Group Testing under Sum Observations
for Heavy Hitter Detection

Chao Wang, Qing Zhao, Chen-Nee Chuah

Abstract

We introduce a variation of the classic group testing problem referred to as group testing under sum observations. In this new formulation, when a test is carried out on a group of items, the result reveals not only whether the group is contaminated, but also the number of defective items in the tested group. We establish the optimal nested test plan within a minimax framework that minimizes the total number of tests for identifying all defective items in a given population. This optimal test plan and its performance are given in closed forms. It guarantees to identify all \( d \) defective items in a population of \( n \) items in \( O(d \log_2(n/d)) \) tests. This new formulation is motivated by the heavy hitter detection problem in traffic monitoring in Internet and general communication networks. For such applications, it is often the case that a few abnormal traffic flows with exceptionally high volume (referred to as heavy hitters) make up most of the traffic seen by the entire network. To detect the heavy hitters, it is more efficient to group subsets of flows together and measure the aggregated traffic rather than testing each flow one by one. Since the volume of heavy hitters is much higher than that of normal flows, the number of heavy hitters in a group can be accurately estimated from the aggregated traffic load.

Index Terms—Group testing, heavy hitter detection, anomaly detection, traffic measurements.

I. INTRODUCTION

A. Classic Group Testing

The group testing problem is concerned with identifying defective items in a given population. A test can be carried out over a group of items, and a binary test result is obtained, indicating whether the...
tested group contains any defective items. The objective is a test plan that minimizes the number of tests required for identifying all defective items.

The problem was first motivated by the practice of screening draftees with syphilis during World War II, and the idea of testing pooled blood samples from a group of people (rather than testing each person one by one) was initiated by Robert Dorfman [1]. In Dorfman’s test plan, draftees are tested in groups with a suitable size. If a group is tested positive, its members are tested one by one to identify the infected individual(s). An improvement to Dorfman’s test plan was proposed by Sterrett in 1957 [2]. The improvement suggests that once an infected person from a group is identified, the rest of the group is again tested together.

General formulations of and rigorous attacks on group testing were pioneered by Sobel and Groll in their paper published in 1959 [3]. Sobel and Groll adopted a probabilistic model on the defective items and focused on the problem of minimizing the expected number of tests. This formulation of group testing is later known as probabilistic group testing (PGT). Recognizing the intractability of the optimal solution to the general problem, Sobel and Groll introduced a class of special test plans referred to as the nested test plans. In a nested test plan, once a test reveals a defective group, the next test must be on a proper subset of this group. Sobel and Groll characterized implicitly the optimal nested test plan with a pair of recursive formulas and solved them numerically. They also established several asymptotic (as the population size approaches infinity) properties of the optimal nested test plan.

The counterpart to PGT is the combinatorial group testing (CGT) formulated and studied by Li [4] and Katona [5]. In CGT, there are $n$ items among which $d$ are defective. There is no probabilistic knowledge on the defective sets, and the objective is to minimize the number of tests in the worst case (i.e., a minimax formulation rather than a Bayesian formulation as in PGT) [6].

Group testing algorithms can be classified as adaptive or non-adaptive. Adaptive test plans are sequential in nature: which group to test next depends on the outcome of the previous tests. The studies in [3]–[5] mentioned above all focus on adaptive test plans. Non-adaptive group testing is a one-stage problem in which all actions are determined offline. Non-adaptive test plans are often represented by matrices [7], [8].

The classic group testing formulation has seen a wide range of applications, including chemical apparatus leakage detection [3], multiaccess communications [9]–[11], idle channel detection in the radio spectrum [12], compressed sensing [13], network tomography [14] and anomaly detection [15], [16]. In particular, non-adaptive group testing has been widely applied to DNA sequencing and DNA library screening [7], [17]–[19].
B. Group Testing under Sum Observations

In this paper, we introduce a variation of the classic group testing problem. Referred to as group testing under sum observations, this new formulation adopts a finer observation model in which a test reveals not only whether the tested group is contaminated, but also the number of defective items in the tested group. We focus on the combinatorial formulation of the problem and develop adaptive nested test plans. While the combinatorial nature of the problem generally leads to exponential complexity and the analytical form of the optimal nested test plan for the classic CGT is still an open problem, we show that for this new formulation of group testing, the optimal nested test plan can be obtained in closed form for any $n$ (the size of the population) and $d$ (the number of defective items). The closed forms for the optimal nested test plan as well as the optimal number of required tests have clear and appealing patterns (see Table I). They show that the optimal test plan guarantees to identify all $d$ defective items in a population of size $n$ in $O(d \log_2(n/d))$ tests, which is logarithmic with the population size and linear with the number of defective items. The reduction in the number of tests over the naive approach of testing items individually is thus significant, especially in applications when the population size is large but the defective items are few.

This new formulation of group testing is motivated by heavy hitter detection for traffic monitoring and anomaly detection in the Internet and other communication networks. For Internet traffic, it is a common observation that a small percentage of high-volume flows (referred to as heavy hitters) account for most of the total traffic [20]. In particular, it was shown in [21] that the top (in terms of volume) 9% of flows make up 90.7% of the total traffic over the Internet. Quickly identifying the heavy hitters is thus crucial to network stability and security. However, the large number of Internet flows makes individual monitoring extremely inefficient if not impossible. The idea of measuring aggregated traffic for heavy hitter detection has been considered (see, for example, [21]). In this paper, we explore a new group testing formulation for heavy hitter detection. The new sum observation model stems from the fact that the difference between the average traffic rates of heavy hitters and normal flows is large, which allows for accurate estimation of the number of heavy hitters from the aggregated traffic load as demonstrated by simulation examples given in Section [\(\checkmark\)]

\[1\] This assumes that $d \leq n/2$. For $d > n/2$, the number of tests is $O((n-d) \log_2(n/(n-d)))$. 
II. PROBLEM FORMULATION

In this section, we formulate the problem of group testing under sum observations. We focus on the combinatorial framework. Under the CGT formulation, we are given a population of $n$ items, each labelled with a unique ID. It is known that among these $n$ items, $d$ are defective. We assume that $1 \leq d \leq n - 1$ to avoid the trivial scenarios of $d = 0$ and $d = n$ and use $(n, d)$ to denote a specific CGT problem. The prior knowledge of $d$ takes the place of the prior probability of an item being defective under the PGT formulation. In Sec. IV, we address scenarios where the number of defective items is not known a priori.

For a given test plan $\pi$, the number of tests required by $\pi$ to identify all $d$ defective items in a population of size $n$ depends on which $d$ items are defective. Let $N_{\pi}(n, d; D)$ denote the number of tests required by $\pi$ when the $d$ defective items are given in the set $D$. Note that $n$ and $d$ are known while $D$ is unknown and is what the test plan needs to identify. Under the combinatorial formulation, the performance of a test plan is determined by the worst instant of $D$ among all subsets with size $d$. The performance of $\pi$, denoted by $N_{\pi}(n, d)$, is thus given by

$$N_{\pi}(n, d) = \max_{D \subset (n), |D| = d} N_{\pi}(n, d; D),$$

where $(n)$ denote the entire population.

Our objective is an optimal nested test plan $\pi^*$ given by

$$\pi^* = \arg \min_{\pi \in \Pi} N_{\pi}(n, d),$$

where $\Pi$ denotes the family of all admissible nested test plans. To simplify the notation, the performance of the optimal nested test plan $\pi^*$ is denoted by $N(n, d)$ (rather than $N_{\pi^*}(n, d)$), which will also be referred to as the optimal number of tests for identifying $d$ defective items in the population.

For any test plan, the first test must be on a proper subset of the population. Assume that the first test is on a subset of size $m$ ($m = 1, \ldots, n$) and the number of defective items in this subset is found to be $d_1$ ($d_1 = \max\{0, d + m - n\}, \ldots, \min\{m, d\}$). For a nested test plan, it then faces two CGT problems: $(m, d_1)$ and $(n - m, d - d_1)$. This is due to the fact that in a nested test plan, once a test reveals a defective group, the next test must be on a proper subset of this group. Considering the minimax nature of the CGT formulation, we arrive at the following recursive formula for the optimal number of tests:

$$N(n, d) = \min_{m=1, \ldots, m_{d_1} = \max\{0, d + m - n\}, \ldots, \min\{m, d\}} \{1 + N(m, d_1) + N(n - m, d - d_1)\}. \quad (1)$$

Define

$$M(n, d) = \arg \min_{m=1, \ldots, m_{d_1} = \max\{0, d + m - n\}, \ldots, \min\{m, d\}} \{1 + N(m, d_1) + N(n - m, d - d_1)\} \quad (2)$$
as the optimal size of the first group test for the CGT problem \((n,d)\). The values of \(M(n,d)\) for all \(n\) and \(d\) thus fully specifies the optimal nested test plan with its performance given by \(N(n,d)\). If there are multiple values of the group size \(m\) that achieve the minimum value in (2), we will set \(M(n,d)\) to be the minimum of such group sizes. A smaller group size is often desirable in practical applications.

The above recursive equations (1) and (2) specify an integer optimization problem. Analytical solutions to integer optimization are generally intractable, and the optimal nested test plan for the classic CGT under binary observations is still open. However, we show in the next section that the CGT under sum observations admits a clean solution in closed form.

### III. Optimal Nested Test Plan in Closed Form

The theorem below characterizes the optimal nested test plan \(M(n,d)\) and its performance \(N(n,d)\) in closed forms. We focus on cases with \(d \leq \frac{n}{2}\). Since identifying all \(d\) defective items is equivalent to identifying all \((n-d)\) normal items, we readily have \(N(n,d) = N(n,n-d)\), and the optimal nested test plan for \((n,d)\) is the same as that for \((n,n-d)\). For the rest of the paper, we assume \(d \leq \frac{n}{2}\) unless otherwise noted.

**Theorem 1:** For a CGT problem \((n,d)\) with \(d \leq \frac{n}{2}\), we have

\[
N(n,d) = (l + 1)d + k - 1,
\]

\[
M(n,d) = n - 2^l(d + k - 1),
\]

where

\[
l = \lceil \log_2(n/d) \rceil - 1,
\]

\[
k = \lceil n/2^l \rceil - d.
\]

The proof of Theorem 1 is involved and is omitted here.

To see the patterns of \(N(n,d)\) and \(M(n,d)\), we list them in TABLE I as sequences in \(n\) for a fixed \(d\) with each sequence starting at \(n = 2d\). From (3) and (4), the starting value of the sequences are given by \(N(2d,d) = 2d - 1\) and \(M(2d,d) = 1\). The remaining of the sequences has the following pattern. Consider first \(N(n,d)\) for a fixed \(d\). The sequence (except the first value) can be partitioned into frames, with each frame consisting of \(d\) equal-length segments. The segment length in the \(l\)th \((l = 1,2,...)\) frame is \(2^l\). The optimal number \(N(n,d)\) of tests takes the same value within a segment and increases by 1 from one segment to the next. The values of \(M(n,d)\) in each segment of the \(l\)th frame is simply given by
1, 2, \ldots, 2^l. With the help of TABLE I, we can also better understand the closed forms given in (3) and (4). In particular, \( l \) and \( k \) defined in (5) and (6) are, respectively, the indexes of the frames and segments.

Next, we show the order of the optimal number of tests in terms of \( n \) and \( d \). We can rewrite (3) as follows:

\[
N(n,d) = \lceil \log_2 \frac{n}{d} \rceil \cdot d + \left\lfloor \frac{n}{d} \right\rfloor - d - 1,
\]

where \( l \) is given in (5). It is easy to see that the second term \( P_2 \) is bounded in \( n \). We subsequently conclude that the optimal nested test plan guarantees to identify all \( d \) defective items in a population of \( n \) in
Fig. 1. The logarithmic order with $n$ of $N(n,d)$ ($d = 5, n \geq 10$).

$O(d \log_2 (n/d))$ tests, which is logarithmic with $n$ and linear with $d$. Fig. 1 illustrates the scaling behavior of $N(n,d)$ in $n$.

IV. CGT WITHOUT PRIOR KNOWLEDGE

We have so far focused on the standard CGT formulation which assumes a prior knowledge on the total number of defective items in the given population. For applications where this prior knowledge is unavailable, the first question is how to start the first test: for any population size $n$, should the first test be carried over the entire population or a proper subset of the population with the size potentially depending on $n$? In the theorem below, we show that within the class of nested test plan, the optimal action in the first step is to test the entire population. The first test will then reveal the total number $d$ of defective items, and the problem is reduced to a CGT of $(n,d)$.

Theorem 2: For a given population with any size $n$, within the class of nested test plans, the optimal action in the first step is to test the entire population.

V. APPLICATION TO HEAVY HITTER DETECTION

In this section, we consider the application of the optimal nested test plan developed in Section III and IV to the heavy hitter detection problem. Consider a network consisting of $n$ flows, among which $d$ are heavy hitters. We assume that each flow is an independent Poisson process with rate $\lambda_0$ for normal
flows and $\lambda_1$ for heavy hitters. Define
\[
\rho = \frac{d}{n}, \quad \eta = \frac{d\lambda_1}{d\lambda_1 + (n-d)\lambda_0}
\] 
(8)
as the fraction of heavy hitters in terms of number of flows and the total traffic volume, respectively. Over the Internet, we typically have $\rho$ around 10% to 20% and $\eta$ around 80% to 90%.

To apply the group testing plan to the heavy hitter detection problem, we need to estimate the number of heavy hitters from measurements of aggregated traffic rate. Assume that $m$ flows are aggregated together and $T$ measurements are taken. Based on the independent Poisson assumption on each flow, the aggregated traffic is again Poisson distributed with rate $d_1\lambda_1 + (m-d_1)\lambda_0$. From the $T$ measurements of the aggregated traffic, we can obtain a Maximum Likelihood (ML) estimate of the number $d_1$ of heavy hitters among this group of $m$ flows. This estimated $d_1$ will be used as the outcome of this group
test, which will determine the size of the next group test based on the optimal nested test plan given in Theorem 1.

Fig. 2 shows the simulation examples on the two types of detection errors for various values of $\rho$ and $\eta$. For convenience, we keep $\lambda_0 = 1$, so that the value of $\lambda_1$ will change with $\rho$ and $\eta$ according to (8). It is obvious that, for fixed $\rho$, when $\eta$ increases, the probabilities of missed detection and false alarm will both decrease. When $\eta$ is fixed, the groups with smaller $\rho$ will have lower detection errors. It’s easy to see from (8) that, when $\rho$ decreases and $\eta$ increases, the distance between $\lambda_1$ and $\lambda_0$ will become larger. It makes the estimation of $d_1$ more accurate. That’s why the probabilities of detection error will decrease. On the other hand, by comparing Fig. 2(a)(b) with (c)(d), we can see when the value of $T$ in each test changes from 1 to 2, the estimation error of $d_1$ becomes smaller with more measurements, therefore the final detection errors will decrease significantly.

VI. CONCLUSIONS

A variation of the classic group testing problem referred to as group testing under sum observations is introduced. Within the combinatorial group testing framework, the optimal nested test plan and its performance are established in closed form. These results find applications in heavy hitter detection for Internet traffic monitoring and anomaly detection.

REFERENCES

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

[2] A. Sterrett, “On the detection of defective members of large populations,” *The Annals of Mathematical Statistics*, vol. 28, no. 4, pp. 1033–1036, 1957.

[3] M. Sobel and P. A. Groll, “Group testing to eliminate efficiently all defectives in a binomial sample,” *Bell System Technical Journal*, vol. 38, no. 5, pp. 1179–1252, 1959.

[4] C. H. Li, “A sequential method for screening experimental variables,” *Journal of the American Statistical Association*, vol. 57, no. 298, pp. 455–477, 1962.

[5] G. O. Katona, “Combinatorial search problems,” *A survey of combinatorial theory*, pp. 285–308, 1973.

[6] D. Du and F. Hwang, *Combinatorial group testing and its applications*. World Scientific, 2nd ed., 2000.

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

[8] D. Du and F. Hwang, *Pooling Design and Nonadaptive Group Testing: Important Tools for DNA Sequencing*. World Scientific, 2006.

[9] J. Wolf, “Born again group testing: Multiaccess communications,” *IEEE Transactions on Information Theory*, vol. 31, pp. 185–191, Mar 1985.
[10] J. K. Wolf, “Principles of group testing and an application to the design and analysis of multi-access protocols,” in *The Impact of Processing Techniques on Communications*, pp. 237–257, Springer, 1985.

[11] T. Berger, N. Mehravari, D. Towsley, and J. Wolf, “Random multiple-access communication and group testing,” *Communications, IEEE Transactions on*, vol. 32, pp. 769–779, Jul 1984.

[12] A. Sharma and C. Murthy, “Group testing based spectrum hole search for cognitive radios,” *IEEE Transactions on Vehicular Technology*, vol. PP, no. 99, pp. 1–1, 2014.

[13] 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, pp. 7057–7067, Oct 2011.

[14] M. Cheraghchi, A. Karbasi, S. Mohajer, and V. Saligrama, “Graph-constrained group testing,” *IEEE Transactions on Information Theory*, vol. 58, pp. 248–262, Jan 2012.

[15] M. T. Thai, Y. Xuan, I. Shin, and T. Znati, “On detection of malicious users using group testing techniques,” in *Distributed Computing Systems, 2008. ICDCS ’08. The 28th International Conference on*, pp. 206–213, IEEE, 2008.

[16] S. Khattab, S. Gobriel, R. Melhem, and D. Mosse, “Live baiting for service-level DoS attackers,” in *INFOCOM 2008. The 27th Conference on Computer Communications*. IEEE, April 2008.

[17] W. J. Bruno, E. Knill, D. J. Balding, D. Bruce, N. Doggett, W. Sawhill, R. Stallings, C. C. Whittaker, and D. C. Torney, “Efficient pooling designs for library screening,” *Genomics*, vol. 26, no. 1, pp. 21–30, 1995.

[18] D. Balding, W. Bruno, D. Torney, and E. Knill, “A comparative survey of non-adaptive pooling designs,” in *Genetic mapping and DNA sequencing*, pp. 133–154, Springer, 1996.

[19] M. Farach, S. Kannan, E. Knill, and S. Muthukrishnan, “Group testing problems with sequences in experimental molecular biology,” in *Compression and Complexity of Sequences 1997. Proceedings*, pp. 357–367, Jun 1997.

[20] K. Thompson, G. Miller, and R. Wilder, “Wide-area internet traffic patterns and characteristics,” *IEEE Network*, vol. 11, pp. 10–23, Nov 1997.

[21] W. Fang and L. Peterson, “Inter-AS traffic patterns and their implications,” in *Global Telecommunications Conference, 1999. GLOBECOM ’99*, vol. 3, pp. 1859–1868 vol.3, 1999.