of L and require an exact estimate of L (whereas in most practical scenarios only an*

upper-bound for L is known), and (iii) constants associated with the big-O notation are very high (of the order of 10^5).

Proof of Theorem 1. Our goal is to show that the randomly generated matrices $\mathsf{M}^{(c)}$ and $\mathsf{M}^{(s)}$ satisfy, with high probability, the conditions of Lemma 1 for

$$e \triangleq (1+\delta)\mu_1 = (1+\delta)q(1-\alpha)M, \tag{6}$$

where δ is a parameter determined later. We denote by $P_{e,1}$ and $P_{e,2}$ the probability that Conditions (1) and (2) of Lemma 1 are violated, respectively. We start by analyzing $P_{e,1}$. Let N_j denote the number of flips in column j. More specifically, for each column $j \in \mathcal{L}$, we define $P_{e,1}^{(j)} = \mathbb{P}[N_j > e]$ to be the probability that column j has more than e flips from $\mathsf{M}^{(c)}$ to $\mathsf{M}^{(s)}$. From the random generation of $\mathsf{M}^{(c)}$ and $\mathsf{M}^{(s)}$, we have

$$\begin{aligned}\mathbb{P}[\mathsf{M}_{i,j}^{(c)}=1,\mathsf{M}_{i,j}^{(s)}=0] &= \mathbb{P}[\mathsf{M}_{i,j}^{(c)}=1]\cdot\mathbb{P}[\mathsf{M}_{i,j}^{(s)}=0 \mid \mathsf{M}_{i,j}^{(c)}=1]\\ &= q(1-\alpha),\end{aligned}$$

and hence we get $\mu_1 \triangleq \mathbb{E}[N_j] = q(1-\alpha)M$. Then, by using the Chernoff bound, we get

$$\begin{aligned}P_{e,1}^{(j)} &= \mathbb{P}[N_j > e] = \mathbb{P}[N_j > (1+\delta)\mu_1]\\ &\le \exp\left(-\frac{\delta^2\mu_1}{2+\delta}\right) = \exp\left(-\frac{\delta^2 q(1-\alpha)M}{2+\delta}\right).\end{aligned} \tag{7}$$

Therefore, using the union bound, we arrive at

$$P_{e,1} \le \sum_{j\in\mathcal{L}} P_{e,1}^{(j)} \le L\exp\left(-\frac{\delta^2 q(1-\alpha)M}{2+\delta}\right). \tag{8}$$

Next, we bound $P_{e,2}$. For every $j \in [n]\setminus\mathcal{L}$, we define R_j as the number of tests that include item j, but all the defective items are either not selected or are inactive, that is,

$$R_j = \left|\left\{\, i : \mathsf{M}_{i,j}^{(c)}=1, \mathsf{M}_{i,\mathcal{L}}^{(s)}=0 \,\right\}\right|. \tag{9}$$

Recall that Condition (2) in Lemma 1 requires that $R_j \ge e+1$, for every $j \in [n]\setminus\mathcal{L}$. We define $P_{e,2}^{(j)} = \mathbb{P}[R_j \le e]$. Since the entries of $\mathsf{M}^{(c)}$ and $\mathsf{M}^{(s)}$ are generated independently, we have

$$\begin{aligned}\mathbb{P}[\mathsf{M}_{i,j}^{(c)}=1,\mathsf{M}_{i,\mathcal{L}}^{(s)}=0] &= \mathbb{P}[\mathsf{M}_{i,j}^{(c)}=1]\cdot\mathbb{P}[\mathsf{M}_{i,\mathcal{L}}^{(s)}=0 \mid \mathsf{M}_{i,j}^{(c)}=1]\\ &= \mathbb{P}[\mathsf{M}_{i,j}^{(c)}=1]\cdot\mathbb{P}[\mathsf{M}_{i,\mathcal{L}}^{(s)}=0]\\ &= \mathbb{P}[\mathsf{M}_{i,j}^{(c)}=1]\cdot\prod\nolimits_{j\in\mathcal{L}}\mathbb{P}[\mathsf{M}_{i,j}^{(s)}=0],\end{aligned} \tag{10}$$

where the second equality follows from independence since $j \notin \mathcal{L}$. Now, note that

$$\begin{aligned}\mathbb{P}[\mathsf{M}_{i,j}^{(s)}=0] &= \mathbb{P}[\mathsf{M}_{i,j}^{(s)}=0,\mathsf{M}_{i,j}^{(c)}=0] + \mathbb{P}[\mathsf{M}_{i,j}^{(s)}=0,\mathsf{M}_{i,j}^{(c)}=1]\\ &= \mathbb{P}[\mathsf{M}_{i,j}^{(c)}=0] + \mathbb{P}[\mathsf{M}_{i,j}^{(c)}=1]\cdot\mathbb{P}[\mathsf{M}_{i,j}^{(s)}=0|\mathsf{M}_{i,j}^{(c)}=1]\\ &= (1-q)+q(1-\alpha) = 1-q\alpha.\end{aligned} \tag{11}$$

Plugging (11) into (10), we arrive at

$$\mathbb{P}[\mathsf{M}_{i,j}^{(c)}=1,\mathsf{M}_{i,\mathcal{L}}^{(s)}=0] = q(1-q\alpha)^L.$$

Hence, we have that

$$\mu_2 \triangleq \mathbb{E}[R_j] = (1-q\alpha)^L qM \ge (1-\theta\alpha)qM, \tag{12}$$

where the inequality follows from the Bernoulli's inequality $(1-q\alpha)^L \ge 1-Lq\alpha$ and the fact that $q = \frac{\theta}{L}$. Similar to the previous case, the error probability $P_{e,2}^{(j)} := \mathbb{P}[R_j \le e]$ can be bounded by the Chernoff bound. However, in order to use the Chernoff bound for $P_{e,2}^{(j)}$, we should have $e < \mu_2$ (where e is defined in (6)). If we can choose the parameters (δ,θ) such that $\beta \triangleq \frac{\mu_2 - e}{\mu_2} > 0$, then we have that

$$\begin{aligned}P_{e,2}^{(j)} &= \mathbb{P}[R_j \le e] = \mathbb{P}[R_j \le (1-\beta)\mu_2]\\ &\overset{(a)}{\le} \exp\left(-\frac{\beta^2\mu_2}{2}\right) = \exp\left(-\frac{(\mu_2-e)^2}{2\mu_2}\right)\\ &\overset{(b)}{\le} \exp\left(-\frac{((1-\theta\alpha)qM-e)^2}{2(1-\theta\alpha)qM}\right) \overset{(c)}{\le} \exp\left(-\frac{M}{L}\gamma\right),\end{aligned}$$

where (a) follows from the Chernoff bound, (b) holds since $f(x) = \exp\left(-\frac{(x-e)^2}{2x}\right)$ is a decreasing function of x and $\mu_2 \ge (1-\theta\alpha)\,qM$ from (12), and finally (c) holds for $\gamma \triangleq \frac{\theta}{2}\left((1-\theta\alpha)-(1+\delta)(1-\alpha)\right)^2$ because $q = \frac{\theta}{L}$ and $1-\theta\alpha \le 1$. Finally, using the union bound, we get

$$P_{e,2} \le \sum_{j\in[n]\setminus\mathcal{L}} P_{e,2}^{(j)} \le n\exp\left(-\frac{M}{L}\gamma\right). \tag{13}$$

Now, we set[1] $\theta = \frac{1}{4}$ and $\delta = \frac{\alpha}{2}$. Then, from (12), we obtain

$$\begin{aligned}&\mu_2 - e \ge (1-\theta\alpha)qM - e\\ &= \left(\left(1-\frac{\alpha}{4}\right)-\left(1+\frac{\alpha}{2}\right)(1-\alpha)\right)qM = \left(\frac{\alpha}{4}+\frac{\alpha^2}{2}\right)qM > 0,\end{aligned} \tag{14}$$

and thus, we have $\beta > 0$, and the Chernoff bound is valid.

Let us analyze the two bounds on the error probabilities $P_{e,1}$ in (8) and on $P_{e,2}$ in (13), for $M = \frac{cL\log(n)}{\alpha^2}$, where $c > 128$ is a constant. First, note that since $\alpha \le \frac{1}{2}$, we have that

$$\frac{\delta^2\theta(1-\alpha)}{(2+\delta)} \ge \frac{\frac{\alpha^2}{4}\cdot\frac{1}{4}\cdot\frac{1}{2}}{\frac{9}{4}} = \frac{\alpha^2}{72}.$$

Therefore, since $L \le n$ we get

$$P_{e,1} \le L\exp\left(-\frac{\delta^2(1-\alpha)}{2+\delta}\cdot\frac{\theta}{L}\cdot\frac{cL\log(n)}{\alpha^2}\right) \le n^{1-\frac{c}{72}}. \tag{15}$$

Also, similar to (14), we get $\gamma = \frac{1}{8}\left(\frac{\alpha}{4}+\frac{\alpha^2}{2}\right)^2 \ge \frac{\alpha^2}{128}$. Thus,

$$P_{e,2} \le n\exp\left(-\frac{cL\gamma\log(n)}{\alpha^2 L}\right) \le n^{1-\frac{c}{128}}. \tag{16}$$

From (15) and (16) it can be readily seen that, for $c > 128$, both $P_{e,1}$ and $P_{e,2}$ go to 0 as $n \to \infty$. In other words, the randomly generated matrices $\mathsf{M}^{(c)}$ and $\mathsf{M}^{(s)}$ will satisfy the conditions of Lemma 1, with high probability, and hence, the e-distance decoder can identify the vector $\boldsymbol{x}$, with a vanishing error probability. This concludes the proof of Theorem 1. □

Proof of Corollary 1. In Theorem 1, we showed that for $\theta = \frac{1}{4}$ and $\delta = \frac{\alpha}{2}$ we have

$$\mathbb{P}[\hat{\boldsymbol{X}} \ne \boldsymbol{x} | \boldsymbol{X} = \boldsymbol{x}] \le L\exp\left(-\frac{\alpha^2}{72L}M\right) + n\exp\left(-\frac{\alpha^2}{128L}M\right),$$

[1] In [10], the value of θ was set to $\theta = \frac{\alpha}{8}$.

for any fixed L-sparse vector $\boldsymbol{x}$. Now, for the universal scenario, we need to show that a common pair of $(\mathsf{M}^{(c)}, \mathsf{M}^{(s)})$ is able to decode all L-sparse vectors $\boldsymbol{x}$. Since there are $\binom{n}{L}$ of such vectors, using the union bound we obtain

$$\begin{aligned}
&\sum_{\substack{\boldsymbol{x}\in\{0,1\}^n\\ \|\boldsymbol{x}\|_0=L}} \mathbb{P}[\hat{\boldsymbol{X}}\neq\boldsymbol{x}|\boldsymbol{X}=\boldsymbol{x}]\\
&\le \binom{n}{L}\left[L\exp\left(-\frac{\alpha^2}{72L}M\right)+n\exp\left(-\frac{\alpha^2}{128L}M\right)\right]\\
&\overset{(a)}{\le} n\left(\frac{n\exp(1)}{L}\right)^L\left[\exp\left(-\frac{\alpha^2}{72L}M\right)+\exp\left(-\frac{\alpha^2}{128L}M\right)\right]\\
&\overset{(b)}{\le} 2n\left(\frac{n\exp(1)}{L}\right)^L\exp\left(-\frac{\alpha^2}{128L}\frac{cL^2\log(n/L)}{\alpha^2}\right)\\
&\overset{(c)}{\le} 2\left(\frac{n}{L}\right)^L\left(\frac{n\exp(1)}{L}\right)^L\left(\frac{n}{L}\right)^{-\frac{cL}{128}}=2\left(\frac{\left(\frac{n}{L}\right)^{\frac{c}{128}-2}}{\exp(1)}\right)^{-L}, \quad (17)
\end{aligned}$$

where in (a) we used $\binom{n}{L}\le(n\exp(1)/L)^L$, in (b) we upper bounded the first exponential term by the second one, and (c) holds since $f(x)=(n/x)^x-n\ge 0$ for $1\le x\le\frac{n}{e}$, and the fact that $L\le n/e$ (for sufficiently large n, since $L=o(n)$). Note that the right-hand side of (17) goes to 0 as $n\to\infty$, if $c>256$. This completes the proof of Corollary 1. □

IV. Probabilistic Group Testing: Tests With Fixed Number of Tested Items

Here, we seek to design a group testing matrix $\mathsf{M}^{(c)}$ such that: (i) each of its rows has the same number of ones, that is, $|\mathsf{supp}(\mathsf{M}^{(c)}_{i,:})| = |\mathsf{supp}(\mathsf{M}^{(c)}_{j,:})|$ for all $i,j\in[n]$, and (ii) it has the same (order of) number of rows as in Theorem 1 (per-instance scenario) and Corollary 1 (universal scenario). The following construction guarantees the fixed group size property, and Theorem 2 shows that the error probability of identifying the vector $\boldsymbol{x}$ using the resulting matrices is vanishing.

Probabilistic construction for $\mathsf{M}^{(c)}$ and $\mathsf{M}^{(s)}$ with fixed-size groups: *Let t be the fixed group size, i.e., the number of ones in each row of $\mathsf{M}^{(c)}$, which will be determined later. Then, each row $\mathsf{M}^{(c)}_{i,:}$ of $\mathsf{M}^{(c)}$ is chosen uniformly at random and independent of the other rows, from the $\binom{n}{t}$ possible rows having exactly t ones. The sampling matrix $\mathsf{M}^{(s)}$ is then generated randomly, by applying the operation in* (2) *on each entry of $\mathsf{M}^{(c)}$, independent of the other entries.*

Theorem 2. *Assume the above construction for the contact and sampling matrices $\mathsf{M}^{(c)}$ and $\mathsf{M}^{(s)}$. Then, with a fixed pool size $t=\Theta\left(\frac{n}{L}\right)$, $M=O(L\log(n)/\alpha^2)$ rows for the per-instance scenario and $M=O(L^2\log(n/L)/\alpha^2)$ rows for the universal scenario suffice to decode the vector $\boldsymbol{x}$ using the distance decoder with a proper parameter, and the probability of error $P_e=\mathbb{P}[\hat{\boldsymbol{X}}\neq\boldsymbol{x}|\boldsymbol{X}=\boldsymbol{x}]$ goes to 0 as $n\to\infty$.*

Proof of Theorem 2. The core of the proof is similar to that of Theorem 1. However, due to the construction used here, the entries of $\mathsf{M}^{(c)}$ in each row are *not* independent anymore. Let $q\triangleq\frac{\theta}{L}$ and set $t=nq=\frac{n\theta}{L}$, where θ is a design parameter. We show that with proper choices of θ and e for the distance decoder, the conditions of Lemma 1 hold with overwhelming probability. First, for a column $j\in\mathcal{L}$, we have that

$$\begin{aligned}
\mathbb{P}[\mathsf{M}^{(c)}_{i,j}=1,\mathsf{M}^{(s)}_{i,j}=0]&=\mathbb{P}[\mathsf{M}^{(c)}_{i,j}=1]\cdot\mathbb{P}[\mathsf{M}^{(s)}_{i,j}=0\;\big|\;\mathsf{M}^{(c)}_{i,j}=1]\\
&=\frac{\binom{n-1}{t-1}}{\binom{n}{t}}(1-\alpha)=\frac{t}{n}(1-\alpha)=q(1-\alpha),
\end{aligned}$$

and hence, for $\mu_1=\mathbb{E}[|\{i:\mathsf{M}^{(c)}_{i,j}=1,\mathsf{M}^{(s)}_{i,j}=0\}|]$ we have

$$\mu_1=\sum_{i=1}^{M}\mathbb{P}[\mathsf{M}^{(c)}_{i,j}=1,\mathsf{M}^{(s)}_{i,j}=0]=Mq(1-\alpha).$$

Note that we have used linearity of the expectation to overcome the correlation between the columns of $\mathsf{M}^{(c)}$. Hence, the inequalities in (7) and (8) also hold for the fixed-group-size setting.

Next, for every $j\in[n]\setminus\mathcal{L}$, we define $\mu_2=\mathbb{E}[R_j]$, where $R_j=\big|\{i:\mathsf{M}^{(s)}_{i,\mathcal{L}}=0,\mathsf{M}^{(c)}_{i,j}=1\}\big|$. Then, we obtain

$$\begin{aligned}
&\mathbb{P}[\mathsf{M}^{(s)}_{i,\mathcal{L}}=0,\mathsf{M}^{(c)}_{i,j}=1]=\mathbb{P}[\mathsf{M}^{(c)}_{i,j}=1]\cdot\mathbb{P}[\mathsf{M}^{(s)}_{i,\mathcal{L}}=0|\mathsf{M}^{(c)}_{i,j}=1]\\
&=\mathbb{P}[\mathsf{M}^{(c)}_{i,j}=1]\sum_{\mathcal{J}\subseteq\mathcal{L}}\mathbb{P}[\mathsf{M}^{(s)}_{i,\mathcal{J}}=0,\mathsf{M}^{(c)}_{i,\mathcal{J}}=1,\mathsf{M}^{(c)}_{i,\mathcal{L}\setminus\mathcal{J}}=0|\mathsf{M}^{(c)}_{i,j}=1]\\
&=\mathbb{P}[\mathsf{M}^{(c)}_{i,j}=1]\sum_{\mathcal{J}\subseteq\mathcal{L}}\Big\{\mathbb{P}[\mathsf{M}^{(s)}_{i,\mathcal{J}}=0|\mathsf{M}^{(c)}_{i,\mathcal{J}}=1]\\
&\qquad\times\mathbb{P}[\mathsf{M}^{(c)}_{i,\mathcal{J}}=1,\mathsf{M}^{(c)}_{i,\mathcal{L}\setminus\mathcal{J}}=0|\mathsf{M}^{(c)}_{i,j}=1]\Big\}\\
&=\sum_{\ell=0}^{L}\sum_{\substack{\mathcal{J}\subset\mathcal{L}\\|\mathcal{J}|=\ell}}\Big\{\mathbb{P}[\mathsf{M}^{(s)}_{i,\mathcal{J}}=0|\mathsf{M}^{(c)}_{i,\mathcal{J}}=1]\\
&\qquad\times\mathbb{P}[\mathsf{M}^{(c)}_{i,\mathcal{J}\cup\{j\}}=1,\mathsf{M}^{(c)}_{i,\mathcal{L}\setminus\mathcal{J}}=0]\Big\}\\
&\overset{(a)}{=}\sum_{\ell=0}^{L}\binom{L}{\ell}(1-\alpha)^\ell\frac{\binom{n-L-1}{t-1-\ell}}{\binom{n}{t}}\overset{(b)}{\ge}\sum_{\ell=0}^{L}\binom{L}{\ell}(1-\ell\alpha)\frac{\binom{n-L-1}{t-1-\ell}}{\binom{n}{t}}\\
&=\frac{t}{n}\left(\sum_{\ell=0}^{L}\binom{L}{\ell}\frac{\binom{n-L-1}{t-1-\ell}}{\binom{n-1}{t-1}}-\sum_{\ell=0}^{L}\binom{L}{\ell}\ell\alpha\frac{\binom{n-L-1}{t-1-\ell}}{\binom{n-1}{t-1}}\right)\\
&=q\left(1-\alpha L\frac{t-1}{n-1}\sum_{\ell=1}^{L}\binom{L-1}{\ell-1}\frac{\binom{n-L-1}{t-1-\ell}}{\binom{n-2}{t-2}}\right)\\
&=q\left(1-\alpha L\frac{t-1}{n-1}\right)\overset{(c)}{\ge}q\left(1-\alpha L\frac{t}{n}\right)=q(1-\alpha\theta), \quad (18)
\end{aligned}$$

where (a) holds since: (i) there are ℓ (independent) flips in the columns in $\mathcal{J}$, and (ii) among the t ones in row i, $\ell+1$ of them should lie in the columns in $\mathcal{J}\cup\{j\}$, and the remaining $t-1-\ell$ ones can be in any position except those in $\mathcal{L}\cup\{j\}$; (b) follows from the Bernoulli's inequality; and (c) follows from the fact that $\frac{t-1}{n-1}\le\frac{t}{n}$ whenever $t\le n$. Using the bound in (18), we have that $\mu_2\ge Mq(1-\alpha\theta)$ which is the same bound as in (12). The rest of the proof is therefore the same as the proof of Theorem 1. Moreover, $\theta=\frac{1}{4}$ and $\delta=\frac{\alpha}{2}$ still work for this fixed-size case. This further implies that the random contact matrix $\mathsf{M}^{(c)}$ with $t=\frac{n\theta}{L}=\frac{n}{4L}$ ones in each row satisfies the conditions of Lemma 1, with high probability, and hence $\mathsf{M}^{(c)}$ together with a properly parameterized distance decoder, can estimate $\boldsymbol{x}$ with an overwhelming probability. The proof for the universal scenario follows similar lines as those of the per-instance scenario. This concludes the proof of Theorem 2. □

REFERENCES

[1] R. Dorfman, "The detection of defective members of large populations," *Ann. Math. Statist.*, vol. 14, no. 4, pp. 436–440, 12 1943.

[2] C. M. Verdun, T. Fuchs, P. Harar, D. Elbrächter, D. S. Fischer, J. Berner, P. Grohs, F. J. Theis, and F. Krahmer, "Group testing for sars-cov-2 allows for up to 10-fold efficiency increase across realistic scenarios and testing strategies," *Frontiers in Public Health*, vol. 9, 2021. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpubh.2021.583377

[3] A. Solanki, M. Cardone, and S. Mohajer, "Non-colluding attacks identification in distributed computing," in *2019 IEEE Information Theory Workshop (ITW)*, 2019, pp. 1–5.

[4] D. Z. Du and F. K. Hwang, "Combinatorial group testing and its applications," 1993.

[5] C. L. Chan, P. H. Che, S. Jaggi, and V. Saligrama, "Non-adaptive probabilistic group testing with noisy measurements: Near-optimal bounds with efficient algorithms," 2011. [Online]. Available: https://arxiv.org/abs/1107.4540

[6] A. Mazumdar, "Nonadaptive group testing with random set of defectives," *IEEE Transactions on Information Theory*, vol. 62, no. 12, pp. 7522–7531, 2016.

[7] A. Barg and A. Mazumdar, "Group testing schemes from codes and designs," *IEEE Transactions on Information Theory*, vol. PP, pp. 1–1, 08 2017.

[8] H. A. Inan, P. Kairouz, M. Wootters, and A. Ozgur, "On the optimality of the kautz-singleton construction in probabilistic group testing," in *2018 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 2018, pp. 188–195.

[9] F. K. Hwang, "Group testing with a dilution effect," *Biometrika*, vol. 63, no. 3, pp. 671–673, 1976. [Online]. Available: http://www.jstor.org/stable/2335750

[10] M. Cheraghchi, A. Hormati, A. Karbasi, and M. Vetterli, "Group testing with probabilistic tests: Theory, design and application," *IEEE Trans. on Inf. Theory*, vol. 57, no. 10, pp. 7057–7067, 2011.

[11] A. Mazumdar and S. Mohajer, "Group testing with unreliable elements," in *2014 52nd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 2014, pp. 1–3.

[12] G. K. Atia and V. Saligrama, "Boolean compressed sensing and noisy group testing," *IEEE Transactions on Information Theory*, vol. 58, pp. 1880–1901, 2009.

[13] J. Scarlett, "Noisy adaptive group testing: Bounds and algorithms," *IEEE Transactions on Information Theory*, vol. 65, 03 2018.

[14] S. Cai, M. Jahangoshahi, M. Bakshi, and S. Jaggi, "Efficient algorithms for noisy group testing," *IEEE Transactions on Information Theory*, vol. 63, no. 4, pp. 2113–2136, 2017.

[15] H. Jiang, H. Ahn, and X. Li, "Group testing with consideration of the dilution effect," *Mathematics*, vol. 10, no. 3, 2022. [Online]. Available: https://www.mdpi.com/2227-7390/10/3/497

[16] G. Arpino, N. Grometto, and A. S. Bandeira, "Group testing in the high dilution regime," in *2021 IEEE International Symposium on Information Theory (ISIT)*, 2021, pp. 1955–1960.

[17] X. Cheng, S. Jaggi, and Q. Zhou, "Generalized group testing," *IEEE Transactions on Information Theory*, vol. 69, no. 3, pp. 1413–1451, 2023.

[18] N. Tan and J. Scarlett, "Improved bounds and algorithms for sparsity-constrained group testing," *ArXiv*, vol. abs/2004.03119, 2020.

[19] V. Gandikota, E. Grigorescu, S. Jaggi, and S. Zhou, "Nearly optimal sparse group testing," *IEEE Transactions on Information Theory*, vol. 65, no. 5, pp. 2760–2773, 2019.

[20] O. Gebhard, M. Hahn-Klimroth, O. Parczyk, M. Penschuck, M. Rolvien, J. Scarlett, and N. Tan, "Near-optimal sparsity-constrained group testing: Improved bounds and algorithms," *IEEE Transactions on Information Theory*, vol. 68, no. 5, pp. 3253–3280, 2022.

[21] S. Hong, H. Yang, and J. Lee, "Hierarchical group testing for byzantine attack identification in distributed matrix multiplication," *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 3, pp. 1013–1029, 2022.

[22] L. Flodin and A. Mazumdar, "Probabilistic group testing with a linear number of tests," in *2021 IEEE International Symposium on Information Theory (ISIT)*, 2021, pp. 1248–1253.