On the Security of Individual Data*

János Demetrovics[1], Gyula O.H. Katona[2], and Dezső Miklós[2]

[1] Computer and Automation Institute,
Hungarian Academy of Sciences
Kende u. 13-17, H-1111, Hungary
dj@ilab.sztaki.hu
[2] Alfréd Rényi Institute of Mathematics,
Hungarian Academy of Sciences
Budapest P.O.B. 127 H-1364 Hungary
{ohkatona,dezso}@renyi.hu

Abstract. We will consider the following problem in this paper:
Assume there are n numeric data $\{x_1, x_2, \ldots, x_n\}$ (like salaries of n individuals) stored in a database and some subsums of these numbers are disclosed by making public or just available for persons not eligible to learn the original data. Our motivating question is: at most how many of these subsums may be disclosed such that none of the numbers $x_1, x_2, \ldots, x_n$ can be uniquely determined from these sums. These types of problems arise in the cases when certain tasks concerning a database are done by subcontractors who are not eligible to learn the elements of the database, but naturally should be given some data to fulfill there task. In database theory such examples are called *statistical databases* as they are used for statistical purposes and no individual data are supposed to be obtained using a restricted list of SUM queries. This problem was originally introduced by Chin and Ozsoyoglu [1], originally solved by Miller *et al.* [5] and revisited by Griggs [4].
It turned out [5] that the problem is equivalent to the following question: If there are n real, non-zero numbers $X = \{x_1, x_2, \ldots, x_n\}$ given, what is the maximum number of 0 subsums of it, that is, what is the maximum number of the subsets of X whose elements sum up to 0. This approach, together with the Sperner theorem shows that no more than $\binom{n}{n/2}$ subsums of a given set of secure data may be disclosed without disclosing at least one of the data, which upper bound is sharp as well.
However, it is natural to assume that the disclosed subsums of the original elements of the database will contain only a limited number of elements, say at most k (in the applications databases are usually huge, while the number of operations is in most of the cases limited). We have now the same question: at most how many of these subsums of at most k members may be given such that none of the numbers $x_1, x_2, \ldots, x_n$ can be uniquely determined from these sums. The main result of this paper gives an upper bound on this number, which turns out to be sharp if we

* The work was supported by the Hungarian National Foundation for Scientific Research grant numbers 37846 and 42706 and the European Community's Centre of Excellence Grant numbers ICA1-CT-2000-70009 and ICA1-CT-2000-70025.

allow subsums of only k or $k-1$ members and asymptotically sharp in case of subsums of at most k members.

1 Introduction

The security of statistical databases has been studied for a long time. In this case the database is only used to obtain statistical information and therefore no individual data is supposed to be obtained as a result of the performed queries. Of course, the user is not allowed to query individual records, still, using only statistical types of queries, it might be possible to make inferences about the individual records. Several authors investigated earlier the possibility of introducing restriction for the prevention of database compromise, which include data and response perturbation, data swapping, random response queries, etc. One of the natural restrictions is to allow only SUM queries, that is queries which return the sum of the attributes corresponding to a set of individuals characterized by *characteristic formula*. For more detailed explanation of these terms see Denning [2,3]. In all of these cases it was assumed and will be assumed throughout of this paper as well that outside user or attacker do not have any further information about the database, only the answers to the SUM queries (e.g. they don't know about any functional dependencies).

Chin and Ozsoyoglu [1] introduced an Audit Expert mechanism for the prevention of database compromise with SUM queries. Later Miller *et al.* [5] determined the maximum number of SUM queries for this mechanism, which is $\binom{n}{\lfloor n/2 \rfloor}$. For example, in the database below one can ask the sum of the salaries of the individuals chosen the same number ($i = 0, 1, 2, 3$) of them from both of the sets {Bush, Carter, Clinton} and {Johnson, Kennedy, Nixon, Reagan}. In such a way one will chose $\binom{3}{0} \times \binom{4}{0} + \binom{3}{1} \times \binom{4}{1} + \binom{3}{2} \times \binom{4}{2} + \binom{3}{3} \times \binom{4}{3} = 1 + 3 \times 4 + 3 \times 6 + 1 \times 4 = 35 = \binom{7}{3}$ queries. Clearly, the given database and the one obtained from this one by lowering the salaries of {Bush, Carter, Clinton} by 1000 and increasing the salaries of {Johnson, Kennedy, Nixon, Reagan} by 1000 will give exactly the same answer to these queries and therefore no individual salary can be exactly calculated from this set of questions.

Table 1. Sample Database

Name	*Salary*
Bush	250000
Carter	180000
Clinton	220000
Johnson	120000
Kennedy	100000
Nixon	140000
Reagan	160000

A natural restriction of the above question is the restriction of the size of the SUM queries, that is assuming that the sums may involve at most or exactly k members. E.g., if in the above database we only consider SUM queries summing up 3 data, a possible scheme of them without compromising the database is to ask the some of the salaries of 3 gentlemen, two chosen from the set {Bush, Carter, Clinton, Johnson, Kennedy} and one from the set {Nixon, Reagan}. Therefore altogether $\binom{5}{2} \times \binom{2}{1} = 20$ queries are made, and, again, by increasing the salaries of {Bush, Carter, Clinton, Johnson, Kennedy} and decreasing the salaries of {Nixon, Reagan} with the double of that amount shows that no individual data can be gained from this set of statistical queries.

The main results of the recent paper, presented is Section 3, Theorems 3.1 and 3.2 answer the questions about the maximal possible SUM queries when either only a given number of data can be summed any time or when the number of the data involved in any SUM queries is bounded above. The first question is solved completely — that is a construction of the possible sequence of queries, the number of them equal to the obtained upper bound, is given — assuming (what can be quite natural in the real use of databases) that the number of records is much larger then the allowed number of them in the SUM queries. The second case is answered asymptotically.

In Section 2 we will carry on a sequence of transformations of the original questions, most of the repeated (or simply referred to) the transformations done by Chin and Ozsoyoglu [1] and Miller *et al.* [5,6] to formulate the exact mathematical questions to be solved in Section 3. In Section 4, we will draw the conclusions to answer the original statistical database questions.

2 Deriving the Mathematical Problems

Let us be given n real numbers $\{x_1, x_2, \ldots, x_n\}$ (like salaries of n individuals in the sample database) stored in a database. A possible SUM query is to ask $\sum_{i \in A} x_i$ for some $A \subset X = \{1, 2, 3, \ldots, n\}$ and we would like to maximize the number of these queries (maybe with some other side constraints) such that they will not determine any of the original data x_i's. That is we would like to give a sequence of subsets of X, $\mathcal{A} = \{A_1, A_2, \ldots, A_m\}$, maximize m, such that the sums $\left\{\sum_{i \in A_j} x_i : 1 \leq j \leq m\right\}$ do not determine any of the x_i's. We will only consider restricted type of attacks, that is methods to calculate the values of x_i's from the known sums, namely linear combinations. However, the upper bound proven for this restricted type of attacks will turn out to be sharp for the general case as well. In Section 4 we will give constructions of databases together with the sequence of SUM queries such that their number will be equal to the obtained maximum (if we only assume linear combination attacks) and the different databases (all individual data will be pairwise different) will both give the same answer to these SUM queries.

To formulate the problem in another, for our investigation more suitable way, consider the n dimensional vector space over the real numbers, $\mathbf{R}^n$, and the unit

vectors $\mathbf{e}_i = \{\overset{1}{0}, \overset{2}{0}, \ldots, \overset{i}{1}, \ldots, \overset{n}{0}\}$, $i = 1, 2, \ldots, n$. Denote the characteristic vectors of the subsets A_i by $\mathbf{v}_i$, that is $\mathbf{v}_i = (v_{i1}, v_{i2}, \ldots, v_{ij}, \ldots v_{in})$, where $v_{ij} = 1$ iff $x_j \in A_i$, 0 otherwise. With this setting, we are looking for the maximum number of $\mathbf{v}_i$'s such that none of the unit vectors $\mathbf{e}_i$'s are in $\langle \mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m \rangle$, the subspace spanned by the m characteristic vectors of the subsets determining the members of the subsums. This can easily be seen with the following straightforward lemma.

Lemma 2.1 *Let* $\mathbf{x}$ *denote the vector* $\{x_1, x_2, \ldots, x_n\}$ *and for given sequence of SUM queries with characteristic vectors* $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m$ *consider the vectors* $\mathbf{v}$ *where the value* $\mathbf{vx}$*, the scalar product of the two vectors* $\mathbf{v}$ *and* $\mathbf{x}$*, can be calculated from the values* $\mathbf{v}_i\mathbf{x}$*, that is,* $\mathbf{vx}$ *is uniquely determined by the two vectors* $\mathbf{v}$ *and* $\mathbf{x}$*. Then these vectors will form a subspace of the original vector space equal to* $\langle \mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m \rangle$*.* □

From now on, instead of the sequence of the SUM queries we will consider the subspace spanned by the characteristic vectors. Any question regarding the maximum number of queries with certain property is equivalent to the question of the maximum number of the $(0, 1)$-vectors (with the additional required properties) of the subspace not containing any of the unit vectors.

The following further reduction steps of the problem are originally due to Chin and Ozsoyoglu [1].

Lemma 2.2 *(***Chin and Ozsoyoglu [1]***) If* $\mathbf{V} \subset \mathbf{R}^n$, $\mathbf{e}_i \notin \mathbf{V}$ $1 \leq i \leq n$, $\dim \mathbf{V} \leq (n-1)$*, then there is a* $\mathbf{W} \supset \mathbf{V}$ *such that* $\dim \mathbf{W} = n - 1$, $\mathbf{e}_i \notin \mathbf{W}$ $1 \leq i \leq n$*.* □

Since any n (full) dimensional space would contain all unit vectors $\mathbf{e}_i$, and — by the lemma above — all at most $n-1$ dimensional spaces will be contained by an exactly $n-1$ dimensional space having the required property, we may assume that the subspace giving the maximum possible number of allowed queries is $n-1$ dimensional. Take the matrix of a basis of this subspace and bring it to it normal form

$$\begin{vmatrix} 1 & 0 & \cdots & 0 & a_1 \\ 0 & 1 & \cdots & 0 & a_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & a_{n-1} \end{vmatrix}$$

where none of the a_i's are equal to 0 due to the fact that the unit vectors are not in the subspace.

The subspace spanned by the characteristic vectors of the allowed SUM queries is also spanned by the rows of this matrix. Therefore all of these characteristic vectors are in the subspace spanned by the rows of the matrix. On the other hand, all the 0, 1 vectors being in the subspace spanned by the rows of the matrix do determine a SUM query and (if they satisfy the further assumptions, like the number of 1's is k or at most k) they are allowed, that is the set of them will not compromise the database.

It is easy to see that the only linear combinations of the rows of this matrix yielding $(0,1)$-vectors are those with coefficients $0,1$. Any such linear combination will be a $0,1$ vector if and only if the sum $\sum a_i$ over all i where the corresponding coefficient is 1 is either 0 or 1. Therefore, we have to maximize the number of sums $\sum_{i\in A} a_i = 0$ or 1 where the A's are subsets of $[n-1] = \{1,2,\ldots,n-1\}$. Let us introduce $a_n = -1$ and now consider the sums $\sum_{i\in B} a_i = 0$ where the B's are subsets of $[n] = \{1,2,\ldots,n\}$. Naturally, there is a one-to-one correspondence between these two sets of sums. Therefore, our original question

Problem 2.3 *Determine the maximum possible number of SUM queries over a set of n records without compromising the database.*

is now reduced to the following one:

Problem 2.4 *Given a set of n real numbers $\{a_1, a_2, \ldots, a_n\}$, none of them being equal to 0, determine the maximum number of sums $\sum_{i\in B} a_i = 0$ where the B's are subsets of $[n] = \{1,2,\ldots,n\}$.*

Further, if we assume that the number of elements in the SUM queries are restricted by a size constraint (like at most or exactly k element subsets are only considered), the same restriction will apply to the sums in Problem 2.4.

In Problem 2.4 we omitted the assumption that $a_n = -1$ since any set of n non-zero real numbers $\{y_1, y_2, \ldots, y_n\}$ can be normalized (simply each of them multiplied by the same non-zero number) such that for the resulting vector $\{x_1, x_2, \ldots, x_n\}$ we will have $x_n = -1$ and the set (and therefore the number) of zero sums naturally will not change.

Now consider the set of real numbers $\{a_1, a_2, \ldots, a_n\}$ and the system of subsets of the indices $X = \{1,2,\ldots,n\}$, $\mathcal{B} = \{B_1, B_2, \ldots, B_m\}$ such that the sums $\sum_{i\in B_j} a_i$ are all 0. Let X_1 be the set of indices of the positive a_i's and X_2 be the set of the indices of the negative a_i's. Since none of the a_i's are zero, $X = X_1 \cup X_2$ is a partition of the set X. If we consider two sets B_1 and B_2 from the set $\mathcal{B}$, then since $\sum_{i\in B_1} a_i = \sum_{i\in B_2} a_i = 0$, we have $\sum_{i\in B_1 - B_2} a_i = -\sum_{i\in B_1\cap B_2} a_i = \sum_{i\in B_2 - B_1} a_i$ and therefore the sums at the two ends of this system of equations are equal and so have the same sign. Therefore, it is not possible that $B_1 - B_2 \subset X_1$ and $B_2 - B_1 \subset X_2$ at the same time.

Definition 2.5 *Let $X = X_1 \cup X_2$ be a partition of the finite set X and F and G two subsets of X. We say that F and G are separated by the partition if $F - G \subset X_1$ and $G - F \subset X_2$ does not happen at the same time.*

We also know that all of the sets B_i have the property $B_i \cap X_1 \neq \emptyset$ and $B_i \cap X_2 \neq \emptyset$, since sum of only negative or only positive numbers may not be equal to zero.

Definition 2.6 *We say that the family $\mathcal{F}$ of subsets of the finite set $X = X_1 \cup X_2$ ($X_1 \cap X_2 = \emptyset$, $X_1 \neq \emptyset$, $X_2 \neq \emptyset$) is difference separated (with respect to the partition $X_1 \cup X_2$) if $F - G$ and $G - F$ are separated by the partition for every pair F, G of distinct members of $\mathcal{F}$ and $F \cap X_1 \neq \emptyset, F \cap X_2 \neq \emptyset$ holds for each member.*

We know by now that for any set of SUM queries not compromising the database we can find a set of a set of n real non-zero numbers $\{a_1, a_2, \ldots, a_n\}$, such that the subsets of indices indices $X = \{1, 2, \ldots, n\}$, $\mathcal{B} = \{B_1, B_2, \ldots, B_m\}$ for which the sums $\sum_{i \in B_j} a_i$ are 0 do correspond to the SUM queries, on one side, and form a difference separated family of subsets on the other side. That is, any upper bound on the size of difference separated families (in our case of subsets of given sizes) will give an upper bound on the possible number of SUM queries not compromising the database. In the next section we will derive such upper bounds, while in the last section — Conclusions — with giving specific examples we will also show that this bounds are sharp not only for the size of the difference separated families but for the number of SUM queries as well.

3 The Main Mathematical Theorems

Theorem 3.1 *Let $0 < k, n$ be fixed even integers, $n_0(k) \leq n$, that is n is large relative to k. Let further X be an n-element set with partition $X = X_1 \cup X_2$ where $X_1, X_2 \neq \emptyset$. Suppose that $\mathcal{F}$ is a difference separated family (with respect to $X_1 \cup X_2$) of subsets of size k and $k-1$. Then*

$$|\mathcal{F}| \leq M(n,k) = \binom{\left\lfloor \frac{(n+1)(k-1)}{k} \right\rfloor}{k-1} \left(n - \left\lfloor \frac{(n+1)(k-1)}{k} \right\rfloor \right). \tag{1}$$

Theorem 3.2 *Let $0 < k, n$ be fixed even integers, $n_0(k) \leq n$, that is n is large relative to k. Let further X be an n-element set with partition $X = X_1 \cup X_2$ where $X_1, X_2 \neq \emptyset$. Suppose that $\mathcal{F}$ is a difference separated family (with respect to $X_1 \cup X_2$) of subsets of size at most k. Then*

$$|\mathcal{F}| \leq M(n,k) + M(n,k-2) + M(n,k-4) + \ldots.$$

Remark. Theorem 3.1 is sharp, since choosing X_1 to have $\left\lfloor \frac{(n+1)(k-1)}{k} \right\rfloor$ elements and taking all $k-1$-element subsets of X_1 combining them with all possible 1-element sets in X_2, the number of sets will be exactly $M(n,k)$. On the other hand this construction obviously satisfies the conditions. Theorem 3.2, however is not necessarily sharp, since the obvious generalization of the construction above does not always work. It is, however, asymptotically sharp, since $M(n,k-2) + M(n,k-4) + \ldots = O(n^{k-3})$ while $M(n,k)$ is of order $k-1$.

The proof will be given by a sequence of lemmas.

Assume that a cyclic ordering is fixed both in X_1 and X_2. We say that the pair (A,B) of subsets $A \subset X_1, B \subset X_2$ is an (a,b)-*interval-pair* if A and B are intervals in the respective X and $|A| = a, |B| = b$. The *middle pair* of the (a,b)-interval-pair (A,B) is (x,y) where x is the $\lfloor \frac{a+1}{2} \rfloor$th element of A and y is the $\lfloor \frac{b+1}{2} \rfloor$th element of B.

Lemma 3.3 *Suppose that (A,B) and (C,D) are (a,b) and (c,d)-interval-pairs, respectively, where $a+b = c+d$ (a,b,c,d are positive integers). If their middle*

pairs coincide then either $A \cup B - C \cup D \subset X_1$ *and* $C \cup B - A \cup B \subset X_2$ *or* $A \cup B - C \cup D \subset X_2$ *and* $C \cup B - A \cup B \subset X_1$ *hold.*

Proof. It is easy to see that if $c \leq a$ then $C \subset A$. The inequality $d \geq b$ is a consequence, this implies $B \subset D$. Hence we have $A \cup B - C \cup D \subset X_1$ and $C \cup B - A \cup B \subset X_2$. The case $c > a$ is analogous. □

Lemma 3.4 *Suppose that* (A, B) *and* (C, D) *are* (a, b) *and* (c, d)*-interval-pairs, respectively, where* $a+b = c+d\pm 1$ *(*a, b, c, d *are positive integers). If their middle pairs coincide then either* $A \cup B - C \cup D \subset X_1$ *and* $C \cup B - A \cup B \subset X_2$ *or* $A \cup B - C \cup D \subset X_2$ *and* $C \cup B - A \cup B \subset X_1$ *hold.*

Proof. The previous proof can be repeated. □

Lemma 3.5 *Let* $\mathcal{G}$ *be a family of difference separated intervals with respect to* $X_1 \cup X_2$ *with members of size* $j-1$ *and* j $(2 \leq j)$*. Then* $|\mathcal{G}| \leq n_1 n_2$ *holds.*

Proof. The members of $\mathcal{G}$ must have different middle pairs by Lemmas 3.2 and 3.3. The number of possible middle pairs is $n_1 n_2$. □

Introduce the following definition:

$$M(n_1, n_2; k) = \max\left(\binom{n_1}{i}\binom{n_2}{j}\right) \tag{2}$$

where the maximum is taken for all $0 < i \leq n_1, 0 < j \leq n_2, i + j = k$.

Lemma 3.6 *Let* $X = X_1 \cup X_2$, $X_1 \cap X_2 = \emptyset$, $|X_1| = n_1$, $|X_2| = n_2$. *If* $\mathcal{F}$ *is a family of difference separated sets of sizes* ℓ *and* $\ell - 1$ *then*

$$|\mathcal{F}| \leq M(n_1, n_2; \ell) \tag{3}$$

holds.

Proof. The number of four-tuples $(\mathcal{C}_1, \mathcal{C}_2, A, B)$ will be counted in two different ways, where $\mathcal{C}_i$ is a cyclic permutation of X_i $(i = 1, 2)$, $A = F \cap X_1$ and $B = F \cap X_2$ holds for some $F \in \mathcal{F}$ and they form an interval-pair for these cyclic permutations.

Let first fix A and B and count the number of cyclic permutations where they are intervals. $\mathcal{C}_1$ can be chosen in $|A|!(n_1 - |A|)!$ many ways, the same applies for B, therefore the number of four-tuples is

$$\sum |A|!(n_1 - |A|)!|B|!(n_2 - |B|)! \tag{4}$$

where the summation is taken for all $A = F \cap X_1, B = F \cap X_2, F \in \mathcal{F}$.

Fix now the cyclic permutations. The subfamily of $\mathcal{F}$ consisting of the interval-pairs for these permutations will be denoted by $\mathcal{G}$. It is a family of difference separated intervals. The application of Lemma 3.4 gives that the numbers of pairs A, B for any given pair of cyclic permutations is at most $n_1 n_2$. Since

the number of permutations is $(n_1-1)!(n_2-1)!$, the number of four-tuples in question is at most $n_1!n_2!$. Compare it with (4):

$$\sum |A|!(n_1-|A|)!|B|!(n_2-|B|)! \le n_1!n_2!.$$

Elementary operations lead to

$$\sum \frac{1}{\binom{n_1}{|A|}\binom{n_2}{|B|}} \le 1 \tag{5}$$

where $A = F \cap X_1, B = F \cap X_2, 0 < |A| \le n_1, 0 < |B| \le n_2, |A|+|B| = \ell$ or $\ell-1$. Since $M(n_1,n_2;\ell-1) \le M(n_1,n_2;\ell)$ holds, by the definition of $M(n_1,n_2;\ell)$ (5) implies

$$\frac{|\mathcal{F}|}{M(n_1,n_2;\ell)} \le 1$$

proving (3). □

Lemma 3.7 *Suppose* $1 \le i < \ell \le n, \ell - i < n - n_1$. *Then*

$$\binom{n_1}{i}\binom{n-n_1}{\ell-i} \le \binom{\left\lfloor \frac{(n+1)i}{\ell} \right\rfloor}{i}\binom{n-\left\lfloor \frac{(n+1)i}{\ell} \right\rfloor}{\ell-i}.$$

Proof. Compare two consecutive expressions:

$$\binom{n_1}{i}\binom{n-n_1}{\ell-i} \le \binom{n_1+1}{i}\binom{n-n_1-1}{\ell-i}.$$

After carrying out the possible cancellations

$$\frac{n-n_1}{n-n_1-\ell+i} \le \frac{n_1+1}{n_1-i+1}$$

is obtained what is equivalent to

$$n_1+1 \le \frac{(n+1)i}{\ell}.$$

Hence

$$\binom{n_1}{i}\binom{n-n_1}{\ell-i}$$

takes on its maximum (with fixed i and ℓ) at

$$\left\lfloor \frac{(n+1)i}{\ell} \right\rfloor.$$

□

Lemma 3.8 *Suppose* $0 < \frac{\ell}{2} \le i \le \ell - 1$. *If* $n_0(\ell) \le n$ *then*

$$\binom{\left\lfloor \frac{(n+1)i}{\ell} \right\rfloor}{i}\binom{n-\left\lfloor \frac{(n+1)i}{\ell} \right\rfloor}{\ell-i} < \binom{\left\lfloor \frac{(n+1)(i+1)}{\ell} \right\rfloor}{(i+1)}\binom{n-\left\lfloor \frac{(n+1)(i+1)}{\ell} \right\rfloor}{\ell-i-1}.$$

Proof. After carrying out the possible cancellations,

$$\frac{\left(n-\left\lfloor\frac{(n+1)i}{\ell}\right\rfloor\right)\dots\left(n-\left\lfloor\frac{(n+1)(i+1)}{\ell}\right\rfloor+1\right)}{(\ell-i)\left(\left\lfloor\frac{(n+1)i}{\ell}\right\rfloor-i\right)}<$$

$$\frac{\left\lfloor\frac{(n+1)(i+1)}{\ell}\right\rfloor\dots\left(\left\lfloor\frac{(n+1)i}{\ell}\right\rfloor+1\right)}{(i+1)\left(n-\left\lfloor\frac{(n+1)(i+1)}{\ell}\right\rfloor-\ell+i+1\right)}$$

is obtained what is equivalent to

$$\frac{\left(n-\left\lfloor\frac{(n+1)i}{\ell}\right\rfloor\right)\dots\left(n-\left\lfloor\frac{(n+1)(i+1)}{\ell}\right\rfloor+1\right)}{\left\lfloor\frac{(n+1)(i+1)}{\ell}\right\rfloor\dots\left(\left\lfloor\frac{(n+1)i}{\ell}\right\rfloor+1\right)}<$$

$$\frac{(\ell-i)\left(\left\lfloor\frac{(n+1)i}{\ell}\right\rfloor-i\right)}{(i+1)\left(n-\left\lfloor\frac{(n+1)(i+1)}{\ell}\right\rfloor-\ell+i+1\right)} \tag{6}$$

Consider the left hand side as a product of quotients: the quotient of the first factors, the second factors, etc., in the numerator and the denominator, respectively. The first of these quotients is the largest one, it is less than 1, their number is at least $\frac{n+1}{\ell}$. Therefore the left hand side in (6) can be replaced by

$$\left(\frac{n-\left\lfloor\frac{(n+1)i}{\ell}\right\rfloor}{\left\lfloor\frac{(n+1)(i+1)}{\ell}\right\rfloor}\right)^{\frac{n+1}{\ell}}.$$

A further increase is

$$\left(\frac{n+1-\frac{(n+1)i}{\ell}}{\frac{(n+1)(i+1)}{\ell}}\right)^{\frac{n+1}{\ell}}=\left(\frac{\ell-i}{i+1}\right)^{\frac{n+1}{\ell}}.$$

Here $\ell-i<i+1$ by the assumption of the lemma, therefore the left hand side of (6) tends exponentially to 0 when n tends to infinity. On the other hand, the right hand side of (6) tends to

$$\frac{(\ell-i)i}{(i+1)(\ell-i-1)},$$

proving the inequality. □

Remark. This lemma seems to be true for small values of n, too, we have technical difficulties to prove it.

Proof of Theorem 3.1. By Lemma 3.6 $|\mathcal{F}|$ cannot exceed the largest product $\binom{n_1}{i}\binom{n-n_1}{k-i}$ where $1\le n_1<n, 1\le i\le n_1, 1\le k-i\le n-n_1$. By symmetry $\frac{\ell}{2}\le i$ can be supposed. Lemma 3.7 gives the best n_1. This upper estimate in Lemma 3.7 can be increased by Lemma 3.8 until we arrive to the largest possible value of i: $\ell-1$. This is the desired upper estimate. □

Proof of Theorem 3.2. Apply Theorem 3.1 with $k, k-2, k-4, \dots$ and sum the obtained upper estimates. □

4 Conclusion

By the argument of Section 2 and the results of Section 3, if $n_0(k) < n$, at most $M(n,k)$ (see (1) for exact value) SUM queries of size k can be asked about a set of data $x_1, x_2, \ldots x_n$ without disclosing at least one of the values x_i. On the other hand, this bound is not only sharp for the mathematical questions asked in the previous section, but also for the original problem. Assume n equal real numbers divided into two parts: B_1 of size $\left\lfloor \frac{(n+1)(k-1)}{k} \right\rfloor$ and B_2 of size $\left(n - \left\lfloor \frac{(n+1)(k-1)}{k} \right\rfloor\right)$. Take all subsums of this numbers of k elements such that $k-1$ are chosen from set B_1 and 1 from set B_2. Now increase all of the elements of B_1 by 1 and decrease all of the elements of B_2 by $k-1$ (assume that originally the common value was not -1 neither $k-1$) and take exactly the same subsums. The answers to these queries are the same in both of cases, that is these answers do not disclose any of the values x_i's.

It is also interesting to mention that the bound $M(n,k)$ is only constant time smaller than the absolute upper bound $\binom{n}{k}$ for the number of SUM queries of size k, and much bigger than the somewhat more general solution of $\binom{n/2}{k/2}^2$, which is $\sqrt{k}$ time smaller than $\binom{n}{k}$. One can get a construction yielding the $\binom{n/2}{k/2}^2$ bound simply using the above method but dividing the set of the values into two equal size subsets and picking always the same number of elements from both sides. For example, if $n = 20$ and $k = 10$, then $\binom{n}{k} = 184756$, $M(n,k) = 97240$ and $\binom{n/2}{k/2}^2 = 63504$.

The bound for the case when the SUM queries may have any number of elements less than or equal to k is most probably not sharp. At the same time, similar to the case in the mathematical theorem, an asymptotically good construction can be given simply taking the above construction for the case when all sums have exactly k elements (see Remark after Theorem 3.2.)

References

1. CHIN, F.Y., OZSOYOGLU, G., Auditing and inference control in statistical databases, *IEEE Transactions on Software Engineering* **SE-8**(1982) 574–582.
2. DENNING, D.E., in "Cryptography and Data Security", Addison-Wesley, Sydney, 1982.
3. DENNING, D.E., SCHLORER, J., Inference controls for statistical databases, *Computer* (1983), 69–82.
4. GRIGGS, J.R., Concentrating subset sums at k points, *Bull. Inst. Comb. Applns.* **20**(1997) 65–74.
5. MILLER, M., ROBERTS, I., SIMPSON I., Application of Symnmetric Chains to an Optimization Problem in the Security of Statistical Databases, *Bull. ICA* **2**(1991) 47–58.
6. MILLER, M., ROBERTS, I., SIMPSON I., Preventation of Relative Compromise in Statistical Databases Using Audit Expert, *Bull. ICA* **10**(1994) 51–62.