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

Thach V. Bui, *Member, IEEE,* Mahdi Cheraghchi, *Senior Member, IEEE,* and Isao Echizen, *Member, IEEE*

***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. The parameter $g = u - \ell - 1$ is called *the gap*. In this paper, we focus on the case $g > 0$, i.e., threshold group testing with a gap. Note that the results presented here are also applicable to the case $g = 0$; however, the results are not as efficient as those in related work. Currently, a few reported studies have investigated 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 more accurate 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 and other tasks such as constructing cover-free families or learning hidden graphs. Specifically, we present a better exact upper bound on the number of tests for disjunct matrices compared with that in related work. The third and fourth contributions are a reduced exact upper bound on the number of tests and a reduced asymptotic bound on the decoding time for identifying defective items in a noisy setting on test outcomes. The fifth contribution is a simulation on the number of tests of the resulting improvements for previous work and the proposed theorems.

***Index Terms*—Non-adaptive threshold group testing with a gap, combinatorial mathematics, algorithms, sparse recovery.**

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 [2], 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 [3]. 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. This model is denoted as (n, d, ℓ, u)-TGT. 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 [4] or generalized group testing with inhibitors [5]. Like previous reports such as [3], [6]–[9], 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 [7], [10], [11].

In general, TGT is more complicated than CGT even for trivial testing since instead of testing all individuals as in CGT, all groups of a certain size (depending on the threshold parameters) have to be tested. It is intuitively obvious that the outcome of a test on a certain subset of items in TGT has less information than one in CGT. For example, if the outcome of a test on a subset of items is negative, we can be sure that there are no defectives in the subset if the test was done under the CGT setting, whereas the subset has up to $u-1$ defectives if the test was done under the TGT setting.

We illustrate TGT for two thresholds ($u = 10$ and $\ell = 2$) versus CGT in Fig. 1. The black and red dots represent negatives and defectives, respectively. A subset containing defectives and/or negatives is a blue circle containing black and/or red dots. The outcome of a test on a subset of items

Manuscript received May 26, 2020; revised March 21, 2021; accepted August 1, 2021. Thach V. Bui was supported in part by Vietnam National University Ho Chi Minh City (VNU-HCM) under Grant No. NCM2019-18-01. M. Cheraghchi's research was partially supported by the National Science Foundation under Grant No. CCF-2006455. Isao Echizen was partially supported by JSPS KAKENHI Grants JP16H06302, JP18H04120, JP20K23355, JP21H04907, and JP21K18023, and by JST CREST Grants JPMJCR18A6 and JPMJCR20D3, Japan. The material in this paper was presented in part at the 2020 IEEE International Symposium on Information Theory [1].

Thach V. Bui was with the Department of Information Engineering, University of Padova, Padova, Italy. He is now with the Department of Computer Science, National University of Singapore, Singapore, on leave from the Faculty of Information Technology, University of Science, VNU-HCMC, Ho Chi Minh City, Vietnam (e-mail: bvthach@fit.hcmus.edu.vn).

Mahdi Cheraghchi is with the Department of EECS, University of Michigan, Ann Arbor, MI 48109, USA (e-mail: mahdich@umich.edu).

Isao Echizen is with the National Institute of Informatics, Tokyo, 101-8430, Japan and with the Department of Information and Communication Engineering, University of Tokyo, Tokyo 113-8654, Japan (e-mail: iechizen@nii.ac.jp).

is positive ($+$) or negative ($-$). In CGT ("Classical" in the figure), 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 TGT with two thresholds u and ℓ ("Threshold" in the figure), 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.

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, it 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 [12], networking [13], and neuroscience [5]. Recently, group testing seems to be an efficient way to economically and quickly identify infected persons during the coronavirus pandemic of 2020–2021 [14], [15].

NAGT can be represented by a (binary) measurement matrix in which each row and each column represent a test and an item, respectively. An entry in the matrix at row i and column j that equals 1 naturally means that item j belongs to test i; and an entry that equals 0 means otherwise. For every group testing problem, we have a few possible cases:

1) The "for all" model (the worst case): we have a single measurement matrix and the same matrix has to recover any set of up to d defectives. Note that the matrix can be randomized or explicit, but once we have the matrix, the same matrix has to work correctly for any configuration of up to d defectives.
2) The "for each" model: for every fixed set of defectives, when we sample a measurement matrix, with high probability (whp) from the measurement outcomes, we can reconstruct the defectives.
3) The "average case" model: the matrix can be explicit or random, and the defectives are chosen randomly (often uniform, or iid). Then, whp over all randomness involved, we should be able to recover the defectives from the measurements.

NAGT generally refers to the "for all" model; otherwise, the model is specified. The focus of the work reported here is on NATGT (Non-Adaptive Threshold Group Testing), which is TGT associated with NAGT (with the "for all" model).

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 $\Omega(d \ln n)$ [12], which is theoretically optimal. The decoding algorithm is usually included in the test design. For NAGT, Porat and Rothschild [16] first proposed explicit non-adaptive 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 [17]–[20]. Using probabilistic methods, Cai et al. [21] 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. [22] 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 [23].

From the genesis of TGT, Damaschke [3] showed that the set of defective items can be identified with up to g false positives (i.e., negative items are identified as defective items) and g false negatives (i.e., defective items are identified as negative items) by using $\binom{n}{u}$ non-adaptive tests. Chen et al. [4] gave an upper bound on the number of tests: $t(n, d, u; z] = O\left(z \left(\frac{d+u}{u}\right)^u \left(\frac{d+u}{d}\right)^d (d+u) \ln \frac{n}{d+u}\right)$, where $\lfloor (z-1)/2 \rfloor$ is usually referred to as the maximum number of errors in the test outcomes. Cheraghchi [6] 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. [24] 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 on decoding algorithms for NATGT with a gap and with the "for all" or "for each" model. By using models for the gap and considering the "for each" model, Chan et al. [8] set that the number of defective items to exactly d, $u = o(d)$, 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 $\epsilon \in (0, 1)$. Recently, by setting $d = O(n^\beta)$ for $\beta \in (0, 1)$ and $u = o(d)$, Reisizadeh et al. [25] use $\Theta(\sqrt{u} d \ln^3 n)$ tests to identify all defective items in time $O(u^{1.5} d \ln^4 n)$ whp with the aid of a $O(u \ln n) \times \binom{n}{u}$ look-up matrix, which is unfeasible when n or u is large. To the best of our knowledge, the first and only work to tackle the "for all" model in NATGT with a decoding algorithm is that by Chen and Fu [9]. They proposed schemes for finding the defective items using $t(n, d-\ell, u; z]$ tests in time $O(n^u \ln n)$. However, the decoding time becomes impractical as n or u increases.

We consider here the potential use of threshold group testing as a tool to tackle the problems in designing tests for detecting viral infections [2], [15], [26] and chemical screening [3]. Damaschke [3] introduced threshold group testing with some potential applications for chemical screening, without presenting a concrete application. Back to the work of Dorfman [2], even for standard blood tests, the uncertainty in deciding the outcome of a test on a pool of blood samples remains problematic in practice. As mentioned in the first paragraph of this section, the outcome of a test on a pool of blood samples is positive if the pool contains at least one syphilitic sample and negative otherwise. However, in practice, before deciding the outcome of a test, we must get *a reference value* associated with the test. A next procedure is to set a threshold such that the outcome of a test is positive if its reference value is larger than or equal to the threshold and negative otherwise. Due to the presence of impurities in blood sample pools,

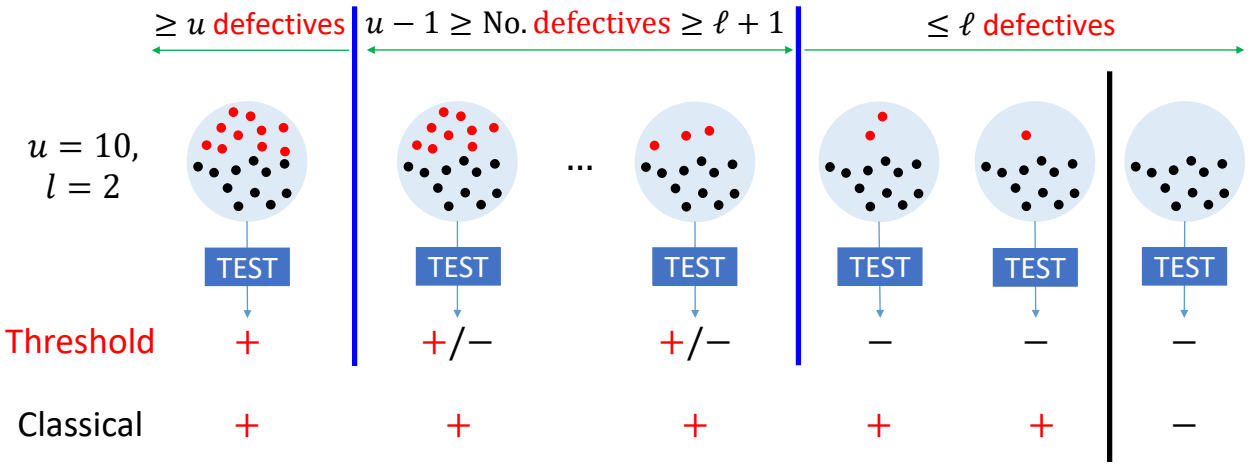


Fig. 1: Illustration of threshold group testing for $u = 10$ and $\ell = 2$ versus classical group testing.

it is difficult to set a unique threshold. Dorfman suggested setting it to be the average of the impurities in the separate samples. This would result in three ranges for the threshold: *positive*, *negative*, and *inconclusive*. If the threshold is in the positive (negative) range, the outcome of a test is positive (negative) if its reference value is larger (smaller) than or equal to the threshold. If the threshold is in the inconclusive range, it is uncertain to decide whether the test outcome is positive or negative. This is exactly what TGT with a gap tries to capture. In 2014, Emad and Milenkovic [26] introduced "semi-quantitative group testing" (SQGT) to tackle a model for quantitative polymerase chain reaction (qPCR) tests. Since TGT is a special case of SQGT, it can also be used in qPCR tests. The work of Gabrys et al. [15] motivated the application of TGT to reverse transcription PCR (RT-PCR) or quantitative PCR (qPCR) tests for viral infections such as Covid-19. The fluorescence values captured by the PCR process has different levels, and again one can assign a positive range, a negative range, and an inconclusive range in a manner similar to the work of Dorfman.

A. Contributions

The focus of this work is TGT with a gap; i.e., $g = u - \ell - 1 > 0$. Note that the results here are also applicable to the no-gap case, i.e., $g = 0$; however, the no-gap case should be treated separately to attain efficient solutions, as explained in [7], [10], [11].

The first contribution, which is summarized in Theorem 3, is correction of the decoding complexity analysis by Chen and Fu [9]. Their inaccurate analysis in decoding complexity resulted in much smaller decoding complexity than the actual one.

The second contribution is a better exact upper bound on the number of tests of $(n, d, u; z]$-disjunct matrices (defined later). We significantly reduce the upper bound on the number of tests for constructing disjunct matrices compared with the work of Chen et al. [4]. The basic idea is that instead of using a hypergraph to generate a disjunct matrix as Chen et al. did, we directly generate a random disjunct matrix. This improvement paves the way to improved results not only in group testing, but also in other fields such as graph learning [27] and cover-free family construction [28].

The third and fourth contributions are a reduced exact upper bound on the number of tests and a reduced asymptotic bound on the decoding time for identifying defective items in a noisy setting on test outcomes compared with the state-of-the-art work of Chen and Fu [9]. The number of tests is directly reduced by using a better upper bound on the number of tests (the second contribution). The basic idea for reducing decoding time is to pick subsets of potential defectives such that each subset contains at least $\ell + 1$ defectives and then return the union of these subsets as an approximate defective set. To attain a better approximate defective set (at the cost of a longer decoding time), the approximate defective set derived as described above is taken as the input to the existing algorithm in [9].

Suppose there are up to $\lfloor (z-1)/2 \rfloor$ erroneous outcomes. Let S' be the approximate defective set returned by decoding procedure. Two sets $S \setminus S'$ and $S' \setminus S$ are referred to as the sets of false negatives and false positives, respectively. Chen and Fu [9] 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 $\alpha = k \ln \frac{\mathrm{e}n}{k} + u \ln \frac{\mathrm{e}k}{u}$, 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) $|S' \setminus S| \leq gw$ and $|S \setminus S'| \leq g$, where $w = \left(\left\lfloor \frac{|S|}{\ell+1} \right\rfloor + u - 1\right) g$.
3) $|S' \setminus S| \leq g$ and $|S \setminus S'| \leq 2g$.

The decoding complexities of these three cases are always smaller than the one (after correction) proposed by Chen and Fu [9].

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

arguments presented here.

The contributions are summarized in Theorems 3, 4, 6, 7, and 8 and illustrated in Fig. 2 (except Theorem 3). The ovals, lines, parallelograms, and rectangles represent start or end point, connectors showing relationships between the representative shapes, inputs or outputs, and processes, respectively. The dash-dot line represents a comment on the representative shapes. The blue arrows represent the previous schemes while the other arrows represent our proposed theorems.

B. Comparison

The one proposed theorem for the number of tests and three proposed non-adaptive algorithms are compared with previous ones in Table I. Our proposed algorithms are error-tolerant and their decoding algorithms are deterministic. Note that Ahlswede et al. [24] 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 theorems is smaller than with the ones proposed by Chen and Fu [9] and Chen et al. [4]. Note that the upper bounds on the number of tests with Chen and Fu's scheme and Chen et al.'s scheme are equal, and so are our proposed theorems. 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 [6] 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 [6] 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. [24]. 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.

The number of tests could be significantly reduced by setting more conditions on g, u, and d, but such conditions would likely make any proposed scheme impractical. Moreover, the previous schemes that followed this approach do not take into account erroneous outcomes. When the Bernoulli model is applied to the gap, i.e., the number of defectives in a test is between the thresholds, the outcome is positive/negative with probability 0.5. Setting $u = o(d)$ and error precision $\epsilon > 0$, Chan et al. [8] achieved a small number of tests $O\left(\ln(1/\epsilon)\cdot d\sqrt{\ell\ln n}\right)$ while Reisizadeh et al. [25] attained $\Theta(\sqrt{u}d\ln^3 n)$ tests. When a linear model is applied to the gap, i.e., the number of defectives in a test is between the thresholds, the probability of a positive outcome linearly increases with the number of defectives. The number of tests with a linear model is $O(g^2 n\ln n + n\ln(1/\epsilon))$ [8].

Once $g = 0$, D'yachkov et al. [10] and Cheraghchi [6] 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 [11] 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 gap g, 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. [25]: $O(u^{1.5}d\ln^4 n)$. 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. [8] 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 constraints: 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. [8] and Reisizadeh et al. [25] impractical. Like Chen and Fu [9], 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 algorithms (summarized in Theorems 6, 7, 8) are much lower than the one proposed by Chen and Fu [9] (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 $w = \left(\left\lfloor\frac{|S|}{\ell+1}\right\rfloor + u - 1\right)g$. 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 frequently used notations are listed in Table II.

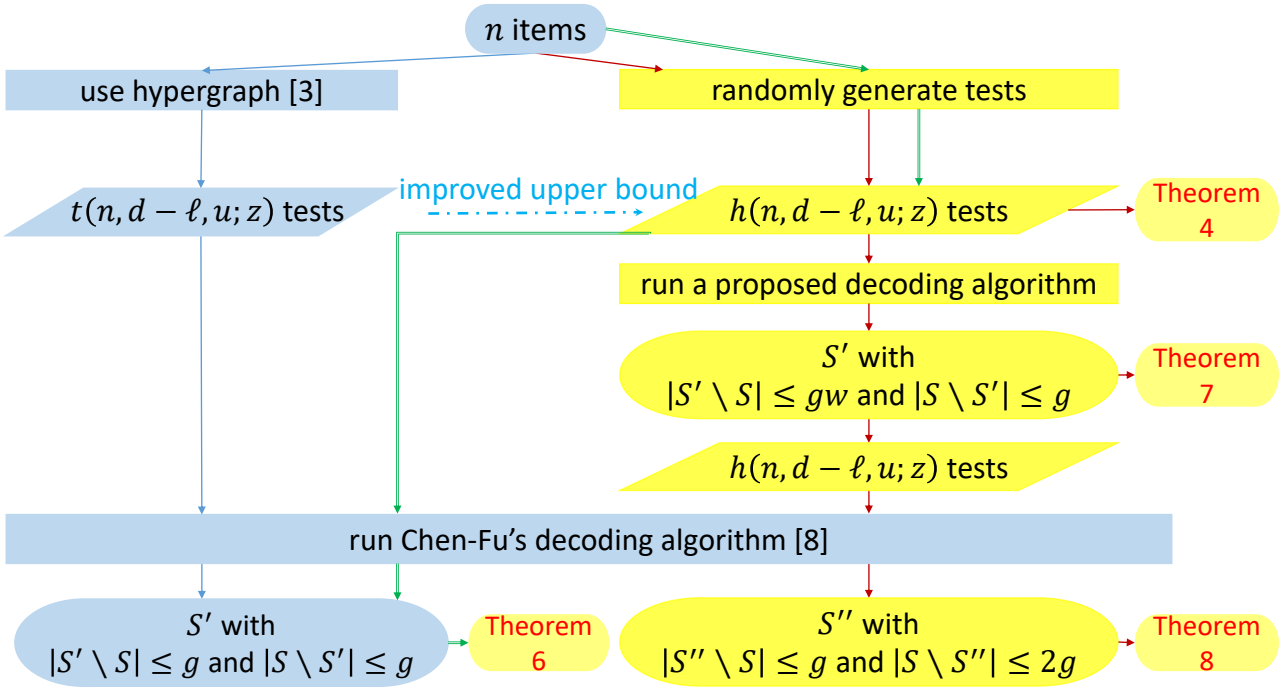


Fig. 2: Flow chart illustrating how contributions were attained in this work (excluding Theorem 3). Flow is from top to bottom. Each output can be reached by following consistent arrow color. To avoid misunderstanding, S'' is used instead of S' for Theorem 8. Both notations represent approximate defective sets recovered after running decoding algorithms. Set S is the true defective set. Parameters $t(n, d-\ell, u; z]$, $h(n, d-\ell, u; z]$, and w are defined in Table I.

B. 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.

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 $|\mathbf{x}| = \sum_{i=1}^{n} x_i \leq d$. 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

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

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 $\mathcal{T}$.

C. Disjunct matrices

Disjunct matrices are a powerful tool to tackle the threshold group testing problem [6], [9], [11]. They were first introduced by Kautz and Singleton [29] as *superimposed codes* and then generalized by Stinson and Wei [28] and D'yachkov et al. [30]. 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 $\mathcal{M}$ 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.,*

Scheme	No. of defectives	Thresholds	No. of items (n)	Model on gap interval	Error tolerance	Number of tests t	Decoding time (Decoding complexity)	Defective set recovered	Decoding type
Ahlswede et al. [24]	$d=\ell+u$	$\ell<u\leq d$	$\geq d$	No	×	$O(u2^{2u}\log n)$	×	×	×
Chen et al. [4]	$\leq d$	$\ell<u\leq d$	$\geq d$	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 [6]	$\leq d$	$\ell<u\leq d$	$\geq d$	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)	$\leq d$	$\ell<u\leq d$	$\geq\frac{(d+u)^2}{u}$	**No**	z	$h(n,d-\ell,u;z]$ $=O\left((1+\frac{z}{\alpha})\cdot\left(\frac{k}{u}\right)^u\left(\frac{k}{d-\ell}\right)^{d-\ell}k\ln\frac{n}{k}\right)$	×	×	×
Chan et al. [8]	$d=o(n)$	$\ell<u=o(d)$	$\omega(d)$	Bernoulli Linear	×	$O\left(\ln\frac{1}{\epsilon}\cdot d\sqrt{\ell}\ln n\right)$ $O(g^2d\ln n+d\ln\frac{1}{\epsilon})$	$O(n\ln n+n\ln\frac{1}{\epsilon})$ $O(g^2n\ln n+n\ln\frac{1}{\epsilon})$	$S'\equiv S$	Rnd.
Reisizadeh et al. [25]	$d=O(n^\beta)$ for $0<\beta<1$	$\ell<u=o(d)$	$O(d^{1/\beta})$	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	$S'\equiv S$	Rnd.
Chen and Fu [9] (**more accurate** in Theorem 3)	$\leq d$	$\ell<u\leq d$	$\geq\frac{(d+u)^2}{u}$	No	z	$t(n,d-\ell,u;z]$	$O\left(t(n,d-\ell,u;z]\times u\left(\binom{n}{u}+(d-u)\binom{n-u}{g+1}\binom{d-1}{g}\binom{d}{u}\right)\right)$	$\lvert S'\setminus S\rvert\leq g$ $\lvert S\setminus S'\rvert\leq g$	Det.
Proposed 1 (Theorem 6)	$\leq d$	$\ell<u\leq d$	$\geq\frac{(d+u)^2}{u}$	**No**	z	$h(n,d-\ell,u;z]$	$O\left(h(n,d-\ell,u;z]\times u\left(\binom{n}{u}+(d-u)\binom{n-u}{g+1}\binom{d-1}{g}\binom{d}{u}\right)\right)$	$\lvert S'\setminus S\rvert\leq g$ $\lvert S\setminus S'\rvert\leq g$	**Det.**
Proposed 2 (Theorem 7)	$\leq d$	$\ell<u<d$	$\geq\frac{e^2(d+u)^2}{u}$	**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)	$\leq d$	$\ell<u<d$	$\geq\frac{e^2(d+u)^2}{u}$	**No**	z	$h(n,d-\ell,u;z]$	$O\left(h(n,d-\ell,u;z]\cdot u\cdot\left(\binom{n}{u}+(d-u)\binom{w+d-u}{g+1}\binom{d-1}{g}\binom{d}{u}\right)\right)$	$\lvert S'\setminus S\rvert\leq g$ $\lvert S\setminus S'\rvert\leq 2g$	**Det.**

TABLE I: Comparison of proposed theorems with previous ones. A × symbol 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 true defective set, respectively. We define $k=d-\ell+u$, $\alpha=k\ln\frac{en}{k}+u\ln\frac{ek}{u}$, $w=(\lfloor|S|/(\ell+1)\rfloor+u-1)\,g$, and $0\leq p<1$. Parameters $t(n,d-\ell,u;z]$ and $h(n,d-\ell,u;z]$ are defined in rows 2 and 4 as well as in (2) and (6), respectively.

Notation	Description
n	Number of items
d	Maximum number of defective items
$\mathbf{x}=(x_1,\ldots,x_n)^T$	Binary representation of n items
ℓ	Lower bound in non-adaptive (n,d,ℓ,u)-TGT model
u	Upper bound in non-adaptive (n,d,ℓ,u)-TGT model
$g=u-\ell-1$	Gap between ℓ and u
$S=\{j_1,j_2,\ldots,j_{\lvert S\rvert}\}$	Set of defective items; cardinality of S is $\lvert S\rvert\leq d$
$N=[n]=\{1,\ldots,n\}$	Set of n items
$\otimes_{\ell,u}$	Operation related to non-adaptive (n,d,ℓ,u)-TGT (to be defined later)
$\mathcal{T}_{i,*}$	Row i of matrix $\mathcal{T}$
$\mathcal{T}_{*,j}$	Column j of matrix $\mathcal{T}$
$\mathcal{M}_{i,*}$	Row i of matrix $\mathcal{M}$
$\mathcal{M}_{*,j}$	Column j of matrix $\mathcal{M}$

TABLE II: Notations frequently used in this paper.

$\left\lfloor\frac{(z-1)}{2}\right\rfloor$ *is usually referred to as the error tolerance.*

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

$$\mathcal{M}=\begin{bmatrix} \ldots & \overbrace{\ldots & \ldots}^{r} & \ldots & \overbrace{\ldots & \ldots}^{d} & \ldots \\ \ldots & 1 & 1 & \ldots & 0 & 0 & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ \ldots & 1 & 1 & \ldots & 0 & 0 & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \end{bmatrix}\begin{matrix} \\ \text{the 1st specific row} \\ \\ \text{the } z\text{th specific row} \\ \\ \\ \end{matrix}$$

Chen et al. [4] gave an upper bound on the number of rows for $(n,d,u;z]$-disjunct matrices as follows.

Theorem 1. *[4, 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*

$$\begin{aligned} t(n,d,u;z] &= z\left(\frac{k}{u}\right)^u\left(\frac{k}{d}\right)^d\left[1+k\left(1+\ln\left(\frac{n}{k}+1\right)\right)\right] \\ &= O\left(z\left(\frac{k}{u}\right)^u\left(\frac{k}{d}\right)^d k\ln\frac{n}{k}\right) \\ &= O(z\cdot t(n,d,u;1]). \end{aligned} \tag{2}$$

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

A. Preliminaries

To clarify the basis of our proposed algorithms, we review Chen and Fu's work [9] which is the first and only work tackling the "for all" model in NATGT with a gap and with a decoding algorithm. They proposed schemes for finding the defective items using $t(n,d-\ell,u;z]$ tests in time $O(n^u\ln n)$. However, the decoding time becomes impractical as n or u increases. The intuition of Chen and Fu algorithm is to initialize an approximate S' of size u such that the outcome of the test on S' is positive. The algorithm then proceeds to increase the size of S' such that the cardinality of S' is not larger than the maximum number of defectives, i.e., d, and the outcome of a test on every subset of u items in S' is positive.

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. Let $[n] = \{1, 2, \ldots, n\}$ be vertex set V. Suppose that a set of edges F contains all u-subsets of S and a fraction of all or all u-subsets of every S'. We can convert threshold group testing with a gap into the problem of reconstructing a hidden graph H in $\mathbb{H} = (V, \mathsf{F})$ that is u-complete with respect to some 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 [3].

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. In other words, the number of false positives and false negatives are up to g.

Given measurement matrix $\mathcal{M}$, Chen and Fu supposed 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, the authors obtained 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. 3. More precisely, the algorithm first initializes set S_1 consisting of the u vertices belonging to an edge of the family F. A new set S_{i+1} is then created such that $|S_{i+1}| = |S_i| + 1$. Set S_{i+1} is made equal to set $(S_i \cup A_i) \setminus B_i$ by selecting set A_i of $g+1$ elements in $V \setminus S_i$ and set B_i of g elements in S_i such that $\mathbb{H}$ is u-complete with respect to $(S_i \cup A_i) \setminus B_i$. 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. *[9, 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*

$$t(n, d-\ell, u; z] \times u\binom{n}{u} + (d-u)\binom{n-u}{g+1}\binom{d-1}{g}\binom{d}{u} \tag{3}$$

Algorithm 1 [Algorithm 2 [9]] $\mathrm{Decoding}_1(\mathbf{y}, \mathcal{M})$: Decoding procedure for non-adaptive (n, d, ℓ, u)-TGT with up to e erroneous outcomes.

Input: Outcome vector $\mathbf{y}$, a $(n, d-\ell, u; z = 2e+1]$-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 $(g+1)$-subset A_i in $V(\mathbb{H}) \setminus S_i$ and a g-subset 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 $S' = S_m$.

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

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

The complexity of the theorem above is attained by taking the 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 *inaccurate* in general. A detailed analysis is given in the Appendix. Here we present a more accurate version of Theorem 2.

Theorem 3 (A more accurate version of Theorem 4.4 in [9])**.** *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*

$$\begin{aligned} &O\left(t(n, d-\ell, u; z] \times u\left(\binom{n}{u}\right.\right. \\ &\left.\left. +(d-u)\binom{n-u}{g+1}\binom{d-1}{g}\binom{d}{u}\right)\right) \\ =&O\left(z\left(\frac{k}{u}\right)^u \left(\frac{k}{d-\ell}\right)^{d-\ell} k \ln\frac{n}{k}\right. \\ &\left. \times u\left(\binom{n}{u} + (d-u)\binom{n-u}{g+1}\binom{d-1}{g}\binom{d}{u}\right)\right), \end{aligned} \tag{4}$$

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

C. Example for Algorithm 1

In this section, we demonstrate Algorithm 1 by setting $n = 6, d = 4, \ell = 0, u = 2$, and $z = 1$. This means that $g = u - \ell - 1 = 1$ and $e = 0$. We assume that the defective items are 1, 2, 4, and 5; i.e., the input vector is $\mathbf{x} = (1, 1, 0, 1, 1, 0)^T$. The true defective set is therefore $S = \{1, 2, 4, 5\} = \mathsf{supp}(\mathbf{x})$.

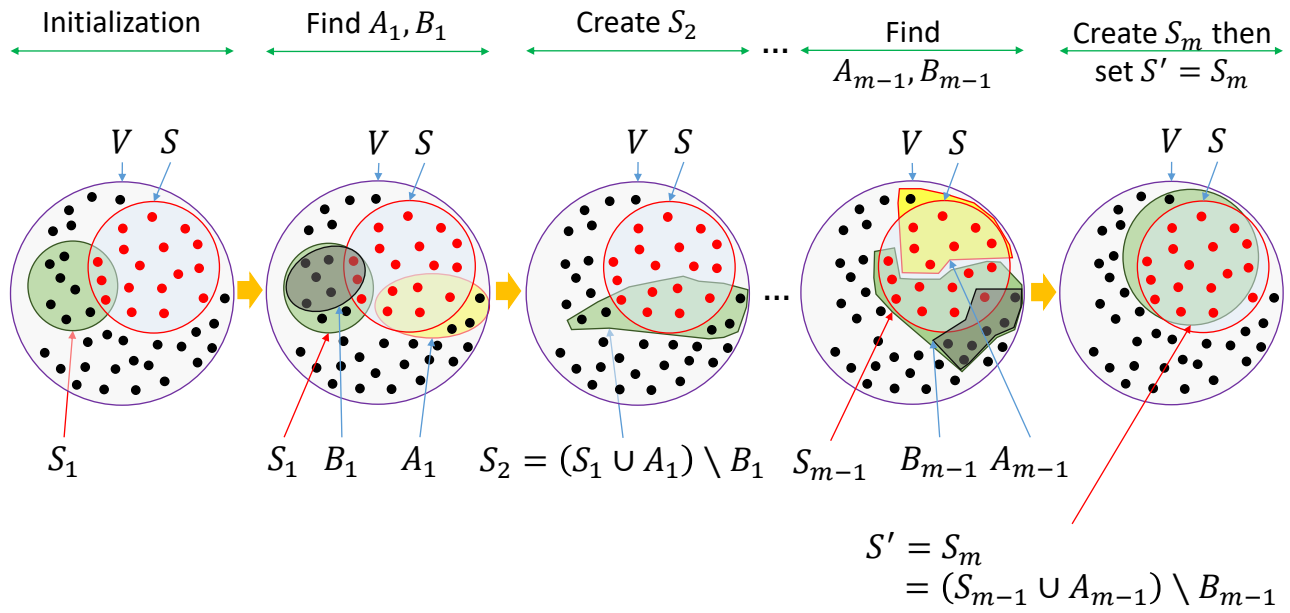


Fig. 3: Illustration of finding an approximate defective set S' of the defective set S such that $|S' \setminus S| \leq g$ and $|S \setminus S'| \leq g$ for Algorithm 1. We set $g = 7, u = 10$, and $\ell = u - g - 1 = 2$.

The $(n = 6, d - \ell = 4, u = 2; z = 1]$-disjunct matrix $\mathcal{M}$ is as follows:

$$\mathcal{M} = \begin{bmatrix} 1&1&0&0&0&0 \\ 1&0&1&0&0&0 \\ 1&0&0&1&0&0 \\ 1&0&0&0&1&0 \\ 1&0&0&0&0&1 \\ 0&1&1&0&0&0 \\ 0&1&0&1&0&0 \\ 0&1&0&0&1&0 \\ 0&1&0&0&0&1 \\ 0&0&1&1&0&0 \\ 0&0&1&0&1&0 \\ 0&0&1&0&0&1 \\ 0&0&0&1&1&0 \\ 0&0&0&1&0&1 \\ 0&0&0&0&1&1 \\ 1&0&1&0&1&0 \\ 0&1&1&0&1&1 \\ 0&0&1&1&0&1 \\ 1&0&1&1&0&1 \\ 1&0&1&0&0&1 \end{bmatrix}, \mathbf{y} = \mathcal{M} \otimes_{0,2} \mathbf{x} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 0 \\ 1 \\ 1 \\ 1 \\ 0 \\ 1 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 1 \\ 1 \\ 0 \\ 1 \\ 0 \end{bmatrix}. \quad (5)$$

We assume that the observed vector is $\mathbf{y}$, as in (5). Algorithm 1 proceeds as follows. In Step 1, hypergraph $\mathbb{H} = (V, \mathsf{F})$ is constructed with the set of vertexes $V = [6] = \{1, 2, 3, 4, 5, 6\}$. A search is made for all 2-subsets $X \in V$ in order to form the set of edges F such that $t_0^{\mathcal{M}}(X) \leq e = 0$. From (5), we get $\mathsf{F} = \{\{1,2\}, \{1,4\}, \{1,5\}, \{2,3\}, \{2,4\}, \{2,5\}, \{3,5\}, \{4,5\}, \{5,6\}\}$[1].

Step 2 starts with an initial 2-subset $S_1 = \{1, 2\}$. All possible cases are checked to obtain some 2-subset A_1 $(g + 1 = 2)$ in $V \setminus S_1 = \{3, 4, 5, 6\}$, which is some element of $\{\{3,4\}, \{3,5\}, \{3,6\}, \{4,5\}, \{4,6\}, \{5,6\}\}$, and a 1-subset B_1 $(g = 1)$ in S_1, which is some element of $\{\{1\}, \{2\}\}$, such that $\mathbb{H}$ is 2-complete with respect to $(S_1 \cup A_1) \setminus B_1$. Since $A_1 = \{3, 5\}$ and $B_1 = \{1\}$ ensure that the condition holds, set $S_2 = (S_1 \cup A_1) \setminus B_1 = \{2, 3, 5\}$.

Since $|S_2| = 3 < 4 = d$, we continue Step 2 by choosing a 2-subset $A_2 \subseteq V \setminus S_2 = \{1, 4, 6\}$ and a 1-subset $B_2 \subseteq S_2$. The lists of potential subsets for A_2 and B_2 are $\{\{1,4\}, \{1,6\}, \{4,6\}\}$ and $\{\{2\}, \{3\}, \{5\}\}$, respectively. We choose $A_2 = \{1, 4\}$ and $B_2 = \{3\}$ because $\mathbb{H}$ is 2-complete with respect to $(S_2 \cup A_2) \setminus B_2$. Set $S_3 = (S_2 \cup A_2) \setminus B_2 = \{1, 2, 4, 5\}$. Since $|S_3| = 4 \geq 4 = d$, the algorithm stops and output $S' = S_3 = \{1, 2, 4, 5\}$. In this case, the approximate defective set S' is identical to the true defective set S.

IV. Improved Upper Bounds on the Number of Tests for Disjunct Matrix

In this section, we present better exact upper bounds on the number of tests compared to the one in Theorem 1.

A. First result

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.

[1]We delineate this example to ensure understanding. Since we use $u = 2$, hypergraph $\mathbb{H}$ becomes a normal graph in which an edge consists of two vertices. However, once $u \geq 3$, an edge in $\mathbb{H}$ consists of at least three vertices. Graph $\mathbb{H}$ is thus no longer a normal graph.

Theorem 4. *Let $2 \le u \le d < k = d+u \le n$ be integers with $(d+u)^2/u \le n$. 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*

$$h(n, d, u; z] = O\left(\left(1 + \frac{z}{\alpha}\right) \cdot \left(\frac{k}{u}\right)^u \left(\frac{k}{d}\right)^d k \ln \frac{n}{k}\right). \quad (6)$$

Proof: Consider a randomly generated $h \times n$ matrix $\mathcal{G} = (g_{ij})_{1 \le i \le h, 1 \le j \le n}$ 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 $|S_2| = d$, we 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 by *a good event*. The probability that the good event happens is:

$$q = p^u (1-p)^d. \quad (7)$$

Set $\alpha = k \ln \frac{en}{k} + u \ln \frac{ek}{u}$ and $\beta = 1 - 2/\alpha$. It is obvious that $0 < \alpha, \beta$. We then set $z = (1-\delta)qh$, where $0 < \delta < 1$. We will later prove that there always exists δ 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 $\mathsf{X}_i = 0$ 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. We get $\mu = E[\mathsf{X}] = \sum_{i=1}^h E[\mathsf{X}_i] = qh$.

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

$$\begin{aligned}\Pr[\mathsf{X} \le z] &= \Pr[\mathsf{X} \le (1-\delta)\mu] \\ &\le \exp\left(-\frac{\delta^2 \mu}{2}\right) = \exp\left(-\frac{\delta^2 qh}{2}\right).\end{aligned}$$

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

$$\begin{aligned}g(p, h, u, d, n) &= \binom{n}{d+u}\binom{d+u}{u} \Pr[\mathsf{X} \le z] \\ &\le \binom{n}{k}\binom{k}{u} \exp\left(-\frac{\delta^2 qh}{2}\right). \quad (8)\end{aligned}$$

To ensure the existence of 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

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

For this to hold, it suffices that

$$\begin{aligned}&\binom{n}{k}\binom{k}{u} \le \left(\frac{en}{k}\right)^k \left(\frac{ek}{u}\right)^u < \exp\left(\frac{\delta^2 qh}{2}\right) \quad (9) \\ \iff \quad &h > \frac{2}{\delta^2} \cdot \frac{1}{q} \cdot \left(k \ln \frac{en}{k} + u \ln \frac{ek}{u}\right) \\ \iff \quad &> \frac{2}{\delta^2} \cdot \left(\frac{k}{u}\right)^u \left(\frac{k}{d}\right)^d \left(k \ln \frac{en}{k} + u \ln \frac{ek}{u}\right). \quad (10)\end{aligned}$$

In the above, we have (9) because $\binom{a}{b} \le \left(\frac{ea}{b}\right)^b$. Since $p = \frac{u}{k}$, from (10), if we set

$$\begin{aligned}h &= h(n, d, u; z] \\ &= \frac{3}{\delta^2} \cdot \frac{1}{q} \cdot \left(k \ln \frac{en}{k} + u \ln \frac{ek}{u}\right) \\ &= \frac{3}{\delta^2} \cdot \frac{1}{q} \cdot \alpha, \text{ where } \alpha = k \ln \frac{en}{k} + u \ln \frac{ek}{u}, \quad (11) \\ &= \frac{3}{\delta^2} \cdot \left(\frac{k}{u}\right)^u \left(\frac{k}{d}\right)^d \cdot \left(k \ln \frac{en}{k} + u \ln \frac{ek}{u}\right),\end{aligned}$$

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 (11), we have:

$$z = (1-\delta)qh = (1-\delta) \cdot \frac{3\alpha}{\delta^2} \quad (12)$$

$$\iff \quad z\delta^2 + 3\alpha\delta - 3\alpha = 0 \quad (13)$$

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

$$\delta = \frac{-3\alpha + \sqrt{9\alpha^2 + 12\alpha z}}{2z} = \frac{\sqrt{3\alpha}\left(\sqrt{3\alpha + 4z} - \sqrt{3\alpha}\right)}{2z}. \quad (14)$$

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

$$\begin{aligned}f(3\alpha + 4z) - f(3\alpha) &= \sqrt{3\alpha + 4z} - \sqrt{3\alpha} \\ &= 4z \cdot f'(b) = 4z \cdot \frac{1}{2\sqrt{b}} = \frac{2z}{\sqrt{b}}. \quad (15)\end{aligned}$$

Combine with (14), we get

$$\delta = \frac{\sqrt{3\alpha}\left(\sqrt{3\alpha + 4z} - \sqrt{3\alpha}\right)}{2z} = \frac{\sqrt{3\alpha}}{2z} \cdot \frac{2z}{\sqrt{b}} = \sqrt{\frac{3\alpha}{b}}. \quad (16)$$

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

$$\frac{1}{\delta^2} = \frac{b}{3\alpha} \in \left(1, 1 + \frac{4z}{3\alpha}\right). \quad (17)$$

Therefore, the number of tests required is

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

■

Since α is always larger than $4/3$, $4z/(3\alpha)$ 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.

Discussion of number of tests for TGT and CGT: With the same settings for n, d, and the maximum number of erroneous outcomes $\lfloor (z-1)/2 \rfloor$, what are the similarities and differences for the number of tests between TGT and CGT? We first transform (6):

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

In Corollary 19 [18], Ngo, Porat, and Rudra show that the number of tests needed to handle $\lfloor (z-1)/2 \rfloor$ erroneous outcomes is $O(d^2 \log n) + O(zd) = t(n,d,1;1] + O\left(\frac{z}{d \log n}\right) \cdot t(n,d,1;1]$. It is well known that $t(n,d,1;1] = O(d^2 \log n)$ is the achievable bound on the number of tests for the noiseless setting ($z = 1$). The authors prove that we only need $O\left(\frac{z}{d \log n}\right) \cdot t(n,d,1;1]$ additional tests to handle up to $\lfloor (z-1)/2 \rfloor$ erroneous outcomes instead of using $z \times t(n,d,1;1]$. The result for Theorem 4 shares this property. Since $t(n,d,u;1]$ is the achievable number of tests for the noiseless setting in TGT, we need only $O\left(\frac{z}{\alpha}\right) \cdot t(n,d,u;1]$ additional tests to handle up to $\lfloor (z-1)/2 \rfloor$ erroneous outcomes instead of $z \times t(n,d,u;1]$ as in Theorem 1.

B. Second result

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

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

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

Proof: By using the same construction and arguments in the proof in Theorem 4 until (10), if we set

$$\begin{aligned}
h &= h(n,d,u;z] \\
&= \left\lfloor \frac{2}{\delta^2} \cdot \frac{1}{q} \cdot \left(k \ln \frac{en}{k} + u \ln \frac{ek}{u}\right) \right\rfloor + 1 \\
&= \left\lfloor \frac{2}{\delta^2} \cdot \left(\frac{k}{u}\right)^u \left(\frac{k}{d}\right)^d \cdot \left(k \ln \frac{en}{k} + u \ln \frac{ek}{u}\right) \right\rfloor + 1 \\
&= O\left(\frac{1}{\delta^2} \cdot \left(\frac{k}{u}\right)^u \left(\frac{k}{d}\right)^d \cdot k \ln \frac{n}{k}\right) \\
&= O\left(\frac{1}{(1-\delta) k \ln \frac{n}{k}} \cdot z \left(\frac{k}{u}\right)^u \left(\frac{k}{d}\right)^d \cdot k \ln \frac{n}{k}\right) \qquad (18) \\
&= O\left(\frac{1}{(1-\delta) k \ln \frac{n}{k}}\right) \cdot t(n,d,u;z],
\end{aligned}$$

then $g(p,h,u,d,n) < 1$, where $t(n,d,u;z]$ is defined in (2); i.e., there exists an $(n,d,u;z]$-disjunct matrix of size $h \times n$. Equation (18) is obtained because

$$\begin{aligned}
&\frac{2(1-\delta)}{\delta^2} \cdot \left(k \ln \frac{en}{k} + u \ln \frac{ek}{u}\right) \\
\le z &= (1-\delta) q h \qquad (19) \\
&= (1-\delta) q \left(\left\lfloor \frac{2}{\delta^2} \cdot \frac{1}{q} \cdot \left(k \ln \frac{en}{k} + u \ln \frac{ek}{u}\right) \right\rfloor + 1\right) \\
&= \Theta\left(\frac{1-\delta}{\delta^2} \cdot k \ln \frac{n}{k}\right) \\
&\le \frac{2(1-\delta)}{\delta^2} \cdot \left(k \ln \frac{en}{k} + u \ln \frac{ek}{u}\right) + 1. \qquad (20)
\end{aligned}$$

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

$$\begin{aligned}
&h(n,d,u;z] \\
=&\left\lfloor \frac{2}{\delta^2} \cdot \frac{1}{q} \cdot \left(k \ln \frac{en}{k} + u \ln \frac{ek}{u}\right) \right\rfloor + 1, \text{ where } \frac{1}{q} = \left(\frac{k}{u}\right)^u \left(\frac{k}{d}\right)^d \\
\le&\frac{2}{\delta^2} \cdot \frac{1}{q} \cdot \left(k \ln \frac{en}{k} + u \ln \frac{ek}{u}\right) + 1 \\
<&\frac{2}{\delta^2} \cdot \frac{1}{q} \cdot 2k \ln \frac{n}{k}. \qquad (21)
\end{aligned}$$

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

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

which is derived from the condition in (19). Combining (22) and (21), we always get $h(n,d,u;z] < t(n,d,u;z]$ if

$$\begin{aligned}
\frac{2}{\delta^2} \cdot \frac{1}{q} \cdot 2k \ln \frac{n}{k} \le& \frac{2(1-\delta)}{\delta^2} \cdot \Big(k \ln \frac{en}{k} \\
&+ u \ln \frac{ek}{u}\Big) \cdot \frac{1}{q} \cdot k \ln \frac{n}{k} \\
\iff \qquad \delta \le& 1 - \frac{2}{k \ln \frac{en}{k} + u \ln \frac{ek}{u}} = \beta.
\end{aligned}$$

Since $0 < \delta \le \beta = 1 - 2/\alpha$, the quantity $2(1-\delta)/\delta^2 \cdot \alpha$ goes from $4/\beta^2$ to infinity. Moreover, from (19) and (20), we have $z \in \left[\frac{2(1-\delta)}{\delta^2} \cdot \alpha, \frac{2(1-\delta)}{\delta^2} \cdot \alpha + 1\right]$, where $\alpha = k \ln \frac{en}{k} + u \ln \frac{ek}{u}$. Therefore, z can range from $\lceil 4/\beta^2 \rceil$ to $+\infty$. In other words, for any integer $z \ge 4/\beta^2 + 1$, we can find a corresponding δ in the interval $(0, \beta]$ such that $z = (1-\delta) q h$. ■

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

Here we present a reduced exact upper bound on the number of tests and a reduced asymptotic bound on the decoding time for identifying defective items in a noisy setting on test outcomes compared with the state-of-the-art work of Chen and Fu [9].

A. First proposed algorithm

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 [9], 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 $\alpha = k \ln \frac{en}{k} + u \ln \frac{ek}{u}$. 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* (6)*. Moreover, the decoding complexity is*

$$O\left(h(n, d-\ell, u; z] \times u \left(\binom{n}{u} + (d-u)\binom{n-u}{g+1}\binom{d-1}{g}\binom{d}{u}\right)\right). \tag{23}$$

B. Second proposed algorithm

We can see that the complexity of the decoding algorithm in the theorem above remains relatively high due to the second operator in (23). To reduce the decoding complexity, one can relax the conditions on $|S' \setminus S|$ but keep the same condition on $|S \setminus S'|$. In other words, we accept more false positives while keeping the same condition on the maximum number of false negatives.

The main idea is to reduce the redundancy of u-subsets created by the $(n, d-\ell, u; z]$-disjunct matrix in Algorithm 1. Since every u-subset $X^+ \subseteq S$ must be in F, the total number of such X^+ is $\binom{|S|}{u}$. In fact, we need only up to $\zeta = \left\lfloor \frac{|S|}{u} \right\rfloor$ disjoint u-subsets X^+s in F to form S if $|S|$ is divisible by u. Therefore, we can use a simple procedure, i.e., collect ζ disjoint u-subsets in F, to form S. However, it is uncertain whether each u-subset we collected is truly a u-subset of S because it may contain only $\ell+1$ defective items. Moreover, $|S|$ may not be divisible by u. As a result, the set formed, says S', may not be identical to S. To remedy this drawback, we propose adding one more step: add the remaining defective items in $S \setminus S'$ into S' until S' is not extendible.

The above strategy is formalized in the following theorem which is associated with Algorithm 2.

Theorem 7. *Let $\ell, 0 < g, 2 \le u = \ell + g + 1 < d < k = d - \ell + u \le n$ be integers with $\mathrm{e}^2(d+u)^2/u \le n$. Set $\alpha = k \ln \frac{en}{k} + u \ln \frac{ek}{u}$. 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 $|S' \setminus S| \le \left(\left\lfloor \frac{|S|}{\ell+1} \right\rfloor + u - 1\right) g \le \left(\frac{d}{\ell+1} + u - 1\right) g$ 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* (6)*. Moreover, the decoding complexity is*

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

Algorithm 2 $\mathrm{Decoding}_2(\mathbf{y}, \mathcal{M})$: 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. $|S' \setminus S| \le \left(\left\lfloor \frac{|S|}{\ell+1} \right\rfloor + u - 1\right) g$ and $|S \setminus S'| \le g$.

1: Construct a family F such that a u-subset $X \subseteq [n]$ is an edge in F iff $t_0^{\mathcal{M}}(X) \le e$, 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 items more than S_i. As an initial S_1, we select all u vertices of an arbitrary edge. To find S_{i+1} for $i \ge 1$, 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 $i \ge r$, we check all possible cases to attain some u-subset $A_i \in \mathsf{F} \setminus \{A_1, \ldots, A_{i-1}\}$ such that $|S_i \cup A_i| \ge |S_i| + g + 1$. 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 $S' = S_m$.

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.

1) Visualization: Steps 2 and 3 of Algorithm 2 are depicted in Fig. 4 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, i.e., $|X \cap S| \geq \ell + 1$. Moreover, every u-subset $X^+ \subseteq S$ is in F. Since the proof is identical to the proof of Lemma 4.1 in [9], we omit it here.

Since every u-subset $X^+ \subseteq S$ must be in F, there exists $\left\lfloor \frac{|S|}{u} \right\rfloor \leq \zeta \leq \left\lceil \frac{|S|}{u} \right\rceil$ disjoint u-subsets $X_1^+, \ldots, X_\zeta^+$ in F such that $|S \setminus \cup_{j=1}^{\zeta} X_j^+| \leq u - 1$.

In Step 2, if another disjoint u-subset $A_1 \in \mathsf{F}$ ($|S_1 \cap A_1| = 0$) is found, set $S_2 = S_1 \cup A_1$. 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, 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 $|S \setminus S_r| \leq u - 1$. We 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 $|S_r \cup A_r| = |S_r| + |A_r| = |S_r| + u$; 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 $|S' \setminus S| \leq \left(\left\lfloor \frac{|S|}{\ell+1} \right\rfloor + u - 1 \right) g$ and $|S \setminus S'| \leq g$ when the algorithm terminates.

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 $|S' \setminus S| \leq \left(\left\lfloor \frac{|S|}{\ell+1} \right\rfloor + u - 1 \right) g$ and $|S \setminus S'| \leq g$, we have $|S'| \leq \left(\left\lfloor \frac{|S|}{\ell+1} \right\rfloor + u - 1 \right) g + d$. 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 $|S_i| \leq |S'| \leq \left(\left\lfloor \frac{|S|}{\ell+1} \right\rfloor + u - 1 \right) g + d$, the complexity of Algorithm 2 is:

$$\begin{aligned}
&h(n, d-\ell, u; z] \cdot u\binom{n}{u} \\
&+ \left(\left\lfloor \frac{|S|}{\ell+1} \right\rfloor + u - 1 \right) \left(\left(\left\lfloor \frac{|S|}{\ell+1} \right\rfloor + u - 1 \right) g + d \right) \times uf \\
&= h(n, d-\ell, u; z] \cdot u\binom{n}{u} + us(gs + d)f, \qquad (24)
\end{aligned}$$

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

$$\begin{aligned}
us(gs + d)f &\leq u(d + u)(g(d + u) + d)\binom{n}{u} \\
&< (d + u)^2(g + 1) \cdot u\binom{n}{u}, \qquad (25)
\end{aligned}$$

and

$$\begin{aligned}
&h(n, d-\ell, u; z] \cdot u\binom{n}{u} \\
&\geq \left(1 + \frac{d-\ell}{u}\right)^u \left(1 + \frac{u}{d-\ell}\right)^{d-\ell} (d + g + 1) \ln \frac{n}{k} \cdot u\binom{n}{u} \\
&\geq 4(g + 1) \left(1 + \frac{d-\ell}{u}\right)^u \left(1 + \frac{u}{d-\ell}\right)^{d-\ell} \cdot u\binom{n}{u}, \qquad (26)
\end{aligned}$$

because $h(n, d-\ell, u; z] \geq \left(1 + \frac{d-\ell}{u}\right)^u \left(1 + \frac{u}{d-\ell}\right)^{d-\ell} (d + g + 1) \ln \frac{n}{k}$ as in Theorem 4, $d \geq u \geq g + 1$ 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:

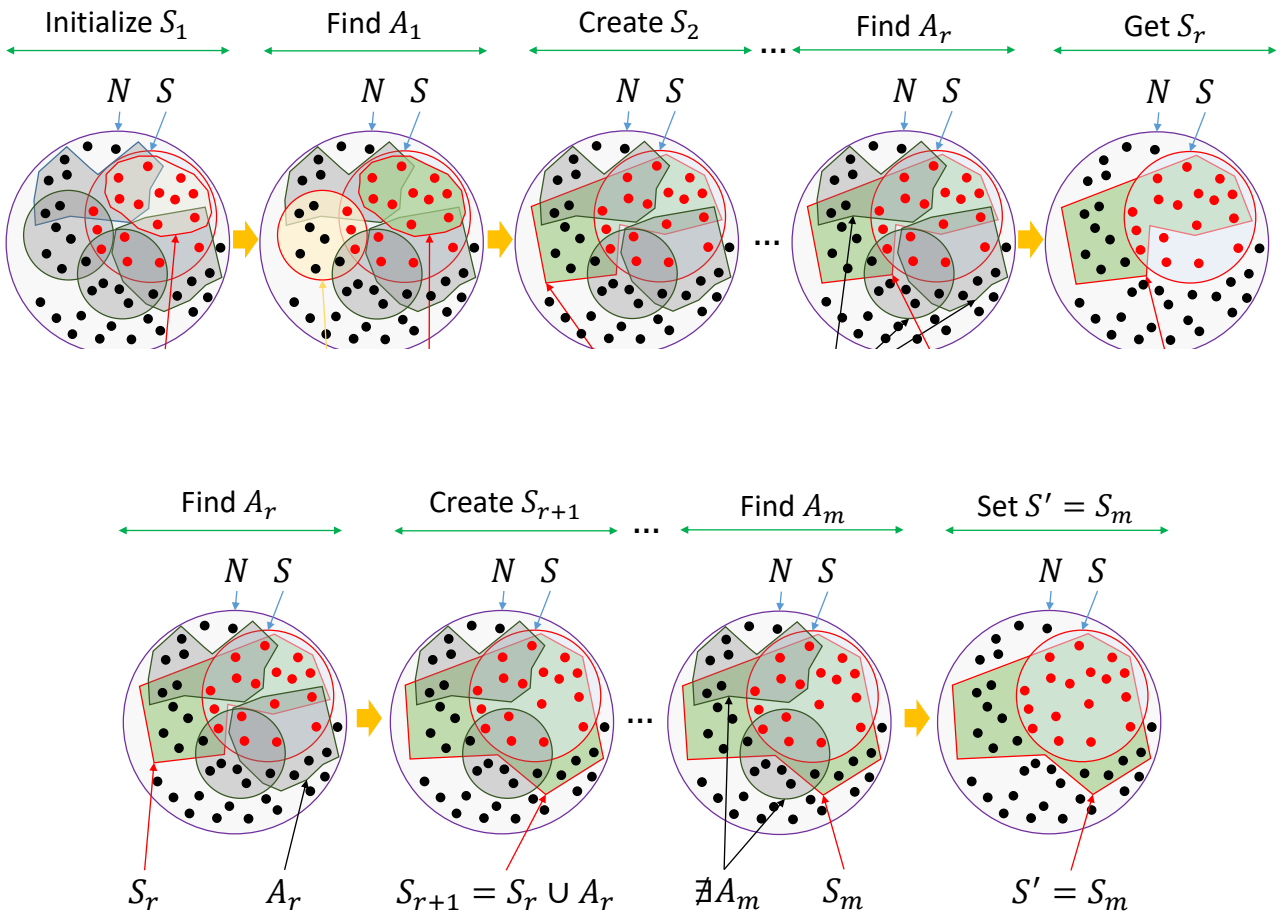
$$\begin{aligned}
(d + u)^2(g + 1) \cdot u\binom{n}{u} \leq& 4(g + 1) \left(1 + \frac{d-\ell}{u}\right)^u \\
&\times \left(1 + \frac{u}{d-\ell}\right)^{d-\ell} \cdot u\binom{n}{u} \qquad (27) \\
\iff d + u \leq& 2 \left(1 + \frac{d-\ell}{u}\right)^{u/2} \left(1 + \frac{u}{d-\ell}\right)^{(d-\ell)/2}
\end{aligned}$$

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

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

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 (25), (26), and (27), we get

$$us(gs + d)f \leq h(n, d-\ell, u; z] \cdot u\binom{n}{u},$$



(b) Step 3

Fig. 4: Illustration of finding an approximate defective set S' of defective set S such that $|S' \setminus S| \leq \left(\left\lfloor \frac{|S|}{\ell+1} \right\rfloor + u - 1\right) g$ and $|S \setminus S'| \leq g$ with $g = 7, u = 10$, and $\ell = u - g - 1 = 2$ for Algorithm 2.

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

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

4) Example for Algorithm 2: We demonstrate Algorithm 2 by the same settings used to demonstrate Algorithm 1 (Section III-C): $n = 6, d = 4, \ell = 0, u = 2, g = u - \ell - 1 = 1, z = 1, e = 0$, and $S = \{1, 2, 4, 5\}$. Input vector $\mathbf{x}$, $(n = 6, d - \ell = 4, u = 2; z = 1]$-disjunct matrix $\mathcal{M}$, and outcome vector $\mathbf{y}$ are as in (5). Note that the condition $\mathrm{e}^2(d+u)^2/u \leq n$ does not hold though Algorithm 2 still works well for this example.

Algorithm 2 proceeds as follows. In Step 1, a family F of 2-subsets $X \subseteq [n]$ is constructed such that $t_0^{\mathcal{M}}(X) \leq e = 0$. From (5), we get $\mathsf{F} = \{\{1,2\},\{1,4\},\{1,5\},\{2,3\},\{2,4\},\{2,5\},\{3,5\},\{4,5\},\{5,6\}\}$.

In Step 2, an initial 2-subset S_1 belonging to F is arbitrarily chosen. Without loss of generality, set $S_1 = \{1, 2\}$. The next phase is to find a 2-subset $A_1 \in \mathsf{F}$ such that $|S_1 \cup A_1| = |S_1| + u = 2+2 = 4$. By exhausted searching in F, one of candidates is $\{3, 5\}$. Set $A_1 = \{3, 5\}$ and $S_2 = S_1 \cup A_1 = \{1, 2, 3, 5\}$. Because there does not exist any $A_2 \in \mathsf{F} \setminus \{A_1\}$ such that $|S_2 \cup A_2| = |S_2| + 2$, Step 2 stops here because S_2 is not extendible.

Since set S_2 returned by Step 2 may not contain some of the defective items when $|S \setminus S_2| > g$, Step 3 exhaustively searches for them in order to produce the final approximate set S', where $|S \setminus S'| \leq g$. It searches for a 2-subset A_2 in $\mathsf{F} \setminus \{A_1\}$ such that $|S_2 \cup A_2| = |S_2| + g + 1 = 3 + 1 + 1 = 5$. Luckily, such an A_2 does not exist, $S' = S_2 = \{1, 2, 3, 5\}$ is output.

Note that the approximate defective set S' is not identical to S as it is in Section III-C. However, set S' is *indistinguishable* from S because $|S \setminus S'| = 1 = g \leq g$ and $|S' \setminus S| = 1 = g \leq g$ [3].

C. Third proposed algorithm

Our main idea here is to combine Algorithms 1 and 2. It is obvious that $|S' \setminus S| \leq \left(\left\lfloor \frac{|S|}{\ell+1} \right\rfloor + u - 1\right) g$ 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 that in Theorem 7 but lower than that in Theorem 6. The conditions on $|S' \setminus S|$ and $|S \setminus S'|$, i.e., the number of false positives and the number of false negatives, 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 $\mathrm{Decoding}_3(\mathbf{y}, \mathcal{M})$: 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 $|S \setminus S'| \leq 2g$.

1: Set $V = \mathrm{Decoding}_2(\mathbf{y}, \mathcal{M})$.
2: Construct hypergraph $\mathbb{H} = (V, \mathsf{F})$ where a u-subset $X \subseteq V$ 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.
3: 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 $(g+1)$-subset A_i in $V(\mathbb{H}) \setminus S_i$ and a g-subset 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', i.e., $|S' \cap S| \geq |S| - 2g$, is at least $|S| - 2g$. We summarize this result as follows.

Theorem 8. *Let $\ell, 0 < g, 2 \leq u = \ell + g + 1 < d < k = d - \ell + u \leq n$ be integers with $\mathrm{e}^2(d+u)^2/u \leq n$. Let z be a positive integer and S be the defective set with $|S| \leq d$. Set $w = \left(\left\lfloor \frac{|S|}{\ell+1} \right\rfloor + u - 1\right) g$ 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* (6)*. Moreover, the decoding complexity is*

$$O\left(h(n, d-\ell, u; z] \cdot u \cdot \left(\binom{n}{u} + (d-u)\binom{w+d-u}{g+1}\binom{d-1}{g}\binom{d}{u}\right)\right). \quad (28)$$

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

1) Correctness: From Theorem 7, we get $|V \setminus S| \leq \left(\left\lfloor \frac{|S|}{\ell+1} \right\rfloor + u - 1\right) g$ 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, it suffices to prove $|P \setminus S'| \leq g$. The property is then straightforwardly attained because $P \subseteq S$ and $|P| \geq |S| - g$. Assume for the sake of contradiction 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 $|V| \leq \left(\left\lfloor \frac{|S|}{\ell+1} \right\rfloor + u - 1\right) g + d = w + d$, 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 second paragraph of Appendix, 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

$$\begin{aligned}
&h(n, d-\ell, u; z] \cdot u\binom{n}{u} + uh(n, d-\ell, u; z] \times \binom{w+d}{u} \\
&+ (d-u)\binom{w+d-u}{g+1}\binom{d-1}{g}\binom{d}{u} \times uh(n, d-\ell, u; z] \\
&= h(n, d-\ell, u; z] \\
&\times u\left(\binom{n}{u} + (d-u)\binom{w+d-u}{g+1}\binom{d-1}{g}\binom{d}{u}\right) \quad (29) \\
&= h(n, d-\ell, u; z] \times u\left(\binom{n}{u}\right. \\
&\left. + (d-u)\binom{\left(\left\lfloor \frac{|S|}{\ell+1} \right\rfloor + u - 1\right) g + d - u}{g+1}\binom{d-1}{g}\binom{d}{u}\right).
\end{aligned}$$

Equation (29) is attained if we suppose that $w + d = \left(\left\lfloor \frac{|S|}{\ell+1} \right\rfloor + u - 1\right) g + d \leq u\left(\frac{d}{\ell+1} + u - 1\right) + d \leq n$. This condition is practical because n is much larger than d.

3) Example for Algorithm 3: We demonstrate Algorithm 3 using the same settings as before: $n = 6, d = 4, \ell = 0, u = 2, g = u - \ell - 1 = 1, z = 1, e = 0$ and $S = \{1, 2, 4, 5\}$. Input vector $\mathbf{x}$, $(n = 6, d-\ell = 4, u = 2; z = 1]$-disjunct matrix $\mathcal{M}$, and outcome vector $\mathbf{y}$ are as in (5). Note that the conditions $\mathrm{e}^2(d+u)^2/u \leq n$ and $w + d = \left(\left\lfloor \frac{|S|}{\ell+1} \right\rfloor + u - 1\right) g + d \leq n$ do not hold though Algorithm 3 still works well for this example.

Algorithm 3 proceeds as follows. In Step 1, the set $($ vertices $V = \{1, 2, 3, 5\}$ is first obtained as described Section V-B4. Our task now is to construct a hypergrap $\mathbb{H} = (V, \mathsf{F})$ using that set. Note that the original set $($ vertices, $[n] = [6] = \{1, 2, 3, 4, 5, 6\}$, is here reduced to V Using the same procedure described in Section III-C, Step searches for all 2-subsets $X \subseteq V$ in order to form a s of edges F such that $t_0^{\mathcal{M}}(X) \leq e = 0$. From (5), we g $\mathsf{F} = \{\{1, 2\}, \{1, 5\}, \{2, 3\}, \{2, 5\}, \{3, 5\}\}$.

Step 3 starts with an initial 2-subset $S_1 = \{1, 2\}$ and checl all possible cases to obtain some 2-subset A_1 in $V \setminus S_1$ $\{3, 5\}$, which is $\{3, 5\}$, and a 1-subset B_1 in S_1, which is some element of $\{\{1\}, \{2\}\}$, such that $\mathbb{H}$ is 2-complete with respect to $(S_1 \cup A_1) \setminus B_1$. Since $A_1 = \{3, 5\}$ and $B_1 = \{1\}$ ensure that the condition holds, set $S_2 = (S_1 \cup A_1) \setminus B_1 = \{2, 3, 5\}$. Next, a 2-subset $A_2 \subseteq V \setminus S_2 = \{1\}$ and a 1-subset $B_2 \subseteq S_2$ are chosen. Since $|V \setminus S_2| = 1 < 2 = u$, there does not exis such an A_2; i.e., S_2 is not extendible. Step 3 thus stops an outputs $S' = S_2 = \{2, 3, 5\}$.

In this example, approximate defective set S' satisfies th two conditions $|S' \setminus S| \leq g$ and $|S \setminus S'| \leq 2g$ in Theorem because $|S' \setminus S| = |\{3\}| = 1 = g \leq g$ and $|S \setminus S'| = |\{1, 4\}| = 2 = 2g \leq 2g$.

VI. Simulation

We visualized (upper bounds on) the number of tests fo threshold group testing with a gap using five parameter 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 theorems, Chen et al.'s scheme, and Chen and Fu's scheme.

Since 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. Similarly, since the numbers of tests with the four proposed theorems (Theorems 4, 6, 7, 8) are identical, we only consider the number of tests in Theorem 4. The two schemes are visualized in Figures 5–6. The red and green lines represent for Theorem 4 and Chen et al.'s scheme, respectively.

Since Chan et al. [8] and Reisizadeh et al. [25] 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 numbers of tests for Theorem 4 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. 5 for $d = 20$ and Fig. 6 for $d = 100$ and $d = 1000$, the number of tests with Theorem 4 was the smallest for all settings compared to Chen et al.'s scheme.

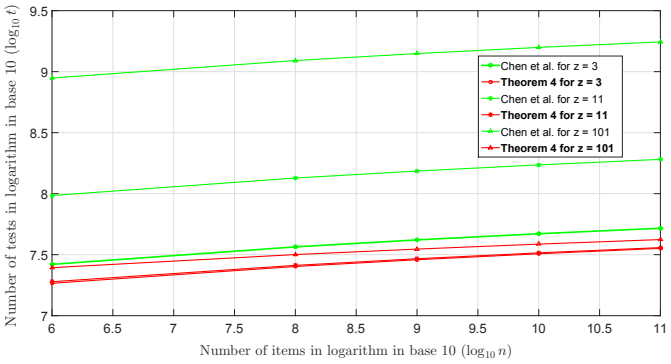
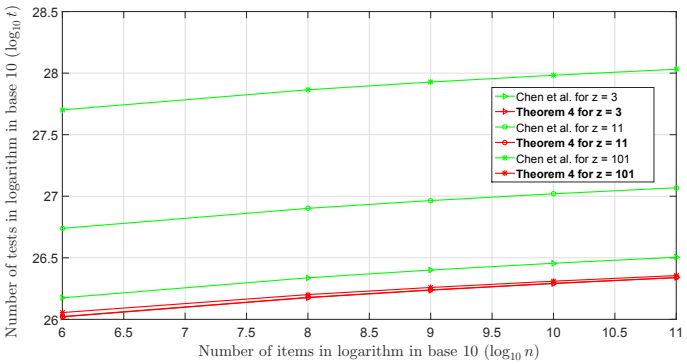
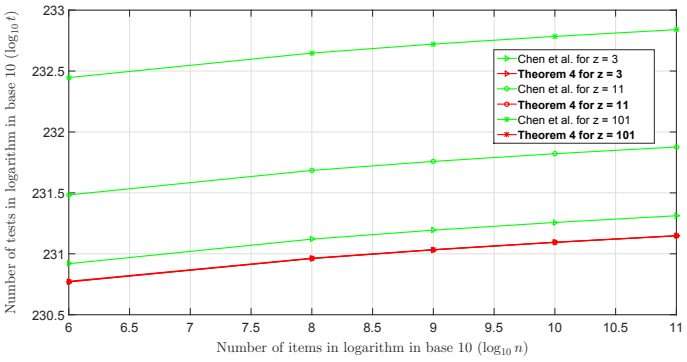


Fig. 5: Upper bounds on the number of tests versus number of items in logarithm in base 10 for $d = 20$, $z = \{3, 11, 101\}$, and $n = \{10^6 = 1\text{M}, 10^8 = 10\text{M}, 10^9 = 1\text{B}, 10^{10} = 10\text{B}, 10^{11} = 100\text{B}\}$ for Chen et al.'s scheme and Theorem 4.



(a) $d = 100$.



(b) $d = 1000$.

Fig. 6: 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 scheme and Theorem 4.

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.

When $d = 20$ (Fig. 5), 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 Theorem 4.

In summary, the results of simulation match those of our

analysis in Section IV-A: the upper bound on the number of tests in Theorem 4 is always smaller than the one in Theorem 1 for any positive z.

VII. Conclusion

In this paper, we have presented a novel construction scheme for disjunct matrices that is better than the construction proposed by Chen et al. [4]. 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 more accurate theorem for Chen and Fu's scheme [9], three proposed algorithms on improving non-adaptive encoding and decoding algorithms for threshold group testing as well as simulation for verifying our arguments throughout this work.

Appendix

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 [9]. Their inaccurate 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 $u \le |S_i| \le d-1$ 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 $t_0^{\mathcal{M}}(Z) \le e$ 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 $\binom{\alpha+1}{u} \le \binom{d}{u}$, 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 *inaccurate*. Since Step 2 is repeated up to $d-u$ times, the complexity of executing this step is

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

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 in which $u \ge 2$, $d = 2u$, and $u = g+1$, i.e., $\ell = 0$. We have:

$$\begin{aligned}
&(d-u)\binom{n-u}{g+1}\binom{d-1}{g}\binom{d}{u} \\
&= u\binom{n-u}{u}\binom{2u-1}{u-1}\binom{2u}{u} \\
&= u \cdot \frac{(n-u)(n-u-1)\dots(n-u-(u-1))}{u!} \\
&\quad \cdot \frac{u}{2u(2u-u+2)}\binom{2u}{u}\cdot\binom{2u}{u} \\
&> \frac{(n-2u+1)^u}{u!}\cdot\frac{u}{2(u+2)}\binom{2u}{u}^2 \\
&> \frac{(n-2u+1)^u}{u!}\cdot\frac{u}{2(u+2)}\left(\frac{1.08444}{2\mathrm{e}^{1/(8u)}\sqrt{u}}\cdot 2^{2u}\right)^2, \quad (30) \\
&> \frac{(n-2u+1)^u}{u!}\cdot\frac{1}{2(u+2)}\cdot\left(\frac{1.08444}{2\mathrm{e}^{1/(8\times 2)}}\right)^2\cdot 16^u \\
&> \frac{(n-2u+1)^u}{u!}\cdot\frac{1}{7(u+2)}\cdot 16^u,
\end{aligned}$$

and

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

where (30) 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 $u \ge 2$ (Corollary 2.9 in [31]). Consider the following inequality:

$$\begin{aligned}
&\frac{(n-2u+1)^u}{u!}\cdot\frac{1}{7(u+2)}\cdot 16^u \ge \frac{n^u}{u!} \\
\iff\quad & 1-\frac{1}{16}\cdot(7(u+2))^{1/u} \ge \frac{2u-1}{n}. \quad (31)
\end{aligned}$$

Since $(7(u+2))^{1/u}$ is a decreasing function of u and $u \ge 2$, for (31) to hold, it suffices that

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

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

$$\begin{aligned}
(d-u)\binom{n-u}{g+1}\binom{d-1}{g}\binom{d}{u} &> \frac{(n-2u+1)^u}{u!}\cdot\frac{1}{7(u+2)}\cdot 16^u \\
&\ge \frac{n^u}{u!} > \binom{n}{u}.
\end{aligned}$$

In summary, the complexity in (3) is inaccurate.

Acknowledgment

The authors thank Jonathan Scarlett at the National University of Singapore for his constructive and insightful comments on an early version of this paper. We also thank Roghayyeh Haghvirdinezhad for her comments on Fig. 2. The authors would like to thank the anonymous reviewers for their invaluable comments on an earlier draft of this work.

References

[1] T. V. Bui, M. Cheraghchi, and I. Echizen, "Improved non-adaptive algorithms for threshold group testing with a gap," in *Proc. IEEE International Symposium on Information Theory, ISIT 2020, Los Angeles, CA, USA, June 21-26, 2020*, pp. 1414–1419, IEEE, 2020.

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

[3] P. Damaschke, "Threshold group testing," in *General Theory of Information Transfer and Combinatorics* (R. Ahlswede, L. Bäumer, N. Cai, H. K. Aydinian, V. M. Blinovsky, C. Deppe, and H. Mashurian, eds.), vol. 4123 of *Lecture Notes in Computer Science*, pp. 707–718, Springer, Berlin, Heidelberg, 2006.

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

[5] 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.
[6] M. Cheraghchi, "Improved constructions for non-adaptive threshold group testing," *Algorithmica*, vol. 67, no. 3, pp. 384–417, 2013.
[7] G. D. Marco, T. Jurdzinski, D. R. Kowalski, M. Rózanski, and G. Stachowiak, "Subquadratic non-adaptive threshold group testing," *J. Comput. Syst. Sci.*, vol. 111, pp. 42–56, 2020.
[8] C. L. Chan, S. Cai, M. Bakshi, S. Jaggi, and V. Saligrama, "Stochastic threshold group testing," in *Proc. IEEE Information Theory Workshop, ITW 2013, Sevilla, Spain, September 9-13, 2013*, pp. 1–5, IEEE, 2013.
[9] H. Chen and H. Fu, "Nonadaptive algorithms for threshold group testing," *Discret. Appl. Math.*, vol. 157, no. 7, pp. 1581–1585, 2009.
[10] A. G. D'yachkov, V. V. Rykov, C. Deppe, and V. S. Lebedev, "Superimposed codes and threshold group testing," in *Information Theory, Combinatorics, and Search Theory - In Memory of Rudolf Ahlswede* (H. K. Aydinian, F. Cicalese, and C. Deppe, eds.), vol. 7777 of *Lecture Notes in Computer Science*, pp. 509–533, Springer, Berlin, Heidelberg, 2013.
[11] T. V. Bui, M. Kuribayashi, M. Cheraghchi, and I. Echizen, "Efficiently decodable non-adaptive threshold group testing," *IEEE Trans. Inf. Theory*, vol. 65, no. 9, pp. 5519–5528, 2019.
[12] D. Du, F. K. Hwang, and F. Hwang, *Combinatorial group testing and its applications*, vol. 12. Singapore: World Scientific, 2000.
[13] A. G. D'yachkov, N. Polyanskii, V. Y. Shchukin, and I. Vorobyev, "Separable codes for the symmetric multiple-access channel," *IEEE Trans. Inf. Theory*, vol. 65, no. 6, pp. 3738–3750, 2019.
[14] N. Shental, S. Levy, V. Wuvshet, S. Skorniakov, B. Shalem, A. Ottolenghi, Y. Greenshpan, R. Steinberg, A. Edri, R. Gillis, *et al.*, "Efficient high-throughput sars-cov-2 testing to detect asymptomatic carriers," *Science advances*, vol. 6, no. 37, p. eabc5961, 2020.
[15] R. Gabrys, S. Pattabiraman, V. Rana, J. Ribeiro, M. Cheraghchi, V. Guruswami, and O. Milenkovic, "Ac-dc: Amplification curve diagnostics for covid-19 group testing," *arXiv preprint arXiv:2011.05223*, 2020.
[16] E. Porat and A. Rothschild, "Explicit nonadaptive combinatorial group testing schemes," *IEEE Trans. Inf. Theory*, vol. 57, no. 12, pp. 7982–7989, 2011.
[17] 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, SODA 2010, Austin, Texas, USA, January 17-19, 2010* (M. Charikar, ed.), pp. 1126–1142, SIAM, 2010.
[18] H. Q. Ngo, E. Porat, and A. Rudra, "Efficiently decodable error-correcting list disjunct matrices and applications - (extended abstract)," in *Automata, Languages and Programming - 38th International Colloquium, ICALP 2011, Zurich, Switzerland, July 4-8, 2011, Proceedings, Part I* (L. Aceto, M. Henzinger, and J. Sgall, eds.), vol. 6755 of *Lecture Notes in Computer Science*, pp. 557–568, Springer, Berlin, Heidelberg, 2011.
[19] M. Cheraghchi, "Noise-resilient group testing: Limitations and constructions," *Discret. Appl. Math.*, vol. 161, no. 1-2, pp. 81–95, 2013.
[20] T. V. Bui, M. Kuribayashi, T. Kojima, R. Haghvirdinezhad, and I. Echizen, "Efficient (nonrandom) construction and decoding for non-adaptive group testing," *J. Inf. Process.*, vol. 27, pp. 245–256, 2019.
[21] S. Cai, M. Jahangoshahi, M. Bakshi, and S. Jaggi, "Efficient algorithms for noisy group testing," *IEEE Trans. Inf. Theory*, vol. 63, no. 4, pp. 2113–2136, 2017.
[22] 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," *IEEE Trans. Inf. Theory*, vol. 67, no. 3, pp. 1559–1570, 2021.
[23] M. Aldridge, O. Johnson, and J. Scarlett, "Group testing: An information theory perspective," *Found. Trends Commun. Inf. Theory*, vol. 15, no. 3-4, pp. 196–392, 2019.
[24] R. Ahlswede, C. Deppe, and V. S. Lebedev, "Bounds for threshold and majority group testing," in *Proc. IEEE International Symposium on Information Theory Proceedings, ISIT 2011, St. Petersburg, Russia, July 31 - August 5, 2011* (A. Kuleshov, V. M. Blinovsky, and A. Ephremides, eds.), pp. 69–73, IEEE, 2011.
[25] A. Reisizadeh, P. Abdalla, and R. Pedarsani, "Sub-linear time stochastic threshold group testing via sparse-graph codes," in *Proc. IEEE Information Theory Workshop, ITW 2018, Guangzhou, China, November 25-29, 2018*, pp. 1–5, IEEE, 2018.
[26] A. Emad and O. Milenkovic, "Semiquantitative group testing," *IEEE Trans. Inf. Theory*, vol. 60, no. 8, pp. 4614–4636, 2014.
[27] H. Abasi, N. H. Bshouty, and H. Mazzawi, "Non-adaptive learning of a hidden hypergraph," *Theor. Comput. Sci.*, vol. 716, pp. 15–27, 2018.
[28] D. R. Stinson and R. Wei, "Generalized cover-free families," *Discret. Math.*, vol. 279, no. 1-3, pp. 463–477, 2004.
[29] W. H. Kautz and R. C. Singleton, "Nonrandom binary superimposed codes," *IEEE Trans. Inf. Theory*, vol. 10, no. 4, pp. 363–377, 1964.
[30] A. G. D'yachkov, P. A. Vilenkin, D. C. Torney, and A. J. Macula, "Families of finite sets in which no intersection of sets is covered by the union of s others," *J. Comb. Theory, Ser. A*, vol. 99, no. 2, pp. 195–218, 2002.
[31] 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.

Thach V. Bui received the B.Sc. degree in computer science of the honor program from the Faculty of Information Technology, University of Science, VNU-HCM, Vietnam, in 2012 and the Ph.D. degree in informatics from The Graduate University of Advanced Studies (SOKENDAI), Kanagawa, Japan affiliated with the National Institute of Informatics, Tokyo, Japan, in 2019. He started his first postdoc as a postdoctoral researcher at the University of Padova, Italy (2019-2020) working in computational biology. He is currently a research fellow at the National University of Singapore, Singapore. His research interest includes group testing, computational biology, information theory, and neuroscience.

Mahdi Cheraghchi (S'05–M'10–SM'16) is an Assistant Professor of Computer Science and Engineering at the University of Michigan–Ann Arbor. He has been on the faculty of Imperial College London from 2015 to 2019, where he maintains an honorary appointment. After completing his Ph.D. degree, he was affiliated as a post-doctoral researcher with the University of Texas at Austin (2010-11), Carnegie Mellon University (2011-13), MIT (2013-14), and the University of California, Berkeley (2015). He obtained his M.Sc. and Ph.D. degrees in computer science from Ecole Polytechnique Fédérale de Lausanne (EPFL), in 2005 and 2010, respectively, and the B.Sc. degree in computer engineering from Sharif University of Technology in 2004.

Dr. Cheraghchi is broadly interested in theoretical computer science, and his research so far has mainly focused on the interconnections between information and coding theory and theoretical computer science, sparse recovery and high-dimensional geometry, information-theoretic privacy and security, and approximation algorithms.

Isao Echizen received B.S., M.S., and D.E. degrees from the Tokyo Institute of Technology, Japan, in 1995, 1997, and 2003, respectively. He joined Hitachi, Ltd. in 1997 and until 2007 was a research engineer in the company's systems development laboratory. He is currently a director and a professor of the Information and Society Research Division, the National Institute of Informatics (NII), a director of the Global Research Center for Synthetic Media, the NII, and a professor in the Department of Information and Communication Engineering, Graduate School of Information Science and Technology, The University of Tokyo, Japan. He was a visiting professor at the Tsuda University, Japan, at the University of Freiburg, Germany, and at the University of Halle-Wittenberg, Germany.

He is currently engaged in research on multimedia security and multimedia forensics. He currently serves as a research director in CREST FakeMedia project, Japan Science and Technology Agency (JST). He received the Best Paper Award from the IPSJ in 2005 and 2014, the Fujio Frontier Award and the Image Electronics Technology Award in 2010, the One of the Best Papers Award from the Information Security and Privacy Conference in 2011, the IPSJ Nagao Special Researcher Award in 2011, the DOCOMO Mobile Science Award in 2014, the Information Security Cultural Award in 2016, and the IEEE Workshop on Information Forensics and Security Best Paper Award in 2017. He was a member of the Information Forensics and Security Technical Committee and the IEEE Signal Processing Society. He is the Japanese representative on IFIP TC11 (Security and Privacy Protection in Information Processing Systems), a member-at-large of board-of-governors of APSIPA, and an editorial board member of the IEEE Transactions on Dependable and Secure Computing and the EURASIP Journal on Image and Video processing.