Efficient probabilistic group testing based on traitor tracing

Citation for published version (APA):
Laarhoven, T. M. M. (2013). *Efficient probabilistic group testing based on traitor tracing*. (arXiv.org; Vol. 1307.2438 [cs.IT]). s.n.

Document status and date:
Published: 01/01/2013

Document Version:
Publisher's PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:

- A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
- The final author version and the galley proof are versions of the publication after peer review.
- The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:

www.tue.nl/taverne

Take down policy
If you believe that this document breaches copyright please contact us at:

openaccess@tue.nl

providing details and we will investigate your claim.

Download date: 04. Oct. 2023

Efficient Probabilistic Group Testing Based on Traitor Tracing

Thijs Laarhoven*

July 10, 2013

arXiv:1307.2438v1 [cs.IT] 9 Jul 2013

Abstract

Inspired by recent results from collusion-resistant traitor tracing, we provide a framework for constructing efficient probabilistic group testing schemes. In the traditional group testing model, our scheme asymptotically requires $T \sim 2K \ln N$ tests to find (with high probability) the correct set of K defectives out of N items. Several other models are also considered, such as some noisy group testing and threshold group testing models, but we emphasize that this framework can be applied to many other variants of the classical model as well, both in adaptive and in non-adaptive settings.

1 Introduction

1.1 Group testing

Suppose a large population has to be tested for a certain illness, to determine which people are ill. One way to do this is to take blood samples from each person and test these samples one by one. However, if only few people are ill, many tests are wasted on non-infected people. It may then be advantageous to test bigger pools of several blood samples with *group tests*. If one of the tested people in a pool is ill, the test will come back positive and further tests are required, but if the test comes back negative, we may conclude that none of the people in the tested group are ill and many tests are saved. Group testing concerns the efficient identification of a small subset of K bad items (defectives) hidden among N total items, using the aforementioned group tests. The goal of group testing is to minimize the number of group tests T required to identify the defective items, by carefully choosing the groups to be tested.

Adaptive group testing In 1943, Dorfman [15] published a seminal paper studying practical ways of testing many blood samples of soldiers for syphilis, which is widely regarded as the first work on group testing. In the decades to follow, a lot of research was done in the area of *adaptive* group testing, where many sequential rounds of testing are considered, and the selection of samples for the next pool may be influenced by the results of the previous group test. In this adaptive setting, using a binary search, $T = K\lceil\log_2 N\rceil$ tests suffice to detect K defectives in a sample of size N. Up to a constant factor, this number of tests is optimal.

Non-adaptive group testing For practical and economical reasons, the focus of later work in group testing shifted more towards the *non-adaptive* setting, where many tests are run in parallel in one or few rounds. With certain combinatorial designs it is possible to find all K defectives in one round with $T = O(K^2 \log(N/K))$ tests [16], while a lower bound of $T = \Omega(K^2 \log N/\log K)$ [17] shows that this number of tests is nearly optimal, when one round of tests is done and when the group testing algorithm always has to identify the correct subset of defectives. If we allow for a small positive probability ε of not detecting the right set of defectives, then even in one round of tests, $T = O(K \log N)$ parallel tests suffice to isolate all defectives with high probability. Together with the lower bound of $T \geq K \log_2 N$ for large N [27], this shows that $T = \Theta(K \log N)$ is optimal. Chan et al. [8] recently gave a computationally efficient algorithm that belongs in the latter category that uses $T = eK \log(N/\varepsilon)$ tests to get a success probability of at least $1 - \varepsilon$.

Variants Besides the pure group testing model, variants have also been studied, such as noisy group testing [3, 4, 8, 10, 11, 28] and threshold group testing [9, 13, 24]. In these models a positive outcome of a test is not equivalent to at least one defective being present in the tested group, as there may be a small probability of making a mistake in the testing procedure, or because the test might not come back positive if very few defectives are present in the tested group. The trivial adaptive group testing algorithm of doing a binary search does not work in these models, and so even finding an efficient adaptive group testing scheme in these models not so easy.

1.2 Collusion-resistant traitor tracing

A completely different, but in fact closely related area of research is that of collusion-resistant traitor tracing. To protect digital content from unauthorized redistribution, copyright holders embed watermarks in the content such that, if an illegal copy is made and distributed, the watermark can be linked to the responsible user. Things become more complicated when several pirates collude, and start mixing their copies to create a new pirated copy of the content that does not match any of their copies of the content exactly. If in some segment of the data all pirates receive the same watermarked version, the *marking assumption* says that they are forced to output this version of the content. However, if they receive several different versions, they may choose any of them to output. *Traitor tracing* concerns assigning watermarks to N users in such a way that, even if K users mix their copies as described above, we may still be able to find the colluders. The goal of traitor tracing is to minimize the number of segments T needed to trace (part of) the coalition, by carefully choosing which watermarked versions of each segment to send to each user.

*T. Laarhoven is with the Department of Mathematics and Computer Science, Eindhoven University of Technology, The Netherlands.
E-mail: mail@thijs.com.

Static (non-adaptive) traitor tracing Work on traitor tracing started only in the late 20th century. In many of the early constructions, the number of segments required was polynomial in N, until Boneh and Shaw [7] gave an efficient construction that uses $T = O(K^4 \log(N/\varepsilon))$ segments to find at least one of the colluders with high probability in the static (non-adaptive) setting. Upper and lower bounds on T were further improved until in 2003, Tardos [31] showed that $T = O(K^2 \log(N/\varepsilon))$ segments are both necessary and sufficient. In the same paper he presented an efficient scheme that achieves this lower bound up to a constant factor. Later research focused on establishing the exact (asymptotic) lower bound [19], which turned out to be $T \gtrsim 2K^2 \ln N$, and decreasing the upper bounds by improving Tardos' scheme [6, 20, 26, 29], which eventually lead to an asymptotic bound of $T \sim \frac{1}{2}\pi^2 K^2 \ln(N/\varepsilon)$.

Dynamic (adaptive) traitor tracing While the above results are based on the static setting where the assignment of watermarks is fixed in advance, some work was also done on dynamic (adaptive) schemes. Besides the well-known deterministic scheme of Fiat and Tassa [18] which requires a large bandwidth, Tassa [32] constructed a low bandwidth dynamic scheme with a length of $O(K^4 \log N)$. Recently, Laarhoven et al. [22] gave a more efficient dynamic scheme where the number of segments is only $O(K^2 \log N)$, and in [21] a trade-off construction was given to build schemes that require a higher bandwidth but with a smaller constant T. In these schemes, all colluders are caught with high probability, whereas in non-adaptive schemes, at least one is caught with high probability.

Relation to group testing Oosterwijk et al. [26] recently considered optimizing Tardos' scheme to the scenario where the pirate strategy is known, e.g., when the pirates always randomly choose one of their versions (the *interleaving* attack) or when the pirates always output the same watermarked version if at least one of them received this version (the *all*-1 attack). The latter pirate strategy corresponds to getting an output of 1 if and only if at least one traitor is present in the set of users who received a 1. This directly corresponds to the traditional group testing game, where the test output is a 1 (positive) if at least one of the defectives has a 1 (is included in the test). Group testing can be seen as a special case of traitor tracing with a specific pirate strategy, and so traitor tracing results that are tailored specifically against certain strategies (such as those from [26]) may also be useful to group testing.

1.3 Contributions

In this paper, we will show that combining and improving several of the aforementioned results from traitor tracing [22, 23, 26, 31] leads to a group testing framework that can deal with many different group testing models efficiently. The resulting group testing algorithms we present are computationally efficient and, for sufficiently large K, require fewer tests than many known algorithms from the literature. For large N, the number of tests required in our schemes scales as follows, depending on the model. Here r is a noise-parameter, which informally corresponds to the probability of not getting the expected result.

- Traditional group testing: $T \sim 2K \ln N$.
- Noisy group testing (dilution): $T \sim 2K \ln N/(1 - r)$.
- Noisy group testing (additive): $T \sim 2K \ln N/(1 - \sqrt{2r})$.
- Noisy group testing (subtractive): $T \sim 2K \ln N/(1 - r)$.
- Noisy group testing (symm.): $T \sim 2K \ln N/(1 - \sqrt{2r})$.
- Threshold group testing (majority): $T \sim \pi K \ln N$.
- Threshold group testing (Bernoulli gap): $T \sim 4K \ln N$.
- Threshold group testing (linear gap): $T \sim 2K^2 \ln N$.
- Threshold group testing (unknown gap): $T \sim 2K^2 \ln N$.

These asymptotics apply to both adaptive and non-adaptive group testing, but the first order terms are considerably smaller in adaptive group testing than in non-adaptive group testing. Although we have worked out the details for several models, this paper also aims to provide a framework to efficiently deal with *any* group testing model. For instance, for threshold group testing with small gaps we did not provide explicit formulas, but one may derive them as we will explain below.

Besides these improvements and this framework, one goal of this paper is to further stimulate a cooperation between the areas of group testing and traitor tracing, as these areas are surprisingly similar. Much work has been done in both areas in similar directions (combinatorial designs, probabilistic analysis, information-theoretic lower bounds), and although the connection between the two areas has been made a few times before (e.g., [12, 25, 30]), a further exchange of ideas may lead to improved results in both areas.

The outline of this paper is as follows. In Sect. 2 we provide the aforementioned framework to deal with arbitrary group testing models. Then, in Sect. 3, 4, and 5 we apply our results to some previously considered models and present our results. Finally, in Sect. 6 we conclude by mentioning an important open problem in traitor tracing that might be of interest to the group testing community. All proofs and many details are omitted due to space limitations, but will appear in the full version.

2 Score-based group testing

In this section, we will look at a framework for probabilistic group testing with average-case errors. We will cover both adaptive and non-adaptive group testing. Before introducing this framework, we first introduce some more notation. We write X to denote the group testing matrix, or code matrix, indicating which items are included in which tests. We denote its length by T, which we will also call the code length. We denote test outcomes with y. Throughout, we will generally index items with j and tests with i, i.e., y_i is the outcome of the ith test, and $X_{j,i} = 1$ if and only if item j is included in the ith test. Finally, we write ε_1 for an upper bound on the probability that one or more non-defective items are marked as defective by our algorithm (getting one or more *false positives*), and ε_2 for an upper bound on the probability that some defective item is not marked defective (a *false negative*).

2.1 Non-adaptive group testing

In 2003, Tardos [31] introduced a collusion-resistant traitor tracing scheme, which he showed to be order-optimal in the number of segments needed. In group testing terminology, this scheme relies on assigning test scores to items based on the results of each test, such that if we add up all test scores for each item, defective items will eventually get much higher scores

Constructing the group test matrix X:

- For each i, j, set $X_{j,i} = 1$ with probability p.

Finding the defectives, given the test results y:

- For each i, j, calculate a *score* $S_{j,i} = h(X_{j,i}, y_i)$.
- For each j, compute the total score $S_j = \sum_i S_{j,i}$.
- Mark item j as defective if and only if $S_j > Z$.

Figure 1: The general outline of non-adaptive score-based group testing. The parameters p, h, Z, and T depend on the model and will be discussed later.

than non-defective items. Given a certain probability p, a score function h, and a threshold Z, this scheme works as described in Fig. 1. Here i refers to the ith test, and j refers to the jth item. [1]

For the time being we develop the theory for a generic score function h, but it is generally chosen such that it assigns positive scores to matches ($X_{j,i} = y_i$) and negative scores to differences, and gives large positive (negative) scores to the matches (differences) that were the least likely. For each test, the expected score for a non-defective item is usually 0, while for defective items it is strictly positive. Therefore, by running sufficiently many tests, with high probability we are able to distinguish between the scores of non-defective items (which have mean 0) and the scores of defective items (which have a large positive mean).

To analyze the performance of score-based schemes, we need to estimate the probabilities that (a) a non-defective item is still marked as defective, and (b) a defective item is not marked as defective. To do this, first note that for each item j, the scores for each test i are independently and identically distributed. For convenience, let us introduce the following notations for the mean and variance of the scores of non-defective items and defective items for each test. Below, we omit subscripts i on y, and we use x ($\tilde{x}$) to denote the symbol $X_{j,i}$ for non-defectives (defectives). Throughout, we will consistently use tildes to indicate variables corresponding to defective items.

$$\mu = \mathbb{E}[h(x, y)], \qquad \tilde{\mu} = \mathbb{E}[h(\tilde{x}, y)], \tag{1}$$

$$\sigma^2 = \mathrm{Var}[h(x, y)], \qquad \tilde{\sigma}^2 = \mathrm{Var}[h(\tilde{x}, y)]. \tag{2}$$

Now, the total score for an item j is given by $S_j = \sum_i S_{j,i}$, where $S_{j,i} = h(X_{j,i}, y_i)$. This is a sum of many i.i.d. random variables, and due to the Central Limit Theorem, for large T we expect S_j to be approximately normally distributed with mean μT ($\tilde{\mu} T$) and variance $\sigma^2 T$ ($\tilde{\sigma}^2 T$). So if we look at the average score per test $S_j^* = \frac{S_j}{T}$, non-defective items (defective items) will have a mean of μ ($\tilde{\mu}$) and a standard deviation of $\sigma^* = \frac{\sigma}{\sqrt{T}}$ ($\tilde{\sigma}^* = \frac{\tilde{\sigma}}{\sqrt{T}}$), as shown in Fig. 2. Therefore, when $\mu < \tilde{\mu}$ and σ and $\tilde{\sigma}$ are sufficiently small, increasing T will make both curves more narrow, and allow us to distinguish between the two curves with high probability. Working out the details, this leads to the following result about T and Z. The proof, as well as many other details, can be found in the full version of this paper.

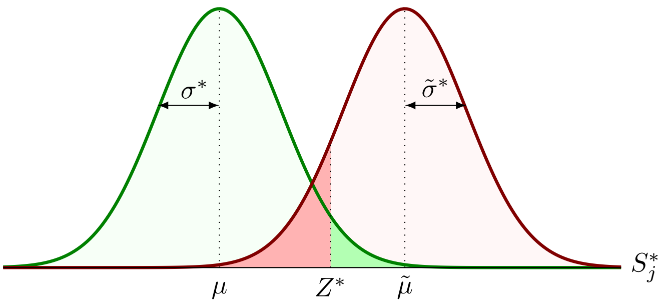


Figure 2: The Gaussian approximation of the score curves $S_j^* = S_j/T$, for non-defectives (left) and defectives (right). The means μ and $\tilde{\mu}$ do not depend on T, but $\sigma^* = \frac{\sigma}{\sqrt{T}}$ and $\tilde{\sigma}^* = \frac{\tilde{\sigma}}{\sqrt{T}}$ decrease when T increases. For sufficiently large T, choosing Z^* appropriately between μ and $\tilde{\mu}$ guarantees that the left (right) marked area has size at most $\frac{\varepsilon_2}{K}$ ($\frac{\varepsilon_1}{N}$).

Theorem 1. *Suppose we use the score-based non-adaptive group testing scheme described in Fig. 1, and the average item scores for each item follow a perfect Gaussian curve. Then, to guarantee that (i) a non-defective item is marked defective with probability at most $\frac{\varepsilon_1}{N}$, and (ii) a defective is marked as non-defective with probability at most $\frac{\varepsilon_2}{K}$, the following parameters suffice:*

$$T = \frac{2}{(\tilde{\mu} - \mu)^2} \left[\sigma \sqrt{\ln\left(\frac{N}{\varepsilon_1}\right)} + \tilde{\sigma} \sqrt{\ln\left(\frac{K}{\varepsilon_2}\right)} \right]^2, \tag{3}$$

$$Z = \left(\frac{\sigma \tilde{\mu} \sqrt{\ln\left(\frac{N}{\varepsilon_1}\right)} + \tilde{\sigma} \mu \sqrt{\ln\left(\frac{K}{\varepsilon_2}\right)}}{\sigma \sqrt{\ln\left(\frac{N}{\varepsilon_1}\right)} + \tilde{\sigma} \sqrt{\ln\left(\frac{K}{\varepsilon_2}\right)}} \right) \cdot T. \tag{4}$$

In particular, it then follows that with probability at least $1 - \varepsilon_1$, all non-defectives are correctly classified as non-defective, and with probability at least $1 - \varepsilon_2$, all defective items are correctly marked as defective.

For notational convenience, let us write

$$A = \frac{2\sigma^2}{(\tilde{\mu} - \mu)^2}, \quad B = \frac{\tilde{\sigma}}{\sigma}, \quad \eta_{\text{non}} = \sqrt{\frac{\ln(K/\varepsilon_2)}{\ln(N/\varepsilon_1)}}, \tag{5}$$

so that the formula for the parameter T in Thm. 1 can be concisely expressed as

$$T = A \ln\left(\frac{N}{\varepsilon_1}\right) \left[1 + B\eta_{\text{non}}\right]^2. \tag{6}$$

We generally have $\eta_{\text{non}} \leq 1$, while for small K and large N, the value of η will be very small. In fact, for $K = N^{o(1)}$ and $N \to \infty$ we have $\eta_{\text{non}} = o(1)$, leading to the following corollary.

Corollary 1. *Suppose that $K = N^{o(1)}$, that ε_1 and ε_2 are fixed, and that $B = O(1)$. Then, for large N we have*

$$T \sim A \ln N, \qquad Z \sim A\tilde{\mu} \ln N. \tag{7}$$

To minimize the number of tests, we are therefore mostly aiming to minimize the value of A. This parameter depends on the choices of p and h, and the model of how the test result y is produced.

[1] In traitor tracing, one actually has to use varying values of p for different tests, as otherwise an adversary could gain knowledge about p and use this knowledge to build a strong attack that can beat the system. However, in group testing there is no adversary, and therefore there is no reason to vary p for different tests.

Constructing X and finding the defectives:
For each $i = 1, \dots, T$, sequentially do the following.
(Initially $\mathcal{A} = \{1, \dots, N\}$ and $S_j(0) = 0$ for all j.)

- For each $j \in \mathcal{A}$, set $X_{j,i} = 1$ with probability p.
- Run the test, and obtain the test output y_i.
- For each $j \in \mathcal{A}$, do the following:
    - Compute $S_{j,i} = h(X_{j,i}, y_i)$.
    - Update $S_j(i) = S_j(i-1) + S_{j,i}$.
    - Mark j as defective if $S_j(i) > Z$.
- Remove all items marked defective from $\mathcal{A}$.

Figure 3: How to adapt the non-adaptive score-based group testing schemes to the adaptive setting, and gain the factor $\sqrt{K}$ in the first order error term.

2.2 Adaptive group testing

The procedure described in Fig. 1 can be adapted to the adaptive setting by making the following small modification: instead of only marking items defective if their *final* scores exceed Z, we mark an item defective (and do not include it in any of the remaining group tests) as soon as its score exceeds the threshold Z. This modification was recently proposed in [22] to build efficient adaptive traitor tracing schemes from Tardos' non-adaptive scheme, but can also be used to make score-based group testing work even more efficiently. The modified scheme is presented in Fig. 3.

It was shown in [22, 23] that with this modification, proving that the *average* defective item score exceeds Z is roughly enough to prove that *all* defective items are found. This means that instead of looking at scores of *single* defective items, we should now look at the *average* score of all defective items. Compared to the right curve in Fig. 2, this curve has the same mean $\tilde{\mu}$, but because it is an averaged score over K individual scores, the normalized standard deviation σ^* will be $\sqrt{K}$ times smaller. This leads to the following result, a proof of which can be found in the full version.

Theorem 2. *Suppose that we use the score-based adaptive group testing scheme described in Fig. 3, and suppose that the average item scores of all items follow a perfect Gaussian curve. Then, to guarantee that (i) a non-defective item is marked defective with probability at most ε_1/N, and (ii) a defective item is not marked defective with probability at most ε_2/K, the following parameters suffice:*

$$T = \frac{2}{(\tilde{\mu} - \mu)^2}\left[\sigma\sqrt{\ln\left(\frac{N}{\varepsilon_1}\right)} + \frac{\tilde{\sigma}}{\sqrt{K}}\sqrt{\ln\left(\frac{K}{\varepsilon_2}\right)}\right]^2 \qquad (8)$$

$$Z = \left(\frac{\sigma\tilde{\mu}\sqrt{\ln\left(\frac{N}{\varepsilon_1}\right)} + \frac{\tilde{\sigma}}{\sqrt{K}}\mu\sqrt{\ln\left(\frac{K}{\varepsilon_2}\right)}}{\sigma\sqrt{\ln\left(\frac{N}{\varepsilon_1}\right)} + \frac{\tilde{\sigma}}{\sqrt{K}}\sqrt{\ln\left(\frac{K}{\varepsilon_2}\right)}}\right) \cdot T. \qquad (9)$$

Similar to the non-adaptive group testing setting, we now write $\eta_{\text{ada}} = \sqrt{\frac{\ln(K/\varepsilon_2)}{K\ln(N/\varepsilon_1)}}$ so that the formula for the parameter T in Thm. 2 can be concisely expressed as

$$T = A\ln\left(\frac{N}{\varepsilon_1}\right)\left[1 + B\eta_{\text{ada}}\right]^2. \qquad (10)$$

The parameter η_{ada} is generally really small due to the factor $\sqrt{K}$. So without making any assumptions on K and N, we may already claim that for large K and/or N, the parameter η_{ada} will go to 0.

Corollary 2. *Suppose that ε_1 and ε_2 are fixed, and that $B = O(1)$. Then, for large N, we have*

$$T \sim A\ln N, \qquad Z \sim A\tilde{\mu}\ln N \qquad (11)$$

To summarize, the asymptotics of T will generally be the same as in the non-adaptive model, but the convergence to this limit will be much faster due to the extra factor $\sqrt{K}$. Also, as noted in [22], the actual number of tests needed to find all defectives is generally much less than the theoretical upper bounds suggest.

2.2.1 Dealing with unknown K

In [22], a scheme is discussed to effectively deal with adaptive scenarios where the number of defectives is not known in advance (the *universal* Tardos scheme), while maintaining equivalent asymptotics on T. This roughly comes down to using several thresholds Z, and the same idea may also be applied to adaptive group testing with an unknown number of defectives. For details, see [22, Sect. V].

2.2.2 Reducing the number of stages

In [22], a setting somewhere between non-adaptive and adaptive traitor tracing is also discussed (the *weakly dynamic* Tardos scheme), and how one could adapt the adaptive scheme to such a setting effectively. Translating those results to group testing, the same asymptotics on T hold even if the number of stages is reduced to $O(K)$ (with $O(T/K)$ tests in each stage). But reducing the number of rounds does lead to larger first order terms and larger practical code lengths. For details, see [22, Sect. IV].

2.3 Optimal score functions h

Recently, Oosterwijk et al. [26] studied the score functions used in traitor tracing, and showed that if the attack strategy of the pirates is known, then the score function h that minimizes A is given as follows. This choice of h is such that it is both *centered* ($\mu = 0$) and *quasi-normalized* ($\sigma^2 = \tilde{\mu}$).

Lemma 1. *[26, Cor. 6] The optimal, centered ($\mu = 0$) and quasi-normalized ($\sigma^2 = \tilde{\mu}$) score function h that minimizes A under the Gaussian assumption is given by*

$$h(x, y) = \frac{1}{K}\left.\frac{\partial \ln\left(p_{y|p_0,p_1}\right)}{\partial p_x}\right|_{p_1 = 1 - p_0 = p}, \qquad (12)$$

where $p_y = P(y_i = y)$ and $p_x = P(X_{j,i} = x)$.

For several attack strategies explicit formulas for h were derived in [26], some of which we will encounter later. For one particular strategy they obtained a score function that turned out to achieve capacity in the non-adaptive traitor tracing game.

2.4 Optimal probabilities p

Once the model (in traitor tracing: attack) is fixed, we can now compute the optimal score function h as described above, and we are almost done. To finalize the scheme, we then only need to choose a parameter p. Since the parameters A and B, and therefore a Gaussian-based estimate of the code length T, can be explicitly computed as a function of p, what remains is a straightforward optimization of p minimizing the estimate of T. Asymptotically, as shown in Cor. 1 and 2, we would like to choose p so as to minimize A, but in practice there is a trade-off between minimizing A and minimizing B. We will further discuss this below.

3 Traditional group testing

With the framework in place, we are ready to start building group testing schemes in arbitrary models, and we will discuss the results in the next few sections. We will naturally start with the most often considered, traditional group testing model, where the outcome of a test is positive if and only if at least one defective item is present in the tested pool. We will first give a scheme based on a straightforward optimization of h and p, and then discuss how the score function can be slightly refined in this particular model, leading to smaller constants T.

3.1 The direct approach

First, Oosterwijk et al. [26, Cor. 22] showed that the following centered and quasi-normalized score function is optimal in the ordinary group testing model, in that it minimizes A under the Gaussian assumption.

$$h(x,y) = \begin{cases} +p/(1-p) & (x,y)=(0,0) \\ -p(1-p)^{K-1}/(1-(1-p)^K) & (x,y)=(0,1) \\ -1 & (x,y)=(1,0) \\ +(1-p)^K/(1-(1-p)^K) & (x,y)=(1,1) \end{cases} \tag{13}$$

Using this score function, we can compute the parameters A and B as a function of p, and find the optimal value of p that minimizes T. For arbitrary values of K, these parameters are somewhat ugly functions of p, but the optimization of p is just a straightforward procedure. We will hide these less pretty details in a full version, and focus on the cleaner asymptotics of T here. Note that "asymptotics" here refers to considering large K, although the results may already provide good approximations of the actual value of T when K is small.

First, as is well known in group testing, one generally has to use small values of p and sparse matrices X. It is generally assumed that $p = \frac{\alpha}{K}$ for some α which is constant or almost constant in K. Using the same parametrization here, we obtain the following asymptotics for the code length constants A and B:

$$A = \frac{2(e^\alpha - 1)}{\alpha} K + O(1), \tag{14}$$

$$B = \sqrt{\frac{1}{e^\alpha - 1}} + O\left(\frac{1}{K}\right). \tag{15}$$

Here, it should be noted that the leading term of A is a strictly increasing function of α, while the leading term of B is strictly decreasing as α increases. There is a clear trade-off here between A and B, and the optimal choice of α depends on the exact set of parameters K, N, ε_1 and ε_2. We mention two simple choices of α and therefore p, and their respective code lengths:

$$p = \frac{1}{K} \Rightarrow T = 2(e-1)K \ln\left(\frac{N}{\varepsilon_1}\right)\left[1 + \frac{\eta}{\sqrt{e-1}}\right]^2, \tag{16}$$

$$p = \frac{\ln 2}{K} \Rightarrow T = 2K \log_2\left(\frac{N}{\varepsilon_1}\right)[1+\eta]^2. \tag{17}$$

If we focus on the regime of large K, we see that $\alpha \to 0$ is optimal to minimize A, in which case we get

$$T = 2K \ln\left(\frac{N}{\varepsilon_1}\right)\left(1 + O\left(\alpha + \frac{\eta}{\sqrt{\alpha}}\right)\right). \tag{18}$$

Setting $\alpha = O(\eta^{2/3})$ balances the order terms, and leads to a first order term of the order $O(\eta^{2/3})$. But the important thing to note here is the leading term of T:

$$T \sim 2K \ln N. \tag{19}$$

For sufficiently large K, this improves upon results of Chan et al. [8]. It has to be noted that in their schemes, there are never any false positives (i.e., $\varepsilon_1 = 0$), which is not true with the above construction. But if a small margin of error is present anyway (e.g., due to errors in the testing procedure), marked items may have to be tested individually anyway to confirm that the items are defective. In that case, allowing $\varepsilon_1 > 0$ makes sense. Note that the asymptotics of T are only a factor $2\ln 2 < 1.39$ above the information-theoretic lower bound [27, Thm. 2].

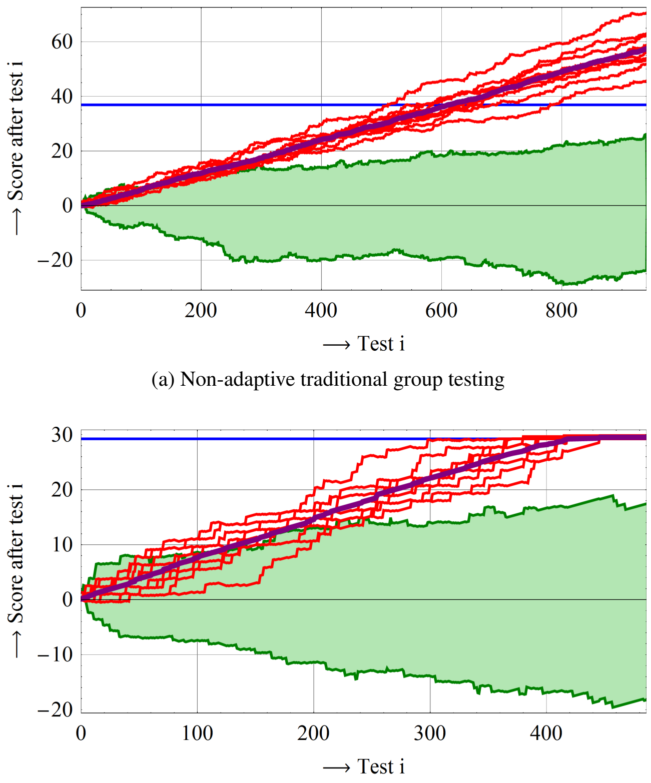
To give an idea of how the scheme actually works, an example is given in Fig. 4 with toy parameters $K = 10$, $N = 1000$, and $\varepsilon_{1,2} = 10^{-2}$. For the non-adaptive scheme, optimizing p then gives us $p \approx 0.091$ leading to $T \approx 941$ and $Z \approx 37$, while for the adaptive scheme we get $p \approx 0.055$ with $T \approx 486$ and $Z \approx 29$. [2]

3.2 Fine-Tuning the Score Function

Taking a step back from the score-based construction and looking at the group testing model, we know that if an item is included in a test ($x = 1$) while the test result is negative ($y = 0$), this item is *certainly* not defective. So in those cases, instead of assigning this item a somewhat negative score of -1 (which may not be enough to guarantee that the item is not marked defective), we may also assign items a score of $-\infty$ when they are included in a test which comes back negative. So we may fine-tune h by setting $h(1,0) = -\infty$. Then in each segment with probability $q = p(1-p)^K$ a non-defective item is assigned a score of $-\infty$, so with probability $1-(1-q)^T$ a non-defective has a score of $-\infty$ after T segments. Setting $p = \frac{1}{K}$ maximizes the latter probability, as was previously noted in [8], and eventually leads to an asymptotic code length of

$$T \sim \frac{2e(e-1)K \ln N}{2e-1} \approx 2.11 K \ln N. \tag{20}$$

[2] Note that in noiseless group testing, a trivial binary search leads to a much better adaptive scheme with $\varepsilon_{1,2} = 0$ and $T = K\lceil \log_2 N \rceil = 100$ tests. However, in noisy settings (Sect. 4) and several other models (Sect. 5) such trivial algorithms do not exist, and in those cases our adaptive construction may also be of interest.



(a) Non-adaptive traditional group testing

(b) Adaptive traditional group testing

Figure 4: An example of the score-based scheme in action, in the non-adaptive setting and in the adaptive setting. The green marked area shows the range of all non-defective item scores, the red lines show the scores of defective items, the horizontal blue line shows the threshold Z, and the purple bold line shows the average score of defective items. As one can see, in non-adaptive group testing we need this purple line to really exceed Z, while in the adaptive setting it suffices to let this average hit Z.

So also for $p = \frac{1}{K}$ we end up with improved asymptotics for T, compared to the $T = eK \ln\left(\frac{N}{\varepsilon_1}\right)$ of [8].

4 Noisy group testing

We saw in Sect. 3.2 that we may use the fact that the result of a test is never positive when one of the defective items is present in the pool, to fine-tune the score function and find all defectives even more efficiently. However, such certainties generally do not exist, as tests may have a small probability of not returning the correct result. Here we discuss two noisy group testing models previously considered in the literature, and show what the asymptotics on T become.

4.1 Dilution model

In the dilution model [3, 4, 10, 11, 28], we assume that a test may not come back positive even if a defective item is present in the tested pool, because this defective item may be *inactive* with a small probability r. This means that the probability that a defective item contributes a 1 to the test result is now not p, but $p' = p(1 - r)$. In this model, optimizing h leads to the centered and quasi-normalized score function given in Table 1. To minimize A, we again need to take α close to 0, in which case the asymptotic code length becomes

$$T \sim \frac{2K \ln N}{1 - r} \,. \tag{21}$$

This is somewhat comparable to a result of [4] which has a factor $\frac{1}{(1-r)^2} \approx \frac{1}{1-2r}$ in the denominator.

4.2 Additive model

Another commonly considered model is that of additive noise [3, 4, 28], where the final extraction of the test result may not always be correct. In particular, we assume that we are in the ordinary group testing model, but the output y may also be 1 with probability r if no defectives were actually present in the test. For this model, after fine-tuning h we get the score function given in Table 1. Again, for small r, the optimum choice of α (minimizing A) will be close to 0. However, in this case α will not approach 0 for fixed $r > 0$, as choosing $\alpha \ll r$ leads to large values of A. In fact, to optimize the leading term of A for fixed $r > 0$, one should choose α to satisfy $e^{\alpha}(1 - \alpha) = 1 - r$, which approximately corresponds to $\alpha \approx \sqrt{2r} + O(r)$. For the asymptotics of the code length we then get

$$T \sim \frac{2K \ln N}{1 - \sqrt{2r} + O(r)} \,. \tag{22}$$

Note that Atia and Saligrama [4] showed that a code length of the order $O(\frac{K \log N}{1-r})$ is already sufficient, and that our result, although practical, does not achieve this bound.

4.3 Subtractive model

If instead of flipping zeros to ones as in the additive model, a one is flipped to a zero with probability r (which one may call the *subtractive model*), the optimal asymptotic code length would be comparable to the dilution model with parameter r, as shown in Table 1. In particular, we should again let $\alpha \to 0$ to minimize A, and the asymptotic code length then becomes $T \sim \frac{2K \ln N}{1-r}$. Although this model seems very similar to the additive model, in this case the asymptotic code length is better when r is small. Together with the results from the dilution model, this motivates an earlier remark of [28] that dealing with false positives is apparently harder than dealing with false negatives.

4.4 Symmetric model

Finally, if the noisy "channel" that the test outcome passes through is symmetric [8], i.e. if any test result y may be wrong with probability r, then doing the above analysis leads to asymptotics comparable to the additive model, as shown in Table 1. In particular, we should not let $\alpha \to 0$ but choose $\alpha \approx \sqrt{2r}$ to minimize A, and we again get an asymptotic code length of $T \sim \frac{2K \ln N}{1-\sqrt{2r}+O(r)}$.

5 Threshold group testing

Finally, a model that has also been considered before is *threshold* group testing [9, 13], where a test result may only be positive if sufficiently many defective items are present in the tested pool. We will restrict our attention to the case where the test result is a (non-deterministic) function only of the number of defectives in a tested pool. This means that all positives are treated symmetrically, and the test result does not depend on how many non-defectives were present in the tested group. In all models, it is assumed that: if at most ℓ defectives are present in a test, the output will be negative; if at least u defectives are present, the test result is positive; and if the number of defectives β in a group test lies between $\ell + 1$ and $u - 1$, the result depends on the specific model.

5.1 Majority group testing

This model was introduced in [24], and considers the case where $y = 1$ if and only if more than half the defectives are present in the tested pool. This corresponds to $\ell = \frac{K-1}{2}$ and $u = \frac{K+1}{2}$. In this case, the score function h becomes a mess, but not if we immediately set p to its optimal value, which we find to be $p = \frac{1}{2}$. In that case, the score function reduces to the trivial function of $+1$ for matches and -1 for differences, as shown in Table 1. This score function is centered ($\mu = 0$) and normalized ($\sigma^2 = 1$), and working out the details for large K leads to an asymptotic code length of

$$T \sim \pi K \ln N. \tag{23}$$

Interpolating between the ordinary group testing model and majority group testing, one may expect that if $\ell = u - 1$ with $0 < \ell < \frac{K-1}{2}$, the optimal value of p is around $\frac{\ell}{K}$ and the asymptotics on T are between $2K \ln N$ and $\pi K \ln N$.

5.2 Bernoulli model

The Bernoulli gap model was previously considered in [9], and says that if the number of defectives in a pool is between $\ell + 1$ and $u - 1$, the probability that the test outcome is positive equals $q = \frac{1}{2}$. We will focus on the extreme case of $\ell = 0$ and $u = K$, although a similar analysis may be done for other values of ℓ and u. First, the optimal (centered, quasi-normalized) score function follows from [26, Cor. 22] and is given in Table 1. As in ordinary group testing, the optimal value of p lies close to 0, and for large K the asymptotic scaling of T is given by

$$T \sim 4K \ln N. \tag{24}$$

This can be generalized to arbitrary values of q, by replacing the 4 above by $\frac{2}{1-q}$. And again, interpolating between several results, if the gap between ℓ and u decreases, we expect the constant T to go down from $4K \ln N$ to $2K \ln N$ if $\ell \to u = K$ or $u \to \ell = 0$, and from $4K \ln N$ to $\pi K \ln N$ if $\ell, u \to \frac{K}{2}$.

5.3 Linear model

In the linear gap model [9, 14], the probability of the test result to be positive scales linearly with the number of defectives in the tested pool. We will again only focus on the case of an extreme gap ($\ell = 0$ and $u = K$) for ease of computation. First, the optimal centered and normalized score function follows from [26, Prop. 9] and is given in Table 1. As shown in [26, Prop. 10], for this model we have $\tilde{\mu} = \frac{1}{K}$ regardless of p, so the best we can do is choose p such that $\tilde{\sigma}^2$ is minimized. This leads to $p = \frac{1}{2}$ and $\tilde{\sigma}^2 = 1 - \frac{1}{K^2}$, and the asymptotic code length becomes

$$T \sim 2K^2 \ln N. \tag{25}$$

For large N this slightly improves upon a previous result of Del Lungo et al. [14], who gave an adaptive scheme with a code length of $T \sim 2K^2 \log_2 N > 2.88K^2 \ln N$.

5.4 Unknown model

Finally, if we assume that the output will be a 0 if no defectives are present, the output is 1 if all defectives are present, and we do not know what happens when *some* defectives are included in the test, then we are back at the traitor tracing game. For this game it is known that in the non-adaptive setting, the capacity-achieving choice is to use the same score function as in the linear gap model, but to vary p for each test by independently drawing it each time from the arcsine distribution (with distribution function $F(p) = \frac{2}{\pi} \arcsin \sqrt{p}$ on $(0, 1)$). This leads to the so-called interleaving defense, discussed in [23, 26], and leads to an asymptotic code length of

$$T \sim 2K^2 \ln N. \tag{26}$$

This asymptotic result is the same as in the linear gap model, which motivates why the linear gap model is the hardest group testing model to deal with.

6 Conclusion

In this paper we considered a new framework for probabilistic non-adaptive and adaptive group testing schemes, based on combining several results from traitor tracing. This lead to efficient group testing schemes for various models.

Although in this work we applied results from traitor tracing to group testing, one may wonder whether something can be done in the other direction as well. With the recent traitor tracing result of [26] achieving capacity in the non-adaptive traitor tracing game, the latter game seems kind of "solved". For the adaptive traitor tracing game one important open question remains, which is establishing the adaptive (dynamic) traitor tracing capacity. Not much is known about this yet, but perhaps combining previous techniques from adaptive group testing [1, 5] and non-adaptive traitor tracing [19] may bring us closer to a solution.

Acknowledgment

The author is grateful to Jan-Jaap Oosterwijk for many valuable suggestions and insightful discussions. The author further thanks Jeroen Doumen, Antonino Simone, Boris Škorić, and Benne de Weger for their valuable comments.

References

[1] M. Aldridge, "Adaptive Group Testing as Channel Coding with Feedback," *IEEE International Symposium on Information Theory (ISIT)*, pp. 1832–1836, 2012.

[2] N. Alon, V. Guruswami, T. Kaufman, and M. Sudan, "Guessing Secrets Efficiently via List Decoding," *ACM Transactions on Algorithms*, vol. 3, no. 4, pp. 1–16, 2007.

[3] G. K. Atia and V. Saligrama, "Noisy Group Testing: An Information Theoretic Perspective," *47th Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 355–362, 2009.

[4] G. K. Atia and V. Saligrama, "Boolean Compressed Sensing and Noisy Group Testing," *IEEE Transactions on Information Theory*, vol. 58, no. 3, pp. 1880–1901, 2012.

[5] L. Baldassini, O. Johnson, and M. Aldridge, "The Capacity of Adaptive Group Testing," *arXiv*, 2013.

[6] O. Blayer and T. Tassa, "Improved Versions of Tardos' Fingerprinting Scheme," *Designs, Codes and Cryptography*, vol. 48, no. 1, pp. 79–103, 2008.

[7] D. Boneh and J. Shaw, "Collusion-Secure Fingerprinting for Digital Data," *IEEE Transactions on Information Theory*, vol. 44, no. 5, pp. 1897–1905, 1998.

[8] C. L. Chan, P. H. Che, S. Jaggi, and V. Saligrama, "Non-adaptive probabilistic group testing with noisy measurements: Near-optimal bounds with efficient algorithms," *49th Allerton Conference on Communication, Control, and Computing*, pp. 1832–1839, 2011.

[9] C. L. Chan, S. Cai, M. Bakshi, S. Jaggi, and V. Saligrama, "Near-Optimal Stochastic Threshold Group Testing," *arXiv*, 2013.

[10] M. Cheraghchi, A. Hormati, A. Karbasi, and M. Vetterli, "Compressed Sensing with Probabilistic Measurements: A Group Testing Solution," *47th Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 30–35, 2009.

[11] M. Cheraghchi, A. Hormati, A. Karbasi, and M. Vetterli, "Group Testing with Probabilistic Tests: Theory, Design and Application," *IEEE Transactions on Information Theory*, vol. 57, no. 10, pp. 7057–7067, 2011.

[12] C. J. Colbourn, D. Horsley, and V. R. Syrotiuk, "Frameproof Codes and Compressive Sensing," *48th Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 985–990, 2010.

[13] P. Damaschke, "Threshold Group Testing," *General Theory of Information Transfer and Combinatorics*, LNCS vol. 4123, pp. 707–718, 2006.

[14] A. Del Lungo, G. Louchard, C. Marini, and F. Montagna, "The Guessing Secrets Problem: A Probabilistic Approach," *Journal of Algorithms*, vol. 55, pp. 142–176, 2005.

[15] R. Dorfman, "The Detection of Defective Members of Large Populations," *The Annals of Mathematical Statistics*, vol. 14, no. 4, pp. 436–440, 1943.

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

[17] A. G. D'yachkov, V. V. Rykov, and A. M. Rashad, "Superimposed distance codes," *Problems of Control and Information Theory*, vol. 18, no. 4, pp. 237–250, 1989.

[18] A. Fiat and T. Tassa, "Dynamic Traitor Tracing," *Journal of Cryptology*, vol. 14, no. 3, pp. 354–371, 2001.

[19] Y.-W. Huang and P. Moulin, "On the Saddle-Point Solution and the Large-Coalition Asymptotics of Fingerprinting Games," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 1, pp. 160–175, 2012.

[20] T. Laarhoven and B. de Weger, "Optimal Symmetric Tardos Traitor Tracing Schemes," *Designs, Codes and Cryptography*, 2012.

[21] T. Laarhoven, J.-J. Oosterwijk, and J. Doumen, "Dynamic Traitor Tracing for Arbitrary Alphabets: Divide and Conquer," *IEEE Workshop on Information Forensics and Security (WIFS)*, 2012.

[22] T. Laarhoven, J. Doumen, P. Roelse, B. Škorić, and B. de Weger, "Dynamic Tardos Traitor Tracing Schemes," *IEEE Transactions on Information Theory*, vol. 59, no. 7, pp. 4230–4242, 2013.

[23] T. Laarhoven, "Dynamic Traitor Tracing Schemes, Revisited," *Submitted*, 2013.

[24] V. S. Lebedev, "Separating Codes and a New Combinatorial Search Model," *Problems of Information Transmission*, vol. 46, no. 1, pp. 1–6, 2010.

[25] P. Meerwald and T. Furon, "Group Testing Meets Traitor Tracing," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4204–4207, 2011.

[26] J.-J. Oosterwijk, B. Škorić, and J. Doumen, "A Capacity-Achieving Simple Decoder for Bias-Based Traitor Tracing Schemes", *Submitted. A preliminary version appeared in 1st ACM Workshop on Information Hiding and Multimedia Security (IH&MMSec)*, pp. 19–28, 2013.

[27] A. Sebő, "On Two Random Search Problems," *Journal of Statistical Planning and Inference*, vol. 11, no. 1, pp. 23–31, 1985.

[28] D. Sejdinovic and O. Johnson, "Note on Noisy Group Testing: Asymptotic Bounds and Belief Propagation Reconstruction," *48th Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 998–1003, 2010.

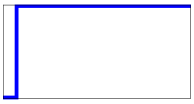
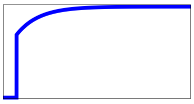
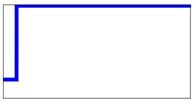
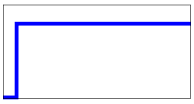
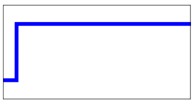
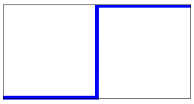
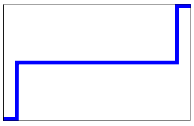
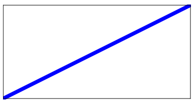
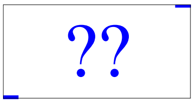
[29] B. Škorić, S. Katzenbeisser, and M. U. Celik, "Symmetric Tardos Fingerprinting Codes for Arbitrary Alphabet Sizes," *Designs, Codes and Cryptography*, vol. 46, no. 2, pp. 137–166, 2008.

[30] D. R. Stinson, T. van Trung, and R. Wei, "Secure Frameproof Codes, Key Distribution Patterns, Group Testing Algorithms and Related Structures," *Journal of Statistical Planning and Inference*, vol. 86, no. 2, pp. 595–617, 2000.

[31] G. Tardos, “Optimal Probabilistic Fingerprint Codes,” in *35th ACM Symposium on Theory of Computing (STOC)*, pp. 116–125, 2003.

[32] T. Tassa, “Low Bandwidth Dynamic Traitor Tracing Schemes,” *Journal of Cryptology*, vol. 18, no. 2, pp. 167–183, 2005.

Table 1: Optimal parameter choices for several group testing models, together with the resulting asymptotics on T. The plots on the left sketch $P(y = 1 \mid \beta)$ against β, where β is the number of defectives included in a group test, for each of the given models. The different cases for h always correspond to $(x, y) = (0, 0)$, $(0, 1)$, $(1, 0)$, and $(1, 1)$ respectively.

Model	**Optimal score function h**	**Optimal p**	**Asymptotics T**
Standard model 	$h(x,y) = \begin{cases} +p/(1-p) \\ -p(1-p)^{K-1}/(1-(1-p)^K) \\ -\infty \\ +(1-p)^K/(1-(1-p)^K) \end{cases}$	$p = \dfrac{O(\eta^{2/3})}{K}$ $\Big(p = \dfrac{1}{K},$	$T \sim 2K \ln N$ $T \sim 2.11K \ln N\Big)$
Noise: Dilution 	$h(x,y) = \begin{cases} +p(1-r)/(1-p(1-r)) \\ -p(1-r)(1-p(1-r))^{K-1}/(1-(1-p(1-r))^K) \\ -1+r/(1-p(1-r)) \\ +(1-p(1-r))^{K-1}(1-p)(1-r)/(1-(1-p(1-r))^K) \end{cases}$	$p = \dfrac{O(\eta^{2/3})}{K}$	$T \sim \dfrac{2K \ln N}{1-r}$
Noise: Additive 	$h(x,y) = \begin{cases} +p/(1-p) \\ -p(1-p)^{K-1}(1-r)/(1-(1-p)^K(1-r)) \\ -\infty \\ +(1-p)^K(1-r)/(1-(1-p)^K(1-r)) \end{cases}$	$p \approx \dfrac{\sqrt{2r}}{K}$	$T \sim \dfrac{2K \ln N}{1-\sqrt{2r}}$
Noise: Subtractive 	$h(x,y) = \begin{cases} +p(1-p)^{K-1}(1-r)/(r+(1-p)^K(1-r)) \\ -p(1-p)^{K-1}/(1-(1-p)^K) \\ -1+r/(r+(1-p)^K(1-r)) \\ +(1-p)^K/(1-(1-p)^K) \end{cases}$	$p = \dfrac{O(\eta^{2/3})}{K}$	$T \sim \dfrac{2K \ln N}{1-r}$
Noise: Symmetric 	$h(x,y) = \begin{cases} +p(1-p)^{K-1}(1-2r)/((1-p)^K(1-2r)+r) \\ -(1-p)^{K-1}p(1-2r)/(1-r-(1-p)^K(1-2r)) \\ -1+r/((1-p)^K(1-2r)+r) \\ (1-2r)/((1-p)^{-K}(1-r)-(1-2r)) \end{cases}$	$p \approx \dfrac{\sqrt{2r}}{K}$	$T \sim \dfrac{2K \ln N}{1-\sqrt{2r}}$
Threshold: Majority 	$h(x,y) = \begin{cases} +1 \\ -1 \\ -1 \\ +1 \end{cases}$	$p = \dfrac{1}{2}$	$T \sim \pi K \ln N$
Threshold: Bernoulli 	$h(x,y) = \big(p^{K-1} + (1-p)^{K-1}\big) \times \begin{cases} +p/(1-p^K+(1-p)^K) \\ -p/(1+p^K-(1-p)^K) \\ -(1-p)/(1-p^K+(1-p)^K) \\ +(1-p)/(1+p^K-(1-p)^K) \end{cases}$	$p = \dfrac{O(\eta^{2/3})}{K}$	$T \sim 4K \ln N$
Threshold: Linear 	$h(x,y) = \begin{cases} +p/(1-p) \\ -1 \\ -1 \\ +(1-p)/p \end{cases}$	$p = \dfrac{1}{2}$	$T \sim 2K^2 \ln N$
Threshold: Unknown 	$h(x,y,p) = \begin{cases} +p/(1-p) \\ -1 \\ -1 \\ +(1-p)/p \end{cases}$	$p \sim F,$ $F(p) = \dfrac{2}{\pi} \arcsin \sqrt{p}$	$T \sim 2K^2 \ln N$