*mathematics*

MDPI

*Article*

# Theoretical Bounds on Performance in Threshold Group Testing Schemes †

**Jin-Taek Seong**

Department of Convergence Software, Mokpo National University, Muan 58554, Korea; jtseong@mokpo.ac.kr

† This paper is an extended version of our paper published in ICEIC 2020, Barcelona, Spain, 19–22 January 2020.

Received: 26 March 2020; Accepted: 20 April 2020; Published: 21 April 2020


**Abstract:** A threshold group testing (TGT) scheme with lower and upper thresholds is a general model of group testing (GT) which identifies a small set of defective samples. In this paper, we consider the TGT scheme that require the minimum number of tests. We aim to find lower and upper bounds for finding a set of defective samples in a large population. The decoding for the TGT scheme is exploited by minimization of the Hamming weight in channel coding theory and the probability of error is also defined. Then, we derive a new upper bound on the probability of error and extend a lower bound from conventional one to the TGT scheme. We show that the upper and lower bounds well match with each other at the optimal density ratio of the group matrix. In addition, we conclude that when the gaps between the two thresholds in the TGT framework increase, the group matrix with a high density should be used to achieve optimal performance.

**Keywords:** defective samples; lower bound; probability of error; threshold group testing; upper bound

## 1. Introduction

Group testing (GT) introduced by Dorfman has been used in a wide range of fields from computer science to biology. A new application area of GT is Compressed Sensing [1], which has recently attracted attention from research communities, and this is a variant of GT. In general, GT has been a research field for handling undetermined problems to identify a subset of defective samples within a large population. The fundamental idea of GT problems comes from as follows: it is assumed that the number of defective samples is very sparse, and even if you select several samples at once, there may be no defect samples among them. Eventually, you can find defective samples without having to individually inspect all samples. In other words, GT is an inverse problem of knowing the original input states through a subset of parameters.

To date, GT has exploited a wide range of applications in biology [2,3], communication theory [4–7], signal processing [8], computer science [9–11], and mathematics [12]. The use of fundamental GTs extend to error correction code [13], identifying available multiple access channel [14,15], recovering sparse signals [1,8], detecting malicious attacks in security networks [16], testing good quality of products [17], and many others. Recently, there has been a move to more precisely study the performance of GT, and furthermore for noiseless and noisy frameworks nearly optimal performance has been presented in [18–20].

First, Dorfman developed the GT during the midst of World War II [21]. At this time, syphilis erupted in the army and the U.S. government faced a situation where it was necessary to quickly find soldiers infected with syphilis. This led to the active participation of the U.S. government to develop an early

GT model to find syphilis soldiers. However, the syphilis test, which was expensive and used a long diagnostic time, was not enough to test all soldiers. Thus, GT has emerged as a new method of syphilis testing. Suppose the number of soldiers infected with syphilis is very small compared to the total number of soldiers. Indeed, this is a plausible and persuasive hypothesis. In this case, the results may be negative even if the blood of several soldiers is mixed and tested for syphilis at once. The number of tests could be reduced because several blood samples were mixed without syphilis testing individually. In addition, quick testing is allowed. This is the background in which GT models have emerged. Since then, it has been exploited and applied in various research fields based on the core ideas of this GT.

The initial GT is performed by the following way. First, mix blood samples from several soldiers to see if they respond to syphilis. When the result is positive, at least one soldier in the group is infected with syphilis. Conversely, if negative, all the blood samples pooled in the syphilis group can be confirmed not to be infected with syphilis. Such syphilis tests are possible because most soldiers are not infected with syphilis and only a few soldiers are infected with syphilis. Here the problem of GT is mainly focused on two issues. First, it is about how to choose a subset of defective samples to be included in one group. The second is about which identifying the defective samples should be used to find a set of defective samples among a plurality of samples. One to classify between many GT schemes is the way how tests are performed. One way is to perform the tests all at once by a predefined method. This is called nonadaptive GT. That is, all tests are conducted simultaneously in a predetermined manner, and the results of one test are independent of the other. Adaptive GT, however, can be used to design more test pools by using one test result to design another [2]. In general, adaptive GTs allow for fewer tests, but for most practical applications, nonadaptive GTs are preferred because they can perform all tests simultaneously. This is because it reduces the time required for testing.

Various GT models have been introduced so far. In the conventional GT [21], the output result is positive when more than one defective sample is included in the group. Therefore, if all samples included in the test are normal, the result is negative, otherwise it is positive. Quantitative GT [2] is a variant of GT. In the quantitative GT model, the output results are designed to show the number of all defective samples included in one test. This is different from the conventional GT. The model called Threshold Group Testing (TGT) [22] is a variant of GT with two thresholds. TGT takes two thresholds and determines whether the output result is negative or positive. In this model, the test result is positive if there are defective samples larger than a predetermined upper threshold in the group to be tested. Conversely, if the number of defective samples in the group is less than a lower threshold, it is negative. The main feature of this model is that it is designed to produce results that are neither negative nor positive. If there is the number of defective samples between the two thresholds, the result is randomly output with the same probability of being positive and negative. Both the upper and lower thresholds are preset when designing the TGT model, and the difference as called the gap between both them affects performance. Recently, a summary of the number of tests required and the decoding complexity of the TGT schemes including two thresholds can be found in more detail in [23,24]. The comparison with the number of tests and the decoding complexity is out of the main scope in this paper.

So far, most of bounds on performance presented in TGT problems have shown meaning results, such as improving encoding and decoding way [23,24], archiving near-optimal number of tests [25], and enhancing robust and efficient designs [26,27] in TGT schemes. However, there is a lack of research on how to design the group matrix and how defective rate of a set of all the samples affects successful decoding on performance. In this paper, it is to find out the answer to the question of how much density a group matrix should be designed to identifying defects for good performance. In addition, we see how the gap affects good design of the group matrix in TGT schemes. This was the motivation for this paper. The main goal of this paper is to clarify the relationship between the density of the group matrix and the defective rate of the signal.

In this paper, we consider a TGT framework with lower and upper thresholds which are the boundaries between positive and negative results. And we derive the lower and upper bounds for finding a set of defective samples out of a large of samples in TGTs. To this end, we exploit minimization of Hamming weight in channel coding theory. And we define probability of error for our TGT decoding scheme. We obtain new upper bounds on the probability of error which are well matched lower bounds from the information-theoretic approach. We show that the upper and lower bounds coincide with each other at the optimal density ratio of the group matrix. In addition, when the gap becomes large, it is necessary to design a group matrix having a high density to obtain optimal TGT performance. Throughout our results, we conclude how the design parameters of the TGT frameworks can affect performance. This is a main contribution of this paper. Next we will define a scheme of TGT problems.

## 2. Threshold Group Testing Framework

### *2.1. Problem Description*

This section describes the TGT problem in more detail. Fist, let $\mathbf{x}$ be a binary vector of size $N$ which has defective and normal samples. Namely, $\mathbf{x} = (x_1, x_2, \cdots, x_N)^T$ where $x_i$ denotes the $i$th entry of $\mathbf{x}$. Each entry of $\mathbf{x}$ refers to the state of the sample by using binary, that is, whether it is defective or normal. the $i$th defective sample is expressed by $x_i = 1$. If the sample is not defective, it is expressed as $x_i = 0$. All the entries are identically independent distributed (i.i.d.) from the following Bernoulli probability distribution,

$$\Pr(x_i = \theta) = \begin{cases} 1-\delta & \text{if } \theta = 0, \\ \delta & \text{if } \theta = 1, \end{cases} \tag{1}$$

where $\delta := \frac{K}{N}$ is the defective rate which in general is very small. $K$ is the number of defective samples in $\mathbf{x}$. In practice, the defective rate is assumed as a very small value, e.g., syphilis infection rate. Since the input signal $\mathbf{x}$ is generated from the probability distribution as defined in (1), it can be seen that the number of 1's of an instance signal $\mathbf{x}$ is 0 to $N$. However, when the length of $\mathbf{x}$ is large, the number of 1's is close to $K$. Let $\mathcal{L}_{k_1}$ be the set of the signals with the number of $k_1$ ones in $\mathbf{x}$, then $\|\mathbf{x}\|_0 = k_1$ and $\binom{N}{k_1} = |\mathcal{L}_{k_1}|$ where $\|\cdot\|_0$ is the Hamming weight and $|\cdot|$ is the cardinality of the set. We define the set $\mathcal{L}$ of the signals as $\mathcal{L} = \bigcup_{k_1=0}^{N} \mathcal{L}_{k_1}$, then its size is $|\mathcal{L}| = \sum_{k_1=0}^{N} \binom{N}{k_1} = 2^N$ which corresponds to the total number of binary input signals with size $N$.

Next, we define a group matrix $\mathbf{A}$. This matrix consists of $M$ rows and $N$ columns. The role of the group matrix in the TGT model is to express which elements of $\mathbf{x}$ to test for each group. Each entry $A_{ji}$ of the group matrix indicates whether the $i$th sample $x_i$ is included in the $j$th group test. If so, $A_{ji} = 1$. Otherwise, $A_{ji} = 0$. Each row is a set of the input samples participating in the corresponding testing. All the entries of the group matrix $\mathbf{A}$ are i.i.d. from the following Bernoulli probability distribution,

$$\Pr(A_{ji} = \theta) = \begin{cases} 1-\gamma & \text{if } \theta = 0, \\ \gamma & \text{if } \theta = 1, \end{cases} \tag{2}$$

where $\gamma$ is the density ratio of the group matrix. High density ratio means that when designing a TGT model, there are many samples participating in tests, so it is expensive and complicated. In this sense, it is necessary to clarify the relation between the defective rate $\delta$ and the density ratio $\gamma$ in order to design the GT schemes effectively.

We describe the output of the TGT scheme. The output TGT is mathematically described in two steps. First, the input signal defined above (1) is projected linearly into the group matrix generated from (2). Let $\mathbf{s}$

be the testing vector from the linear combination of $\mathbf{x}$ and $\mathbf{A}$. Therefore, it is expressed as the product of the input signal and the group matrix as follows:

$$\mathbf{s} = \mathbf{A}\mathbf{x}. \tag{3}$$

Since all the entries of $\mathbf{x}$ and $\mathbf{A}$ are binary, the each entry of the testing vector $\mathbf{s}$ with size $M$ has a nonnegative integer from 0 to $N$, thus, $\mathbf{s} \in \{0, 1, 2, \cdots, N\}^M$. Let $s_j$ be the $j$th entry of the testing vector $\mathbf{s}$. From the linear combination of (3), the entry $s_j$ is obtained as $s_j = \sum_{i=1}^{N} A_{ji} x_i$. The next step is to map the non-negative integer $s_j$ to the binary output of the TGT model using two thresholds. Let $L$ and $U$ be defined as lower and upper thresholds, respectively. And both thresholds $L$ and $U$ are nonnegative integers.

Suppose $f$ is the decision function that transforms the testing vector $\mathbf{s}$ into the binary output vector $\mathbf{y}$, whose values are positive or negative. The input parameters of the function $f$ are the vector $\mathbf{s}$ and the two thresholds $L$ and $U$. Therefore, for function $f$, we express,

$$\mathbf{y} = f(\mathbf{s}, L, U). \tag{4}$$

As defined in the TGT model [22], using the function $f$, if the entry $s_j$ is greater than or equal to the upper threshold $U$, the corresponding output $y_j$ is positive as $y_j = 1$. If $s_j$ is less than or equal to $L$, the output is negative, $y_j = 0$. Due to the nature of the TGT scheme, $s_j$ may be between two thresholds $L$ and $U$. In this case, the output $y_j$ for $s_j$ is random. That is, the function $f$ determines 0 or 1 with the same probability for the negative and positive outputs. Therefore, the function $f$ is defined as follows using the input parameters $s_j$, $L$, and $U$.

$$f(s_j, L, U) = \begin{cases} 0 & \text{if } s_j \leq L, \\ 0 \text{ or } 1 & \text{if } L < s_j < U, \\ 1 & \text{if } s_j \geq U. \end{cases} \tag{5}$$

Suppose both thresholds are preset and constant. Let the difference between the two thresholds be the gap as $G = U - L - 1$. In general, for the TGT schemes the gap is positive, $G > 0$. In the conventional GT, $G = 0$ since $L = 0$ and $U = 1$, so that this determines exactly the two output results, positive and negative.

The aim of the TGT scheme is to find an unknown signal $\mathbf{x}$ from a group matrix $\mathbf{A}$ and a corresponding vector $\mathbf{y}$. So far, the main research direction of GT problems is to determine the number of tests $M$ needed for successfully finding the defective samples of the input signal $\mathbf{x}$. Next, we define the probability of error on successful decoding to derive a lower and an upper bound of the performance.

### *2.2. Definition on Probability of Error*

In this section, we define the probability of error on finding defective samples of $\mathbf{x}$ in a TGT framework for given parameters, i.e., $N$, $K$, and $M$. Priori defining the probability of error, we classify the input signal $\mathbf{x}$ as a set of signals with the number of ones in $\mathbf{x}$.

We assume that an estimator for decoding in our TGT framework is to find a feasible solution $\hat{\mathbf{z}}$ using the minimization of Hamming weight as follows,

$$\hat{\mathbf{z}} = \arg\min ||\mathbf{z}||_0 \quad \text{subject to } f(\mathbf{A}\mathbf{z}) = f(\mathbf{A}\mathbf{x}), \tag{6}$$

where $\mathbf{z} \in \mathcal{L}$ is a feasible signal. Let $k_2$ be the number of ones in $\mathbf{z}$ as $k_2 = ||\mathbf{z}||_0$, so that $k_2 \leq k_1$. We define the error as occurring when a feasible solution $\hat{\mathbf{z}}$ decided by the minimization rule (6) is not equal to an instance signal $\mathbf{x}$ which is desired, $f(\mathbf{A}\hat{\mathbf{z}}) = f(\mathbf{A}\mathbf{x})$ but $\mathbf{x} \neq \hat{\mathbf{z}}$. Let $\mathcal{E}_0(\mathbf{x}, \hat{z}) := \{\mathbf{A} : \mathbf{x} \neq \hat{\mathbf{z}}\}$ be

the exact error event of this decoder as a function of the group matrix $\mathbf{A}$. This error event $\mathcal{E}_0$ is a subset of the following feasible error event $\mathcal{E}$ since a feasible signal $\mathbf{z}$ is a potential candidate of estimated signals. We define the feasible error event $\mathcal{E}$ as follows,

$$\mathcal{E}(\mathbf{x}, \mathbf{z}) := \{\mathbf{A} : \mathbf{x} \neq \mathbf{z}, f(\mathbf{A}\mathbf{z}) = f(\mathbf{A}\mathbf{x})\}. \tag{7}$$

Note that $\mathcal{E}_0(\mathbf{x}, \hat{z}) \subseteq \mathcal{E}(\mathbf{x}, \mathbf{z})$. Let $\Pr(\mathcal{E}_0)$ and $\Pr(\mathcal{E})$ be the probability of error for both events $\mathcal{E}_0(\mathbf{x}, \hat{z})$ and $\mathcal{E}(\mathbf{x}, \mathbf{z})$, respectively. And then, the following inequality is satisfied as $\Pr(\mathcal{E}_0) \leq \Pr(\mathcal{E})$. The probability of error $P_e := \Pr(\mathcal{E}_0)$ is upper bounded by

$$\begin{aligned} P_e &\leq \Pr(\mathcal{E}) \\ &= \frac{1}{|\mathcal{L}|} \sum_{\mathbf{x} \in \mathcal{L}} \sum_{\mathbf{z} \in \mathcal{L}, \mathbf{z} \neq \mathbf{x}} \Pr(\mathbf{A} \in \mathcal{E}(\mathbf{x}, \mathbf{z}) \mid (\mathbf{x}, \mathbf{z})). \\ &= \frac{1}{|\mathcal{L}|} \sum_{\mathbf{x} \in \mathcal{L}} \sum_{\mathbf{z} \in \mathcal{L}, \mathbf{z} \neq \mathbf{x}} \Pr(f(\mathbf{A}\mathbf{z}) = f(\mathbf{A}\mathbf{x}) \mid (\mathbf{x}, \mathbf{z})). \end{aligned} \tag{8}$$

In a brute-force approach, enumeration of individual probabilities of feasible error events in (8) is almost intractable because $|\mathcal{L}|$ is typically very large. This brute-force approach can be avoided with what will be described subsequently next.

## 3. Theoretic Bounds on Performance

### *3.1. Lower Bound*

We aim to obtain a theoretic lower bound through this section. First, we investigate related works for some lower bounds. In [2], the author presented a lower bound on the probability of error by using an information-theoretic approach. That is, if we use a sequential algorithm of a conventional GT framework, the minimum number of tests for finding $K$ defective samples is achieved as

$$M \geq \log_2 |\mathcal{X}|, \tag{9}$$

where $\mathcal{X}$ denotes the set of the sample space. This bound is obtained from the fact that for each group test, $\mathcal{X}$ is divided into two disjoint subsets. Moreover, each subset corresponds to one of the two possible groups in the conventional GT schemes. As you know, the lower bound itself is unachievable in small problems. Note that since it must be realizable by some GT problems, the splitting of the set $\mathcal{X}$ is not random.

We use Fano's inequality [28] to find the lower bound on the probability of error for the TGT framework. Fano's inequality is a theorem for the relation between conditional entropy of input signals and output results and the probability of error for any decoder. In addition, in coding theory, Fano's inequality has been used to derive converse proofs. The following provides an important clue that clarifies the bounds of inferences. In our earlier work [29], we presented the following lower bound on performance in the TGT scheme which is an extension of the previous work [30]. The following is the lower bound on th probability of error for the TGT scheme which is referred to as Theorem 1. Full proof of this theorem is detailed in [29].

**Theorem 1** ([29]). *For any group matrix from* (2)*, any decoding scheme defined in* (6)*, and* $0 < \delta \leq 1/2$*, the probability of error* $P_e$ *for the TGT scheme is lower bounded by*

$$P_e \geq H_b(\delta) - \frac{M}{N} H_b(p) - \frac{1}{N} \tag{10}$$

*where* $H_b(\cdot)$ *denotes the binary entropy,* $p$ *is the probability that each entry* $y_j$ *has 0.*

The right side of (10) means the minimum probability of error. In other words, it indicates the lower bound on the probability of error even if any decoder is used. Namely, the bound on the right side of (10) must be negative in order for that the probability of error is vanished. Thus, in the TGT schemes, the following necessary condition is satisfied for perfect decoding,

$$M > \frac{NH_b(\delta) - 1}{H_b(p)} \geq NH_b(\delta) - 1 \tag{11}$$

where the second inequality comes from $0 \leq H_b(p) \leq 1$. The result of (11) has the following meaning. The number of tests required to successfully reconstruct all defective samples in a TGT problem must meet the following necessary condition: $M > NH_b(\delta)$. This condition is the same as that of [2] obtained from the information-theoretic approach. It also matches the necessary condition for successfully decoding of the conventional GT scheme. Second, if the following condition is satisfied with $H_b(p) = 1$, the minimum number of tests $M$ is achieved. The binary entropy is maximized as the probability $p$ is 0.5. Therefore, for the maximum binary entropy, the following (12) must be satisfied.

Next, we clarify results of (12) on how much effect on design parameters for TGT frameworks with respect to defective rate $\delta$, density ratio $\gamma$, the lower and upper thresholds, $L$ and $U$. Figure 1 shows the number of tests for different density ratios of the group matrix and gaps between lower and upper thresholds in TGT. As shown in Figure 1, the larger the gap, the more we can design a group matrix with a density ratio over a wider range required for the minimum number of tests. In fact, this can be understood intuitively. Because of the large gaps between two lower and upper thresholds, we can only separate between positive and negative results by performing TGT with more samples. In other words, there is a greater probability that the results of TGT would be negative or random if not enough samples are involved in TGT. Figure 2 shows the number of tests in TGT frameworks when the gaps are equal to 0, $G = 0$, but the two thresholds vary. Overall, to keep the minimum number of tests, we need to use a larger density ratio of the group matrix as the two thresholds increase. It also allows us to use a group matrix with a wider density ratio when we use large thresholds.

$$\frac{1}{2} = \sum_{s_1=0}^{L} \binom{N}{s_1} (\delta\gamma)^{s_1} (1-\delta\gamma)^{N-s_1} + \frac{1}{2} \sum_{s_1=L+1}^{U-1} \binom{N}{s_1} (\delta\gamma)^{s_1} (1-\delta\gamma)^{N-s_1} \tag{12}$$

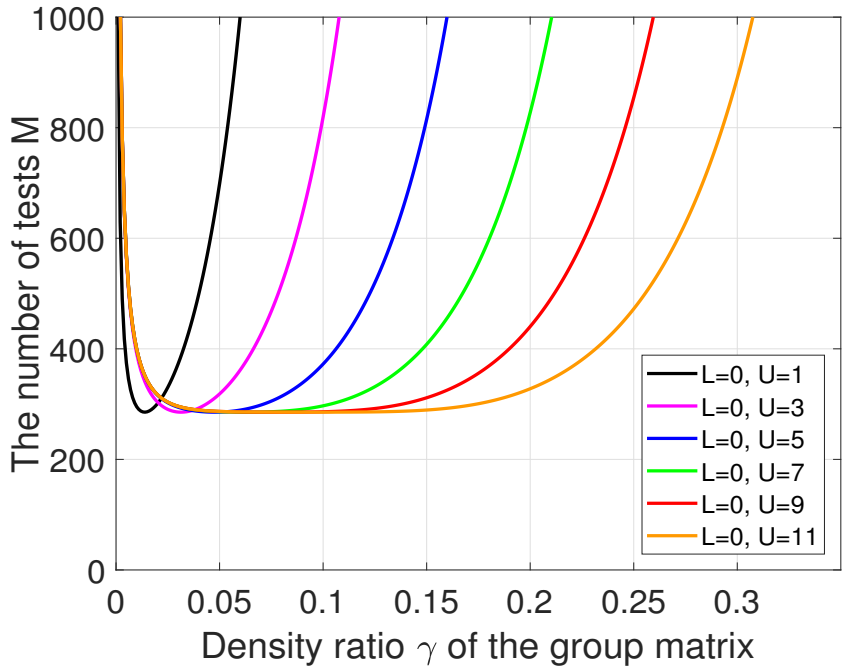


**Figure 1.** The number of tests $M$ with respect to different gaps equal to $G = 0, 2, 4, 6, 8$, and 10, and density ratio of the group matrix at $N = 1000$, $\delta = 0.05$ (evaluated by probability of error at $10^{-5}$).

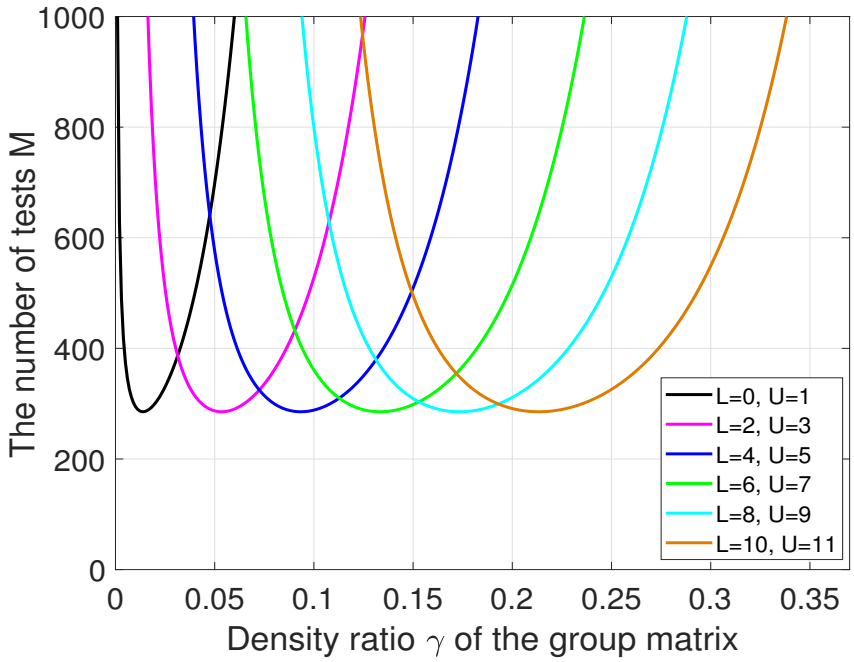


**Figure 2.** The number of tests $M$ with respect to two thresholds, and density ratio of the group matrix at $N = 1000$, $\delta = 0.05$ (evaluated by probability of error at $10^{-5}$).

### *3.2. Upper Bound*

In this section, we explain the upper bound on performance of TGT. In [19], as a function of the number of tests, an upper bound on the probability of error was obtained. This upper bound is the same as the theoretical lower bound. In this section, we aim to obtain the upper bound by using the minimization of Hamming weight defined in (6). Basically, the minimization of Hamming weight is NP hard, but it allows us to analyze the performance of TGT frameworks on how we can construct the group matrix and how good the decoding algorithm can improve compared to other algorithms. As far as we know, the proposed upper bound is a new approach which is simple and clear for the evaluation of the performance on TGT.

Let us recall the upper bound on probability of error we have defined in (8). Now we aim to drive the upper bound as written in (8). The basic ideas to tackle this bound can be thought of as follows. We first think of the same error pattern. The next step is to find the probability for that error pattern. Finally, we can obtain the total probability by collecting all the individual probabilities with the same error pattern. To find the same error pattern, consider the following: two probabilities are the same, i.e., $\Pr(f(\mathbf{A}_j\mathbf{x}) = f(\mathbf{A}_j\mathbf{z}_1)) = \Pr(f(\mathbf{A}_j\mathbf{x}) = f(\mathbf{A}_j\mathbf{z}_2))$ such that $(\mathbf{z}_1 \neq \mathbf{z}_2) \in \mathcal{L}$ and $\|\mathbf{z}_1\|_0 = \|\mathbf{z}_2\|_0 = k_2$ where $\mathbf{A}_j$ is the $j$th row of the group matrix $\mathbf{A}$. In other words, two probabilities for $\mathbf{z}_1$ and $\mathbf{z}_2$ having the same Hamming weights are the same (for further detail, we will prove in below). And then, we add the individual probabilities with the same Hamming weights with respect to two vectors $\mathbf{x}$ and $\mathbf{z}$. And we count the number of vectors with the same probability.

**Theorem 2.** *Given any unknown signal in* (1) *and any group matrix in* (2)*, using arbitrary decoder defined in* (6)*, the probability of error $P_e$ is bounded by*

$$P_e \leq \frac{1}{|\mathcal{L}|} \sum_{k_1=0}^{N} \sum_{k_2=0, \mathbf{x}\neq\mathbf{z}}^{k_1} \binom{N}{k_2}\binom{N}{k_1} \delta^{k_1}(1-\delta)^{N-k_1} \Pr\Big(f(\mathbf{A}\mathbf{x}) = f(\mathbf{A}\mathbf{z})\Big) \tag{13}$$

**Proof of Theorem 2.** The conditional probability in (8) with given condition of $k_1$ and $k_2$ Hamming weights for $\mathbf{x}$ and $\mathbf{z}$, can be rewritten by the independent rows as follows,

$$\Pr\Big(f(\mathbf{A}\mathbf{x}) = f(\mathbf{A}\mathbf{z})\Big) = \prod_{j=1}^{M} \Pr\Big(f(\mathbf{A}_j\mathbf{x}) = f(\mathbf{A}_j\mathbf{z})\Big) \tag{14}$$

Let $P := \Pr\big(f(\mathbf{A}_j\mathbf{x}) = f(\mathbf{A}_j\mathbf{z})\big)$. We therefore take into account the probability for the $j$th entry of $\mathbf{y}$, and look at the probability in more detail,

$$P = \Pr\Big(f(\mathbf{A}_j\mathbf{x}) = 0\Big) \Pr\Big(f(\mathbf{A}_j\mathbf{z}) = 0\Big) + \Pr\Big(f(\mathbf{A}_j\mathbf{x}) = 1\Big) \Pr\Big(f(\mathbf{A}_j\mathbf{z}) = 1\Big) \tag{15}$$

where equality holds between the left and right sides of (15) for $f(\mathbf{A}_j\mathbf{x}) = f(\mathbf{A}_j\mathbf{z})$, i.e., $0 = 0$ and $1 = 1$. Now we find out two probabilities as follows,

$$\begin{aligned} \Pr\Big(f(\mathbf{A}_j\mathbf{x}) = 0\Big) \Pr\Big(f(\mathbf{A}_j\mathbf{z}) = 0\Big) = &\left[\Pr\Big(\sum_{i=1}^{N} A_{ji}x_i \leq L\Big) + \frac{1}{2}\Pr\Big(L < \sum_{i=1}^{N} A_{ji}x_i < U\Big)\right] \\ &\times \left[\Pr\Big(\sum_{i=1}^{N} A_{ji}z_i \leq L\Big) + \frac{1}{2}\Pr\Big(L < \sum_{i=1}^{N} A_{ji}z_i < U\Big)\right] \end{aligned} \tag{16}$$

And then,

$$\begin{aligned}\Pr\left(f(\mathbf{A}_j\mathbf{x})=1\right)\Pr\left(f(\mathbf{A}_j\mathbf{z})=1\right) &= \left[\Pr\left(\sum_{i=1}^{N}A_{ji}x_i\geq U\right)+\frac{1}{2}\Pr\left(L<\sum_{i=1}^{N}A_{ji}x_i<U\right)\right]\\ &\times\left[\Pr\left(\sum_{i=1}^{N}A_{ji}z_i\geq U\right)+\frac{1}{2}\Pr\left(L<\sum_{i=1}^{N}A_{ji}z_i<U\right)\right]\end{aligned} \tag{17}$$

where $\mathbf{x}$ and $\mathbf{z}$ have $k_1$ and $k_2$ Hamming weights, respectively, and so that (16) and (17) hold.

Given Hamming weight and two thresholds, we obtain the following conditional probability for $\mathbf{x}$,

$$\begin{aligned}\Pr\left(L<\sum_{i=1}^{N}A_{ji}x_i<U\middle|\|\mathbf{x}\|_0=k_1\right) &= \Pr\left(L<\sum_{i=1}^{k_1}A_{ji}<U\right)\\ &= \sum_{d=L+1}^{U-1}\binom{k_1}{d}\gamma^d(1-\gamma)^{k_1-d}\end{aligned} \tag{18}$$

Using (18), we find out all the probabilities for the given $\mathbf{x}$ and $\mathbf{z}$ in (16) and (17). Next, the probability with $k_1$ Hamming weight for $\mathbf{x}$ is defined as

$$\Pr\left(\|\mathbf{x}\|_0=k_1\right)=\binom{N}{k_1}\delta^{k_1}(1-\delta)^{N-k_1} \tag{19}$$

This is the end of the proof for Theorem 2. □

For a special case with $k_1=k_2=K$ that we know exactly $K$ defective samples in advance,

$$\begin{aligned}P_e &\leq \binom{N}{K}\Pr\left(f(\mathbf{A}\mathbf{x})=f(\mathbf{A}\mathbf{z})\right)=\binom{N}{K}P^M\\ &\leq 2^{NH_b(K/N)+M\log_2 P}\end{aligned} \tag{20}$$

where recall that $P:=\Pr\left(f\left(\mathbf{A}_j\mathbf{x}\right)=f\left(\mathbf{A}_j\mathbf{z}\right)\right)$ where $0.5\leq P\leq 1$. In order for vanishing the probability of error $P_e$ in the right side of (20), the following condition holds as $N\to\infty$:

$$\begin{aligned}M &> \frac{NH_b(K/N)}{\log_2 P^{-1}}\\ &\geq NH_b(K/N)\end{aligned} \tag{21}$$

The minimum number of tests $M$ required for finding defective samples can be obtained when the probability $P$ is 0.5. This result is exactly the same as the lower bound in (11).

Figure 3 shows the plot of comparisons of the upper and the lower bounds with different thresholds $L$ and $U$ for $N=1000$, $\delta=0.05$, and $G=0$. This figure is drawn from the expression given in such that for $K=50$, the number of tests $M$ is obtained from the probability of error being lower than $10^{-5}$. One interesting point of this result is that there is an optimal density ratio of the group matrices to obtain the minimum number of tests. In addition, we see that our proposed upper bounds are well matched in comparison with the lower bounds from the information-theoretic theory. One more fact from Figure 3 is that as the two thresholds $L$ and $U$ increase while the gaps are constant, the group matrix should be denser to successfully find defective samples with only a small number of tests. This is an important result. For example, if the two thresholds are small, we have to generate more sparser group matrices.

Otherwise, the performance of the TGT framework would be worse. One of the meaning findings shown in Figure 3 is that the acceptable range of the density ratio mainly depends on the lower and upper thresholds. In other words, when the two thresholds are small, a narrow range of the density ratio can be suitable. This characteristic is one of the factors to consider when designing TGT frameworks.

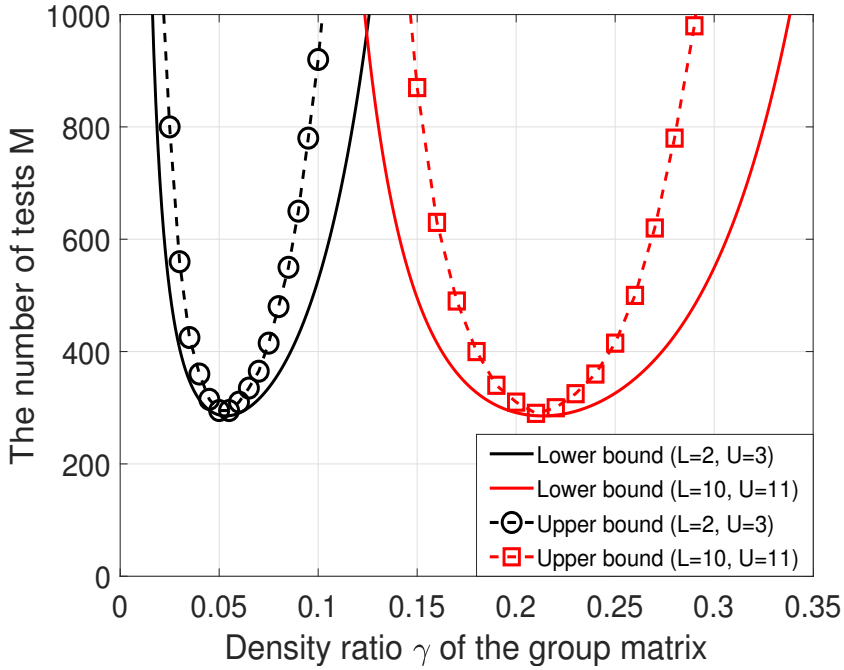


**Figure 3.** Comparisons of the upper and lower bounds (evaluated at $10^{-5}$) with different thresholds $L$ and $U$ for $N = 1000$, $\delta = 0.05$, and $G = 0$: solid lines indicate the lower bounds and marked dashed lines indicate the upper bounds.

Figure 4 shows how the gap affects performance on the number of tests where we use $N = 1000$ and $\delta = 0.05$ evaluated at the probability of error of $10^{-5}$. As shown in Figure 4, the proposed upper bounds have the same values as the lower bounds at the optimum density ratio around, but it is slightly different in the region outside the optimal range. From Figure 4, we observe that the larger the gaps $G$, the greater the density ratio of the group matrix and the better the performance. Higher density ratio of the group matrix in TGTs are a cautious approach due to increasing computational burden. Figure 5 shows comparisons of the upper and lower bounds for $N = 1000$ with the optimal density ratios of the group matrices. There is no difference between both bounds over whole range of defective rates. Namely, our lower and upper bounds well match with each other.

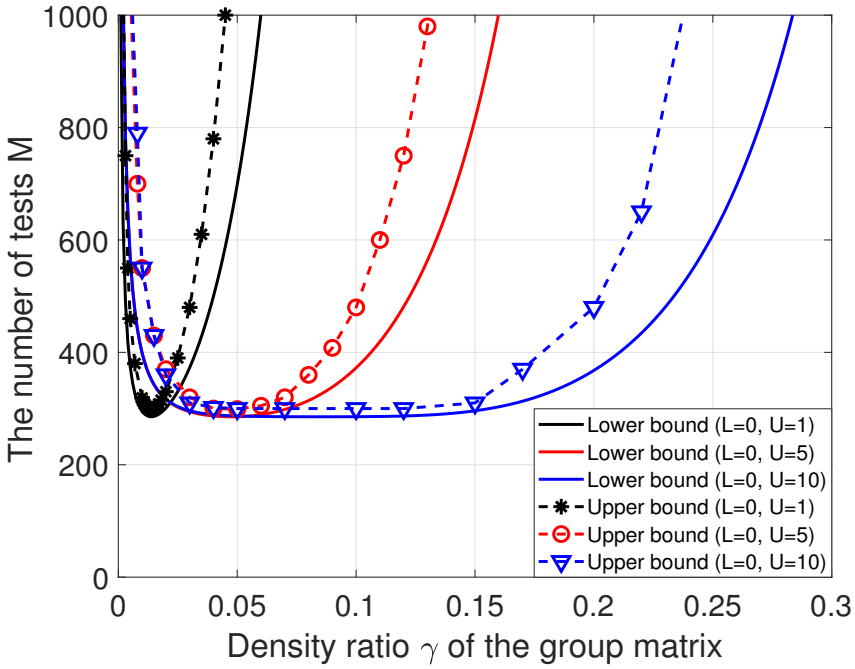


**Figure 4.** Comparisons of the upper and lower bounds (evaluated at $10^{-5}$) with different gaps $G$ for $N = 1000$ and $\delta = 0.05$: solid lines indicate the lower bounds and marked dashed lines indicate the upper bounds.

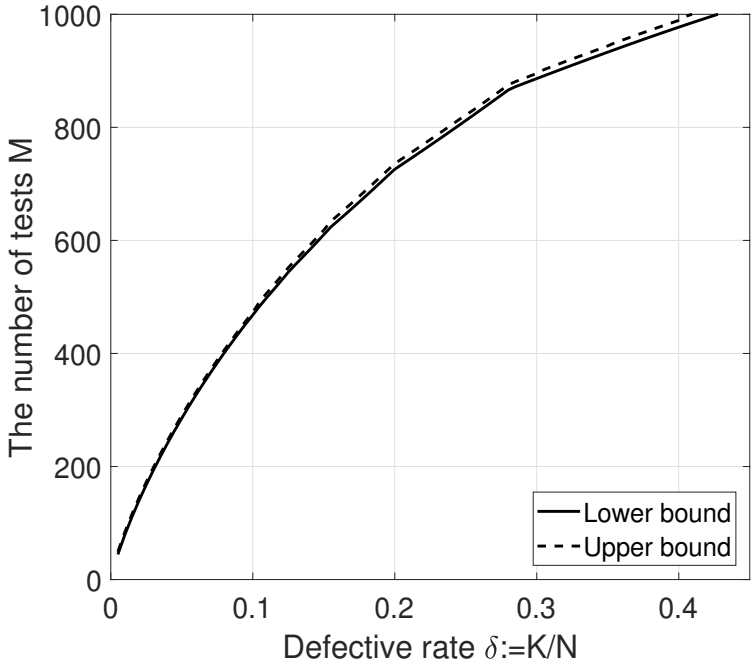


**Figure 5.** Comparisons of the upper and lower bounds (evaluated at $10^{-5}$) for $N = 1000$: solid lines indicate the lower bounds and dashed lines indicate the upper bounds.

## 4. Conclusions

In this paper, we considered a TGT framework with lower and upper thresholds which are the boundaries between positive and negative results. In addition, we derived the lower and upper bounds for finding defective samples out of a large of samples in TGTs. To this end, we exploited the minimization of the Hamming weight in channel coding theory. Moreover, we defined the probability of error for our decoding scheme. We found the new upper bounds on the probability of error which are well matched with the lower bounds obtained from the information-theoretic bound. We showed that the upper and lower bounds coincide with each other at the optimal density ratio of the group matrix. In addition, we observed that when the gaps between the two thresholds in the TGT scheme increase, the group matrix with high density should be used to achieve optimal performance. Through our results, we concluded how the design parameters of the TGT frameworks can affect the performance.

**Author Contributions:** Funding acquisition, J.-T.S.; methodology, J.-T.S.; supervision, J.-T.S.; writing–original draft, J.-T.S. The author has read and agreed to the published version of the manuscript.

**Funding:** This paper was supported by the National Research Foundation of Korea (NRF) grant funded by the korean government (NRF-2017R1C1B5075823).

**Conflicts of Interest:** The author declares no conflict of interest.

## References

1. Donoho, D.L. Compressed Sensing. *IEEE Trans. Inf. Theory* **2006**, *52*, 1289–1306. [CrossRef]
2. Du, D.-Z.; Hwang, F.-K. *Pooling Designs and Nonadaptive Group Testing: Important Tools for DNA Sequencing*; World Scientific: Singapore, 2006.
3. Bar-Lev, S.K.; Kleiner, I.; Perry, D.; Stadje, W. Recycled incomplete identification procedures for blood screening. *Eur. J. Oper. Res.* **2017**, *259*, 330–343. [CrossRef]
4. Tsybakov, A.; Likhanov, P. Packet communication on a channel without feedback. *Probl. Inf. Transm.* **1983**, *19*, 69–84.
5. Wolf, J.K. Born again group testing: multi-access communications. *IEEE Trans. Inf. Theory* **1984**, *31*, 185–191. [CrossRef]
6. Anderson, P.-O. *Superimposed Codes for the Euclidean Channel*; Linkoping University; Linkoping, Sweden, 1994.
7. Fan, P.Z.; Darnell, M.; Honary, B. Superimposed codes for the multiaccess binary adder channel. *IEEE Trans. Inf. Theory* **1995**, *41*, 1178–1182. [CrossRef]
8. Candes, E.; Romberg, J.; Tao, T. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inf. Theory* **2006**, *52*, 489–509. [CrossRef]
9. Amiri, E.; Tardos, G. High rate fingerprinting codes and fingerprinting capacity. In Proceedings of the 20th ACM-SIAM Sympos, Discrete Algorithms, New York, NY, USA, 4–6 January 2009.
10. Barg, A.; Blakley, G.R.; Kabatiansky, G.A. Digital fingerprinting codes: Problem statements, constructions, identification of traitors. *IEEE Trans. Inf. Theory* **2003**, *49*, 852–865. [CrossRef]
11. Desmedt, Y.; Duif, N.; Tilborg, V.H.; Wang, H. Bounds and constructions for key distribution schemes. *Adv. Math. Commun.* **2009**, *3*, 273–293. [CrossRef]
12. Colbourn, C.J.; Keri, G.; Rivas Soriano, R.P.; Schlage-Puchta, J.-C. Covering and radius-covering arrays: constructions and classification. *Discret. Appl. Math.* **2010**, *158*, 1158–1180. [CrossRef]
13. Jnr, E.A.; Key, J.D. *Designs and Their Codes*; Cambridge University Press: Cambridge, England, 1992.
14. Dyachkov, A.G.; Rykov, V.V. A coding model for a multiple-access adder channel. *Probl. Inf. Transm.* **1981**, *17*, 94–104.
15. Bar-David, I.; Plotnik, E.; Rom, R. Forward collision resolution—A technique for random multiple-access to the adder channel. *IEEE Trans. Inf. Theory* **1993**, *39*, 1671–1675. [CrossRef]

16. Laarhoven, T. Efficient probabilistic group testing based on traitor tracing. In Proceedings of the 51st Annual Allerton Conference on Communication, Control, and Computing, Monticello, IL, USA, 2–4 October 2013.
17. Bar-Lev, S.K.; Boneh, A.; Perry, D. Incomplete identification models for group-testable items. *Nav. Res. Logist.* **1990**, *37*, 647–659. [CrossRef]
18. Ganditota, V.; Grigorescu, E.; Jaggi, S.; Zhou, S. Nearly Optimal Sparse Group Testing. *IEEE Trans. Inf. Theory* **2019**, *65*, 2760–2773. [CrossRef]
19. Chan, C.L.; Che, P.H.; Jaggi, S.; Saligrama, V. Non-adaptive probabilistic group testing with noisy measurements: near-optimal bounds with efficient algorithms. In Proceedings of the 49th Annual Allerton Conference on Communication, Control, and Computing, Monticello, IL, USA, 28–30 September 2011.
20. Scarlett, J. Noisy Adaptive Group Testing: Bounds and Algorithms. *IEEE Trans. Inf. Theory* **2019**, *65*, 3646–3661. [CrossRef]
21. Dorfman, R. The Detection of Defective Members of Large Populations. *Ann. Math. Stat.* **1943**, *14*, 436–440. [CrossRef]
22. Damaschke, P. Threshold group testing. In *General Theory of Information Transfer and Combinatorics*; Springer: Berlin/Heidelberg, Germany, 2006; Volume 4123, pp. 707–718.
23. Bui, T.V.; Kuribayashi, M.; Cheraghchi, M.; Echizen, I. Efficiently Decodable Non-Adaptive Threshold Group Testing. *IEEE Trans. Inf. Theory* **2019**, *65*, 5519–5528. [CrossRef]
24. Bui, T.V.; Kuribayashi, M.; Cheraghchi, M.; Echizen, I. Improved encoding and decoding for non-adaptive threshold group testing. *arXiv* **2019**, arXiv:1901.02283.
25. Chan, C.L.; Cai, S.; Bakshi, M.; Jaggi, S.; Saligrama, V. Near-Optimal Stochastic Threshold Group Testing. In Proceeding of the 2013 IEEE Information Theory Workshop, Sevilla, Spain, 9–13 September 2013.
26. Chen, H.; Bonis, A.D. An almost optimal algorithm for generalized threshold group testing with inhibitors. *J. Comput. Biol.* **2011**, *18*, 851–864. [CrossRef]
27. De Marco, G.; Jurdzinski, T.; Rozanski, M.; Stachowiak, G. Subquadratic non-adaptive threshold group testing. *Fundam. Comput. Theory* **2017**, 177–189.
28. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*; Wiley: Hoboken, NJ, USA, 2009.
29. Seong, J.-T. A Bound for Finding Defective Samples in Threshold Group Testing. In Proceedings of the 2020 International Conference on Electronics, Information, and Communication (ICEIC), Bacelona, Spain, 19–22 January 2020.
30. Seong, J.-T. Density of Pooling Matrices vs. Sparsity of Signal of Group Testing Frameworks. *IEICE Trans. Inf. Syst.* **2019**, *E102*, 1081–1084. [CrossRef]