Improved non-adaptive algorithms for threshold group testing with a gap

Thach V. Bui*, Mahdi Cheraghchi†, and Isao Echizen‡§
*University of Science, VNU-HCMC, Ho Chi Minh City, Vietnam.
†University of Michigan, Ann Arbor MI, USA.
‡National Institute of Informatics, Tokyo, Japan.
§The University of Tokyo, Tokyo, Japan.
Email: *bvthach@fit.hcmus.edu.vn, †mahdich@umich.edu, ‡§iechizen@nii.ac.jp.

Abstract

The basic goal of threshold group testing is to identify up to d defective items among a population of n items, where d is usually much smaller than n. The outcome of a test on a subset of items is positive if the subset has at least u defective items, negative if it has up to ℓ defective items, where $0 \leq \ell < u$, and arbitrary otherwise. This is called threshold group testing with a gap. There are a few reported studies on test designs and decoding algorithms for identifying defective items. Most of the previous studies have not been feasible because there are numerous constraints on their problem settings or the decoding complexities of their proposed schemes are relatively large. Therefore, it is compulsory to reduce the number of tests as well as the decoding complexity, i.e., the time for identifying the defective items, for achieving practical schemes.

The work presented here makes five contributions. The first is a corrected theorem for a non-adaptive algorithm for threshold group testing proposed by Chen and Fu. The second is an improvement in the construction of disjunct matrices, which are the main tools for tackling (threshold) group testing. Specifically, we present a better upper bound on the number of tests for disjunct matrices compared to related work. The third and fourth contributions are a reduction in the number of tests and a reduction in the decoding time for identifying defective items in a noisy setting on test outcomes. The fifth contribution is a demonstration of the resulting improvements by simulation for previous work and the proposed schemes.

I. Introduction

Identification of up to d defective items in a large population of n items is the main objective of group testing. Defective items satisfy a specific property while negative (non-defective) items do not. Dorfman [1], an economist who served during World War II, initiated this research direction in an effort to identify syphilitic draftees among a large population of draftees. Rather than testing the draftees one by one, which would have taken much time and money, he proposed pooling the draftees into groups for testing, which is more efficient. Ideally, if there was at least one syphilitic draftee present in the group, the test outcome would be positive. Otherwise, it would be negative. This approach can be generalized by replacing "draftee" with "item," "syphilis" with "a specific property," and "syphilitic draftee" with "defective item." This is classical group testing (CGT) without noise. Formally, in CGT without noise, the outcome of a test on a subset of items is positive if the subset has at least one defective item and negative otherwise. If noise is present, the outcome may flip from positive to negative and vice versa.

A generalization of CGT called *threshold group testing* (TGT) was introduced by revising the definition of the test outcome [2]. In this model, the outcome of a test on a subset of items is positive if the subset has at least u defective items, negative if it has up to ℓ defective items, where $0 \leq \ell < u$, and arbitrary otherwise. The parameter $g = u - \ell - 1$ is called the gap. When $g = 0$, i.e., $\ell = u - 1$, threshold group testing has no gap. When $u = 1$, TGT reduces to CGT. TGT can be considered as a special case of complex group testing [3] or generalized group testing with inhibitors [4]. Like previous reports such as [2], [5]–[8], the focus of this paper is on threshold group testing with a gap, i.e., $g > 0$. Note that the results here are also applicable to the no-gap case ($g = 0$). However, this case should be treated separately to attain efficient solutions as presented in [6], [9], [10].

There are two approaches to designing tests. The first is *adaptive group testing* (AGT) in which the design of a test depends on the designs of the previous tests. This approach usually achieves optimal bounds on the number of tests; however, takes much time. The second is *non-adaptive group testing* (NAGT) which is an alternative solution for AGT. With this approach, all tests are designed independently and can be performed in parallel. Because of the resulting time saving, NAGT has been widely applied in various fields such as computational and molecular biology [11], networking [12], and neuroscience [4]. The focus of the work reported here is on NATGT, which stands for Non-Adaptive Threshold Group Testing, which is TGT associated with NAGT.

There are two main requirements for efficiently tackling group testing: minimize the number of tests and efficiently identify the set of defective items. Lengthy and intensive study of CGT has shown that the number of tests needed for effective use of AGT is $\Omega(d \ln n)$ [11], which is optimal theoretically. The decoding algorithm is usually included in the test design. For NAGT, Porat and Rothschild [13] first proposed using explicit nonadaptive constructions using $O(d^2 \ln n)$ tests with no efficient (sublinear to n) decoding algorithm. To have an efficient decoding algorithm, says $\mathsf{poly}(d, \ln n)$, while keeping the number of tests as small as possible, says $O(d^{1+o(1)} \ln^{1+o(1)} n)$, several schemes have been proposed [14]–[17]. Using probabilistic methods, Cai et al. [18] required only $O(d \ln d \cdot \ln n)$ tests to find defective items in time $O(d(\ln n + \ln^2 d))$. Recently, Bondorf et al. [19] presented a bit mixing coding that achieves asymptotically vanishing error probability with $O(d \log n)$ tests to identify defective items in time $O(d^2 \log d \cdot \log n)$ as $n \to \infty$. For further reading, we recommend readers to refer to the survey in [20].

From the genesis of TGT, Damaschke [2] showed that the set of positive items can be identified with up to g false positives and g false negatives by using $\binom{n}{u}$ non-adaptive tests. Chen et al. [3] gave an upper bound on the number of tests: $t(n, d, u; z] = O\left(z \left(\frac{k}{u}\right)^u \left(\frac{k}{d}\right)^d k \ln \frac{n}{k}\right)$, where $\lfloor (z-1)/2 \rfloor$ is usually referred to as the maximum number of errors in test outcomes. Cheraghchi [5] asserted that this bound is not optimal. Therefore, he reduced it to $O(d^{g+2} \ln(n/d) \cdot (8u)^u) = O(d^{g+2} \ln(n/d))$ tests under the assumption that u is constant, which is asymptotically optimal. When $d = \ell + u$, Ahlswede et al. [21] gave an upper bound on the number of tests, which is $O(u 2^{2u} \log n)$. They also considered the case $d \neq \ell + u$, however, the bound on the number of tests has no constructive approximations for inference.

There have been a few studies of decoding algorithms for NATGT with a gap. Chan et al. [7] set that the number of defective items to exactly d and $u = o(d)$ and used $O\left(\ln \frac{1}{\epsilon} \cdot d\sqrt{u} \ln n\right)$ tests to identify the defective items in time $O(n \ln n + n \ln \frac{1}{\epsilon})$, which is linear to the number of items, where $\epsilon \in (0, 1)$. Chen and Fu [8] proposed schemes for finding the defective items using $t(n, d - \ell, u; z]$ tests in time $O(n^u \ln n)$, which becomes impractical as n or u increases. Since the number of tests is quite large, the decoding time is not efficient for small n. Recently, by setting $d = O(n^\beta)$ for $\beta \in (0, 1)$ and $u = o(d)$, Reisizadeh et al. [22] use $\Theta(\sqrt{u} d \ln^3 n)$ tests to identify all defective items in time $O(u^{1.5} d \ln^4 n)$ w.h.p with the aid of an $O(u \ln n) \times \binom{n}{u}$ look-up matrix, which is unfeasible when n or u is large.

A. Contribution

The focus of this work is TGT with a gap, i.e., $g = u - \ell - 1 \geq 0$. The work presented here makes five contributions. The first is a corrected theorem for a non-adaptive algorithm for threshold group testing proposed by Chen and Fu [8]. The second is a better upper bound on the number of tests of disjunct matrices, which are the main tools for tackling (threshold) group testing. The third and fourth contributions are a reduction in the number of tests and a reduction in the decoding time for identifying defective items in a noisy setting on test outcomes. The fifth contribution is a demonstration of the resulting improvements by simulation for previous work and the proposed schemes.

The first contribution, which is summarized in Theorem 3, is correction of the decoding complexity analysis by Chen and Fu [8]. Their incorrect analysis in decoding complexity resulted in much smaller decoding complexity than the actual one.

Scheme	No. of defectives	Thresholds	Model on gap interval	Error tolerance	Number of tests t	Decoding time (Decoding complexity)	Defective set recovered	Decoding type
Ahlswede et al. [21]	$d = \ell + u$	$\ell < u \leq d$	No	×	$O(u2^{2u} \log n)$	×	×	×
Chen et al. [3]	$\leq d$	$\ell < u \leq d$	No	z	$t(n, d-\ell, u; z] =$ $O\left(z\left(\frac{k}{u}\right)^u \left(\frac{k}{d-\ell}\right)^{d-\ell} k \ln \frac{n}{k}\right)$	×	×	×
Cheraghchi [5]	$\leq d$	$\ell < u \leq d$	No	$O\left(\frac{pd^2 \log \frac{n}{d}}{(1-p)^2}\right)$	$O\left(\frac{d^{g+2} \ln \frac{n}{d}}{(1-p)^2} \cdot (8u)^u\right)$	×	×	×
Proposed 0 (**Theorem 4**)	$\leq d$	$\ell < u \leq d$	No	z	$h(n, d-\ell, u; z]$ $= O\left(\left(1 + \frac{z}{\alpha}\right) \cdot\right.$ $\left.\left(\frac{k}{u}\right)^u \left(\frac{k}{d-\ell}\right)^{d-\ell} k \ln \frac{n}{k}\right)$	×	×	×
Chan et al. [7]	$d = o(n)$	$\ell < u = o(d)$	Bernoulli Linear	×	$O\left(\ln \frac{1}{\epsilon} \cdot d\sqrt{\ell} \ln n\right)$ $O(g^2 d \ln n + d \ln \frac{1}{\epsilon})$	$O(n \ln n + n \ln \frac{1}{\epsilon})$ $O(g^2 n \ln n + n \ln \frac{1}{\epsilon})$	$S' = S$	Rnd.
Reisizadeh et al. [22]	$d = O(n^\beta)$ for $0 < \beta < 1$	$\ell < u = o(d)$	Bernoulli	×	$\Theta(\sqrt{u} d \ln^3 n)$	$O(u^{1.5} d \ln^4 n)$ with a $O(u \ln n) \times \binom{n}{u}$ look-up matrix	$S' = S$	Rnd.
Chen and Fu [8] (**corrected** in Theorem 3)	$\leq d$	$\ell < u \leq d$	No	z	$t(n, d-\ell, u; z]$	$O\left(t(n, d-\ell, u; z] \times u\left(\binom{n}{u}\right.\right.$ $\left.\left.+(d-u)\binom{n-u}{g+1}\binom{d-1}{g}\binom{d}{u}\right)\right)$	$\lvert S' \setminus S \rvert \leq g$ $\lvert S \setminus S' \rvert \leq g$	Det.
Proposed 1 (**Theorem 6**)	$\leq d$	$\ell < u \leq d$	**No**	z	$h(n, d-\ell, u; z]$	$O\left(h(n, d-\ell, u; z] \times u\left(\binom{n}{u}\right.\right.$ $\left.\left.+(d-u)\binom{n-u}{g+1}\binom{d-1}{g}\binom{d}{u}\right)\right)$	$\lvert S' \setminus S \rvert \leq g$ $\lvert S \setminus S' \rvert \leq g$	**Det.**
Proposed 2 (**Theorem 7**)	$\leq d$	$\ell < u \leq d$	**No**	z	$h(n, d-\ell, u; z]$	$O\left(h(n, d-\ell, u; z] \cdot u\binom{n}{u}\right)$	$\lvert S' \setminus S \rvert \leq gw$ $\lvert S \setminus S' \rvert \leq g$	**Det.**
Proposed 3 (**Theorem 8**)	$\leq d$	$\ell < u \leq d$	**No**	z	$h(n, d-\ell, u; z]$	$O\left(h(n, d-\ell, u; z] \cdot u \cdot \left(\binom{n}{u}\right.\right.$ $\left.\left.+(d-u)\binom{w+d-u}{g+1}\binom{d-1}{g}\binom{d}{u}\right)\right)$	$\lvert S' \setminus S \rvert \leq g$ $\lvert S \setminus S' \rvert \leq 2g$	**Det.**

TABLE I: Comparison of four proposed schemes with previous ones. A × means that the criterion does not hold for that scheme. The terms "Randomized" and "Deterministic" are abbreviated to "Rnd." and "Det.". Sets S' and S are the recovered defective set and the true defective set. Set $k = d - \ell + u$, $\alpha = k \ln \frac{en}{k} + u \ln \frac{ek}{u}$, $w = g\left(\left\lfloor \frac{|S|}{\ell+1} \right\rfloor + u - 1\right)$, and $0 \leq p < 1$.

The second contribution is a better upper bound on the number of tests of $(n, d, u; z]$-disjunct matrices. We significantly reduce the upper bound on the number of tests for constructing disjunct matrices compared with the work of Chen et al. [3]. This improvement paves the way to improve results not only in group testing, but also in other fields such as graph learning [23] and cover-free families [24].

The third and fourth contributions are a reduction in the number of tests and a reduction in the decoding time for identifying defective items in a noisy setting on test outcomes. We reduced both the number of tests and the decoding complexity compared with the state-of-the-art work of Chen and Fu [8]. Suppose there are up to $\lfloor (z-1)/2 \rfloor$ erroneous outcomes. Chen and Fu [8] use $t(n, d-\ell, u; z] = O\left(z\left(\frac{k}{u}\right)^u \left(\frac{k}{d-\ell}\right)^{d-\ell} k \ln \frac{n}{k}\right)$ tests to recover a set S' with $|S' \setminus S| \leq g$ and $|S \setminus S'| \leq g$, where $k = d - \ell + u$. By using $h(n, d-\ell, u; z] = O\left(\left(1 + \frac{z}{\alpha}\right) \cdot \left(\frac{k}{u}\right)^u \left(\frac{k}{d-\ell}\right)^{d-\ell} k \ln \frac{n}{k}\right)$ tests where $k = d - \ell + u$ and $\alpha = k \ln \frac{en}{k} + u \ln \frac{ek}{u}$, we can recover a set S' close to the true defective set S as follows:

1) $|S' \setminus S| \leq g$ and $|S \setminus S'| \leq g$.
2) $|S' \setminus S| \leq g$ and $|S \setminus S'| \leq 2g$.
3) $|S' \setminus S| \leq gw$ and $|S \setminus S'| \leq g$, where $w = g\left(\left\lfloor \frac{|S|}{\ell+1} \right\rfloor + u - 1\right)$.

The decoding complexities of these three cases are decreasing and always smaller than the one (after correction) proposed by Chen and Fu [8].

The last contribution is simulation for previous work and our proposed schemes. The results demonstrate the superiority of our proposed theorems over previous ones and validate the arguments presented here.

B. Comparison

The four proposed schemes are compared with previous ones in Table I. Our proposed schemes were error-tolerant and their decoding algorithms are deterministic. Note that Ahlswede et al. [21] also considered the case $d \neq \ell + u$, however, the bound on the number of tests has no constructive approximations for easy inference. Therefore, we do not include that bound in Table I for easy comparison.

1) Number of tests: When there are no models for the gap g, the upper bound on the number of tests with our proposed schemes is smaller than with the ones proposed by Chen and Fu [8] and Chen et al. [3]. Note that the upper bounds on the number of tests with Chen and Fu's scheme and Chen et al.'s scheme are equal. The number of tests $O\left(\frac{d^{g+2}\ln\frac{n}{d}}{(1-p)^2}\cdot(8u)^u\right)$ with the scheme proposed by Cheraghchi [5] can be reduced to $O\left(\frac{d^{g+2}\ln\frac{n}{d}}{(1-p)^2}\right)$ as u is a constant; i.e., the multiplicity $(8u)^u$ can be removed because it is constant. It is essentially the optimal asymptotic number of tests. However, Cheraghchi [5] does not focus on the finite length regime and refining the bounds for that as well as the algorithmic recovery problem. When $d = \ell + u$, a similar number of tests, which is $O(u2^{2u}\log n)$, is attained by Ahlswede et al. [21]. The big O notation is not useful in practice for this case because this multiplicity is extremely large and should not be removed. For example, we have $(8u)^u = 2^{20} = 1,048,576$ when $u = 4$ and $(8u)^u \geq 102,400,000$ when $u \geq 5$. Therefore, in terms of asymptotics, the number of tests with the scheme proposed by Cheraghchi is good as u is constant, although it is extremely large in practice.

By setting more conditions on g, u, and d, it is possible to significantly reduce the number of tests. However, these conditions would likely make any schemes associated with them impractical. Moreover, the previous work follow this approach do not consider erroneous outcomes. We apply the Bernoulli model to the gap; i.e., if the number of defectives in a test is between the thresholds, the outcome is positive with probability $1/2$. With $u = o(d)$ and error precision $\epsilon > 0$, the scheme of Chan et al. [7] can achieve a small number of tests $O\left(\ln\frac{1}{\epsilon}\cdot d\sqrt{\ell}\ln n\right)$ while that of Reisizadeh et al. [22] can attain $\Theta(\sqrt{u}d\ln^3 n)$ tests. We next apply a linear model to the gap; i.e., if the number of defectives in a test is between the thresholds, the probability that the outcome is positive increases linearly as the number of defectives increases. The number of tests in the scheme proposed by Chan et al. [7] is $O(g^2 n\ln n + n\ln\frac{1}{\epsilon})$.

Once $g = 0$, D'yachkov et al. [9] and Cheraghchi [5] show that it is possible to obtain an optimal bound on the number of tests, i.e., $O\left(d^2\ln n\right)$ tests, when u is a constant. Since the objective of this work is to consider the case $g > 0$, we recommend readers, who are interested in the case $g = 0$, to [10] for further reading.

2) Decoding time: Let S' and S be the recovered defective set and the true defective set. For threshold group testing with a gap, S' and S are indistinguishable if $|S'\setminus S| \leq g$ and $|S\setminus S'| \leq g$. Nevertheless, if a model is applied to the gap, $S' \equiv S$ can be attained with some probability. With this approach, the fastest decoding was at with the scheme of Reisizadeh et al. [22]: $O(u^{1.5}d\ln^4 n)$. However, this scheme is based on the assumption that $\ell < u = o(d)$, that the Bernoulli model is applied to the gap, and that an auxiliary look-up matrix of size $O(u\ln n)\times\binom{n}{u}$ is stored somewhere. The need for a look-up matrix makes this scheme an impractical solution. For example, if $n = 10^6$ and $u = 5$, the number of columns in the look-up matrix is more than 8.3 octillion (8.3×10^{27}). Moreover, n and u are more likely larger in practice. The scheme of Chan et al. [7] attains a near-optimal decoding time: $O\left(\ln\frac{1}{\epsilon}\cdot d\sqrt{\ell}\ln n\right)$ or $O(g^2 d\ln n + d\ln\frac{1}{\epsilon})$ for $\epsilon > 0$. However, this decoding time is attained only under certain constrains: the Bernoulli or a linear model is applied to the gap, n and $d = o(n)$ are large enough, and $\ell = o(d)$. This scheme is thus also likely impractical.

The conditions on the gap and on n, ℓ, u, and d make the schemes proposed by Chan et al. [7] and Reisizadeh et al. [22] impractical. Like Chen and Fu [8], we consider the case in which there are no constraints on the gap and $\ell < u \leq d < n$. Our decoding algorithms are deterministic. With the goal of attaining $|S'\setminus S| \leq g$ and $|S\setminus S'| \leq g$, the number of tests and the decoding time with our proposed schemes (summarized in Theorems 6, 7, 8) are much lower than the one proposed by Chen and Fu [8] (summarized in Theorem 3).

There are two terms in the decoding complexity of Theorem 6 (in Proposed 1): $\binom{n}{u}$ and $(d-u)\binom{n-u}{g+1}\binom{d-1}{g}\binom{d}{u}$. To remove the second term, we relax the condition on $|S'\setminus S|$ from $|S'\setminus S| \leq g$ to $|S'\setminus S| \leq wg$, where $w = g\left(\left\lfloor\frac{|S|}{\ell+1}\right\rfloor + u - 1\right)$. This reduces the decoding complexity of Theorem 6 to $O\left(h(n, d-\ell, u; z] \times u\binom{n}{u}\right)$, which is significantly less than the original one in Theorem 6. This result is summarized in Theorem 7 (in Proposed 2).

However, it is clear that the condition $|S' \setminus S| \leq wg$ in Theorem 7 is not as tight as the condition $|S' \setminus S| \leq g$ in Theorem 6. To remedy this drawback, we derived Theorem 8 (in Proposed 3), which slightly increases the decoding complexity while attaining the conditions $|S' \setminus S| \leq 2g$ and $|S \setminus S'| \leq g$.

II. Preliminaries

A. Notations

For consistency, we use capital calligraphic letters for matrices, non-capital letters for scalars, bold letters for vectors, and capital letters for sets. All matrix and vector entries are binary. The main notations are as follows:

1) $n, d, \mathbf{x} = (x_1, \ldots, x_n)^T$: number of items, maximum number of defective items, binary representation of n items.
2) ℓ, u: lower and upper bounds in (n, d, ℓ, u)-NATGT model.
3) $g = u - \ell - 1$: gap between ℓ and u.
4) $S = \{j_1, j_2, \ldots, j_{|S|}\}$: set of defective items; cardinality of S is $|S| \leq d$.
5) $N = [n] = \{1, \ldots, n\}$: set of n items.
6) $\otimes_{\ell,u}$: operation related to (n, d, ℓ, u)-NATGT (to be defined later).
7) $\mathcal{T}_{i,*}, \mathcal{M}_{i,*}, \mathcal{M}_j$: row i of matrix $\mathcal{T}$, row i of matrix $\mathcal{M}$, column j of matrix $\mathcal{M}$.
8) $\mathsf{wt}(\cdot)$: number of non-zero entries in input vector.

B. Disjunct matrices

Disjunct matrices were first introduced by Kautz and Singleton [25] as *superimposed codes* and then generalized by Stinson and Wei [24] and D'yachkov et al. [26].The support set for vector $\mathbf{v} = (v_1, \ldots, v_w)$ is $\mathsf{supp}(\mathbf{v}) = \{j \mid v_j \neq 0\}$. The formal definition of a disjunct matrix is as follows.

Definition 1. *An $m \times n$ binary matrix $\mathcal{T}$ is called an $(n, d, r; z]$-disjunct matrix if, for any two disjoint subsets $S_1, S_2 \subset [n]$ such that $|S_1| = d$ and $|S_2| = r$, there exists at least z rows in which there are all 1's among the columns in S_2 while all the columns in S_1 have 0's, i.e., $\left|\bigcap_{j \in S_2} \mathsf{supp}\left(\mathcal{T}_j\right) \setminus \bigcup_{j \in S_1} \mathsf{supp}\left(\mathcal{T}_j\right)\right| \geq z$. Parameter $\lfloor (z-1)/2 \rfloor$ is usually referred to as the error tolerance.*

Matrix $\mathcal{T}$ can be illustrated as follows.

$$\begin{array}{c|cccccccc|l}
 & & \multicolumn{2}{c}{\overbrace{\hphantom{\cdots\;\cdots}}^{r}} & & \multicolumn{3}{c}{\overbrace{\hphantom{\cdots\;\cdots\;\cdots}}^{d}} & & \\
 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \\
 & \cdots & 1 & 1 & \cdots & 0 & 0 & 0 & \cdots & \text{the 1st specific row} \\
\mathcal{T} = & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \\
 & \cdots & 1 & 1 & \cdots & 0 & 0 & 0 & \cdots & \text{the } z\text{th specific row} \\
 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \\
 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots &
\end{array}$$

Chen et al. [3] gave an upper bound on the number of rows for $(n, d, u; z]$-disjunct matrices as follows.

Theorem 1. *[3, Theorem 3.2] For any positive integers d, u, z, and n with $k = d + u \leq n$, there exists a $t \times n$ $(n, d, u; z]$-disjunct matrix with*

$$\begin{aligned} t(n, d, u; z] &= z \left(\frac{k}{u}\right)^u \left(\frac{k}{d}\right)^d \left[1 + k\left(1 + \ln\left(\frac{n}{k} + 1\right)\right)\right] \\ &= O\left(z \left(\frac{k}{u}\right)^u \left(\frac{k}{d}\right)^d k \ln \frac{n}{k}\right) = O(z \cdot t(n, d, u; 1]). \end{aligned} \tag{1}$$

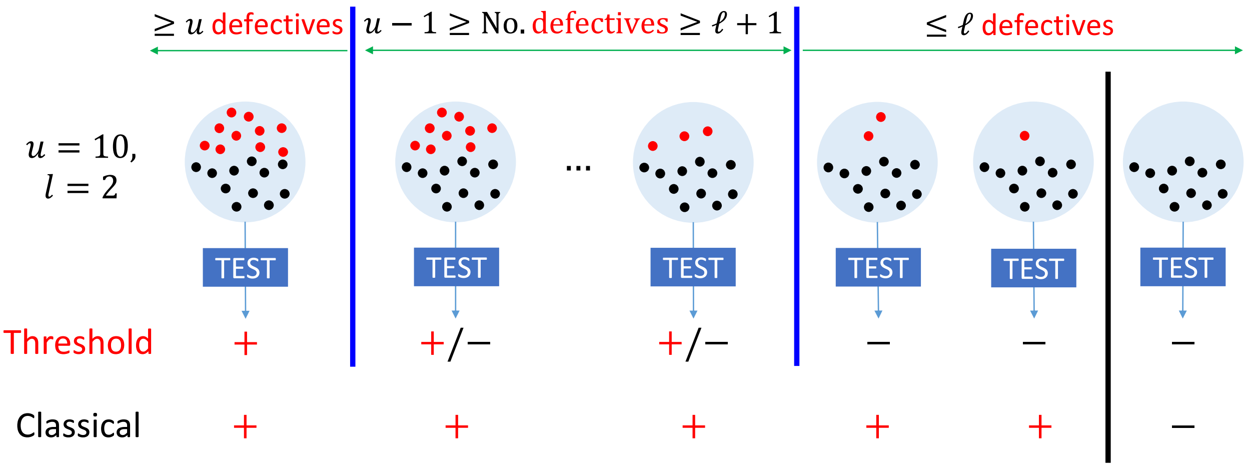


Fig. 1: Illustration of the (n, d, ℓ, u)-TGT model for $u = 10$ and $\ell = 2$ versus the classical group testing.

C. Problem definition

We index the population of n items from 1 to n. Let $[n] = \{1, 2, \ldots, n\}$ and S be the defective set, where $|S| \leq d$. A test is defined by a subset of items $P \subseteq [n]$. A pool with a negative (positive) outcome is called a negative (positive) pool. The outcome of a test on a subset of items is positive if the subset contains at least u defective items, is negative if the subset contains up to ℓ defective items, and arbitrary otherwise. Formally, the test outcome is positive if $|P \cap S| \geq u$, negative if $|P \cap S| \leq \ell$, and arbitrary if $\ell < |P \cap S| < u$. This model is denoted as (n, d, ℓ, u)-TGT. In addition, $g = u - \ell - 1$ is the gap.

Since a test outcome depends on two thresholds, u and ℓ, we illustrate the (n, d, ℓ, u)-TGT model for $u = 10$ and $\ell = 2$ as shown in Fig. 1. The black and red dots represent negatives and defectives, respectively. A subset containing defectives and negatives is a blue circle containing black and/or red dots. The outcome of a test on a subset of items is positive $(+)$ or negative $(-)$. In CGT ("Classical" in Fig 1), the outcome of a test on a subset of items is positive if the subset has at least one red dot, and negative otherwise. In (n, d, ℓ, u)-TGT ("Threshold" in Fig 1), the outcome of a test on a subset of items is positive if the subset has at least $u = 10$ red dots, negative if the subset has up to $\ell = 2$ red dots, and arbitrary otherwise.

We can model non-adaptive (n, d, ℓ, u)-TGT as follows. A $t \times n$ binary matrix $\mathcal{T} = (t_{ij})$ is defined as a measurement matrix, where n is the number of items and t is the number of tests. Vector $\mathbf{x} = (x_1, \ldots, x_n)^T$ is the binary representation vector of n items, where $|\mathbf{x}| \leq d$. An entry $x_j = 1$ indicates that item j is defective, and $x_j = 0$ indicates otherwise. The jth item corresponds to the jth column of the matrix. An entry $t_{ij} = 1$ naturally means that item j belongs to test i, and $t_{ij} = 0$ means otherwise. The outcome of all tests is $\mathbf{y} = (y_1, \ldots, y_t)^T$, where $y_i = 1$ if test i is positive and $y_i = 0$ otherwise. The procedure used to get outcome vector $\mathbf{y}$ is called *encoding*. The procedure used to identify defective items from $\mathbf{y}$ is called *decoding*. Outcome vector $\mathbf{y}$ is given by

$$\mathbf{y} = \mathcal{T} \otimes_{\ell,u} \mathbf{x} = \begin{bmatrix} \mathcal{T}_{1,*} \otimes_{\ell,u} \mathbf{x} \\ \vdots \\ \mathcal{T}_{t,*} \otimes_{\ell,u} \mathbf{x} \end{bmatrix} = \begin{bmatrix} y_1 \\ \vdots \\ y_t \end{bmatrix} \tag{2}$$

where $\otimes_{\ell,u}$ is a notation for the test operation in non-adaptive (n, d, ℓ, u)-TGT; namely, $y_i = \mathcal{T}_{i,*} \otimes_{\ell,u} \mathbf{x} = 1$ if $\sum_{j=1}^n x_j t_{ij} \geq u$, $y_i = \mathcal{T}_{i,*} \otimes_{\ell,u} \mathbf{x} = 0$ if $\sum_{j=1}^n x_j t_{ij} \leq \ell$, and $y_i = \mathcal{T}_{i,*} \otimes_{\ell,u} \mathbf{x} = \{0, 1\}$ if $\ell < \sum_{j=1}^n x_j t_{ij} < u$, for $i = 1, \ldots, t$.

Our objective is to find an efficient encoding and decoding scheme with non-adaptive approach to identify up to d defective items in non-adaptive (n, d, ℓ, u)-TGT. Precisely, our task is to minimize the number of rows in matrix $\mathcal{T}$ and the time for recovering $\mathbf{x}$ from $\mathbf{y}$ by using $\mathcal{T}$.

III. Review and analysis of Chen and Fu's work

A. Preliminaries

To clarify the basis of our proposed schemes, we review Chen and Fu's work [8] here. To facilitate the problem of identifying defectives, the graph search problem is first introduced. Given a vertex set $V = \{1, \ldots, n\}$, the goal is to reconstruct a hidden graph H defined on V by asking queries in the following format: for $U \subseteq V$, the query is "Does a complete graph induced by U contain any edge of H?" In other words, a pool containing all vertices in U is positive if at least one edge of H is also an edge of the complete graph induced by U.

Given a finite set V, a hypergraph $\mathbb{H} = (V, \mathsf{F})$ is a family $\mathsf{F} = \{E_1, E_2, \ldots, E_m\}$ of subsets of V. The elements of V are called vertices, and the subsets E_i's are the edges of the hypergraph $\mathbb{H}$.

A hypergraph is called a u-hypergraph if each edge consists of exactly u vertices. A subset of a set is called a u-subset if it contains exactly u elements of the set. Let W be a subset of V. A hypergraph is u-complete with respect to W if and only if (iff) every u-subset of W is an edge of the hypergraph.

Recall that our objective is to identify a set of defectives S from a given set of items $N = [n]$. Let S' be a set such that $|S' \setminus S| \leq g$ and $|S \setminus S'| \leq g$. Note that there is more than one set S' satisfying these properties. By mapping N as a vertex set V and the set of all u-subsets S's as an edge set F, we can convert threshold group testing with a gap into the problem of reconstructing a hidden graph $\mathbb{H} = (V, \mathsf{F})$ that is u-complete with respect to S'.

B. Main idea

The main idea is to construct a family F such that, for any subset $X \in \mathsf{F}$, $|X| = u$, $|X \cap S| \geq \ell + 1$ and every u-subset $X^+ \subseteq S$ must be in F. An approximate defective set S' is then recovered by using F, where $|S' \setminus S| \leq g$ and $|S \setminus S'| \leq g$. Note that S' is the best defective set that can be recovered [2].

To construct F, an indicator of "false negatives" is introduced. We say that a set X of the columns in a matrix appears in a row if every column in X has a 1 in the row. For a subset X of the columns in matrix $\mathcal{M}$, we define $t_0^{\mathcal{M}}(X)$ to be the number of negative pools in which all columns in X appear. Attaining S' is done by increasing the size of an approximate defective set S' from u until the properties $|S' \setminus S| \leq g$ and $|S \setminus S'| \leq g$ hold.

Given a measurement matrix $\mathcal{M}$, suppose that $\mathbf{y}$ is the outcome vector with up to e erroneous outcomes in non-adaptive (n, d, ℓ, u)-TGT. By setting $\mathcal{M}$ as an $(n, d - \ell, u; 2e + 1]$-disjunct matrix, Chen and Fu proposed a decoding algorithm in which an approximate set S' is attained as shown in Algorithm 1. Step 1 is to construct a family F and a hypergraph $\mathbb{H} = (V, \mathsf{F})$. Step 2 is to attain S' by using $\mathbb{H}$, as illustrated in Fig. 2. More precisely, we first initialize set S_1 consisting of the u vertices belonging to an edge of the family F. We then create a new set S_{i+1} such that $|S_{i+1}| = |S_i| + 1$. By selecting set A_i of g elements from S_i and set B_i of $g + 1$ elements in $V \setminus S_i$ such that $\mathbb{H}$ is u-complete with respect to $(S_i \cap A) \cup B$, set $S_{i+1} = (S_i \cap A) \cup B$. It is obvious that $|S_{i+1}| = |S_i| + 1$. This process stops once either S_i is not extendable or $|S_i| \geq d$. If the process stops when $i = m$, S' is set to S_m.

By using an $(n, d - \ell, u; z = 2e + 1]$-disjunct matrix and Algorithm 1, we can attain an approximate defective set S' as follows.

Theorem 2. *[8, Theorem 4.4] For an (n, d, ℓ, u)-TGT model with at most e erroneous outcomes, there exists a non-adaptive algorithm that successfully identifies some set S' with $|S' \setminus S| \leq g$ and $|S \setminus S'| \leq g$, using no more than $t(n, d - \ell, u; z = 2e + 1]$ tests. Moreover, the decoding complexity is*

$$\begin{aligned} & t(n, d - \ell, u; z] \times u\binom{n}{u} + (d - u)\binom{n - u}{g + 1}\binom{d - 1}{g}\binom{d}{u} \qquad (3) \\ & = O\left(z\left(\frac{k}{u}\right)^u \left(\frac{k}{d - \ell}\right)^{d - \ell} k \ln\frac{n}{k} \cdot u\binom{n}{u}\right), \end{aligned}$$

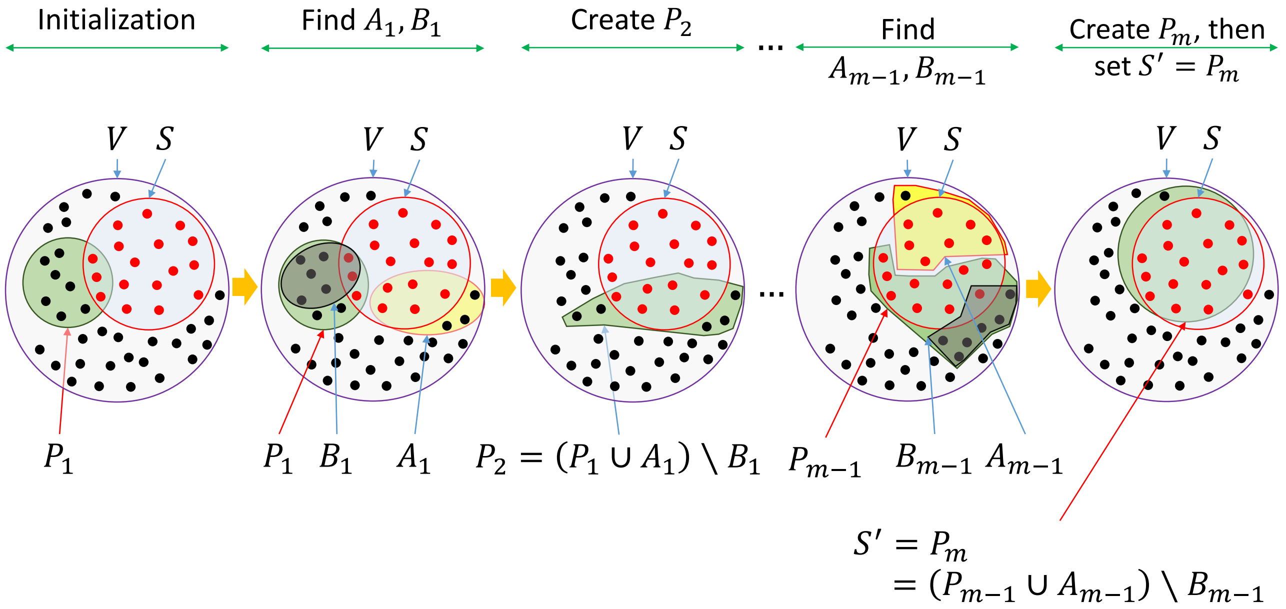


Fig. 2: Illustration of finding an approximate defective set S' of the defective set S such that $|S' \setminus S| \leq g$ and $|S \setminus S'| \leq g$ for Algorithm 1. We set $g = 7, u = 10$, and $\ell = u - g - 1 = 2$.

Algorithm 1 [Algorithm 2 [8]] $\text{Decoding}_1(\mathbf{y}, \mathcal{M})$: Decoding procedure for non-adaptive (n, d, ℓ, u)-TGT with up to e erroneous outcomes.

Input: Outcome vector $\mathbf{y}$, a $(d - \ell, u; z = 2e + 1]$-disjunct matrix $\mathcal{M}$.
Output: Set of defective items S' s.t $|S' \setminus S| \leq g$ and $|S \setminus S'| \leq g$.

1: Construct a hypergraph $\mathbb{H} = (V, \mathsf{F})$, where $V = [n]$ is the vertex set of n items and a u-subset $X \subseteq [n]$ is an edge in F iff $t_0^{\mathcal{M}}(X) \leq e$.

2: We want to establish increasing vertex-sets S_i's, $|S_1| < |S_2| \ldots < |S_m|$, such that the hypergraph $\mathbb{H}$ is u-complete with respect to each S_i. As an initial S_1, we may choose all u vertices of an arbitrary edge. To find S_{i+1} for $i \geq 1$, we check all possible cases to obtain some $(g + 1)$-set A_i in $V(\mathbb{H}) \setminus S_i$ and a g-set B_i in S_i such that $\mathbb{H}$ is u-complete with respect to $(S_i \cup A_i) \setminus B_i$. If such a pair A_i, B_i exists, then set $S_{i+1} = (S_i \cup A_i) \setminus B_i$. Continue this process till either S_m is not extendable or $|S_i| \geq d$. Output the set $S' = S_m$.

where $k = d - \ell + u$.

The complexity of the theorem above is attained by taking sum of the complexities of Steps 1 and 2. Step 1 is done in time $t(n, d - \ell, u; z] \times u\binom{n}{u}$. Step 2 is done in time $(d - u)\binom{n-u}{g+1}\binom{d-1}{g}\binom{d}{u}$, which is *incorrect* in general. We analyze and then correct Theorem 2 in the following section.

C. Analysis

We use the full expression for (3) instead of removing $(d - u)\binom{n-u}{g+1}\binom{d}{g}\binom{d}{u}$ as done by Chen and Fu [8]. Their incorrect analysis in the complexity of Step 2 led to *inaccurate* decoding complexity in Algorithm 1. They presumed that $(d - u)\binom{n-u}{g+1}\binom{d}{g}\binom{d}{u}$ can be reduced to $O(n^{g+1})$, and therefore is smaller than $\binom{n}{u} = O(n^u)$. We first analyze the complexity of Step 2. Let α be the cardinality of S_i. We always have $u \leq |S_i| \leq d - 1$ for $i < m$. The time costs of finding all possible subsets A_i and B_i are $\binom{n-\alpha}{g+1}$ and $\binom{\alpha}{g}$, respectively. One can verify whether "$\mathbb{H}$ is u-complete with respect to $(S_i \cup A_i) \setminus B_i$" if $t_0^{\mathcal{M}}(Z) \leq e$ for every u-subset $Z \subseteq V$. The complexity of the verification is $\binom{\alpha+1}{u} \times u \times t(n, d - \ell, u; z]$. Chen and Fu claimed that this cost is $\binom{\alpha+1}{u} \leq \binom{d}{u}$, which is simply equivalent to the complexity of counting all possibilities of u-subsets in $(S_i \cup A_i) \setminus B_i$. This claim is *incorrect*. Since Step 2 is repeated up to $d-u$ times, the complexity of executing this step is

$$(d-u)\binom{n-\alpha}{g+1}\binom{\alpha}{g}\binom{\alpha+1}{u}u\times t(n,d-\ell,u;z] = O\left(u(d-u)\binom{n-u}{g+1}\binom{d-1}{g}\binom{d}{u}t(n,d-\ell,u;z]\right).$$

We next prove that the quantity $(d-u)\binom{n-u}{g+1}\binom{d-1}{g}\binom{d}{u}$ in (3) should not be removed because it is not always smaller than $\binom{n}{u}$. Let us consider the case $u \geq 2$, $d = 2u$, and $u = g+1$, i.e., $\ell = 0$. We have:

$$\begin{aligned}
&(d-u)\binom{n-u}{g+1}\binom{d-1}{g}\binom{d}{u} \\
&= u\binom{n-u}{u}\binom{2u-1}{u-1}\binom{2u}{u} \\
&= u \cdot \frac{(n-u)(n-u-1)\dots(n-u-(u-1))}{u!} \cdot \frac{u}{2u(2u-u+2)}\binom{2u}{u} \cdot \binom{2u}{u} \\
&> \frac{(n-2u+1)^u}{u!} \cdot \frac{u}{2(u+2)}\binom{2u}{u}^2 \\
&> \frac{(n-2u+1)^u}{u!} \cdot \frac{u}{2(u+2)}\left(\frac{1.08444}{2\mathrm{e}^{1/(8u)}\sqrt{u}} \cdot 2^{2u}\right)^2, && (4) \\
&> \frac{(n-2u+1)^u}{u!} \cdot \frac{1}{2(u+2)} \cdot \left(\frac{1.08444}{2\mathrm{e}^{1/(8\times 2)}}\right)^2 \cdot 16^u \\
&> \frac{(n-2u+1)^u}{u!} \cdot \frac{1}{7(u+2)} \cdot 16^u, && (5)
\end{aligned}$$

and

$$\binom{n}{u} = \frac{n(n-1)\dots(n-(u-1))}{u!} < \frac{n^u}{u!},$$

where (4) is attained by using the inequality $\binom{mu}{u} > 1.08444\mathrm{e}^{-1/(8u)}u^{-1/2}\frac{m^{m(u-1)+1}}{(m-1)^{(m-1)(u-1)}}$ for integers $m > 1$ and $u \geq 2$ (Corollary 2.9 in [27]). Consider the following inequality:

$$\begin{aligned}
& & \frac{(n-2u+1)^u}{u!} \cdot \frac{1}{7(u+2)} \cdot 16^u \geq \frac{n^u}{u!} \\
&\Longleftrightarrow & 1 - \frac{1}{16} \cdot (7(u+2))^{1/u} \geq \frac{2u-1}{n}. && (6)
\end{aligned}$$

Since $(7(u+2))^{1/u}$ is a decreasing function of u and $u \geq 2$, for (6) to hold, it suffices that

$$\begin{aligned}
& & 1 - \frac{1}{16} \cdot (7(u+2))^{1/u} \geq 1 - \frac{\sqrt{28}}{16} = 1 - \frac{\sqrt{7}}{8} \geq \frac{2u-1}{n} \\
&\Longleftrightarrow & n \geq \frac{8(2u-1)}{8-\sqrt{7}}.
\end{aligned}$$

Therefore, when $d = 2u, u = g+1 \geq 2$, and $n \geq \frac{8(2u-1)}{8-\sqrt{7}}$, we always have the following inequality

$$(d-u)\binom{n-u}{g+1}\binom{d-1}{g}\binom{d}{u} > \frac{(n-2u+1)^u}{u!} \cdot \frac{1}{7(u+2)} \cdot 16^u \geq \frac{n^u}{u!} > \binom{n}{u}.$$

In summary, the complexity in (3) is incorrect. A corrected version of Theorem 2 is as follows.

Theorem 3 (Corrected version of Theorem 4.4 in [8])**.** *For an (n,d,ℓ,u)-TGT model with at most e erroneous outcomes, there exists a non-adaptive algorithm that successfully identifies some set S' with*

$|S' \setminus S| \leq g$ and $|S \setminus S'| \leq g$ using no more than $t(n, d-\ell, u; z = 2e+1]$ tests. Moreover, the decoding complexity is

$$O\left(t(n, d-\ell, u; z] \times u\left(\binom{n}{u} + (d-u)\binom{n-u}{g+1}\binom{d-1}{g}\binom{d}{u}\right)\right) \tag{7}$$

$$= O\left(z\left(\frac{k}{u}\right)^u\left(\frac{k}{d-\ell}\right)^{d-\ell} k\ln\frac{n}{k}\cdot u\left(\binom{n}{u} + (d-u)\binom{n-u}{g+1}\binom{d-1}{g}\binom{d}{u}\right)\right),$$

where $k = d - \ell + u$.

IV. Improved upper bounds on the number of tests for disjunct matrix

The upper bound on the number of tests with Theorem 1 is large because of the multiplicity z. We present a better upper bound on the number of tests as follows.

Theorem 4. *Let $2 \leq u \leq d < k = d + u \leq n$ be integers with $(d+u)^2/u \leq n$. Set $\alpha = k\ln\frac{en}{k} + u\ln\frac{ek}{u}$. For any positive integer z, there exists an $h \times n$ $(n, d, u; z]$-disjunct matrix with*

$$h(n, d, u; z] = O\left(\left(1 + \frac{z}{\alpha}\right)\cdot\left(\frac{k}{u}\right)^u\left(\frac{k}{d}\right)^d k\ln\frac{n}{k}\right). \tag{8}$$

Proof. Consider a randomly generated $h \times n$ matrix $\mathcal{G} = (g_{ij})_{1\leq i\leq h, 1\leq j\leq n}$ in which each entry g_{ij} is assigned to 1 with probability p and to 0 with probability $1-p$. For any pair of disjoint subsets $S_1, S_2 \subset [n]$ such that $|S_1| = u$ and $|S_2| = d$, let denote the event that for a row, there are 1's among the columns in S_1 and all 0's among the columns in S_2 on the same row as *a good event*. The probability that the good event happens is:

$$q = p^u(1-p)^d. \tag{9}$$

Set $\alpha = k\ln\frac{en}{k} + u\ln\frac{\mathrm{e}k}{u}$ and $\beta = 1 - 2/\alpha$. It is obvious that $0 < \alpha, \beta$. We then set $z = (1-\delta)qh$, where $0 < \delta$. We will later prove that there always exists $0 < \delta$ which depends on n, u, d, and z such that $z = (1-\delta)qh$. For a pair of disjoint subsets $S_1, S_2 \subset [n]$ such that $|S_1| = u$ and $|S_2| = d$, let $\mathsf{X}_i = 1$ be an event that a good event occurs at row i and $\mathsf{X} = 0$ be an event that a good event does not occur at row i. It is obvious that $\Pr[\mathsf{X}_i = 1] = q$, $\Pr[\mathsf{X}_i = 0] = 1-q$, and $E[\mathsf{X}_i] = q$. Let $\mathsf{X} = \sum_{i=1}^h \mathsf{X}_i$ denote the number of the good events happen for h rows for $i = 1, \ldots, h$. We get $\mu = E[\mathsf{X}] = \sum_{i=1}^h E[\mathsf{X}_i] = qh$.

By using Chernoff's bound, for a fixed S_1 and S_2, the probability that a good event occurs for up to z rows among h rows is

$$\Pr[\mathsf{X} \leq z] = \Pr[\mathsf{X} \leq (1-\delta)\mu] \leq \exp\left(-\frac{\delta^2\mu}{2}\right) = \exp\left(-\frac{\delta^2 qh}{2}\right). \tag{10}$$

Using a union bound, for any pair of disjoint subsets $S_1, S_2 \subset [n]$ with $|S_1| = u$ and $|S_2| = d$, the probability that a good event occurs for no more than z rows among h rows; i.e., the probability that $\mathcal{G}$ is not an $(n, d, u; z]$-disjunct matrix is at most

$$g(p, h, u, d, n) = \binom{n}{d+u}\binom{d+u}{u}\Pr[\mathsf{X} \leq z] \leq \binom{n}{k}\binom{k}{u}\exp\left(-\frac{\delta^2 qh}{2}\right). \tag{11}$$

To ensure that there exists an $(n, d, u, g; z]$-disjunct matrix $\mathcal{G}$, one needs to find p and h such that $g(p, h, u, d, n) < 1$. Set $p = \frac{u}{d+u} = \frac{u}{k}$ and $q = p^u(1-p)^d = \left(\frac{u}{k}\right)^u\left(\frac{d}{k}\right)^d$. We then have

$$g(p, h, u, d, n) \leq \binom{n}{k}\binom{k}{u}\exp\left(-\frac{\delta^2 qh}{2}\right) < 1.$$

For this to hold, it suffices that

$$\binom{n}{k}\binom{k}{u} \leq \left(\frac{\mathrm{e}n}{k}\right)^k \left(\frac{\mathrm{e}k}{u}\right)^u < \exp\left(\frac{\delta^2 qh}{2}\right) \tag{12}$$

$$\iff \qquad h > \frac{2}{\delta^2} \cdot \frac{1}{q} \cdot \left(k \ln\frac{\mathrm{e}n}{k} + u \ln \frac{\mathrm{e}k}{u}\right) = \frac{2}{\delta^2} \cdot \left(\frac{k}{u}\right)^u \left(\frac{k}{d}\right)^d \left(k \ln\frac{\mathrm{e}n}{k} + u \ln \frac{\mathrm{e}k}{u}\right). \tag{13}$$

In the above, we have (12) because $\binom{a}{b} \leq \left(\frac{\mathrm{e}a}{b}\right)^b$. Since $p = \frac{u}{k}$, from (13), if we set

$$\begin{aligned} h = h(n,d,u;z] &= \frac{3}{\delta^2} \cdot \frac{1}{q} \cdot \left(k \ln\frac{\mathrm{e}n}{k} + u \ln \frac{\mathrm{e}k}{u}\right) = \frac{3}{\delta^2} \cdot \frac{1}{q} \cdot \alpha, \text{ where } \alpha = k \ln\frac{\mathrm{e}n}{k} + u \ln \frac{\mathrm{e}k}{u}, \quad (14) \\ &= \frac{3}{\delta^2} \cdot \left(\frac{k}{u}\right)^u \left(\frac{k}{d}\right)^d \cdot \left(k \ln\frac{\mathrm{e}n}{k} + u \ln \frac{\mathrm{e}k}{u}\right), \end{aligned}$$

then $g(p,h,u,w,n) < 1$; i.e., there exists an $(n,d,u;z]$-disjunct matrix of size $h \times n$.

We now calculate δ versus n, d, u, and z. Since $z = (1-\delta)qh$ and $h = \frac{3}{\delta^2} \cdot \frac{1}{q} \cdot \alpha$ in (14), we have:

$$z = (1-\delta)qh = (1-\delta) \cdot \frac{3\alpha}{\delta^2} \tag{15}$$

$$\iff \qquad z\delta^2 + 3\alpha\delta - 3\alpha = 0 \tag{16}$$

Since the left side is a quadratic equation of δ and $\delta > 0$, we can derive

$$\delta = \frac{-3\alpha + \sqrt{9\alpha^2 + 12\alpha z}}{2z} = \frac{\sqrt{3\alpha}\left(\sqrt{3\alpha + 4z} - \sqrt{3\alpha}\right)}{2z}. \tag{17}$$

Let $f(x) = \sqrt{x}$. We have $f(x)$ is continuous on a closed interval $[3\alpha, 3\alpha + 4z]$ and differentiable on the open interval $(3\alpha, 3\alpha + 4z)$. By using the Lagrange's mean value theorem, then there is at least one point $b \in (3\alpha, 3\alpha + 4z)$ such that

$$\begin{aligned} f(3\alpha + 4z) - f(3\alpha) &= \sqrt{3\alpha + 4z} - \sqrt{3\alpha} \\ &= 4z \cdot f'(b) = 4z \cdot \frac{1}{2\sqrt{b}} = \frac{2z}{\sqrt{b}}. \end{aligned} \tag{18}$$

Combine with (17), we get

$$\delta = \frac{\sqrt{3\alpha}\left(\sqrt{3\alpha + 4z} - \sqrt{3\alpha}\right)}{2z} = \frac{\sqrt{3\alpha}}{2z} \cdot \frac{2z}{\sqrt{b}} = \sqrt{\frac{3\alpha}{b}}. \tag{19}$$

Because $b \in (3\alpha, 3\alpha + 4z)$, the following condition is straightfowardly attained

$$\frac{1}{\delta^2} = \frac{b}{3\alpha} \in \left(1, 1 + \frac{4z}{3\alpha}\right). \tag{20}$$

Therefore, the number of tests required is

$$\begin{aligned} h = h(n,d,u;z] &= \frac{3}{\delta^2} \cdot \left(\frac{k}{u}\right)^u \left(\frac{k}{d}\right)^d \cdot \left(k \ln\frac{\mathrm{e}n}{k} + u \ln \frac{\mathrm{e}k}{u}\right) \\ &< 3\left(1 + \frac{4z}{3\alpha}\right) \cdot \left(\frac{k}{u}\right)^u \left(\frac{k}{d}\right)^d \cdot \left(k \ln\frac{\mathrm{e}n}{k} + u \ln \frac{\mathrm{e}k}{u}\right). \end{aligned}$$

□

Since α is always larger than 1, z/α is always smaller than z. It implies that the upper bound on the number of tests in Theorem 4 is always tighter than the one in Theorem 1.

With an addition constrain on z, an alternative version of Theorem 4 can be derived to directly attain a better upper bound on the number of tests compared with the upper bound in Theorem 1.

Theorem 5. *Let $2 \leq u \leq d < k = d + u \leq n$ be integers with $(d+u)^2/u \leq n$. Set $\alpha = k \ln \frac{en}{k} + u \ln \frac{ek}{u}$ and $\beta = 1 - 2/\alpha$. For any integer $z \geq 4/\beta^2 + 1$, there exists an $h \times n$ $(n, d, u; z]$-disjunct matrix with*

$$h(n,d,u;z] = \left\lfloor \frac{2}{\delta^2} \cdot \left(\frac{k}{u}\right)^u \left(\frac{k}{d}\right)^d \cdot \left(k \ln \frac{en}{k} + u \ln \frac{ek}{u}\right) \right\rfloor + 1 = O\left(\frac{1}{\delta^2} \cdot \left(\frac{k}{u}\right)^u \left(\frac{k}{d}\right)^d \cdot k \ln \frac{n}{k}\right)$$
$$< t(n,d,u;z] = z \left(\frac{k}{u}\right)^u \left(\frac{k}{d}\right)^d \left[1 + k\left(1 + \ln\left(\frac{n}{k} + 1\right)\right)\right], \tag{21}$$

where $0 < \delta \leq \beta$.

Proof. By using the same construction and arguments in the proof in Theorem 4 until (13), if we set

$$\begin{aligned} h = h(n,d,u;z] &= \left\lfloor \frac{2}{\delta^2} \cdot \frac{1}{q} \cdot \left(k \ln \frac{en}{k} + u \ln \frac{ek}{u}\right) \right\rfloor + 1 \\ &= \left\lfloor \frac{2}{\delta^2} \cdot \left(\frac{k}{u}\right)^u \left(\frac{k}{d}\right)^d \cdot \left(k \ln \frac{en}{k} + u \ln \frac{ek}{u}\right) \right\rfloor + 1 \\ &= O\left(\frac{1}{\delta^2} \cdot \left(\frac{k}{u}\right)^u \left(\frac{k}{d}\right)^d \cdot k \ln \frac{n}{k}\right) \\ &= O\left(\frac{1}{(1-\delta) k \ln \frac{n}{k}} \cdot z \left(\frac{k}{u}\right)^u \left(\frac{k}{d}\right)^d \cdot k \ln \frac{n}{k}\right) \qquad (22) \\ &= O\left(\frac{1}{(1-\delta) k \ln \frac{n}{k}}\right) \cdot t(n,d,u;z], \end{aligned}$$

then $g(p,h,u,d,n) < 1$, where $t(n,d,u;z]$ is defined in (1); i.e., there exists an $(n,d,u;z]$-disjunct matrix of size $h \times n$. Equation (22) is obtained because

$$\begin{aligned} \frac{2(1-\delta)}{\delta^2} \cdot \left(k \ln \frac{en}{k} + u \ln \frac{ek}{u}\right) \leq z &= (1-\delta) q h \qquad (23) \\ &= (1-\delta) q \left(\left\lfloor \frac{2}{\delta^2} \cdot \frac{1}{q} \cdot \left(k \ln \frac{en}{k} + u \ln \frac{ek}{u}\right) \right\rfloor + 1 \right) \\ &= \Theta\left(\frac{1-\delta}{\delta^2} \cdot k \ln \frac{n}{k}\right) \\ &\leq \frac{2(1-\delta)}{\delta^2} \cdot \left(k \ln \frac{en}{k} + u \ln \frac{ek}{u}\right) + 1. \qquad (24) \end{aligned}$$

We next prove that $h(n,d,u;z] < t(n,d,u;z]$ once $0 < \delta \leq 1 - \frac{2}{k \ln \frac{en}{k} + u \ln \frac{ek}{u}}$. Indeed, we have

$$\begin{aligned} h(n,d,u;z] &= \left\lfloor \frac{2}{\delta^2} \cdot \frac{1}{q} \cdot \left(k \ln \frac{en}{k} + u \ln \frac{ek}{u}\right) \right\rfloor + 1, \text{ where } \frac{1}{q} = \left(\frac{k}{u}\right)^u \left(\frac{k}{d}\right)^d \\ &\leq \frac{2}{\delta^2} \cdot \frac{1}{q} \cdot \left(k \ln \frac{en}{k} + u \ln \frac{ek}{u}\right) + 1 \\ &< \frac{2}{\delta^2} \cdot \frac{1}{q} \cdot 2k \ln \frac{n}{k}. \qquad (25) \end{aligned}$$

This equation is attained because $k\ln\frac{\mathrm{e}n}{k}+u\ln\frac{\mathrm{e}k}{u}<2k\ln\frac{\mathrm{e}n}{k}$ as $(d+u)^2/u\le n$. On the other hand, we have

$$\begin{aligned} t(n,d,u;z] &= z\cdot\frac{1}{q}\cdot\left[1+k\left(1+\ln\left(\frac{n}{k}+1\right)\right)\right] \\ &> z\cdot\frac{1}{q}\cdot k\ln\frac{n}{k}\ge\frac{2(1-\delta)}{\delta^2}\cdot\left(k\ln\frac{\mathrm{e}n}{k}+u\ln\frac{\mathrm{e}k}{u}\right)\cdot\frac{1}{q}\cdot k\ln\frac{n}{k}. \end{aligned} \tag{26}$$

which is derived from the condition in (23). Combining (26) and (25), we always get $h(n,d,u;z]<t(n,d,u;z]$ if

$$\begin{aligned} & \frac{2}{\delta^2}\cdot\frac{1}{q}\cdot 2k\ln\frac{n}{k}\le\frac{2(1-\delta)}{\delta^2}\cdot\left(k\ln\frac{\mathrm{e}n}{k}+u\ln\frac{\mathrm{e}k}{u}\right)\cdot\frac{1}{q}\cdot k\ln\frac{n}{k} \\ \Longleftrightarrow\quad & \delta\le 1-\frac{2}{k\ln\frac{\mathrm{e}n}{k}+u\ln\frac{\mathrm{e}k}{u}}=\beta. \end{aligned}$$

Because of the condition in (24) and $0<\delta\le\beta$, z can range from $4/\beta^2+1$ to $+\infty$. Therefore, for any integer $z\ge 4/\beta^2+1$, we can find a corresponding δ in the interval $(0,\beta]$ such that $z=(1-\delta)qh$. □

V. Improved non-adaptive algorithms for threshold group testing with a gap

A. First proposed scheme

By using the construction of an $(n,d-\ell,u;z]$-disjunct matrix described in Section IV, we can reduce the number of tests for encoding and the decoding time for decoding in TGT with a gap. From Chen and Fu's work [8], if we use the $(n,d-\ell,u;z]$-disjunct matrix described in Theorem 4 as the input to Algorithm 1, the following theorem is derived:

Theorem 6. *Let $\ell, 0<g, 2\le u=\ell+g+1\le d<k=d-\ell+u\le n$ be integers with $(d+u)^2/u\le n$. Set $\alpha=k\ln\frac{\mathrm{e}n}{k}+u\ln\frac{\mathrm{e}k}{u}$. Let z be a positive integer and S be the defective set with $|S|\le d$. For an (n,d,ℓ,u)-TGT model with at most $e=\lfloor(z-1)/2\rfloor$ erroneous outcomes, there exists a non-adaptive algorithm that successfully identifies some set S' with $|S'\setminus S|\le g$ and $|S\setminus S'|\le g$ using no more than $h(n,d-\ell,u;z]$ tests, where $h(n,d-\ell,u;z]$ is defined in* (8)*. Moreover, the decoding complexity is*

$$O\left(h(n,d-\ell,u;z]\times u\left(\binom{n}{u}+(d-u)\binom{n-u}{g+1}\binom{d-1}{g}\binom{d}{u}\right)\right). \tag{27}$$

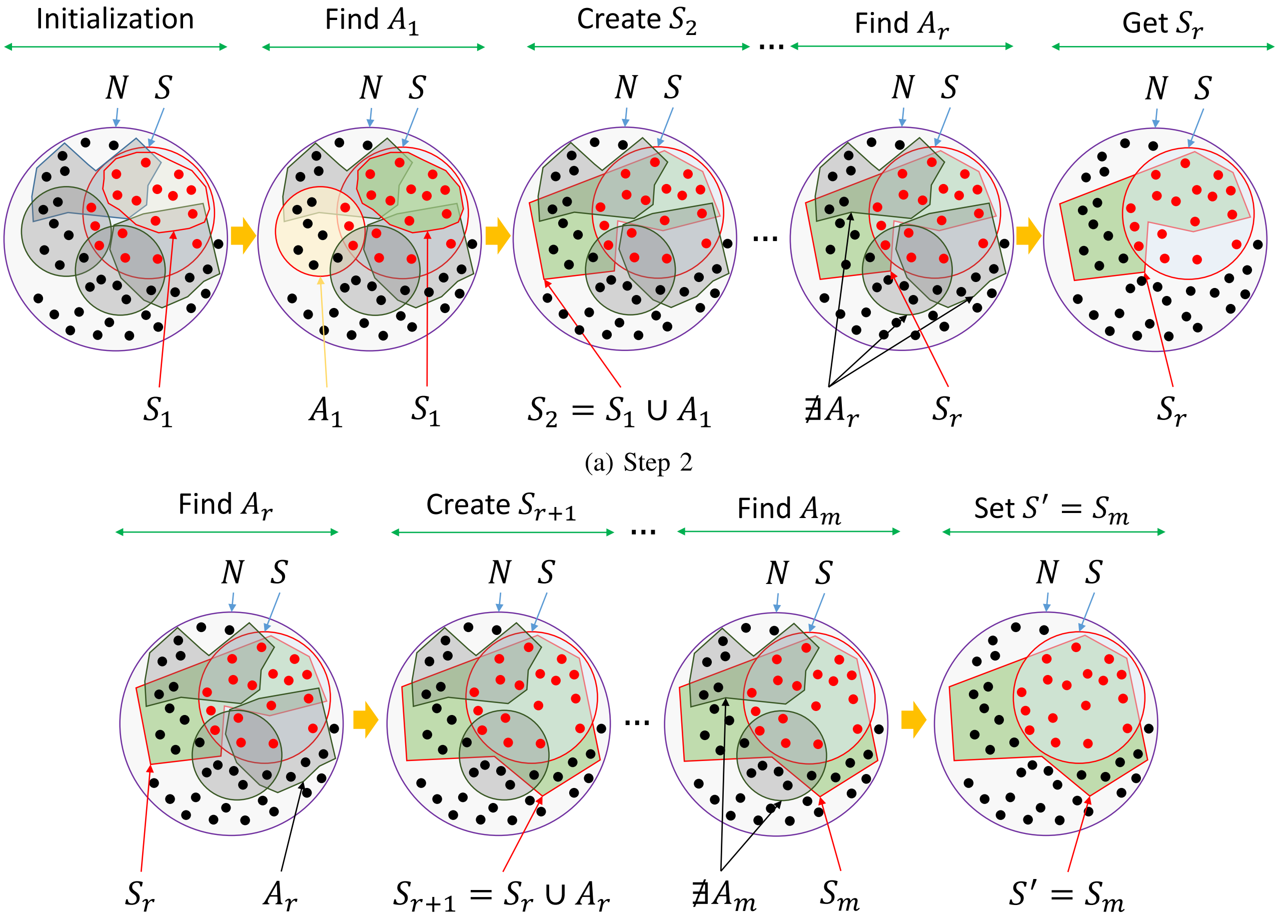
B. Second proposed scheme

We can see that the complexity of the decoding algorithm in the theorem above remains relatively high due to the second operator in (27). To reduce the decoding complexity, one can relax the conditions on $|S'\setminus S|$ and/or $|S\setminus S'|$. This approach is exemplified in the following theorem which is associated with Algorithm 2.

Theorem 7. *Let $\ell, 0<g, 2\le u=\ell+g+1\le d<k=d-\ell+u\le n$ be integers with $\mathrm{e}^2(d+u)^2/u\le n$. Set $\alpha=k\ln\frac{\mathrm{e}n}{k}+u\ln\frac{\mathrm{e}k}{u}$. Let z be a positive integer and S be the defective set with $|S|\le d$. For an (n,d,ℓ,u)-TGT model with at most $e=\lfloor(z-1)/2\rfloor$ erroneous outcomes, there exists a non-adaptive algorithm that successfully identifies some set S' with $|S'\setminus S|\le g\left(\left\lfloor\frac{|S|}{\ell+1}\right\rfloor+u-1\right)\le g\left(\frac{d}{\ell+1}+u-1\right)$ and $|S\setminus S'|\le g$ using no more than $h(n,d-\ell,u;z]$ tests, where $h(n,d-\ell,u;z]$ is defined in* (8)*. Moreover, the decoding complexity is*

$$O\left(h(n,d-\ell,u;z]\cdot u\binom{n}{u}\right).$$

The proof of this theorem is divided into two parts: correctness and decoding complexity. However, we first present visualizations that convey the essence of Algorithm 2.



(a) Step 2

(b) Step 3

Fig. 3: Illustration of finding an approximate defective set S' of defective set S such that $|S' \setminus S| \leq g\left(\left\lfloor \frac{|S|}{\ell+1} \right\rfloor + u - 1\right)$ and $|S \setminus S'| \leq g$ with $g = 7, u = 10$, and $\ell = u - g - 1 = 2$ for Algorithm 2.

1) Visualization: Steps 2 and 3 of Algorithm 2 are depicted in Fig. 3 for $n = 49, g = 7, u = 10, l = u - g - 1 = 2, d = 17$, and $|S| = 17$. There are many u-subsets belonging to F, but we depict only five of them here. Step 2 proceeds as shown in the upper five images as follows. Subset S_1, containing 10 defectives, selected as the initial subset. Scanning every subset of F reveals that A_1 is a subset such that $|S_1 \cup A_1| = |S_1| + |A_1| = 20$. Set $S_2 = S_1 \cup A_1$. The process continues until S_r consists of 13 defectives and 7 negatives. Since there are no u-subsets A_r's in $\mathsf{F} \setminus \{A_1, \ldots, A_{r-1}\}$ such that $|S_r \cup A_r| = |S_r| + |A_r|$, S_r is not extendible.

Step 3 proceeds as shown in the lower four images. Starting with subset S_r, we try to find a u-subset A_r's in $\mathsf{F} \setminus \{A_1, \ldots, A_{r-1}\}$ such that $|S_r \cup A_r| = |S_r| + g + 1$. If A_r exists, a new subset $S_{r+1} = S_r \cup A_r$ is attained. This process is repeated until there are no u-subsets A_m's in $\mathsf{F} \setminus \{A_1, \ldots, A_{m-1}\}$ such that $|S_m \cup A_m| \geq |S_m| + g + 1$. In other words, S_m is not extendible. The algorithm terminates and $S' = S_m$ is attained.

2) Correctness: To prove the correctness of Algorithm 2, we first prove that after Step 1, for every u-subset $X \in \mathsf{F}$, X contains no more than g items not in S. Moreover, every u-subset $X^+ \in S$ is in F. Since the proof is identical to the proof of Lemma 4.1 in [8], we omit it here.

Since every u-subset $X^+ \subseteq S$ must be in F, there exists ζ disjoint u-subsets $X_1^+, \ldots, X_\zeta^+$ in F and $|S \setminus \cup_{j=1}^{\zeta} X_j^+| \leq u - 1$.

In Step 2, another disjoint u-subset A_1 ($|S_1 \cap A_1| = 0$) is found from an initial subset S_1. Then

$S_2 = S_1 \cup A_1$. In general, to find S_{i+1} for $i \geq 1$, all possible cases are checked to attain some u-subset $A_i \in \mathsf{F} \setminus \{A_1, \ldots, A_{i-1}\}$ such that $|S_i \cup A_i| = |S_i| + |A_i| = |S_i| + u$. If A_i exists, i.e., S_i is extendible, then set $S_{i+1} = S_i \cup A_i$. On the other hand, if S_r is not extendible (A_i does not exist), we can infer that $|S \setminus S_r| \leq u - 1$. Assume that $|S \setminus S_r| \geq u$. Select $A_r \subseteq S \setminus S_r$ with $|A_r| = u$. Since $A_r \subseteq S \setminus S_r \in \mathsf{F}$ and $|A_r \cap S_r| = 0$, we get $|S_r \cup A_m| = |S_r| + |A_m| = |S_r| + u$, i.e., S_r is extendible. This contradicts the assumption.

There is a special case that if $|S \setminus S_r| \leq \ell$, this process stops. If $|S \setminus S_r| \leq \ell$, for any $A_r \in \mathsf{F} \setminus \{A_1, \ldots, A_{r-1}\}$, we have $|A_r \cap S_r| \geq 1$ because $|A_r \cap S| \geq \ell + 1$. Therefore, there does not exist $A_r \in \mathsf{F}$ such that $|S_r \cup A_r| = |S_r| + u$ because $|S_r \cup A_r| = |S_r| + |A_r| - |S_r \cap A_r| \leq |S_r| + u - 1 < |S_r| + u$.

Because each A_i can contain exactly $\ell + 1$ defectives in the worst case, Step 2 can run up to $\left\lfloor \frac{|S|}{\ell+1} \right\rfloor$ times.

We now consider Step 3. Subset S_m is not extendible iff there does not exist a u-subset $A_m \in \mathsf{F} \setminus \{A_1, \ldots, A_{m-1}\}$ such that $|S_m \cup A_m| \geq |S_m| + g + 1$. We then must have $|S \setminus S_m| \leq g$. Indeed, let us assume that $|S \setminus S_m| \geq g + 1$. Select $C \subseteq S \setminus S_m$ with $|S| = g + 1$ and $D \subseteq S \setminus C$ with $|D| = \ell$. Such a pair C, D always exists because $|S| \geq u = g + 1 + \ell$. Set $A_m = C \cup D$. Therefore, $|S_m \cup A_m| \geq |S_m| + g + 1$ and $A_m \in \mathsf{F}$. Hence, S_m is extendible, which contradicts the assumption that S_m is not extendible.

We have $|S \setminus S_r| \leq u - 1$ after running Step 2. It follows that Step 3 runs at most $(u - 1)$ times, i.e., $m - r \leq u - 1$, because S_i adds at least one defective for each iteration of Step 3.

In summary, Steps 2 and 3 run up to $\left\lfloor \frac{|S|}{\ell+1} \right\rfloor$ and $(u - 1)$ times, respectively. Because the subset considered at each iteration adds a u-subset having at least $\ell + 1$ defectives and up to g negatives, we have $|S' \setminus S| \leq g \left(\left\lfloor \frac{|S|}{\ell+1} \right\rfloor + u - 1 \right)$ and $|S \setminus S'| \leq g$ when the algorithm terminates.

Algorithm 2 Decoding$_2(\mathbf{y}, \mathcal{M})$: Decoding procedure for non-adaptive (n, d, ℓ, u)-TGT with up to e erroneous outcomes.

Input: Outcome vector $\mathbf{y}$, an $(n, d - \ell, u; z = 2e + 1]$-disjunct matrix $\mathcal{M}$.

Output: Set of defective items S' s.t $|S' \setminus S| \leq g \left(\left\lfloor \frac{|S|}{\ell+1} \right\rfloor + u - 1 \right)$ and $|S \setminus S'| \leq g$.

```
1: Construct a family F such that a $u$-subset $X \subseteq [n]$ is an edge in F iff $t_0^{\mathcal{M}}(X) \leq e$, where $t_0^{\mathcal{M}}(X)$ is
   the number of negative pools in which all columns in $X$ appear when using $\mathcal{M}$ as a measurement
   matrix.
2: We first want to establish increasing vertex-sets $S_i$'s, $|S_1| < |S_2| \ldots < |S_r|$, such that $S_{i+1}$ contains
   exactly $u$ defectives more than $S_i$. As an initial $S_1$, we can select all $u$ vertices of an arbitrary edge.
   To find $S_{i+1}$ for $i \geq 1$, we check all possible cases to attain some $u$-subset $A_i \in \mathsf{F} \setminus \{A_1, \ldots, A_{i-1}\}$
   such that $|S_i \cup A_i| = |S_i| + u$. If $A_i$ exists, then set $S_{i+1} = S_i \cup A_i$. This process is continued until
   $S_r$ is not extendible.
3: We then want to establish increasing vertex-sets $S_i$'s, $|S_{r+1}| < |S_{r+2}| \ldots < |S_m|$, such that $S_{i+1}$
   contains at least one defective item more than $S_i$. To find $S_{i+1}$ for $i \geq r$, we check all possible cases
   to attain some $u$-subset $A_i \in \mathsf{F} \setminus \{A_1, \ldots, A_{i-1}\}$ such that $|S_i \cup A_i| \geq |S_i| + g + 1$. If $A_i$ exists, then
   set $S_{i+1} = S_i \cup A_i$. This process is continued until $S_m$ is not extendible. Output set $S' = S_m$.
```

3) Decoding complexity: Step 1 takes $h(n, d - \ell, u; z] \cdot u\binom{n}{u}$ time. Since every u-subset in F has at least $\ell + 1$ defectives and up to $g = u - \ell + 1$ negatives, the maximum cardinality of F is:

$$f = \sum_{i=\ell+1}^{u} \binom{|S|}{i} \binom{n - |S|}{u - i} < \sum_{i=0}^{u} \binom{|S|}{i} \binom{n - |S|}{u - i} = \binom{n}{u}.$$

Because $|S' \setminus S| \leq g \left(\left\lfloor \frac{|S|}{\ell+1} \right\rfloor + u - 1 \right)$ and $|S \setminus S'| \leq g$, we have $|S'| \leq g \left(\left\lfloor \frac{|S|}{\ell+1} \right\rfloor + u - 1 \right) + d$. Since we scan the family F up to $\left\lfloor \frac{|S|}{\ell+1} \right\rfloor + (u - 1)$ times in both Steps 2 and 3, $|\mathsf{F}| \leq f$, and $|S_i| \leq |S'| \leq$

$g\left(\left\lfloor \frac{|S|}{\ell+1} \right\rfloor + u - 1\right) + d$, the complexity of Algorithm 2 is:

$$\begin{aligned} & h(n, d-\ell, u; z] \cdot u\binom{n}{u} + \left(\left\lfloor \frac{|S|}{\ell+1} \right\rfloor + u - 1\right)\left(g\left(\left\lfloor \frac{|S|}{\ell+1} \right\rfloor + u - 1\right) + d\right) \times uf \\ & = h(n, d-\ell, u; z] \cdot u\binom{n}{u} + us(gs+d)\sum_{i=\ell+1}^{u}\binom{|S|}{i}\binom{n-|S|}{u-i}, \end{aligned} \tag{28}$$

where $s = \left\lfloor \frac{|S|}{\ell+1} \right\rfloor + (u-1) \leq d+u$ and $k = d-\ell+u = d+g+1$.

We have

$$us(gs+d)\sum_{i=\ell+1}^{u}\binom{|S|}{i}\binom{n-|S|}{u-i} \leq u(d+u)(g(d+u)+d)\binom{n}{u} < (d+u)^2(g+1) \cdot u\binom{n}{u}, \tag{29}$$

and

$$\begin{aligned} h(n, d-\ell, u; z] \cdot u\binom{n}{u} & \geq \left(1 + \frac{d-\ell}{u}\right)^{u}\left(1 + \frac{u}{d-\ell}\right)^{d-\ell}(d+g+1)\ln\frac{n}{k} \cdot u\binom{n}{u} \\ & \geq 4(g+1)\left(1 + \frac{d-\ell}{u}\right)^{u}\left(1 + \frac{u}{d-\ell}\right)^{d-\ell} \cdot u\binom{n}{u}, \end{aligned} \tag{30}$$

because $d \geq u \geq g+1$ and $\ln\frac{n}{k} \geq 2$ ($n \geq \mathrm{e}^2(d+u)^2/u > \mathrm{e}^2(d-\ell+u)$). We next consider the following inequality:

$$(d+u)^2(g+1) \cdot u\binom{n}{u} \leq 4(g+1)\left(1 + \frac{d-\ell}{u}\right)^{u}\left(1 + \frac{u}{d-\ell}\right)^{d-\ell} \cdot u\binom{n}{u} \tag{31}$$

$$\iff \qquad d+u \leq 2\left(1 + \frac{d-\ell}{u}\right)^{u/2}\left(1 + \frac{u}{d-\ell}\right)^{(d-\ell)/2} \tag{32}$$

For this inequality to hold, by using Bernoulli's inequality, it suffices that

$$\begin{aligned} & d+u \leq 2\left(1 + \frac{d-\ell}{u} \times \frac{u}{2}\right)\left(1 + \frac{u}{d-\ell} \times \frac{d-\ell}{2}\right) \leq 2\left(1 + \frac{d-\ell}{u}\right)^{u/2}\left(1 + \frac{u}{d-\ell}\right)^{(d-\ell)/2} \\ \iff \quad & d+u \leq \frac{(d-\ell+2)(u+2)}{2} = \frac{du}{2} + (d+u) + 2 - \frac{\ell(u+2)}{2} \\ \iff \quad & \ell(u+2) \leq du + 4. \end{aligned} \tag{33}$$

The last inequality always holds because $\ell(u+2) \leq (u-1)(u+2) < u(u+1) + 4 \leq du + 4$ for $d \geq u+1$. Combining (29), (30), and (31), we get

$$us(gs+d)\sum_{i=\ell+1}^{u}\binom{|S|}{i}\binom{n-|S|}{u-i} \leq h(n, d-\ell, u; z] \cdot u\binom{n}{u},$$

for any $d \geq u+1$ and $n \geq \mathrm{e}^2(d+u)^2/u > \mathrm{e}^2(d-\ell+u)$. Therefore, the decoding complexity of Algorithm 2 is up to

$$h(n, d-\ell, u; z] \cdot 2u\binom{n}{u}.$$

C. Third proposed scheme

Our main idea here is to combine Algorithms 1 and 2. It is obvious that $|S' \setminus S| \leq g\left(\left\lfloor \frac{|S|}{\ell+1} \right\rfloor + u - 1\right)$ in Theorem 7, which is worse than the condition $|S' \setminus S| \leq g$ in Theorem 6. Theorem 7 can be improved to achieve the conditions $|S' \setminus S| \leq g$ and $|S \setminus S'| \leq 2g$ by using the outcome of Algorithm 2 as the input of Algorithm 1. An extension of Algorithm 2 is described in Algorithm 3. The decoding complexity of the improved algorithm is higher than the that in Theorem 7 but lower than that in Theorem 6. The conditions on $|S' \setminus S|$ and $|S \setminus S'|$ are respectively looser than and equal to the corresponding ones in Theorem 7. On the other hand, the conditions on $|S' \setminus S|$ and $|S \setminus S'|$ are equal to and tighter than the corresponding ones in Theorem 6. These comparisons are summarized in Table I.

Algorithm 3 $\text{Decoding}_3(\mathbf{y}, \mathcal{M})$: Decoding procedure for non-adaptive (n, d, ℓ, u)-TGT with up to e erroneous outcomes.

Input: Outcome vector $\mathbf{y}$, a $(d-\ell, u; z = 2e+1]$-disjunct matrix $\mathcal{M}$.
Output: Set of defective items S' s.t $|S' \setminus S| \leq g$ and $|S \setminus S'| \leq 2g$.

```
Set $V = \text{Decoding}_2(\mathbf{y}, \mathcal{M})$.
Construct hypergraph $\mathbb{H} = (V, \mathsf{F})$ where a $u$-subset $X \subseteq V$ is an edge in $\mathsf{F}$ iff $t_0^{\mathcal{M}}(X) \leq e$, where $t_0^{\mathcal{M}}(X)$ is the number of negative pools in which all columns in $X$ appear when using $\mathcal{M}$ as a measurement matrix.
We want to establish increasing vertex-sets $S_i$'s, $|S_1| < |S_2| \ldots < |S_m|$ such that hypergraph $\mathbb{H}$ is $u$-complete with respect to each $S_i$. As an initial $S_1$, we can select all $u$ vertices of an arbitrary edge. To find $S_{i+1}$ for $i \geq 1$, we check all possible cases to attain some $(g+1)$-set $A_i$ in $V(\mathbb{H}) \setminus S_i$ and a $g$-set $B_i$ in $S_i$ such that $\mathbb{H}$ is $u$-complete with respect to $(S_i \cup A) \setminus B$. If such a pair $A_i, B_i$ exists, set $S_{i+1} = (S_i \cup A_i) \setminus B_i$. This process is continued until either $S_m$ is not extendable or $|S_i| \geq d$. Output the set $S' = S_m$.
```

The set S' attained from Algorithm 3 satisfies two properties: $|S \setminus S'| \leq 2g$ and $|S' \setminus S| \leq g$. This can be interpreted to mean that the number of defective items in S' is at least $|S| - 2g$ ($|S' \cap S| \geq |S| - 2g$). We summarize this result as follows.

Theorem 8. *Let $\ell, 0 < g, 2 \leq u = \ell+g+1 \leq d < k = d-\ell+u \leq n$ be integers with $\mathrm{e}^2(d+u)^2/u \leq n$ and $d \geq u+1$. Let z be a positive integer and S be the defective set with $|S| \leq d$. Set $w = g\left(\left\lfloor \frac{|S|}{\ell+1} \right\rfloor + u - 1\right) < g\left(\frac{d}{\ell+1} + u - 1\right)$ and $w + d \leq n$. For an (n, d, ℓ, u)-TGT model with at most $e = \lfloor (z-1)/2 \rfloor$ erroneous outcomes, there exists a non-adaptive algorithm that successfully identifies some set S' with $|S' \setminus S| \leq g$ and $|S \setminus S'| \leq 2g$ using no more than $h(n, d-\ell, u; z]$ tests, where $h(n, d-\ell, u; z]$ is defined in* (8). *Moreover, the decoding complexity is*

$$O\left(h(n, d-\ell, u; z] \cdot u \cdot \left(\binom{n}{u} + (d-u)\binom{w+d-u}{g+1}\binom{d-1}{g}\binom{d}{u}\right)\right). \tag{34}$$

As with the previous one, the proof is divided into two parts: correctness and decoding complexity.

1) Correctness: From Theorem 7, we get $|V \setminus S| \leq g\left(\left\lfloor \frac{|S|}{\ell+1} \right\rfloor + u - 1\right)$ and $|S \setminus V| \leq g$. Set $P = V \cap S$. We always have $|P| \geq |S| - g$ because $|S \setminus V| \leq g$.

Using the same argument as in the first paragraph of Section V-B2, for any u-subset $X \in \mathsf{F}$, we get $|X \cap S| \geq \ell + 1$ and every u-subset $X^+ \subseteq P$ must be in F. Because $V(\mathbb{H})$ is u-complete with respect to $S' = S_m$, we attain $|S' \setminus S| \leq g$.

We now show that $|S \setminus S'| \leq 2g$ once $S' = S_m$ is not extendable or $|S_m| \geq d$. Consider the case $|S'| \geq d$. Since $|S \setminus S'| \leq g$, we get $|S' \cap S| \geq d - g$. This indicates that $|S \setminus S'| \leq g \leq 2g$ because $|S| \leq d$.

It is now adequate to show that if S' is not extendable, then $|S \setminus S'| \leq 2g$. To prove this property, we to prove $|P \setminus S'| \leq g$. The property is then straightforwardly attained because $P \subseteq S$ and $|P| \geq |S| - g$.

Assume that $|P \setminus S'| > g$. Set $A_m \subseteq P \setminus S'$ and $|A_m| = g+1$, and let B_m be any subset with $S' \setminus P \subseteq B_m \subset S'$ and $|B_m| = g$. Subset B_m always exists because $|S' \setminus S| \leq g$ and the initial S' has $u > g$ elements. Therefore, $(S' \cup A_m) \setminus B_m$ is contained in P. It follows that $\mathbb{H}$ is u-complete with respect to $(S' \cup A_m) \setminus B_m$. This contradicts the assumption that S' is not extendable.

In summary, $|S \setminus S'| \leq 2g$ and $|S' \setminus S| \leq g$ are always attained after running Algorithm 3.

2) Complexity: From Theorem 7, the complexity of Step 1 is $h(n, d-\ell, u; z] \cdot u\binom{n}{u}$.

Because $|V| \leq g\left(\left\lfloor \frac{|S|}{\ell+1} \right\rfloor + u - 1\right) + d = w + d$, the complexity of Step 2 is $uh(n, d-\ell, u; z] \times \binom{|V|}{u} \leq uh(n, d-\ell, u; z] \times \binom{w+d}{u}$.

We can verify whether "$\mathbb{H}$ is u-complete with respect to $(S_i \cup A_i) \setminus B_i$" if $t_0^{\mathcal{M}}(Z) \leq e$ for every u-subset $Z \subseteq V$. Using an argument similar to the one described in the first paragraph of Section III-C, we get that the complexity of Step 3 is $(d-u)\binom{w+d-u}{g+1}\binom{d-1}{g}\binom{d}{u} \times uh(n, d-\ell, u; z]$. The total complexity of Algorithm 3 is then at most

$$\begin{aligned}
&h(n, d-\ell, u; z] \cdot u\binom{n}{u} + uh(n, d-\ell, u; z] \times \binom{w+d}{u} \\
&+ (d-u)\binom{w+d-u}{g+1}\binom{d-1}{g}\binom{d}{u} \times uh(n, d-\ell, u; z] \\
&= h(n, d-\ell, u; z] \cdot u \cdot \left(\binom{n}{u} + (d-u)\binom{w+d-u}{g+1}\binom{d-1}{g}\binom{d}{u}\right) \qquad (35) \\
&= h(n, d-\ell, u; z] \cdot u \cdot \left(\binom{n}{u} + (d-u)\binom{g\left(\left\lfloor \frac{|S|}{\ell+1} \right\rfloor + u - 1\right) + d - u}{g+1}\binom{d-1}{g}\binom{d}{u}\right).
\end{aligned}$$

Equation (35) is attained if we suppose that $w + d = g\left(\left\lfloor \frac{|S|}{\ell+1} \right\rfloor + u - 1\right) + d \leq u\left(\frac{d}{\ell+1} + u - 1\right) + d \leq n$. This condition is practical because n is much larger than d.

VI. Simulation

We visualized (upper bounds on) the number of tests for threshold group testing with a gap using five parameters n, d, u, ℓ, and z using simulation. For each fixed z, we derived δ in Theorem 4 accordingly. Since the number of tests with Cheraghchi's scheme and Ahlswede et al.'s scheme is asymptotic while the number of tests with other works is exact, we consider only the other works, which are our proposed schemes, Chen et al.'s scheme, and Chen and Fu's scheme. They are visualized in Figures 4–5. The red lines represent for our proposed schemes.

Since Chan et al. [7] and Reisizadeh et al. [22] used a model for the test outcome when the number of defectives in a test fell between ℓ and u, we do not show the number of tests for their work here. The number of test with Chen and Fu's scheme is equal to the one with Chen et al.'s scheme, we only consider Chen et al.'s scheme here. The numbers of tests for our proposed scheme and Chen et al.'s scheme are plotted in the figures as $\log_{10} t$ versus $\log_{10} n$ for various settings of n, d, u, ℓ, and z, where t is the number of tests.

Parameter z was set to $\{3, 11, 101\}$ corresponding to error tolerance $e = \{1, 5, 50\}$. The number of items n and the maximum number of defectives d were respectively set to $\{10^6 = 1\text{M}, 10^8 = 10\text{M}, 10^9 = 1\text{B}, 10^{10} = 10\text{B}, 10^{11} = 100\text{B}\}$ and $\{20, 100, 1000\}$. Finally, upper threshold u and lower threshold ℓ were respectively set to $0.2d$ and $0.5u = 0.1d$.

As shown in Fig. 4 for $d = 20$ and Fig. 5 for $d = 100$ and $d = 1000$, the number of tests with our proposed scheme was the smallest for all settings compared to Chen et al.'s scheme. More importantly, the number was smaller than the number of items (except for $n = 10^6$) while those with the other schemes were mostly larger than the number of items.

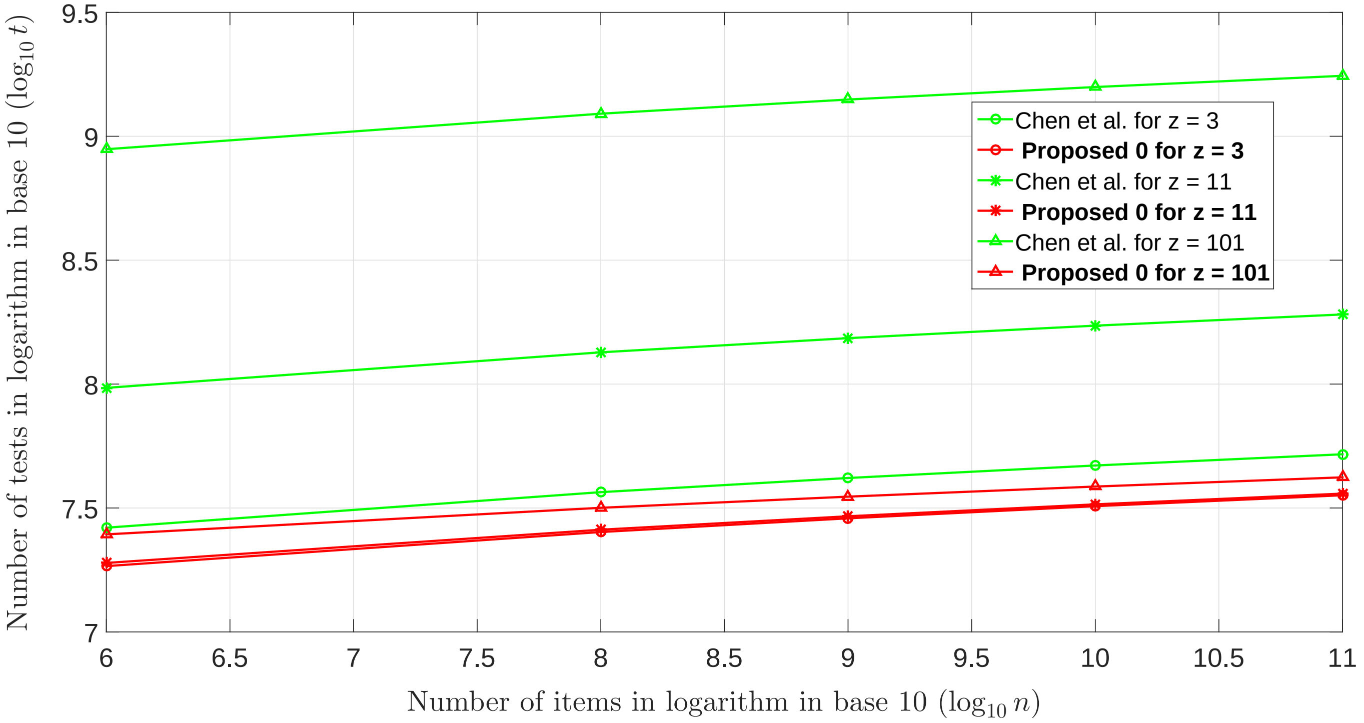
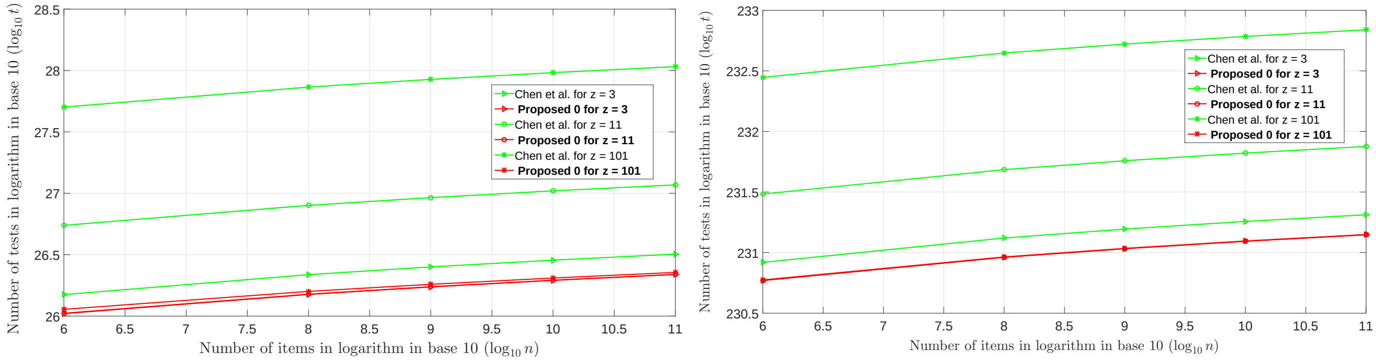


Fig. 4: Upper bounds on the number of tests versus number of items in logarithm in base 10 for $d = 20$, $z = \{3, 11, 101\}$, and $n = \{10^6 = 1\text{M}, 10^8 = 10\text{M}, 10^9 = 1\text{B}, 10^{10} = 10\text{B}, 10^{11} = 100\text{B}\}$ for Chen et al.'s and and our schemes.



(a) $d = 100$.

(b) $d = 1000$.

Fig. 5: Upper bounds on the number of tests versus number of items in logarithm in base 10 for $d = \{100, 1000\}$, $z = \{3, 11, 101\}$, and $n = \{10^6 = 1\text{M}, 10^8 = 10\text{M}, 10^9 = 1\text{B}, 10^{10} = 10\text{B}, 10^{11} = 100\text{B}\}$ for Chen et al.'s and our schemes.

When $d = 20$ (Fig. 4), for a small z, the number of tests with Chen et al.'s scheme was relatively close to ours. However, as z increased, the number of tests with Chen et al.'s scheme quickly diverged from that with our scheme.

VII. Conclusion

In this paper, we have presented a novel construction scheme for disjunct matrices which is better than the construction proposed by Chen et al. [3]. For threshold group testing, Cheraghchi gave a hint that the number of tests can be asymptotically to $O(d^{2+g} \log(n/d) \cdot c_u)$ which is essentially optimal, where

$c_u = (8u)^u$. Therefore, it is an interesting question that whether we can reduce the magnitude of the constant c_u and have a decoding algorithm associated with that number of tests.

We next presented a corrected theorem for Chen and Fu's scheme [8], three proposed schemes on improving non-adaptive encoding and decoding algorithms for threshold group testing as well as simulation for verifying our arguments throughout this work.

VIII. Acknowledgments

The authors thank Dr. Jonathan Scarlett in the Department of Computer Science and Department of Mathematics at the National University of Singapore for his constructive and insightful comments on an early version of this paper.

References

[1] R. Dorfman, "The detection of defective members of large populations," *The Annals of Mathematical Statistics*, vol. 14, no. 4, pp. 436–440, 1943.

[2] P. Damaschke, "Threshold group testing," in *General theory of information transfer and combinatorics*, pp. 707–718, Springer, 2006.

[3] H.-B. Chen, H.-L. Fu, and F. K. Hwang, "An upper bound of the number of tests in pooling designs for the error-tolerant complex model," *Optimization Letters*, vol. 2, no. 3, pp. 425–431, 2008.

[4] T. V. Bui, M. Kuribayashi, M. Cheraghchi, and I. Echizen, "A framework for generalized group testing with inhibitors and its potential application in neuroscience," *arXiv preprint arXiv:1810.01086*, 2018.

[5] M. Cheraghchi, "Improved constructions for non-adaptive threshold group testing," *Algorithmica*, vol. 67, no. 3, pp. 384–417, 2013.

[6] G. De Marco, T. Jurdziński, M. Różański, and G. Stachowiak, "Subquadratic non-adaptive threshold group testing," in *FCT*, pp. 177–189, Springer, 2017.

[7] C. L. Chan, S. Cai, M. Bakshi, S. Jaggi, and V. Saligrama, "Stochastic threshold group testing," in *2013 IEEE Information Theory Workshop (ITW)*, pp. 1–5, IEEE, 2013.

[8] H.-B. Chen and H.-L. Fu, "Nonadaptive algorithms for threshold group testing," *Discrete Applied Math.*, vol. 157, no. 7, pp. 1581–1585, 2009.

[9] A. Dyachkov, V. Rykov, C. Deppe, and V. Lebedev, "Superimposed codes and threshold group testing," in *Information Theory, Combinatorics, and Search Theory*, pp. 509–533, Springer, 2013.

[10] T. V. Bui, M. Kuribayashi, M. Cheraghchi, and I. Echizen, "Efficiently decodable non-adaptive threshold group testing," *IEEE Transactions on Information Theory*, 2019.

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

[12] A. D'yachkov, N. Polyanskii, V. Shchukin, and I. Vorobyev, "Separable codes for the symmetric multiple-access channel," in *2018 IEEE ISIT*, pp. 291–295, IEEE, 2018.

[13] E. Porat and A. Rothschild, "Explicit nonadaptive combinatorial group testing schemes," *IEEE Trans. Inf. Theory*, vol. 57, no. 12, pp. –, 2011.

[14] 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*, pp. 1126–1142, Society for Industrial and Applied Mathematics, 2010.

[15] H. Q. Ngo, E. Porat, and A. Rudra, "Efficiently decodable error-correcting list disjunct matrices and applications," in *ICALP*, pp. 557–568, Springer, 2011.

[16] M. Cheraghchi, "Noise-resilient group testing: Limitations and constructions," *Discrete Applied Mathematics*, vol. 161, no. 1-2, pp. 81–95, 2013.

[17] T. V. Bui, M. Kuribayashi, T. Kojima, R. Haghvirdinezhad, and I. Echizen, "Efficient (nonrandom) construction and decoding for non-adaptive group testing," *Journal of Information Processing*, vol. 27, pp. 245–256, 2019.

[18] S. Cai, M. Jahangoshahi, M. Bakshi, and S. Jaggi, "Grotesque: noisy group testing (quick and efficient)," in *Allerton*, pp. 1234–1241, 2013.

[19] 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," *arXiv preprint arXiv:1904.10102*, 2019.

[20] M. Aldridge, O. Johnson, and J. Scarlett, "Group testing: an information theory perspective," *arXiv preprint arXiv:1902.06002*, 2019.

[21] R. Ahlswede, C. Deppe, and V. Lebedev, "Bounds for threshold and majority group testing," in *2011 IEEE International Symposium on Information Theory Proceedings*, pp. 69–73, IEEE, 2011.

[22] A. Reisizadeh, P. Abdalla, and R. Pedarsani, "Sub-linear time stochastic threshold group testing via sparse-graph codes," in *2018 IEEE Information Theory Workshop (ITW)*, pp. 1–5, IEEE, 2018.

[23] H. Abasi, N. H. Bshouty, and H. Mazzawi, "Non-adaptive learning of a hidden hypergraph," *Theoretical Computer Science*, vol. 716, pp. 15–27, 2018.

[24] D. R. Stinson and R. Wei, "Generalized cover-free families," *Discrete Mathematics*, vol. 279, no. 1-3, pp. 463–477, 2004.

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

[26] A. D'yachkov, P. Vilenkin, D. Torney, and A. Macula, "Families of finite sets in which no intersection of ℓ sets is covered by the union of s others," *Journal of Combinatorial Theory, Series A*, vol. 99, no. 2, pp. 195–218, 2002.

[27] P. Stanica, "Good lower and upper bounds on binomial coefficients," *Journal of Inequalities in Pure and Applied Mathematics*, vol. 2, no. 3, p. 30, 2001.