Strongly Explicit and Efficiently Decodable Probabilistic Group Testing

Huseyin A. Inan and Ayfer Ozgur
Department of Electrical Engineering, Stanford University
Emails: {hinan1, aozgur}@stanford.edu

***Abstract*—We consider the non-adaptive probabilistic group testing problem where d random defective items are identified with high probability from a population of N items by applying t binary group tests. There has been recent progress towards developing explicit and efficiently decodable group testing schemes with $t = \Theta(d \log N)$ tests, which is known to be order-optimal for this setting when $d = O(N^{1-\alpha})$ for some constant $\alpha > 0$. In particular, a recent work develops an explicit scheme while another one develops an efficiently decodable scheme for this setting, both with the order-optimal $t = \Theta(d \log N)$ tests. However, to the best of our knowledge, there is no order-optimal scheme that is both explicit and efficiently decodable. In this paper, we close this gap by introducing the first (strongly) explicit and efficiently decodable construction that is order-optimal for the non-adaptive probabilistic group testing problem.**

I. Introduction

The objective of group testing is to identify a set of d (or less) "defective" items within a large population of N items. A group test is performed on a subset of the items $\{1, \dots, N\}$ and the result of the test is either positive, indicating that the group contains at least one defective item, or negative, indicating that all the items in the group are non-defective. The goal is to design a group testing scheme that contains as few tests t as possible in order to identify the defective items.

The original group testing framework was developed in 1943 by Robert Dorfman [1] to efficiently identify which WWII draftees were infected with syphilis without having to test them individually. In Dorfman's application, items represented draftees and tests represented actual blood tests. Over the years, group testing has found numerous applications in various fields spanning biology [2], medicine [3], machine learning [4], data analysis [5], computer science [6], wireless communication [7]–[11] and signal processing [12].

Group testing strategies can be *adaptive*, where tests are designed sequentially, i.e. the design of each new test depends on the outcomes of the previous tests or *non-adaptive*, where all tests are designed in one shot. A non-adaptive group testing strategy can be represented by a $t \times N$ binary matrix M, where $M_{ij} = 1$ indicates that item j participates in test i. We are interested in non-adaptive schemes in this work and suppress the word non-adaptive in the rest of the paper for brevity. Group testing schemes can also be *combinatorial* [13], [14] or *probabilistic* [15]–[23].

The goal of combinatorial group testing schemes is to recover any set of up to d defective items with zero-error. Combinatorial group testing schemes are known to require at least $t = \min\{N, \Omega(d^2 \log_d N)\}$ tests [24], [25]. Through random designs, one can prove the existence of schemes with $\min\{N, O(d^2 \log N)\}$ tests [14]. In [26], Kautz and Singleton provided a strongly explicit construction[1] that uses $t = \min\{N, O(d^2 \log_d^2 N)\}$ tests. This gap between explicit and randomized group testing constructions was closed in [27], which introduced an explicit construction with $t = \min\{N, O(d^2 \log N)\}$ tests. All these constructions had $O(tN)$ decoding complexity. Another main research line focused on developing low-complexity decoding schemes, particularly motivated by emerging applications involving massive datasets [28]–[30]. A group testing scheme is called efficiently decodable if the decoding rule can identify the defective set in poly(t) time complexity. [29] introduced a randomized construction with $t = O(d^2 \log N)$ and decoding complexity poly(t), and furthermore showed that their construction can be derandomized in the regime $d = O(\log N / \log \log N)$. Later, [30] removed the constraint on d and provided an explicit construction that can be decoded in time poly(t). This progress in combinatorial group testing lead to an explicit and efficiently decodable construction that has the same number of tests as the best known achievability result of $t = \min\{N, O(d^2 \log N)\}$.

In contrast, probabilistic group testing schemes assume a random defective set of size d and allow for an arbitrarily small probability of reconstruction error. It is known that probabilistic group testing schemes require $t = \Omega(d \log N)$ tests and optimal schemes with such minimal number tests exist when $d = O(N^{1-\alpha})$ for some constant $\alpha > 0$ [17]–[19]. However these schemes are randomized, i.e. their existence is established by a probabilistic argument. Recently, there has been progress towards developing explicit and efficiently decodable constructions for probabilistic group testing, analogous to the combinatorial case. [31] presented a strongly explicit scheme with $t = \Theta(d \log^2 N / \log d)$ tests, which is order-optimal when d is proportional to N^α for $\alpha \in (0, 1)$. More recently, [23] showed that Kautz and Singleton's strongly explicit scheme is order-optimal for probabilistic group testing, i.e. achieves $t = \Theta(d \log N)$ in the regime where $d = \Omega(\log^2 N)$. In terms of efficient decoding, [32], [33] and a related approach in [34] introduced randomized

[1] A $t \times N$ group testing matrix is called strongly explicit if any column of the matrix can be constructed in time poly(t). A matrix is called explicit if it can be constructed in time poly(t, N).

This work was supported in part by NSF award NeTS #1817205.

schemes that are efficiently decodable (poly(t) decoding complexity) and require $t = O(d \log d \log N)$ tests. Finally, [35] introduced a randomized construction that is order-optimal, i.e. it uses $t = \Theta(d \log N)$ tests, and at the same time efficiently decodable. However, to the best of our knowledge there is no order-optimal construction that is both explicit and efficiently decodable in the probabilistic group testing framework.

In this paper, we close this gap. We introduce a (strongly) explicit and efficiently decodable construction for probabilistic group testing with $t = \Theta(d \log N)$ tests. This can be seen as the counterpart of the progress made in combinatorial group testing. Indeed, the progression in the probabilistic case is more complete than the combinatorial setting on two fronts: (1) the number of tests $t = \Theta(d \log N)$ is optimal, i.e. matches the lower bound $t = \Omega(d \log N)$; (2) our construction is strongly explicit. Our result builds on ideas from [23] and [35]; we modify the Kautz and Singleton's strongly explicit scheme [26], in a manner similar to [23] but in a different parameter setting, to conform to the decoding method introduced in [35].

II. System Model and Basic Definitions

For any $t \times N$ matrix M, we use M_i to refer to its i'th column and M_{ij} to refer to its (i, j)'th entry. We denote the set of coordinates where M_i has nonzero entries by supp(M_i). For an integer $m \geq 1$, we denote the set $\{1, \ldots, m\}$ by $[m]$. The Hamming weight of a column of M will be simply referred to as the weight of the column.

In the probabilistic group testing setting, there is a random defective set S of size d among the items $[N]$. We define $\mathcal{S}$ as the set of all possible defective sets, i.e., the set of all $\binom{N}{d}$ subsets of $[N]$ with cardinality d, and let S be uniformly distributed over $\mathcal{S}$.[2] The goal is to determine S from the binary measurement vector Y of size t taking the form

$$Y = \bigvee_{i \in S} M_i, \tag{1}$$

where the $t \times N$ binary measurement matrix M indicates which items are included in the test, i.e., $M_{ij} = 1$ if the item j participates in test i. We focus on the noiseless model for the sake of brevity, however, our results also apply to the noisy setting with i.i.d. bit flips in the test measurements. In words, the measurement vector Y is the Boolean OR combination of the columns of the binary measurement matrix M corresponding to the defective items. Note that in this noiseless case, the randomness in the measurement vector Y is only due to the random choice of the defective set.

Given M and Y, a decoding procedure forms an estimate $\hat{S}$ of S. The performance metric we consider in this paper is the average probability of error for *exact recovery*, given by

$$P_e \triangleq \Pr(\hat{S} \neq S).$$

Our goal is to design a group testing scheme, i.e. the matrix M and a decoding strategy that outputs $\hat{S}$ given Y, such that $P_e \to 0$ as $N, d \to \infty$.

[2]This assumption is not critical. Our results carry over to the setting where the defective items are sampled with replacement.

III. Main Results

A. Overview

We first describe our construction coupled with the decoding strategy introduced in [35]. Our construction will be the column-wise concatenation of two binary measurement matrices M^1 and M^2, i.e. $M = \begin{bmatrix} M^1 \\ M^2 \end{bmatrix}$. We will employ the Kautz-Singleton construction [26] for generating M^1 and M^2 will be generated based on M^1 and the erasure correction code introduced in [36]. We point out that both M^1 and M^2 are constructed a priori as we focus on the non-adaptive case.

We first note that since $Y = \bigvee_{i \in S} M_i$, we have supp$(M_i) \subseteq$ supp(Y) for each $i \in S$. Therefore, a naive decoding rule would be to go through each column $j \in [N]$ of M and call item j defective if supp$(M_j) \subseteq$ supp(Y), and non-defective otherwise. If we can ensure that the event

$$\text{supp}(M_j) \not\subseteq \text{supp}(Y) = \cup_{i \in S} \text{supp}(M_i) \quad \forall j \in [N] \backslash S$$

occurs for a fraction of the S's approaching to 1, this decoding rule will output the defective set correctly with vanishing probability of error. However, the decoding complexity is $O(tN)$, which is not poly(t) in the sparse setting (e.g., $d = O(\text{poly}(\log N))$ and $t = \Theta(d \log N)$.

Taking this into account, we design M^1 as follows. We employ the Kautz-Singleton construction [26] (will be described shortly), with potentially much smaller number of columns compared to N. In particular, we will use the Kautz-Singleton construction M^{KS} of size $\Theta(d \log d) \times \Theta(d^3)$, which satisfies the event[3]

$$\text{supp}(M^{\text{KS}}_j) \not\subseteq \cup_{i \in S} \text{supp}(M^{\text{KS}}_i) \quad \forall j \in [\Theta(d^3)] \backslash S \tag{2}$$

with probability approaching to 1. We then define the binary measurement matrix M^1 of size $\Theta(d \log d) \times N$ by concatenating the Kautz-Singleton construction M^{KS} row-wise with itself such that we have N columns[4], i.e. $M^1 = \begin{bmatrix} M^{\text{KS}} & M^{\text{KS}} & \ldots & M^{\text{KS}} \end{bmatrix}$. Note that if $\Theta(d^3) \ll N$ (e.g. $d = O(\text{poly}(\log N))$), same columns repeat multiple times in M^1. This procedure is illustrated in Figure 1.

We define Y^1 as the measurement vector corresponding to the test matrix M^1, i.e., $Y^1 = \bigvee_{i \in S} M^1_i$. We will take Y^1 and apply the naive decoding rule described above to the original matrix M^{KS} resulting from the Kautz-Singleton construction. If all defective items have different columns from the Kautz-Singleton construction and the condition (2) is satisfied, this decoding rule will identify d columns of M^{KS} corresponding to defective items. We will show that the time complexity of this operation is $O(\Theta(\log d) \cdot \Theta(d^3)) = O(d^3 \cdot \log d)$, however, the issue is that the columns of M^{KS} are repeated multiple times in M^1, therefore the identified columns correspond to many items including non-defectives. We need to have an additional mechanism to figure out the defectives among them.

[3]Slightly overloading the notation, S is d random columns over $\Theta(d^3)$ columns here.

[4]For the sake of brevity, we assume $\Theta(d^3) | N$. Violation of this assumption only changes the constants to handle adding/removing dummy columns to get exactly N columns.

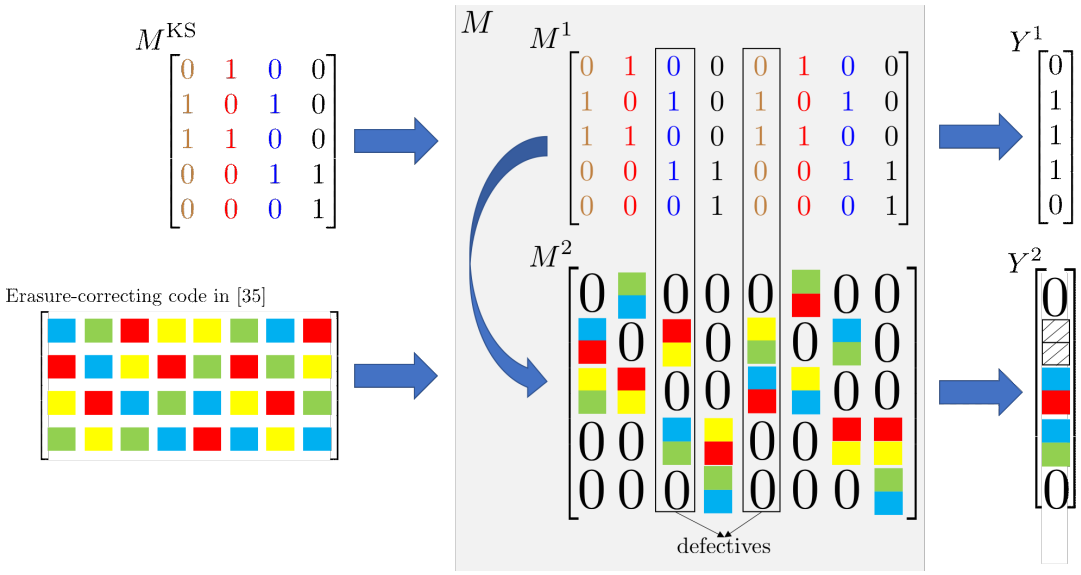

Fig. 1. An example illustrating our construction and the decoding procedure (best viewed in color). There are 8 items and two of them are defectives (item 3 and item 5). In this example, $M^1 = [M^{\text{KS}}\ M^{\text{KS}}]$ and we have $Y^1 = M_3^1 \bigvee M_5^1$. Applying the naive decoding rule over M^{KS} as explained in Section III outputs M_1^{KS} and M_3^{KS} correctly since $\text{supp}(M_2^{\text{KS}}) \not\subseteq Y^1$ and $\text{supp}(M_4^{\text{KS}}) \not\subseteq Y^1$. However, M_1^{KS} belongs to both M_1^1 and M_5^1 and M_3^{KS} belongs to both M_3^1 and M_7^1. Therefore, we need an additional step figuring out the defectives among them. This additional step is performed with the help of the second construction M^2. Each block in the erasure-correcting code represents a symbol from an alphabet $\mathcal{A}$ and each symbol has a binary representation of l bits. In this example, $w = 4$ and $w^{\text{KS}} = 2$, therefore, each entry that is equal to 1 in a column of M^1 is replaced by column-wise concatenated binary representations of $w/w^{\text{KS}} = 2$ symbols ($2 \cdot l$ bits) from the corresponding codeword in the erasure-correcting code. Each zero entry is simply replaced by $2 \cdot l$ zeros. For decoding, we know from the first step that the defective items have M_1^{KS} and M_3^{KS} in their columns. We look at the entries that are equal to 1 for M_1^{KS} and M_3^{KS}. We see that the first entries are intersecting and the second entries are not intersecting. Using Y^2, we see the codeword [Erasure, Erasure, Blue, Red] for M_1^{KS}, which gives us item 5 as defective from the erasure-correcting code. Similarly, using Y^2, we see the codeword [Erasure, Erasure, Blue, Green] for M_3^{KS}, which gives us item 3 as defective from the erasure-correcting code. We exploit the knowledge of our construction to figure out the erasure positions and clean positions for each column decoded in the first step.

To recap, the goal of the first construction M^1 is to figure out which columns of M^{KS} belong to the defectives, however, these columns also belong to some non-defectives in M^1. The second submatrix M^2 allows to resolve this ambiguity and figure out the defective items. Towards this goal, we will require one more condition to be satisfied by the Kautz-Singleton construction M^{KS}. We require that each column in M^1 has a constant weight w^{KS}, and that furthermore

$$|\text{supp}(M_j^{\text{KS}}) \setminus \cup_{i \in S \setminus \{j\}} \text{supp}(M_i^{\text{KS}})| \geq w^{\text{KS}}/2 \quad \forall j \in S, \quad (3)$$

with probability approaching to 1. In words, we ask that for any column in S, at least half of its support is not covered by the union of the support of the remaining columns in S.

In the second submatrix M^2, similar to [35], we utilize the following erasure-correcting code introduced in [36].

Lemma 1 ([36]: Thm. 1, [35]: Lemma 4)**.** *For any $r \in (0,1)$ and arbitrarily small $\epsilon > 0$, there exists an alphabet $\mathcal{A}$ whose size is a constant depending only on ϵ, and a codebook $\mathcal{C}$ (with codeword symbols on $\mathcal{A}$) and associated encoder/decoder pair, such that the following properties hold:*

- *$\mathcal{C}$ has rate r, i.e., the number of codewords is $|\mathcal{A}|^{wr}$, where w is the block length.*
- *The decoder corrects any (worst-case) fraction $1 - r - \epsilon$ of erasures.*
- *The encoding and decoding time are linear in the block length.*

We fix $r = 1/3$ and $\epsilon = 1/6$. The alphabet size is $|\mathcal{A}| = 2^l$ for some constant l. We have N items, therefore, the number of codewords is equal to N, i.e., $|\mathcal{A}|^{wr} = N$, which implies $w = (3/l) \cdot \log N = \Theta(\log N)$. We note that in this construction, each codeword is unique and the decoder can correct any fraction $1 - r - \epsilon = 1/2$ of erasures from the codewords. Finally, M^2 is constructed as follows. We first copy M^1 to M^2. We then replace each bit of M^1 with $\alpha \triangleq (w/w^{\text{KS}}) \cdot l$ number of bits as follows. A zero entry of M^1 is simply replaced by α zeros. The non-zero entries in a given column of M^1 are replaced by the binary representations of the symbols in the corresponding codeword in the erasure-correcting code. The weight of each column of M^1 is w^{KS}, therefore each non-zero entry of M^1 is replaced by the binary representations of w/w^{KS} symbols, i.e. $\alpha = (w/w^{\text{KS}}) \cdot l$ bits. See Fig. 1 for an illustration of this step. We will see that $w^{\text{KS}} = \Theta(\log d)$, hence, M^2 has $\Theta(d \log d) \cdot \Theta(\log N / \log d) \cdot l = \Theta(d \log N)$ rows and has size $\Theta(d \log N) \times N$. Therefore, M is of size $\Theta(d \log N) \times N$, which is order-optimal in the probabilistic group testing.

We explain the purpose of this construction as follows. Assume that using Y^1 and M^{KS}, we correctly identify the columns $M_{i_1}^{\text{KS}}, \ldots, M_{i_d}^{\text{KS}}$ that correspond to the defective items. We define Y^2 as the measurement vector corresponding to the test matrix M^2, i.e., $Y^2 = \bigvee_{i \in S} M_i^2$. If M^{KS} satisfies the condition (3), this allows at least half of the symbols

of the erasure-correcting codeword of each defective item to be observed in Y^2 without any collisions, i.e. in non-intersecting positions, with the symbols of the codewords of the other defective items. The intersecting symbols can simply be regarded as erasures. The locations of the erasures can be identified from the columns $M^{\text{KS}}_{i_1}, \dots, M^{\text{KS}}_{i_d}$ decoded in the first step. Since the erasure-correcting code corrects any fraction 1/2 of erasures, we can identify the defective items without any errors in this step. See Fig. 1 for an illustration.

B. Analysis

We describe the Kautz-Singleton construction employed in Section III-A. Kautz and Singleton provide a construction by converting a Reed-Solomon (RS) code [37] to a binary matrix. We begin with the definition of Reed-Solomon codes.

Definition 1. *Let $\mathbb{F}_q$ be a finite field and $\alpha_1, \dots, \alpha_n$ be distinct elements from $\mathbb{F}_q$. Let $k \leq n \leq q$. The Reed-Solomon code of dimension k over $\mathbb{F}_q$, with evaluation points $\alpha_1, \dots, \alpha_n$ is defined with the following encoding function. The encoding of a message $m = (m_0, \dots, m_{k-1})$ is the evaluation of the corresponding $k-1$ degree polynomial $f_m(X) = \sum_{i=0}^{k-1} m_i X^i$ at all the α_i's:*

$$\text{RS}(m) = (f_m(\alpha_1), \dots, f_m(\alpha_n)).$$

The Kautz-Singleton construction starts with a $[n, k]_q$ RS code with $n = q$ and $N = q^k$. Each q-ary symbol is then replaced by a unit weight binary vector of length q, via "identity mapping" which takes a symbol $i \in [q]$ and maps it to the vector in $\{0, 1\}^q$ that has a 1 in the i'th position and zero everywhere else. Note that the resulting binary matrix will have $t = nq = q^2$ rows (tests).

While the choice $n = q$ is appropriate for the combinatorial group testing problem, we will shortly see that we need to set $q = \Theta(d)$ and $n = \Theta(\log d)$ in our problem. In Section III-A, we required that the Kautz-Singleton construction M^{KS} of size $\Theta(d \log d) \times \Theta(d^3)$ satisfies the conditions (2) and (3) with probability approaching to 1. We show this in the following theorem.

Theorem 1. *The Kautz-Singleton construction with parameters $n = \Theta(\log d)$, $k = 3$, and $q = \Theta(d)$ provides a binary matrix M^{KS} of size $\Theta(d \log d) \times \Theta(d^3)$ with constant weight columns $w^{\text{KS}} = \Theta(\log d)$. When S is the set of d columns chosen uniformly at random among all columns in M^{KS}, the following conditions*

$$\text{supp}(M^{\text{KS}}_j) \not\subseteq \cup_{i \in S} \text{supp}(M^{\text{KS}}_i) \quad \forall j \in [\Theta(d^3)] \backslash S,$$
$$|\text{supp}(M^{\text{KS}}_j) \backslash \cup_{i \in S \backslash \{j\}} \text{supp}(M^{\text{KS}}_i)| \geq w^{\text{KS}}/2 \quad \forall j \in S$$

are satisfied with probability at least $1 - \Theta(1/d)$.

We prove this theorem in Appendix. Building on the construction and decoding procedure introduced in Section III-A and the results of Theorem 1, we show the main result of this work in the following theorem.

Theorem 2. *Under the model introduced in Section II, the strongly explicit construction introduced in Section III-A has $t = \Theta(d \log N)$ tests and satisfies $P_e \leq \Theta(1/d)$ under the decoding procedure introduced in Section III-A. Furthermore, the decoding time is $O(d^3 \cdot \log d + d \cdot \log N)$.*

Proof. We begin with the number of tests of our construction $M = \begin{bmatrix} M^1 \\ M^2 \end{bmatrix}$. We know from Theorem 1 that M^1 is of size $\Theta(d \log d) \times N$ with constant weight columns $w^{\text{KS}} = \Theta(\log d)$. We further know from Section III-A that the block length of the code in Lemma 1 is $w = \Theta(\log N)$ and each symbol has l bits representation for some constant l. M^2 is generated by replacing each bit of M^1 with $(w/w^{\text{KS}}) \cdot l = \Theta(\log N / \log d)$ number of bits. Therefore, M^2 has $\Theta(d \log d) \cdot \Theta(\log N / \log d) = \Theta(d \log N)$ rows and is of size $\Theta(d \log N) \times N$. Hence, M is of size $\Theta(d \log N) \times N$.

We next analyze the error probability. We have a correct decoding based on successfully completing three steps as follows. First, we require that the defective items have different columns in M^1, i.e., they correspond to different columns in the Kautz-Singleton construction M^{KS}. Since M^{KS} is of size $\Theta(d \log d) \times \Theta(d^3)$, the probability of two random columns having the same column in M^1 can be bounded by $1/\Theta(d^3)$. Applying union bound over $\binom{d}{2}$ pairs among d defectives, the probability of error for this requirement is bounded by $\Theta(1/d)$.

Given that all defectives have different columns $M^{\text{KS}}_{i_1}, \dots, M^{\text{KS}}_{i_d}$, applying the naive decoding procedure with Y^1 over the columns of M^{KS} returns these columns with error probability at most $\Theta(1/d)$ from Theorem 1. Furthermore, these columns also satisfy (3). Although $M^{\text{KS}}_{i_1}, \dots, M^{\text{KS}}_{i_d}$ correspond to the defective items, they also correspond to some non-defective items as well.

Finally, for each $j \in [d]$, we take $M^{\text{KS}}_{i_j}$ and find out the non-intersecting entries in its support by comparing with the rest of the columns in $\{M^{\text{KS}}_{i_1}, \dots, M^{\text{KS}}_{i_d}\} \backslash M^{\text{KS}}_{i_j}$. We know from (3) that each column has at least half of its support that is not intersecting with the other defective columns. We take the corresponding non-intersecting symbols from Y^2 and consider the intersecting symbols as simply erasures. The erasure-correcting code is capable of correcting any fraction 1/2 of erasures, therefore, we can identify the defective items with zero-error in this step. Applying union bound over the failure of these three steps, error probability is bounded by $\Theta(1/d)$.

We next discuss the decoding complexity. Regarding the naive decoding rule and checking if $\text{supp}(M^{\text{KS}}_j) \subseteq \text{supp}(Y^1)$ is satisfied for each column $j \in [\Theta(d^3)]$, this can be done by looking at the positions in the support of M^{KS}_j. Since we have $w^{\text{KS}} = \Theta(\log d)$ and M^{KS} has $\Theta(d^3)$ columns, the decoding complexity here is $O(d^3 \cdot \log d)$. In the second step, we first find out the positions in the support of each decoded column that are not intersecting with the support of any other decoded columns. Since $w^{\text{KS}} = \Theta(\log d)$, in total this requires $O(d^2 \cdot \log d)$ time complexity. Finally, given that we know the non-intersecting symbols for each decoded column, the decoding time is linear in the block length $w = \Theta(\log N)$ for the erasure-correcting code, therefore, figuring out the defectives require $O(d \cdot \log N)$ time complexity. Overall, the

decoding time is $O(d^3 \cdot \log d + d \cdot \log N)$. $\square$

We conclude that the construction introduced in Section III-A is strongly explicit and efficiently decodable with order-optimal number of tests $t = \Theta(d \log N)$ and achieves $P_e \to 0$ in the probabilistic group testing framework.

APPENDIX

The proof of Theorem 1: We note that the Kautz-Singleton construction produces a binary matrix M^{KS} with q^k columns and nq rows. Each symbol is mapped to a vector in $\{0,1\}^q$ that has only a single 1, therefore, each column has constant weight $w^{\text{KS}} = n$. We will fix the parameters $n = \Theta(\log d)$, $k = 3$, and $q = \Theta(d)$, therefore, M^{KS} will be of size $\Theta(d \log d) \times \Theta(d^3)$ with constant weight columns $w^{\text{KS}} = \Theta(\log d)$.

Using union bound, the probability of failing either condition in Theorem 1 can be bounded as follows

$$\begin{aligned} P_e &\leq \sum_{j \in [q^3] \setminus S} \Pr\left(\text{supp}(M_j^{\text{KS}}) \subseteq \cup_{i \in S} \text{supp}(M_i^{\text{KS}})\right) \\ &\quad + \sum_{j \in S} \Pr\left(|\text{supp}(M_j^{\text{KS}}) \setminus \cup_{i \in S \setminus \{j\}} \text{supp}(M_i^{\text{KS}})| < n/2\right) \\ &\leq \sum_{j \in [q^3] \setminus S} \Pr\left(|\text{supp}(M_j^{\text{KS}}) \setminus \cup_{i \in S} \text{supp}(M_i^{\text{KS}})| < n/2\right) \\ &\quad + \sum_{j \in S} \Pr\left(|\text{supp}(M_j^{\text{KS}}) \setminus \cup_{i \in S \setminus \{j\}} \text{supp}(M_i^{\text{KS}})| < n/2\right) \\ &\leq \sum_{j \in [q^3]} \Pr\left(|\text{supp}(M_j^{\text{KS}}) \setminus \cup_{i \in S_{[q^3]/\{j\}}} \text{supp}(M_i^{\text{KS}})| < n/2\right) \end{aligned} \tag{4}$$

where in the last equation $S_{[q^3]/\{j\}}$ is uniformly distributed on the sets of size d among the items in $[q^3]/\{j\}$. The inequality (4) holds because $|S \setminus \{j\}| = d - 1$ while $|S_{[q^3] \setminus \{j\}}| = d$, therefore, $\Pr\left(|\text{supp}(M_j^{\text{KS}}) \setminus \cup_{i \in S \setminus \{j\}} \text{supp}(M_i^{\text{KS}})| < n/2\right) \leq \Pr\left(|\text{supp}(M_j^{\text{KS}}) \setminus \cup_{i \in S_{[q^3]/\{j\}}} \text{supp}(M_i^{\text{KS}})| < n/2\right)$.

Fix any n distinct elements $\alpha_1, \alpha_2, \ldots, \alpha_n$ from $\mathbb{F}_q$. We denote $\Psi \triangleq \{\alpha_1, \alpha_2, \ldots, \alpha_n\}$. We note that $|\text{supp}(M_j^{\text{KS}}) \setminus \cup_{i \in S_{[q^3]/\{j\}}} \text{supp}(M_i^{\text{KS}})| < n/2$ occurs if and only if the corresponding symbols of M_j^{KS} are contained in the union of symbols of $S_{[q^3]/\{j\}}$ in the RS code for at least $n/2$ rows in $[n]$. Denoting $f_{m_i}(X)$ as the polynomial corresponding to the column M_i^{KS}, let us define the random set $\Upsilon = \{\alpha \in \Psi : f_{m_j}(\alpha) \in \{f_{m_i}(\alpha) : i \in S_{[q^3]/\{j\}}\}\}$. We then have

$$\begin{aligned} &\Pr\left(|\text{supp}(M_j^{\text{KS}}) \setminus \cup_{i \in S_{[q^3]/\{j\}}} \text{supp}(M_i^{\text{KS}})| < n/2\right) \\ &\qquad = \Pr\left(|\Upsilon| > n/2\right). \end{aligned}$$

We note that

$$\begin{aligned} |\Upsilon| &= |\{\alpha \in \Psi : f_{m_j}(\alpha) \in \{f_{m_i}(\alpha) : i \in S_{[q^3]/\{j\}}\}\}| \\ &= |\{\alpha \in \Psi : 0 \in \{f_{m_i}(\alpha) - f_{m_j}(\alpha) : i \in S_{[q^3]/\{j\}}\}\}| \\ &= |\{\alpha \in \Psi : 0 \in \{f_{m_i}(\alpha) : i \in S'\}\}| \end{aligned}$$

where in the last equality the random set of polynomials $\{f_{m_i}(X) : i \in S'\}$ is generated by picking d nonzero polynomials of degree at most $k-1$ without replacement.

We define the random polynomial $h(X) \triangleq \prod_{i \in S'} f_{m_i}(X)$. Note that, for any $\alpha \in \Psi$ we have $0 \in \{f_{m_i}(\alpha) : i \in S'\} \Leftrightarrow h(\alpha) = 0$. We next bound the number of roots of the polynomial $h(X)$. We will use the following result from [38].

Lemma 2 ([38, Lemma 3.9]). *Let $R_q(l,k)$ denote the set of nonzero polynomials over $\mathbb{F}_q$ of degree at most k that have l distinct roots in $\mathbb{F}_q$. For all powers q and integers l, k,*

$$|R_q(l,k)| \leq q^{k+1} \cdot \frac{1}{l!}.$$

Let r denote the number of roots of a random nonzero polynomial of degree at most $k-1$. We have $\mathbb{E}[r] \leq 1$ since there is exactly one value of m_0 that makes $f_m(X) = 0$ for any fixed X and $m_1, \ldots, m_{k-1}$. Furthermore, using Lemma 2, we get $\mathbb{E}[r^2] \leq \sum_{i=1}^{k-1} \frac{i^2}{i!} = \sum_{i=1}^{k-1} \frac{i}{(i-1)!} < 2e$ where the first inequality is due to $\Pr(r = i) = |R_q(i, k-1)|/q^k \leq 1/i!$ from Lemma 2. Hence we can bound $\mathbb{E}[r^2] < 6$. We denote r_i as the number of roots of the polynomial $f_{m_i}(X)$ and r_h as the number of roots of the polynomial $h(X)$. Note that $r_h \leq \sum_{i \in S'} r_i$. We use the Bernstein concentration bound [39, Proposition 1.4] to $\sum_{i \in S'} r_i$ and obtain

$$\begin{aligned} \Pr\left(\sum_{i \in S'} r_i > 2d\right) &\leq \Pr\left(\frac{1}{d} \sum_{i \in S'} (r_i - \mathbb{E}[r_i]) > 1\right) \\ &\leq \exp\left(-\frac{d}{12 + k(2/3)}\right). \end{aligned}$$

Since we picked $k = 3$, the number of roots of $h(X)$ is bounded by $2d$ with probability at least $1 - \exp(-d/14)$.

Let $\mathcal{A}$ be the event of $h(X)$ having at most $2d$ number of roots. We can bound $\Pr\{|\Upsilon| \geq n/2 | \mathcal{A}\}$ by calculating the probability of having at least $n/2$ symbols from Ψ when we pick $2d$ symbols from $[q]$ uniformly at random without replacement. Let us fix $q = 10d$. Hence, if we ensure $n \leq 4d$, then we have

$$\begin{aligned} \Pr\{|\Upsilon| \geq n/2 | \mathcal{A}\} &\leq \frac{\binom{n}{n/2}\binom{q-n/2}{2d-n/2}}{\binom{q}{2d}} \\ &\overset{(i)}{\leq} 2^n \frac{(10d - n/2)!}{(2d - n/2)! 8d!} \frac{2d! 8d!}{10d!} \\ &= 4^{n/2} \frac{2d(2d-1)\ldots(2d-n/2+1)}{10d(10d-1)\ldots(10d-n/2+1)} \\ &\leq (4/5)^{n/2} \end{aligned}$$

where we use $\binom{n}{n/2} \leq 2^n$ in (i). We then have

$$\begin{aligned} &\Pr\left(|\text{supp}(M_j^{\text{KS}}) \setminus \cup_{i \in S_{[q^3]/\{j\}}} \text{supp}(M_i^{\text{KS}})| < n/2\right) \\ &\qquad \leq \exp\left(-d/14\right) + (5/4)^{-n/2}. \end{aligned}$$

Applying the summation over all $j \in [q^3]$ in (4), we obtain $P_e \leq (10d)^3 \exp(-d/14) + (10d)^3 (5/4)^{-n/2}$. Therefore, the average probability of error can be bounded as $P_e \leq \Theta(1/d)$ by choosing $n = \frac{8}{\log(5/4)} \log d$. The condition $n \leq 4d$ required in the proof is also satisfied under this choice for sufficiently large d.

REFERENCES

[1] R. Dorfman, "The detection of defective members of large populations," *Ann. Math. Statist.*, vol. 14, no. 4, pp. 436–440, Dec. 1943. [Online]. Available: https://doi.org/10.1214/aoms/1177731363

[2] H.-B. Chen and F. K. Hwang, "A survey on nonadaptive group testing algorithms through the angle of decoding," *Journal of Combinatorial Optimization*, vol. 15, no. 1, pp. 49–59, Jan. 2008. [Online]. Available: https://doi.org/10.1007/s10878-007-9083-3

[3] A. Ganesan, S. Jaggi, and V. Saligrama, "Learning immune-defectives graph through group tests," *IEEE Transactions on Information Theory*, vol. 63, no. 5, pp. 3010–3028, May. 2017.

[4] D. Malioutov and K. Varshney, "Exact rule learning via boolean compressed sensing," in *Proceedings of the 30th International Conference on Machine Learning*, vol. 28, no. 3, Jun. 2013, pp. 765–773.

[5] A. C. Gilbert, M. A. Iwen, and M. J. Strauss, "Group testing and sparse signal recovery," in *42nd Asilomar Conference on Signals, Systems and Computers*, Oct. 2008, pp. 1059–1063.

[6] M. T. Goodrich, M. J. Atallah, and R. Tamassia, "Indexing information for data forensics," in *Applied Cryptography and Network Security*. Springer Berlin Heidelberg, 2005, pp. 206–221.

[7] T. Berger, N. Mehravari, D. Towsley, and J. Wolf, "Random multiple-access communication and group testing," *IEEE Transactions on Communications*, vol. 32, no. 7, pp. 769–779, Jul. 1984.

[8] J. Wolf, "Born again group testing: Multiaccess communications," *IEEE Transactions on Information Theory*, vol. 31, no. 2, pp. 185–191, Mar. 1985.

[9] J. Luo and D. Guo, "Neighbor discovery in wireless ad hoc networks based on group testing," in *46th Annual Allerton Conference on Communication, Control, and Computing*, Sep. 2008, pp. 791–797.

[10] H. A. Inan, P. Kairouz, and A. Ozgur, "Sparse group testing codes for low-energy massive random access," in *55th Annual Allerton Conference on Communication, Control, and Computing*, Oct. 2017, pp. 658–665.

[11] H. A. Inan, P. Kairouz, and A. Özgür, "Sparse combinatorial group testing," *IEEE Transactions on Information Theory*, vol. 66, no. 5, pp. 2729–2742, 2020.

[12] A. Emad and O. Milenkovic, "Poisson group testing: A probabilistic model for nonadaptive streaming boolean compressed sensing," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May. 2014, pp. 3335–3339.

[13] H. Q. Ngo and D.-Z. Du, "A survey on combinatorial group testing algorithms with applications to dna library screening," *Discrete mathematical problems with medical applications*, vol. 55, pp. 171–182, 2000.

[14] D.-Z. Du and F. K. Hwang, *Combinatorial group testing and its applications*. World Scientific, 2000, vol. 12.

[15] G. K. Atia and V. Saligrama, "Boolean compressed sensing and noisy group testing," *Information Theory, IEEE Transactions on*, vol. 58, no. 3, pp. 1880–1901, 2012.

[16] D. Sejdinovic and O. Johnson, "Note on noisy group testing: Asymptotic bounds and belief propagation reconstruction," *CoRR*, vol. abs/1010.2441, 2010.

[17] 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," in *2011 49th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, Sept 2011, pp. 1832–1839.

[18] J. Scarlett and V. Cevher, "Phase Transitions in Group Testing," in *ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2016.

[19] M. Aldridge, L. Baldassini, and O. Johnson, "Group testing algorithms: Bounds and simulations," *IEEE Transactions on Information Theory*, vol. 60, no. 6, pp. 3671–3687, June 2014.

[20] O. Johnson, M. Aldridge, and J. Scarlett, "Performance of group testing algorithms with near-constant tests per item," *IEEE Transactions on Information Theory*, vol. 65, no. 2, pp. 707–723, Feb 2019.

[21] J. Scarlett and V. Cevher, "Near-optimal noisy group testing via separate decoding of items," *IEEE Journal of Selected Topics in Signal Processing*, vol. 12, no. 5, pp. 902–915, Oct 2018.

[22] J. Scarlett and O. Johnson, "Noisy non-adaptive group testing: A (near-)definite defectives approach," *IEEE Transactions on Information Theory*, pp. 1–1, 2020.

[23] H. A. Inan, P. Kairouz, M. Wootters, and A. Özgür, "On the optimality of the kautz-singleton construction in probabilistic group testing," *IEEE Transactions on Information Theory*, vol. 65, no. 9, pp. 5592–5603, Sep. 2019.

[24] A. G. D'yachkov and V. V. Rykov, "Bounds on the length of disjunctive codes," *Problemy Peredachi Informatsii*, vol. 18, no. 3, pp. 7–13, 1982.

[25] Z. Furedi, "On r-cover-free families," *Journal of Combinatorial Theory, Series A*, vol. 73, no. 1, pp. 172–173, 1996.

[26] W. Kautz and R. Singleton, "Nonrandom binary superimposed codes," *IEEE Transactions on Information Theory*, vol. 10, no. 4, pp. 363–377, Oct. 1964.

[27] E. Porat and A. Rothschild, "Explicit non-adaptive combinatorial group testing schemes," in *Automata, Languages and Programming*. Springer Berlin Heidelberg, 2008, pp. 748–759.

[28] M. Cheraghchi, "Noise-resilient group testing: Limitations and constructions," in *Fundamentals of Computation Theory*. Springer Berlin Heidelberg, 2009, pp. 62–73.

[29] P. Indyk, H. Q. Ngo, and A. Rudra, "Efficiently decodable non-adaptive group testing," in *Proceedings of the Twenty-first Annual ACM-SIAM Symposium on Discrete Algorithms*, 2010, pp. 1126–1142. [Online]. Available: http://dl.acm.org.stanford.idm.oclc.org/citation.cfm?id=1873601.1873692

[30] H. Q. Ngo, E. Porat, and A. Rudra, "Efficiently decodable error-correcting list disjunct matrices and applications," in *Automata, Languages and Programming*. Springer Berlin Heidelberg, 2011, pp. 557–568.

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

[32] 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, Apr. 2017.

[33] K. Lee, R. Pedarsani, and K. Ramchandran, "Saffron: A fast, efficient, and robust framework for group testing based on sparse-graph codes," in *Information Theory (ISIT), 2016 IEEE International Symposium on*. IEEE, 2016, pp. 2873–2877.

[34] A. Vem, N. T. Janakiraman, and K. R. Narayanan, "Group testing using left-and-right-regular sparse-graph codes," *CoRR*, vol. abs/1701.07477, 2017. [Online]. Available: http://arxiv.org/abs/1701.07477

[35] S. Bondorf, B. Chen, J. Scarlett, H. Yu, and Y. Zhao, "Sublinear-time non-adaptive group testing with $O(k \log n)$ tests via bit-mixing coding," *CoRR*, vol. abs/1904.10102, 2019. [Online]. Available: http://arxiv.org/abs/1904.10102

[36] N. Alon and M. Luby, "A linear time erasure-resilient code with nearly optimal recovery," *IEEE Transactions on Information Theory*, vol. 42, no. 6, pp. 1732–1736, Nov 1996.

[37] I. S. Reed and G. Solomon, "Polynomial codes over certain finite fields," *Journal of the Society for Industrial and Applied Mathematics*, vol. 8, no. 2, pp. 300–304, 1960.

[38] T. Hartman and R. Raz, "On the distribution of the number of roots of polynomials and explicit weak designs," *Random Structures & Algorithms*, vol. 23, no. 3, pp. 235–263, 2003.

[39] R. Bardenet and O.-A. Maillard, "Concentration inequalities for sampling without replacement," *Bernoulli*, vol. 21, no. 3, pp. 1361–1385, 08 2015. [Online]. Available: https://doi.org/10.3150/14-BEJ605