A Variant of Non-Adaptive Group Testing and Its Application in Pay-Television via Internet

Thach Bui, Oanh Nguyen, Van Dang, Nhung Nguyen, Thuc Nguyen

A VARIANT OF NON-ADAPTIVE GROUP TESTING AND ITS APPLICATION IN PAY-TELEVISION VIA INTERNET

Thach V. Bui[1], Oanh K. Nguyen[2], Van H. Dang[1] , Nhung T.H. Nguyen[1,3] and Thuc D. Nguyen[1]

[1] University of Science, Ho Chi Minh City, Vietnam
{bvthach,dhvan,nthnhung,ndthuc}@fit.hcmus.edu.vn *
[2] Saigon Technology University, Ho Chi Minh City, Vietnam
oanh.nguyenkieu@stu.edu.vn
[3] Japan Advanced Institute of Science and Technology, Japan

Abstract. In non-adaptive group testing (NAGT), the time for decoding is a crucial problem. Given an unknown string $x \in \{0,1\}^N$ with at most d ones, the problem is how to determine $x_i = 1$ using as few tests as possible so that x can be decoded as fast as possible. A NAGT can be represented by a $t \times N$ matrix. Although we do not know x, this matrix, which is called d-disjunct matrix, can reconstruct it exactly. In this paper, we consider a general problem, in which x is an array of N non-negative integer elements and has up to d positive integers. From nonrandom construction, we prove that we can decode a d-disjunct matrix, which is built from $[n,k]_q$-Reed-Solomon codes and identity matrix I_q, and recover x defined above in $poly(d) \cdot t \log^2 t + O(d^3 n \log(d \log N))$ with $t = O(d^2 \log^2 N)$. We also discuss this problem when x contains negative integer elements.
Pay-Television internet-based can be applied these results directly. Since the number of customers is very large, our system must be prevented from illegal buyers. This problem is called *traitor tracing*. To the best of our knowledge, this is the first result that raises a variant of NAGT and gets how to trace traitors without using probability.

Keywords: Group Testing, Traitor Tracing, Pay-TV via Internet.

1 Introduction

In the World War II, the authorities in USA enlisted millions of citizens to join the army. At that time, infectious diseases such as gonorrhea, syphilis, are serious problems. The cost for testing who was infected in turn was very expensive and it also took several times. They wanted to detect who was infected as

* This work was financially supported by the KC.01.TN16/11-15, Ministry of Science and Technology (MOST) grant and the National Foundation for Science and Technology Development (NAFOSTED), Vietnam.

fast as possible with the lowest cost. R. Dorfman [19], a statistician worked for United States Army Air Forces, proposed that we got N bloods samples from N citizens and combined groups of blood samples to test. It would help him detect infected/disinfected soldiers as few tests as possible. This idea formed a new research field: Group Testing. However, he did not give an explicit way to detect who was infected. D.-Z. Du and F. K. Hwang [18] gave a *naive* algorithm to solve this problem. If the test is negative, all soldiers, whose blood samples belong to this test, are not infected. Otherwise, at least one is infected. When we know who is not infected, the remaining soldiers are infected. For a formal definition, we represent Group Testing as a $t \times N$ binary matrix M, where each column stands for a sample and each row stands for a test. $M_{ij} = 1$ means the jth sample belongs to the ith test, and vice versa. The N infected/disinfected samples are considered as a vector $X = (x_1\ x_2\ ...\ x_N)^T$, where $x_j = 1$ if and only if (iff) jth sample is infected. An outcome vector, or an outcome of testing, is equal to $C = MX$. It is easy to map $C_i \geq 1$ to ith test which is infected. The time to decode C using naive algorithm is $O(tN)$. The decoding time is very important, however, not be considered for the long time. In 2010, P. Indyk, H.Q. Ngo and A. Rudra [1] proved that we could decode d-disjunct matrix in $poly(d) \cdot t \log^2 t + O(t^2)$. They also showed that these d-disjunct matrices were strongly explicit construction, e.g. any entry in M could be computed in time $poly(t)$. The other critical problem in Group Testing is how to generate d-disjunct matrices. There exists two approaches for this problem: probability ($t = O(d^2 \log N)$) and non-randomness ($t = O(d^2 \log^2 N)$). In many high accurate applications, e.g. cryptography, we can not use random construction because we want to control everything. Therefore, nonrandom construction is very important. In this paper, we only consider the nonrandom construction of d-disjunct matrix. For applications, NAGT can be found in data stream [24], data forensics [23] and DNA library screening [25].

List decoding has developed about 50 years. The initial works by [14] defined what list decoding was and gave some bounds for this code. For more details, please refer to the thesis of V. Guruswami [15]. List decoding has many applications and *traitor tracing* is a one of them [16]. Although A. Silverberg et.al. [16] found the relationship between the traitor tracing and list decoding in 2001, traitor tracing had already raised by Chor B. et. al. [12] seven years ago. Traitor tracing is very useful in systems which have pirated users.

Group testing, list decoding and traitor tracing have a strong relationship. In 2010, Indyk P., Ngo H.Q. and Rudra A. [1] proved that list decoding and group testing could be constructed in the same way. Next year, M. Peter, and T. Furon [3] proved that group testing and traitor tracing could be interchangeable since they had the same goal. However, they used probability to solve their problem. Therefore, the output might contain errors.

Digital television (TV) is widely used and studied [8]-[10]. However, when Pay-TV with cable and satellite TV become more and more popular, Pay-TV via internet is also a promising business using the advantages of broad-band networks. There would be a large number of users at the same time for live pro-

grams such as football matches or music live shows. One of the threats of this system is their customers can share their account with others. This would lead to bandwidth overload in rush hours. In 2002, C. Lobbecke et. al. [6] studied German's internet-based TV market. Although the entrance feasibility (include technology) is not clear, they believed that it will be popular. In 2011, according to [7], the world population was 7 billions. Among it, China's and India's population are 1,345.9 and 1,241.3 millions, respectively. Therefore, if only small fraction (assume 0.1%) of their population use Pay-TV internet-based, the number of users is over 1 million. Therefore, the rising problem is how to prevent and detect illegal users in this system. Shuhui HOU, et al. [11] showed that they could detect k colluders (d traitors in our term) with code length k^3 and support up to about k^4 users (N users in our terms). Using group testing, we can treat this problem better than them.

Our Main Result: In this paper, we present a variant of NAGT, show that the traitor tracing problem can be solved without using probability and illustrate these results through an application in Pay-TV via internet. To the best of our knowledge, this is the first result that raises variants of NAGT and gets how to trace traitors without using probability.

Paper outline: In Section 2, we present some preliminaries and prove the efficient decoding time of the variant of Group Testing. In Section 3, we describe its application in Pay-TV internet-based, compare the efficiency of our proposed solution and raise some open problems. The last Section is conclusion.

2 Preliminaries

2.1 d-disjunct matrix, Reed-Solomon codes and concatenated codes

An $t \times N$ binary matrix M is a *d-disjunct* matrix iff the union of at most d columns does not contain another columns. The rising problem is how to construct matrix M. Kauzt and Singleton [17] had a strongly explicit way to construct a binary superimposed code of order m based on Unique-Decipherable of order $m+1$ (UD_{m+1}) or Zero-False-Drop of order m (ZFD_m). There has a strong relationship between the binary superimposed code and the disjunct matrix. A matrix M being d-disjunct matrix is equivalent to a binary superimposed code of order d.

G. David Froney Jr. [20] presented a basic knowledge about *concatenated codes*. The concatenated codes are constructed by an *outer code* $C_{out} : [q]^{k_1} \to [q]^{n_1}$, where $q = 2^{k_2}$, and a binary *inner code* $C_{in} : \{0,1\}^{k_2} \to \{0,1\}^{n_2}$. Let $C = C_{out} \circ C_{in}$ be a concatenated code. C's size is $(n_1 n_2) \times q^{k_1}$. In [17], the authors chose C_{out} as a q-nary code and C_{in} as an identity matrix.

Reed-Solomon (RS) codes, which were invented by Reed, I.S. and Solomon, G. [21], are the famous codes that are applied in many fields [22]. They are not only q-nary codes but also the *maximum distance separable* codes. A $[n,k]_q$-code C, $1 \le k \le n \le q$, is a subset $C \subseteq [q]^n$ of size q^k. The parameters n, k and q are known as the *block length, dimension* and *alphabet size*. In this model,

we choose C_{out} as $[q-1,k]_q$-RS code and C_{in} as an identity matrix I_q. A d-disjunct matrix ($d = \lfloor \frac{n-1}{k-1} \rfloor$) is achieved from $C_{out} \circ C_{in}$ by putting all $N = q^k$ codewords as columns of the matrix. According to [17], given d and N, if we chose $q = O(d \log N)$, $k = O(\log N)$, the resulting matrix is $t \times N$ d-disjunct, where $t = O(d^2 \log^2 N)$. In 2010, P. Indyk, Hung Q. Ngo and A. Rudra [1] gave a random construction of d-disjunct matrices with $t = O(d^2 \log N)$ and cited to [17] for non-random construction with $t = O(d^2 \log^2 N)$, that can be decoded in $poly(d) \cdot t \log^2 t + O(t^2)$.

2.2 A variant of Group Testing and the connection between Traitor Tracing and Group Testing

In 2011, P. Meerwald and T. Furon [3] pointed out that there exists a connection between Group Testing and Traitor Tracing. Researchers in these fields aim to find very few specific people in a huge population. The authors used probability to illustrate their model. After estimating d (K in [3]), they computed $Score_j$ for jth user and checked whether $Score_j$ was larger than a threshold. If the answer is "Yes", jth user is infected, and vice versa. This procedure is called *single decoder with likelihood ratio test*. Since this method was relied on probability and the threshold, which was "ambiguous" parameter, their result might contain some false positive users. To improve the performance, the authors calculated scores for τ-tuples and reduced significantly the number of tests in practice. However, in general, since the probability of exact recovery is never equals to 1, using this solution may lead us accuse wrong users. Moreover, the time to find infected users is still questionaire for this approach.
The main question in Group Testing is how to identify the positive values in **binary** vector x from $y = Mx$, where $M_{t \times N}$ is d-disjunct and $x = (x_1\ x_2 \cdots x_N)^T$. The value x_i represents ith person. If this individual is infected iff $x_i = 1$. In 2012, T.D. Nguyen, T.V. Bui, V.H. Dang and D. Choi [2] extended this problem turn out: given $M_{t \times N}$ is d-disjunct and $y = Mx$, where $x = (x_1\ x_2 \cdots x_N)^T$ is a **non-negative** vector and $wt(x) = \sum_{i=1, x_i \neq 0}^{N} 1 \leq d$, find x from y and M.
For the convenience, the phrase "the frequency of jth user" and "the frequency of his keys" are equivalent. We also denote I_j is the jth column of I_q and $M_x(y)$ is (x, y) entry of $[n,k]_q$-RS codes. The following theorem describes the efficient decoding time of variant of Group Testing.

Theorem 1. *If any d-disjunct matrix $M_{t \times N}$ is constructed by concatenated codes, which is built from $[n,k]_q$-RS codes and identity matrix I_q, we can recover a non-negative integer vector $x_{N \times 1}$ from $y = Mx$ in $poly(d) \cdot t \log^2 t + O(d^3 n \log(d \log N))$, where $wt(x) \leq d$.*

Proof. According to Corollary C.1 [1], $[n,k]_q$-RS codes are $(d, O(d^2 \log(d \log N)))$-list recoverable codes ($N = q^k$ and $q = d \log N$). They can be decoded in $poly(d,q)$ [5] or $poly(d).t \log^2 t$ [4], and output $\mu = O(d^2 \log(d \log N))$ candidates. We denote the index of μ candidates of x is $\tau = \{\tau_1, \tau_2, \cdots, \tau_t\}$. After that, we split the result vector y into n blocks $C = \{S_1, S_2, \ldots S_n\}$, each block's size is q.

For each block, we decompose this block into set of symbols (in F_q) by the following rule: If $S_i = f_{i_1} I_{i_1} + f_{i_2} I_{i_2} + \ldots + f_{i_l} I_{i_l}$, where $f_{i_k} > 0$ for $k = 1, 2, \ldots, l$, then S_i can be represented as follow: $S_i = \{\{i_1, f_{i_1}\}, \{i_2, f_{i_2}\}, \ldots, \{i_l, f_{i_l}\}\}$, where $i = 1, 2, \cdots, n$. According to [2], the frequency of τ_jth user is $\min\{f_{i_k} : M_i(\tau_j) = i_k \in S_i,\ \forall i = 1, \ldots, n\}$. τ_jth user is infected iff $f_{\tau_j} \neq 0$. Since $|S_i| \leq d$, the time for finding infected users is $dn.O(d^2 \log(d \log N))$. Therefore, the overall time is: $poly(d) \cdot t \log^2 t + O(d^3 n \log(d \log N))$.

3 The application in Pay-TV internet-based and comparison

In [12], Chor B. et. al. proposed the scheme that although server only keeps t keys, it could support N users, where $t \ll N$. The key idea is each user holds *a set of keys* F, $|F| \leq t$. A pirated user will use a set of keys, that may be combined from a small group of traitors. The authors trace the traitors by using probability, which may create error tolerant. To overcome this issue, we propose a method that can find exactly who are traitors and how many times their keys are used.

3.1 Algorithm

In this scenario, at the time we check whether who is traitor, assume that all users log in our system and the pirated users are at most d. A d-disjunct matrix $M_{t \times N}$ is generated by the system. Assume $sum = M \times 1_{N \times 1}$ and $C = 0_{t \times 1}$. Every jth user is represented by a unique column M_j of M. Server stores t keys, denote that $F_{key} = \{k_1, k_2, \cdots, k_t\}$ and jth user stores a subset $F_j = \{k_h : M_{hj} = 1,\ where\ h = 1, \ldots, t\}$ of F_{key}. The key distribution procedure can apply the way that D. Boneh and M.Franklin proposed in [13]. When jth user is authenticated, server will increase the counter $C = C + M_j$. After this procedure, we calculate $trace = C - sum$ and use the Theorem 1 to find who is traitor and the frequency of his keys.

If jth user is disinfected, the frequency of the set F_j is 1. Therefore, if all users are disinfected, the vector counter will be equal to $C = \sum_{j=1}^{N} M_j = sum$. Note that there exists the bijection between M_j and F_j, if any hth user is infected, C will turn out: $C = sum + M_h$. Thus, after authentication phrase, if $trace = C - sum$ is not equal to zero, some users are counterfeited.

3.2 Comparison

In practice, the authors [11] only generate up to $p = 11$ (d in our term), the code length is 1332 and support at most 13431 users. Using MATLAB, we can generate d-disjunct matrices as defined in Section 2 and support the number of users as much as we want. For examples, a matrix that is generated from $[31, 3]_{32}$-RS codes and identity matrix I_{32} can support up to $32^3 = 32768$ users, detect at most $d = \lfloor \frac{31-1}{3-1} \rfloor = 15$ where the code length is $t = 31 \times 32 = 992$.

In theory, since the authors built ACC code from BIBD code, they faced many problems from this approach. D.-Z. Du and F. K. Hwang [18] pointed out that for the same d and N, the code length that was achieved from random construction is always smaller than BIBD construction. Last, in [11], the authors did not show how to find p colluders. In our solution, we satisfy this requirement as well. In the above model, we are only successful if the number of authenticated users is larger than N and all users who are legal must be log in your system. It seems to be practical since users pay money for this service. However, though the number of users is larger than N, some legal users are missing (they do not log in at that time). Hence, there are some illegal users. In this case, how can we identify them when the system that contains both pirated and missing users? The following proposition will answer who traitors are and also be *other variant* of NAGT.

Proposition 2. *Given d-disjunct matrix $M_{t\times N}$ is constructed by concatenated codes, which is built from $[n,k]_q$-RS code and identity matrix I_q, the positive integer vector $x_{N\times 1} = (x_1, x_2, \ldots, x_N)^T$ and binary vector $y_{N\times 1}$ such that: $x_i > wt(y)$ if $x_i > 0$ for $i \in [N]$ and $wt(x) + wt(y) \leq d$. If $z = M(x - y)$, we can identify the index of positive elements of x in time poly(t).*

Proof. Let denote $\gamma = x - y$. Since $wt(x) + wt(y) \leq d$, $|\gamma| \leq d$. Assume the index set of positive elements of x is $I = \{i_1, i_2, \ldots, i_h\}$. Thanks to $x_k > wt(y)$ where $k \in I$, the positive elements of x are also the positive elements of z. The output vector $z = (z_1, z_2, \ldots, z_t)^T$ will be converted in to positive/negative vector by the following rule: ith test is positive iff $z_i > 0$. After this conversion, using the Corollary C.1 in [1] to find the index of positive elements of x.

4 Conclusion

We present for the first time the variant of NAGT and the connection between traitor tracing and group testing without using probability. Simultaneously, we show that these results can be applied in Pay-TV via internet. Our future work aims at lowering the cost of decoding d-disjunct matrices which are constructed randomly, detect who is missing and find the frequency of pirated users in Pay-TV internet-based.

References

1. Indyk, P. and Ngo, H.Q. and Rudra, A. *Efficiently decodable non-adaptive group testing.* Proceedings of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms: 1126–1142, 2010.
2. T.D. Nguyen, T.V. Bui, V.H. Dang and D. Choi. *Efficiently Preserving Data Privacy Range Queries in Two-Tiered Wireless Sensor Networks.* In Ubiquitous Intelligence & Computing and 9th International Conference on Autonomic & Trusted Computing (UIC/ATC), 2012 9th International Conference on, pp. 973-978. IEEE, 2012.
3. P. Meerwald and T. Furon. *Group testing meets traitor tracing.* In Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE Inter. Conference on, pp. 4204-4207.

4. Alekhnovich, Michael. *Linear Diophantine equations over polynomials and soft decoding of Reed-Solomon codes.* In Foundations of Computer Science, 2002. Proceedings. The 43rd Annual IEEE Symposium on, pp. 439-448. IEEE, 2002.
5. F. Parvaresh and A. Vardy. "Correcting errors beyond the Guruswami-Sudan radius in polynomial time." Foundations of Computer Science, 2005. FOCS 2005. 46th Annual IEEE Symposium on. IEEE, 2005.
6. C. Lbbecke and M. Falkenberg. *A framework for assessing market entry opportunities for internet-based TV.* Inter. J. on Media Management 4(2) (2002): 95-104.
7. Population Reference Bureau. (2011). Population data sheet 2011. Washington, DC: Population Reference Bureau.
8. Y.-L. Huang, S. Shieh, F.-S. Ho and J.-C. Wang. *Efficient key distribution schemes for secure media delivery in pay-TV systems.* Multimedia, IEEE Transactions on 6, no. 5 (2004): 760-769.
9. Macq, Benoit M., and J-J. Quisquater. *Cryptology for digital TV broadcasting.* Proceedings of the IEEE 83, no. 6 (1995): 944-957.
10. C. Kim, Y. Hwang, and P. Lee. *Practical pay-TV scheme using traitor tracing scheme for multiple channels.* Information Security Applications (2005): 264-277.
11. S. HOU, T. Uehara, T. Satoh, Y. Morimura, and M. Minoh. *Fingerprinting codes for Internet-based live pay-TV system using balanced incomplete block designs.* IEICE transactions on information and systems 92, no. 5 (2009): 876-887.
12. Chor, B., Fiat, A., & Naor, M. (1994). *Tracing traitors.* In Advances in CryptologyCRYPTO94 (pp. 257-270). Springer Berlin/Heidelberg.
13. D. Boneh, and M. Franklin. *An efficient public key traitor tracing scheme.* In Advances in CryptologyCRYPTO99, pp. 783-783. Springer Berlin/Heidelberg, 1999.
14. Elias, Peter. *List decoding for noisy channels.* Massachusetts Institute of Technology, Research Laboratory of Electronics, 1957.
15. Guruswami, Venkatesan. *List decoding of error-correcting codes: winning thesis of the 2002 ACM doctoral dissertation competition.* Vol. 3282. Springer, 2005.
16. A. Silverberg, J. Staddon, and J. Walker. *Efficient traitor tracing algorithms using list decoding.* Advances in CryptologyASIACRYPT 2001 (2001): 175-192.
17. W.H.Kautz and R.C.Singleton. *Nonrandom binary Superimposed codes.* IEEE Transactions on Information Theory, 10(4): 363-377,1964.
18. Du, Dingzhu, and Frank Hwang. *Combinatorial group testing and its applications.* World Scientific Publishing Company Incorporated, 1993.
19. R.Dorfman. *The detection of defective members of large populations.* The Annals of Mathematical Statistics, 14(4): 436-440, 1943.
20. Froney Jr, G.D. *Concatenated codes.* DTIC Document, 1965.
21. Reed, I.S. and Solomon, G. *Polynomial codes over certain finite fields.* Journal of the Society for Industrial and Applied Mathematics, Vol.8(2): 300–304, 1960.
22. Wicker, S.B. and Bhargava, V.K. *Reed-Solomon codes and their applications.* Wiley-IEEE Press, 1999.
23. M. Goodrich, M. Atallah and R. Tamassia. *Indexing information for data forensics.* Applied Cryptography and Network Security. Springer Berlin/Heidelberg, 2005.
24. G. Cormode and S. Muthukrishnan. *What's hot and what's not: tracking most frequent items dynamically.* In Proceedings of the 22^{th} ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, pp. 296-306. ACM, 2003.
25. Ngo H.Q. & Du D.Z. (2000). *A survey on combinatorial group testing algorithms with applications to DNA library screening.* Discrete mathematical problems with medical applications, 55, 171-182.