Tracing Many Users With Almost No Rate Penalty

Noga Alon [*] Vera Asodi [†]

Abstract

For integers n, $r \geq 2$ and $1 \leq k \leq r$, a family $\mathcal{F}$ of subsets of $[n] = \{1, \ldots, n\}$ is called k-out-of-r multiple user tracing if, given the union of any $\ell \leq r$ sets from the family, one can identify at least $\min(k, \ell)$ of them. This is a generalization of superimposed families ($k = r$) and of single user tracing families ($k = 1$). The study of such families is motivated by problems in molecular biology and communication. In this paper we study the maximum possible cardinality of such families, denoted by $h(n, r, k)$, and show that there exist absolute constants $c_1, c_2, c_3, c_4 > 0$ such that $\min(\frac{c_1}{r}, \frac{c_2}{k^2}) \leq \frac{\log h(n,r,k)}{n} \leq \min(\frac{c_3}{r}, \frac{c_4 \log k}{k^2})$. In particular, for all $k \leq \sqrt{r}$, $\frac{\log h(n,r,k)}{n} = \Theta(1/r)$. This improves an estimate of Laczay and Ruszinkó.

1 Introduction

Let $[n] = \{1, 2, \ldots, n\}$, and let $\mathcal{F} \subseteq 2^{[n]}$ be a family of subsets of $[n]$. $\mathcal{F}$ is called r-superimposed if given the union of up to r sets from $\mathcal{F}$, one can identify all those sets. The problem of determining or estimating $f(n, r)$ - the maximum possible cardinality of an r-superimposed family of subsets of $[n]$ has been considered in various papers [3, 4, 5, 6, 7, 9]. This problem can be posed as a group testing problem, which is motivated by practical problems in molecular biology. Examples include the quality control of DNA chips, closing the remaining gaps in the genome at the end of a sequencing project and clone library screening. For more details see [2] and its references. As shown in [3, 9, 5],

$$\frac{c_1}{r^2} \leq \frac{\log f(n,r)}{n} \leq \frac{c_2 \log r}{r^2},$$

where $c_1, c_2 > 0$ are absolute constants.

A weaker requirement is that, given the union of up to r sets, one will be able to identify at least one of those sets. Such families are called r-single-user tracing superimposed (r-SUT), and were introduced by Csűrös and Ruszinkó [2]. This problem is also motivated by applications in molecular biology, where, for example, a group of DNA sequences that carry relevant genomic information is

[*] Schools of Mathematics and Computer Science, Raymond and Beverly Sackler Faculty of Exact Sciences, Tel Aviv University, Tel Aviv, Israel and IAS, Princeton, NJ 08540, USA. Research supported in part by the Israel Science Foundation, by a USA-Israeli BSF grant, by NSF grant CCR-0324906, by the James Wolfensohn fund and by the State of New Jersey. Email: nogaa@post.tau.ac.il.

[†] School of Computer Science, Raymond and Beverly Sackler Faculty of Exact Sciences, Tel Aviv University, Tel Aviv, Israel. Email: veraa@post.tau.ac.il.

under study, and the objective is to find at least one sequence with this information. Let $g(n,r)$ denote the maximum possible cardinality of an r-SUT family of subsets of $[n]$. The rate of such families is

$$\frac{\log g(n,r)}{n} = \Theta\left(\frac{1}{r}\right).$$

The upper bound was proved in [2] and the lower bound in [1].

Laczay and Ruszinkó introduced in [8] the notion of multiple user tracing families. This is a generalization of both r-superimposed and r-SUT families. For $r \geq 2$ and $1 \leq k \leq r$, a family $\mathcal{F} \subseteq 2^{[n]}$ is called k-out-of-r Multiple User Tracing ($MUT_k(r)$) if given the union of any $\ell \leq r$ sets from $\mathcal{F}$, one can identify at least $\min(\ell, k)$ of them. This problem also has applications in communication, search problems and molecular biology. See [8] for further discussion of such applications.

Let $h(n,r,k)$ denote the maximum possible cardinality of a $MUT_k(r)$ family of subsets of $[n]$. Laczay and Ruszinkó [2] have shown that

$$\frac{1}{5k(8e)^k r} \leq \frac{\log h(n,r,k)}{n} \leq \frac{2}{r}.$$

In this paper we improve their result and show that there exist absolute constants $c_1, c_2, c_3, c_4 > 0$ such that

$$\min\left(\frac{c_1}{r}, \frac{c_2}{k^2}\right) \leq \frac{\log h(n,r,k)}{n} \leq \min\left(\frac{c_3}{r}, \frac{c_4 \log k}{k^2}\right).$$

Note that this determines the maximum possible rate of $MUT_k(r)$ families for all $k \leq \sqrt{r}$ up to a constant factor, and that, somewhat surprisngly, in all this range the rate is $\Theta(\frac{1}{r})$, independently of k.

Throughout the paper log stands for the binary logarithm, and we omit all floor and ceiling signs whenever these are not crucial.

2 The Rate of Multiple User Tracing Families

Definition 1 *Let $r \geq 2$, $1 \leq k \leq r$. A family $\mathcal{F}$ of subsets of $[n]$ is called k-out-of-r multiple-user tracing superimposed ($MUT_k(r)$) if given the union of $\ell \leq r$ sets from $\mathcal{F}$ one can identify at least $\min(k, \ell)$ of these sets. That is, for all $t \geq 2$ and all choices of distinct $\mathcal{F}_1, \ldots, \mathcal{F}_t \subseteq \mathcal{F}$ with $1 \leq |\mathcal{F}_i| \leq r$ for all $1 \leq i \leq t$, the equality*

$$\bigcup_{A \in \mathcal{F}_1} A = \bigcup_{A \in \mathcal{F}_2} A = \ldots = \bigcup_{A \in \mathcal{F}_t} A$$

implies

$$\left|\bigcap_{i=1}^{t} \mathcal{F}_i\right| \geq k$$

Let $h(n,r,k)$ denote the maximum possible cardinality of a $MUT_k(r)$ family of subsets of $[n]$. We prove the following bounds on $h(n,r,k)$.

Theorem 1 *There exist absolute constants $c_1, c_2, c_3, c_4 > 0$ such that for any $r \geq 2$, $1 \leq k \leq r$ and $n \geq \max(100r, 8k^2)$,*

$$\min\left(\frac{c_1}{r}, \frac{c_2}{k^2}\right) \leq \frac{\log h(n,r,k)}{n} \leq \min\left(\frac{c_3}{r}, \frac{c_4 \log k}{k^2}\right).$$

The upper bound simply follows from the known bounds on the maximum possible cardinalities of SUT and superimposed families, since every $MUT_k(r)$ family is also r-SUT and k-superimposed. In the rest of this section we prove the lower bound.

Fix $r \geq 2$, $1 \leq k \leq r$ and $n \geq \max(100r, 8k^2)$. It is known that for every r' and $n' > r'^2$, there exists an r'-superimposed family of subsets of $[n']$ of size $2^{c\frac{n'}{r'^2}}$, where $c > 0$ is an absolute constant. Let $m = \min(2^{\frac{n}{40r}}, 2^{\frac{cn}{8k^2}})$, and let $X = \{1, \ldots, \lfloor \frac{n}{2} \rfloor\}$ and $Y = \{\lfloor \frac{n}{2} \rfloor + 1, \ldots, n\}$. Let $\mathcal{C} = \{C_1, \ldots, C_m\}$ be a $2k$-superimposed family of subsets of X. Now let $p = \frac{1}{r}$, and choose a family $\mathcal{D} = \{D_1, \ldots, D_m\}$ of subsets of Y at random, where the subsets D_i are chosen independently as follows. Every $y \in Y$ is chosen to be in D_i independently with probability p.

Define $F_i = C_i \cup D_i$ for all $1 \leq i \leq m$, and $\mathcal{F} = \{F_1, \ldots, F_m\}$. We next show that with positive probability the family $\mathcal{F}$ is $MUT_k(r)$. Thus, we show that, with positive probability, for all choices of $\mathcal{F}_1, \ldots, \mathcal{F}_t \subseteq \mathcal{F}$ such that $1 \leq |\mathcal{F}_i| \leq r$ for all $1 \leq i \leq t$ and $|\cap_{i=1}^t \mathcal{F}_i| < k$, the unions $\cup_{A \in \mathcal{F}_i} A$ for $1 \leq i \leq t$ are not all equal. To prove this we need the following Propositions.

Proposition 2 *The following holds with probability greater than $\frac{1}{2}$. For all $2k \leq s < 2r$, and for all distinct $A_1, \ldots, A_{k-1}, B_1, \ldots B_{s-k+1} \in \mathcal{D}$, there exists an element $y \in Y$ that belongs to none of the sets A_i, $1 \leq i \leq k-1$, and to exactly one of the sets B_i, $1 \leq i \leq s-k+1$.*

Proof: Fix $2k \leq s < 2r$ and distinct $A_1, \ldots, A_{k-1}, B_1, \ldots B_{s-k+1} \in \mathcal{D}$. The probability that there is no element $y \in Y$ that belongs to none of the sets A_i, $1 \leq i \leq k-1$, and to exactly one of the sets B_i, $1 \leq i \leq s-k+1$, is

$$\begin{aligned}
\left[1 - (s-k+1)p(1-p)^{s-1}\right]^{\frac{n}{2}} &\leq \left[1 - \frac{s-k+1}{r}\left(1 - \frac{1}{r}\right)^{2r-2}\right]^{\frac{n}{2}} \\
&\leq \left(1 - \frac{s-k+1}{r}e^{-2}\right)^{\frac{n}{2}} \\
&\leq e^{-e^{-2}\frac{(s-k+1)n}{2r}} \\
&< 2^{-\frac{3(s-k+1)n}{40r}}.
\end{aligned}$$

Thus, the expected number of choices of distinct $A_1, \ldots, A_{k-1}, B_1, \ldots B_{s-k+1} \in \mathcal{D}$, $2k \leq s < 2r$, for which there is no element $y \in Y$ that belongs to none of the sets A_i, $1 \leq i \leq k-1$, and to exactly one of the sets B_i, $1 \leq i \leq s-k+1$, is at most

$$\sum_{s=2k}^{2r-1} m^s 2^{-\frac{3(s-k+1)n}{40r}} = \sum_{i=0}^{2r-2k-1} m^{i+2k} 2^{-\frac{3(i+k+1)n}{40r}}$$

$$\begin{aligned}
&= m^{2k}2^{-\frac{3n(k+1)}{40r}}\sum_{i=0}^{2r-2k-1} m^i 2^{-\frac{3in}{40r}} \\
&\le 2^{\frac{nk}{20r}} \cdot 2^{-\frac{3nk}{40r}} \sum_{i=0}^{2r-2k-1} 2^{\frac{in}{40r}} \cdot 2^{-\frac{3in}{40r}} \\
&= 2^{-\frac{nk}{40r}} \sum_{i=0}^{2r-2k-1} \left(2^{-\frac{n}{20r}}\right)^i \\
&< 2^{-\frac{nk}{40r}} \cdot \frac{1}{1-2^{-\frac{n}{20r}}} \\
&< \frac{1}{2},
\end{aligned}$$

where the last inequality holds since $n \ge 100r$. Therefore, by Markov's inequality, the probability that there is no choice of $A_1, \ldots, A_{k-1}, B_1, \ldots B_{s-k+1} \in \mathcal{D}$ as above is greater than $\frac{1}{2}$. □

Proposition 3 *The following holds with probability greater than $\frac{1}{2}$. For all distinct $A_1, \ldots, A_r$, $B_1, \ldots B_r \in \mathcal{D}$,*

$$\bigcup_{i=1}^r A_i \not\subseteq \bigcup_{i=1}^r B_i.$$

Proof: Fix distinct $A_1, \ldots, A_r, B_1, \ldots B_r \in \mathcal{D}$. For $y \in Y$, the probability that $y \in \cup_{i=1}^r A_i$ and $y \notin \cup_{i=1}^r B_i$ is

$$\left[1-\left(1-\frac{1}{r}\right)^r\right]\left(1-\frac{1}{r}\right)^r \ge \frac{1}{2}e^{-1}(1-e^{-1}) > 0.1$$

Therefore,

$$Pr\left(\bigcup_{i=1}^r A_i \subseteq \bigcup_{i=1}^r B_i\right) < 0.9^{\frac{n}{2}},$$

and hence the expected number of choices of distinct $A_1, \ldots, A_r$, $B_1, \ldots B_r \in \mathcal{D}$, such that

$$\bigcup_{i=1}^r A_i \subseteq \bigcup_{i=1}^r B_i.$$

is at most

$$m^{2r}0.9^{\frac{n}{2}} \le 2^{\frac{n}{20}}0.9^{\frac{n}{2}} < \frac{1}{2},$$

for $n \ge 100r \ge 200$. Therefore, by Markov's inequality, the probability that there is no choice of $A_1, \ldots, A_r, B_1, \ldots B_r \in \mathcal{D}$ as above is greater than $\frac{1}{2}$. □

Proposition 4 *If $\mathcal{D}$ satisfies the properties in Propositions 2 and 3 then $\mathcal{F}$ is $MUT_k(r)$. Therefore, with positive probability, the family $\mathcal{F}$ is $MUT_k(r)$.*

Proof: Suppose $\mathcal{D}$ satisfies the properties in Propositions 2 and 3. We have to show that for all $\mathcal{F}_1, \ldots, \mathcal{F}_t \subseteq \mathcal{F}$ such that $1 \leq |\mathcal{F}_i| \leq r$ for all $1 \leq i \leq t$ and $| \cap_{i=1}^t \mathcal{F}_i| < k$, the unions $\cup_{A \in \mathcal{F}_i} A$ for $1 \leq i \leq t$ are not all equal. Consider first all such $\mathcal{F}_1, \ldots, \mathcal{F}_t$ for which

$$| \bigcup_{i=1}^t \mathcal{F}_i| < 2k.$$

For all $1 \leq i \leq t$, let $\mathcal{C}_i = \{A \cap X \mid A \in \mathcal{F}_i\}$. Since $\mathcal{C}$ is $2k$-superimposed, and since $|\mathcal{C}_i| < 2k$ for all $1 \leq i \leq t$, all the unions $\cup_{A \in \mathcal{C}_i} A$, $1 \leq i \leq t$, are distinct, and hence all the unions $\cup_{A \in \mathcal{F}_i} A$ are also distinct.

Next consider all $\mathcal{F}_1, \ldots, \mathcal{F}_t$ as above for which

$$2k \leq | \bigcup_{i=1}^t \mathcal{F}_i| = s < 2r.$$

Let $A_1, \ldots, A_{k-1}$ denote the sets in $\cap_{i=1}^t \mathcal{F}_i$, with addition of arbitrary sets from $\cup_{i=1}^t \mathcal{F}_i$, if there are less than $k-1$ sets in $\cap_{i=1}^t \mathcal{F}_i$. Let $B_1, \ldots, B_{s-k+1}$ be all other sets in $\cup_{i=1}^t \mathcal{F}_i \setminus \{A_1, \ldots A_{k-1}\}$. For all $1 \leq i \leq k-1$, let $A'_i = A_i \cap Y$, and for all $1 \leq i \leq s-k+1$, let $B'_i = B_i \cap Y$. Since $A'_1, \ldots, A'_{k-1}, B'_1, \ldots B'_{s-k+1}$ are sets in $\mathcal{D}$, and since $\mathcal{D}$ satisfies the property in Proposition 2, there is an element $y \in Y$ that belongs to none of the A'_i's, and to exactly one of the B'_i's, and thus to none of the A_i's and to exactly one of the B_i's. Since the B_i's are not in $\cap_{j=1}^t \mathcal{F}_j$, there exists $1 \leq \ell \leq t$ such that $\mathcal{F}_\ell$ does not contain this set, and hence $y \notin \cup_{A \in \mathcal{F}_\ell} A$. On the other hand, there is some $1 \leq \ell' \leq t$ for which $\mathcal{F}_{\ell'}$ contains this set. Therefore, $y \in \cup_{A \in \mathcal{F}_{\ell'}} A$, and hence $\cup_{A \in \mathcal{F}_\ell} A \neq \cup_{A \in \mathcal{F}_{\ell'}} A$, as needed.

Now consider the choices of $\mathcal{F}_1, \ldots, \mathcal{F}_t$ for which

$$| \bigcup_{i=1}^t \mathcal{F}_i| \geq 2r.$$

Let $B_1, \ldots, B_r$ denote the sets in $\mathcal{F}_1$, with addition of arbitrary sets from $\cup_{i=1}^t \mathcal{F}_i$ if $|\mathcal{F}_1| < r$. Let $A_1, \ldots, A_r$ be distinct sets in $(\cup_{i=1}^t \mathcal{F}_i) \setminus \{B_1, \ldots, B_r\}$. For all $1 \leq i \leq r$, let $A'_i = A_i \cap Y$ and let $B'_i = B_i \cap Y$. If all the unions $\cup_{A \in \mathcal{F}_i} A$ for $1 \leq i \leq t$ were equal, then we would have

$$\bigcup_{i=1}^r A'_i \subseteq \bigcup_{i=1}^r B'_i.$$

But as $A'_1, \ldots, A'_r, B'_1, \ldots, B'_r$ are distinct sets in $\mathcal{D}$, and $\mathcal{D}$ satisfies the property in Proposition 3, the above cannot hold. Thus, no choice of $\mathcal{F}_1, \ldots, \mathcal{F}_t$ with $| \cup_{i=1}^t \mathcal{F}_i| \geq 2r$ violates the desired property.

The assertion in each of the two propositions holds with probability exceeding 1/2, hence they hold simultaneously with positive probability. This completes the proof of Theorem 1. □

Note that by Theorem 1, the rate of $MUT_k(r)$ families for $k \leq \sqrt{r}$ is determined up to a constant factor, and is independent of k.

Corollary 5 *There are absolute positive constants c_1, c_2 such that for any $r \leq 2$, $1 \leq k \leq \sqrt{r}$ and $n \geq 100r$,*

$$\frac{c_1}{r} \leq \frac{\log h(n,r,k)}{n} \leq \frac{c_2}{r}.$$

References

[1] N. Alon and V. Asodi, Tracing a single user, European J. Combinatorics, to appear.

[2] M. Csűrös and M. Ruszinkó, Single-user tracing and disjointly superimposed codes, IEEE Transactions on Information Theory, vol. 51, no. 4, 1606-1611 (2005).

[3] A. G. Dyachkov and V. V. Rykov, Bounds on the length of disjunctive codes, Problemy Peredachi Informatsii, vol. 18, no. 3, 158-166 (1982).

[4] P. Erdős, P. Frankl, and Z. Füredi, Families of finite sets in which no set is covered by the union of r others, Israel J. Math., vol. 51, 79-89 (1985).

[5] Z. Füredi, A note on r-cover-free families, Journal of Combinatorial Theory Series A, vol. 73, no. 1, 172-173 (1996).

[6] F. K. Hwang and V. T. Sós, Non-adaptive hypergeometric group testing, Stud. Sci. Math. Hungarica, vol. 22, 257-263 (1987).

[7] W. H. Kautz and R. C. Singleton, Nonrandom binary superimposed codes, IEEE Transactions on Information Theory, vol. 10, 363-377 (1964).

[8] B. Laczay and M. Ruszinkó, Multiple user tracing codes, submitted.

[9] M. Ruszinkó, On the upper bound of the size of the r-cover-free families, Journal of Combinatorial Theory Series A, vol. 66, no. 2, 302-310 (1994).