Problem 5: A Strange Network

Team: *Belarus*

Abstract

In this paper we study the problem of recovering code with a number of subcodes known. Parameter α was redetermined, because for initial statement one can consider only binary codes satisfying given condition for any α. Then $T(n, k, \alpha)$ is also obviously redetermined.

In Theorem 1 the function T was proved to be monotonically decreasing by α, i.e.

$$T(n,k,1) \geq T(n,k,2) \geq \ldots \geq T(n,k,(n-1))$$

The following estimation for an arbitrary α was obtained using Theorem 1:

$$T(n,k,\alpha) > C_n^k - C_{n-2}^{k-2}$$

The following particular cases were considered (for all of them the corresponding recovering algorythm was found):

1. $\alpha = n - 1$.In this case $T(n,k,n-1) = C_n^k - C_{n-2}^{k-2} + 1$.
2. $\alpha = n - 2$. In this case $T(n,k,n-2) = C_n^k - C_{n-2}^{k-2} + 1$.
3. $\alpha = 1$, $k = n - 1$. In this case $T(n,n-1,1) = \left[\frac{n}{2}\right] + 2$.

Besides, the problem of non-suitability of pair (n, k) was considered, where pair (n, k) is called *non-suitable* if there exists a pair of codes X, Y of length n such that their multisets of k-subcodes are equal ($T(n,k,\alpha)$ cannot be found or estimated in this case).

Problem №5

First of all, α needs to be redetermined, because for initial statement one can consider only binary codes (since Theorem 1 shows this case gives maximal T) satisfying given condition for any α. Then $T(n, k, \alpha)$ is also obviously redetermined.

Definition. *Characteristic* of code is the number of distinct numbers containing in it.

Note that, without loss of generality, the numbers in code range from 0 to characteristic minus 1. Denote the biggest number in the code by α.

Remark that, by this definition, $1 \le \alpha \le n-1$.

Non-recoverable codes

For arbitrary code X denote the multiset of all k-subcodes of X by $P_k(X)$.

In fact, some of the source codes cannot be recovered at all. Really,

- codes [01] and [10] of length 2 have equal $P_1(X)$ = ([0], [1]).
- codes [0110] and [1001] of length 4 have equal $P_2(X)$ = ([00], [01], [01], [10], [10], [11]).
- codes [0111001] and [1001110] of length 7 have equal $P_3(X)$.

Definition. Call pair (n, k) *non-suitable* if there exists a pair of codes X, Y of length n such that $P_k(X) = P_k(Y)$ ($T(n,k,\alpha)$ cannot be found or estimated in this case). Otherwise the pair is called *suitable*.

Lemma 1. If pair (n,k) is non-suitable then pair $(n,k-1)$ is also non-suitable.

Proof. By definition, if the pair (n,k) is non-suitable, then there exist two codes (X and Y) of length n such that $P_k(X) = P_k(Y)$. Each of the k-subcodes corresponds to the certain multiset of $(k-1)$-subcodes (remove consequently each number from a k-subcode). Join all those $(k-1)$-subcodes in a multiset. On the other hand, each $(k-1)$-subcode appears in the multiset exactly $(n-k+1)$ times (fix a $(k-1)$-subcode and look for all k-subcodes containing it). Therefore, $P_{k-1}(X) = P_{k-1}(Y)$. □

Lemma 2. If pair (n,k) is non-suitable then pair $(n+1,k)$ is also non-suitable.

Proof. By definition, if the pair (n,k) is non-suitable, then there exist two codes (X and Y) of length n such that $P_k(X) = P_k(Y)$. Consider codes $X'=X+[0]$ and $Y'=Y+[0]$ (attach 0 to the right sides of both codes). A k-subcode of X' (Y') is either k-subcode of X (Y) or $(k-1)$-subcode of X (Y) with [0] attached. $P_k(X) = P_k(Y)$, $P_{k-1}(X) = P_{k-1}(Y)$ (by Lemma 1), thus $P_k(X') = P_k(Y')$, and pair $(n+1,k)$ is non-suitable. □

Corollary. If pair *(n,k)* is non-suitable then pair *(n+v,k-u)* is also non-suitable for arbitrary non-negative integers *v* and *u, k-u* > 0.

Recovering of codes

In all other considerations only *suitable* pairs *(n,k)* are considered.

Theorem 1. $T(n,k,1) \geq T(n,k,2) \geq \ldots \geq T(n,k,(n-1))$ for all suitable *n* and *k*.

Proof. Let's prove that $T(n,k,\alpha-1) \geq T(n,k,\alpha)$ for any $\alpha > 1$.

By definition of $T(n, k, \alpha)$ there exist two different codes *X*, *Y* with characteristic *α+1* such that $|P_k(X) \cap P_k(Y)| = T(n,k,\alpha) - 1$.

Since *X* and *Y* are different there exists a position such that the numbers on this position in codes differ (without loss of generality these numbers are 0 and 1). Replace all *α*'s in *X* and *Y* with 1. We obtained different codes *X'* and *Y'* with characteristic *α* such that $|P_k(X') \cap P_k(Y')| \geq T(n,k,\alpha) - 1$. The result follows. □

Claim 1. There exist C_{n-2}^{k-2} subcodes of $X[a_1,\ldots,a_n]$ containing both a_i and a_j for all pairs *(I ,j)* of distinct numbers ($1 \leq i \leq n;\ 1 \leq j \leq n;$).

Proof. Really, if they a_i and a_j are fixed in the subcode, *k-2* of remaining *n-2* numbers can be chosen to complete it. □

Lemma 3. $T(n,k,\alpha) > C_n^k - C_{n-2}^{k-2}$ for all suitable *n* and *k*.

Proof. Suppose all numbers in the code are distinct. Consider the numbers on the two last positions of the source code (a_n and a_{n-1}). If there does not exist a subcode, containing both a_{n-1} and a_n, then we cannot find out the positions of each of them with respect to each other.

There are C_n^k *k*-subcodes, and C_{n-2}^{k-2} subcodes containing both a_{n-1} and a_n (by Claim 1). So, $C_n^k - C_{n-2}^{k-2}$ subcodes are not enough according to the considerations above. Therefore, $T(n,k,n-1) > C_n^k - C_{n-2}^{k-2}$.

By Theorem 1, $T(n,k,\alpha) \geq T(n,k,n-1) > C_n^k - C_{n-2}^{k-2}$. □

Proposition 1. $T(n,k,n-1) = C_n^k - C_{n-2}^{k-2} + 1$.

Proof. For each pair of distinct numbers *(i, j)* there exists at least one subcode containing both a_i and a_j (since there are only $C_n^k - C_{n-2}^{k-2}$ subcodes, containing at most one of that two numbers by Claim 1). That's why we can find out relative order for each pair of numbers *(i, j).* Thus one can find the order of the numbers, and therefore recover the code. □

Proposition 2. $T(n,k,n-2) = C_n^k - C_{n-2}^{k-2} + 1$.

Proof. In this situation exactly one number is duplicated in the code (without loss of generality, 0). For any two non-zero number we know their relative order (by lemma 3). It's left to notice that we can recover the order of non-zero numbers with respect to zeroes in the same way. Really, the numbers situated to the left of zeroes in all subcodes containing zeroes, are situated to the left side of both zeroes in the source code, and so on. □

Theorem 2. $T(n,n-1,1) = \left\lceil \frac{n}{2} \right\rceil + 2$.

Proof.

1) Show that $\left\lceil \frac{n}{2} \right\rceil + 1$ subcodes is not enough to recover the source code.

If $n = 2k$: pair of codes $([\underbrace{00\ldots01}_{k} \mid \underbrace{00..00}_{k}]$ and $[\underbrace{00\ldots00}_{k} \mid \underbrace{10\ldots00}_{k}])$ has $\left\lceil \frac{n}{2} \right\rceil$ common equal subcodes $[\underbrace{00\ldots0}_{k-1}\,1\,\underbrace{0\ldots00}_{k-1}]$ and one common subcode $[\underbrace{00..00}_{2k-1}]$ - $\left\lceil \frac{n}{2} \right\rceil + 1$ common subcodes.

If $n = 2k+1$: pair of codes $([\underbrace{00\ldots0}_{k} |1| \underbrace{11\ldots11}_{k}]$ and $[\underbrace{00\ldots00}_{k} |0| \underbrace{11\ldots11}_{k}])$ has $\left\lceil \frac{n}{2} \right\rceil + 1$ common equal subcodes $[\underbrace{00\ldots0}_{k}\,\underbrace{11..11}_{k}]$ - $\left\lceil \frac{n}{2} \right\rceil + 1$ common subcodes.

2) Now show that $\left\lceil \frac{n}{2} \right\rceil + 2$ subcodes are enough to restore the source code. Denote maximal number of equal subcodes by *M*.

2.1) Suppose M > $\left\lceil \frac{n}{2} \right\rceil$. Let's prove lemma 4

Lemma 4. The source code has *w* equal subcodes if and only if it has a group of *w* equal numbers situated one after another, and the equal subcodes were obtained by excluding numbers from this group.

Proof. Suppose the source code has *w* equal subcodes. Notice that the subcodes are obtained by excluding the same number (w.l.o.g. 0; otherwise the number of zeroes in equal subcodes is different). Now it's left to prove that there are no 1's between the 0's being excluded. Assume the contrary. Then the code can be represented as $X[a_1,\ldots,a_n] = X_1 \cup [0] \cup X_2 \cup [1] \cup X_3 \cup [0] \cup X_4$, where X_i are some (possibly empty) subcodes, and X_2 consists of 0's. By assumption $X_1 \cup [0] \cup X_2 \cup [1] \cup X_3 \cup X_4 = X_1 \cup X_2 \cup [1] \cup X_3 \cup [0] \cup X_4$, therefore the last digit of X_2 is 1 – contradiction. The result follows.

The converse proposition is obviously true. □

According to Lemma 4, the source code has a group of more than $\left\lceil \frac{n}{2} \right\rceil$ going one after another equal numbers, and each of these subcodes is a group of not less than $\left\lceil \frac{n}{2} \right\rceil$ equal numbers. A subcode contains *n-1* numbers in total, so there is not more than one such group in it. If one more number of the same sort is added to the group in subcode, the source code is recovered. In this case at most $\left\lceil \frac{n}{2} \right\rceil + 1$ subcodes are required.

2.1) Let M ≤ $\left\lceil \frac{n}{2} \right\rceil$. It means that there exist two distinct subcodes.

Make a table by writing down all given subcodes one below another, (example for *n=8, k=7*):

0	0	0	1	1	1	1
0	0	0	0	1	1	1
0	0	0	0	1	1	1
0	0	0	0	1	1	1
0	0	0	0	1	1	1
0	0	0	0	1	1	1

Denote by *c(a,i)* the amount of numbers obtained from *a-th* position of the source code in *i-th* column and denote by *d(x,i)* the amount of numbers *x* in *i-th* column.

Claim 2. *c(a,i-1)+c(a,i)=T or c(a,i-1)+c(a,i)=T* for any position *a* of the source code, *1≤i≤n.* Let *c(a,0) and c(a,n)* be equal zero.

Proof. For each chosen position there exists at most one subcode not containing the number from this position of the source code. Thus, the number from this position appears in the table in either *T-1* or *T* cases. Now show that it can only appear in either *i-1-th* or *i*-th column. If the subcode is obtained from the source code by removing a number from position that goes before *i-th,* then this number shifts to (*i-1)-th* column, otherwise it occupies its former position.

Lemma 5. If two distinct subcodes have common fragment at the beginning (end), then the source code also contains this fragment at the beginning (resp. end).

Proof. Assume the contrary, i.e. there exist some positions, such that the source code and these fragments vary in this position. Without loss of generality, consider fragment at the beginning. Let *A* be the number of the first position (counting from the beginning) with different numbers. Suppose that this two subcodes are obtained from the source code by removing numbers from x_1*-th* and x_2*-th* positions ($x_1 < x_2$). Notice that $x_1 \le A \le x_2$, because the subcodes obviously don't differ before x_1*-th* position and after

x_2-*th* position. But the subcode obtained by removing number from x_2-*th* position contains number from *A-th* position of the source code on its *A-th* position. This contradiction finishes the proof.

Lemma 6. If there exist 3 distinct *(n-1)*-subcodes, the source code can be recovered.

Proof. Denote by *a, b, c* position that have been removed from the source code in order to get three given subcodes. Without loss of generality, *a<b<c*. Denote subcodes by *A, B, C* respectively. For each pair of subcodes consider maximal common fragment at the beginning and maximal common fragment at the end. Choose the maximal of the considered fragments (one at the beginning and one at the end). These are the fragments from the end of *A* and *B* codes and fragment from the beginning of *C* and *B* codes (due to ordering introduced below). By Lemma 5, the source code contains both of these fragments. If their union is shorter than the source code, then it equals the source code with number from position *b* removed, furthermore *b* is known. Otherwise the source code is recovered. Code *B* can be detected since it equals the union of the fragments. Code *C* can be detected as *C* and *B* codes have the biggest common fragment from the beginning. Consider the number from *b-th* position of *C* code. It is the number on the *b-th* position of the source code. Unite the two fragments and this number. The source code is now recovered. □

Notice that Lemma 6 is true for codes with any characteristic.

It's left to consider the case when we have exactly two different subcodes. Consider the columns containing both *0's* and *1's.* For each of them consider four inequalities:

$d(0,j-1)+ d(0,j)\geq T-1$

$d(0,j)+ d(0,j+1)\geq T-1$

$d(1,j-1)+ d(1,j)\geq T-1$

$d(1,j)+ d(1,j+1)\geq T-1$

If some of them does not hold, one of possible positions for one of the numbers can be excluded using Claim 2. Thus it can be found out which number occupies *j-th* position of the source code and which occupies *(j-1)-th.* If they all hold, repeat the reasoning for the next column being considered. Let *subcode A* be the subcode containing *(j+1)-th* number of the source code on its *j-th* position and let *subcode B* be a subcode containing *j-th* number of the source code on its *j-th* position. Unite the first *j* numbers of subcode *A* and the last *(n-j)* numbers of subcode *B.* The source code is now recovered. In this case at most $\left\lceil \frac{n}{2} \right\rceil + 1$ subcodes are required.

If for all considered columns four inequalities hold then the whole table of subcodes has the following form (accurate to order of 0 and 1):

0	…	0	1	…	0	1	0	…	0
0	…	1	0	…	1	0	1	…	0
…	…	…	…	…	…	…	…	…	…
0	…	1	0	…	1	0	1	…	0
0	…	1	0	…	1	0	1	…	0

Notice that the first subcode is unique (by Lemma 4, assuming that all groups of *1's* consist of not more than one 1). Thus, one more code is enough for reconstruction using Lemma 6. In this case at most $\left\lceil \frac{n}{2} \right\rceil + 1$ subcodes are required.

Theorem 2 is now proved. □