On the security of individual data

J. Demetrovics [a], G. O. H. Katona [b] and D. Miklós [b,*]

[a] *Computer and Automation Institute, Hungarian Academy of Science, Kende u. 13–17, H-1111 Budapest, Hungary*
E-mail: dj@ilab.sztaki.hu

[b] *Alfréd Rényi Institute of Mathematics, HAS, Budapest P.O.B. 127 H-1364, Hungary*
E-mail: {ohkatona; dezso}@renyi.hu

We will consider the following problem in this paper: Assume that there are n numerical data $\{x_1, x_2, \ldots, x_n\}$ (like salaries of n individuals) stored in a database and some subsums of these numbers are made 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 [1], originally solved by Miller et al. [7] and revisited by Griggs [4, 5]. It was shown in [7] 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. To calculate a subsum, it might need some operations whose number is limited. This is why 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. The goal of the present paper is to determine the maximum number of subsums of size at most k which can be disclosed without making possible to calculate any of the individual data x_i. The maximum is exactly determined for the case when the number of data is much larger than the size restriction k.

Keywords: database, datamining, security, subset sums

AMS subject classification: 68P15, 68R05, 05D05

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 the 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 do not know about any functional dependency).

Let the database consist of n confidential records, which are real numbers $x_1, \ldots, x_n$. Introduce the notations $[n] = \{1, 2, \ldots, n\}$ and $2^{[n]} = \{A \subseteq [n]\}$. The user may choose any subset A of $[n]$ and request the sum $\sum_{i \in A} x_i$. We say that the database is *compromised* if one of the x_i's can be determined from these sums, which are on the disposal of the user. For example, if $|A| = 1$, A will consist of one element, whose value would be obviously disclosed; therefore, only sets of size > 1 are allowed to choose.

Chin and Ozsoyoglu [1] introduced an Audit Expert mechanism for the prevention of database compromise with SUM queries. Audit Expert will keep track of which queries it has been previously answered and decline to answer the next query if it would, together with the previous answers, lead to a compromise of the database. For instance if the first two queries were $\{1, 2\}$ and $\{2, 3\}$ then $\{1, 3\}$ cannot be answered as a third query since

$$x_3 = \frac{1}{2}(x_1 + x_3) + \frac{1}{2}(x_2 + x_3) - \frac{1}{2}(x_1 + x_2)$$

determines x_3 (and also x_1, x_2).

It is easy to see that if the sets A are chosen independently, then the procedure end rather early with a high probability. This fact justifies the other approach. Here the set of possible queries is fixed in advance and the problem is to maximize the number of these queries. Miller et al. [7] 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 1,000 and increasing the salaries of {Johnson, Kennedy, Nixon, Reagan} by 1,000 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).

Miller et al. [7] converted the problem to an extremal problem for matrices and to an extremal problem for subsums of real numbers (see below). Griggs ([4]) observed

Table 1
Sample database.

Name	Salary
Bush	250,000
Carter	180,000
Clinton	220,000
Johnson	120,000
Kennedy	100,000
Nixon	140,000
Reagan	160,000

the close correspondence between their problem and the famous Littlewood–Offord problem ([6]) of combinatorial number theory.

There are many analogous problems, generalizations of the one above. The first one is the case of *relative compromise*, when the difference $x_i - x_j (i \neq j)$ is determined from the answered queries. Again, the maximum number of queries which can be answered without a relative compromise is to be determined. This problem has been solved in [8] with giving the best construction, as well. We call the attention of the interested reader to the survey paper [5] of Griggs which shows many of the similar generalizations and its unifying approach makes easier to understand the theory.

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 k members. E.g., if in the above database we only consider SUM queries summing up three data, a possible scheme of them without compromising the database is to ask the sum of the salaries of three 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 same amount shows that no individual data can be gained from this set of statistical queries.

In Section 2 we will carry on a sequence of transformations of the original problem, most of them repeated (or simply referred to) the transformations done by Chin and Ozsoyoglu [1], Miller et al. [7] and Griggs [5] to formulate the combinatorial problem what is needed to be solved here. It is an extremal problem on a family of at least k-element subsets of an n-element set.

Section 3 is devoted to this combinatorial problem. Theorems 3.1 and 3.7 give the exact solution if n is large relative to k.

In Section 4, we will answer the original statistical database question. We obtain the exact maximum number of the SUM queries involving at most k data without compromising the database in the case when the number of data is much larger than the size k of the restriction.

2. Deriving the combinatorial 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 $A \subseteq [n]$ 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 x_i's. That is we would like to give a family of subsets of X, $\mathcal{A} = \{A_1, A_2, \ldots, A_m\}$, maximize m, such that the sums $\{\sum_{i \in A_j} x_i : 1 \leqslant j \leqslant m\}$ do not determine any of the x_i's. Let the *characteristic vector* of the set $A \subseteq [n]$ be a 0,1 vector of dimension n whose i'th coordinate is 1 iff $i \in A$. The characteristic vector of the one-element $\{i\}$ is denoted by $\mathbf{e}_i = (0, 0, \ldots, 1, \ldots, 0)$. Let further $\mathbf{x} = (x_1, \ldots, x_n)$. Using this terminology, one can say that the user knows the values $\mathbf{v}_j\mathbf{x} (1 \leqslant j \leqslant m)$ but cannot determine the values $\mathbf{e}_i\mathbf{x} (1 \leqslant i \leqslant n)$ from this information. Of course, the user can calculate the linear combinations

$$\sum_{i=1}^{m} \lambda_i \mathbf{v}_i \mathbf{x} = \left(\sum_{i=1}^{m} \lambda_i \mathbf{v}_i \right) \mathbf{x}$$

where λ_i are arbitrary real numbers. These values are *linearly calculable* from the values $\mathbf{v}_i\mathbf{x} (1 \leqslant i \leqslant n)$. This is where we can get rid of the original real numbers x_i, since the linearly calculable values are exactly

$$\left(\sum_{i=1}^{m} \lambda_i \mathbf{v}_i \right) \mathbf{x},$$

that is the inner products of $\mathbf{x}$ and the vectors $\sum_{i=1}^{m} \lambda_i \mathbf{v}_i$. The latter ones are the linear combinations of the characteristic vectors $\mathbf{v}_j (1 \leqslant j \leqslant m)$. The set of such vectors is a subspace $\mathbf{V}$ of $\mathbf{R}^n$. This subspace is called the subspace spanned by the vectors $\mathbf{v}_j (1 \leqslant j \leqslant m)$, and is denoted by $\langle \mathbf{v}_1, \ldots, \mathbf{v}_m \rangle$. Let us formulate these thoughts in a form of a lemma.

Lemma 2.1. Let $\mathbf{x}$ denote the vector $\{x_1, x_2, \ldots, x_n\}$ and for a given family of SUM queries with characteristic vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m$ consider the vectors $\mathbf{v}$ where the value $\mathbf{vx}$ can be linearly calculated from the values $\mathbf{v}_i\mathbf{x}$. Then these vectors will form a subspace of the vector space $\mathbf{R}^n$ equal to $\langle \mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m \rangle$.

The original problem now is reduced to the following one: Find the maximum number of distinct 0,1 vectors $\mathbf{v}_1, \ldots, \mathbf{v}_m$ of dimension n in such a way that $\mathbf{e}_i \notin \langle \mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m \rangle$ holds for every $1 \leqslant i \leqslant n$.

We were here emphasizing that only the linear operations are considered during the trials of finding individual values. The user may, of course use other (say

polynomial) operations over the results of the SUM queries, or of the vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m$, but it does not help much. The maximum number m will be determined under the weaker condition when linear operations are only allowed. However, we will show that this is sufficient, one can construct this many vectors in such a way that no calculation (linear, polynomial, or any) can determine the $\mathbf{e}_i$'s from them. The method is easy: We will give two sets of data, returning the same answer for each of the given SUM queries, but differing from each other in each single entry.

Here we slightly change our viewpoint: Instead of the sequence of the SUM queries we will consider the subspace $\mathbf{V}$ spanned by the characteristic vectors. The question regarding the maximum number of queries satisfying a certain property is equivalent to the question of finding the maximum number of the 0,1 vectors (satisfying the additional property) of a subspace $\mathbf{V}$ not containing any of the unit vectors $\mathbf{e}_i$.

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

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

Since any n (full) dimensional space would contain all unit vectors $\mathbf{e}_i$, we may suppose that the subspace $\mathbf{V}$ giving the maximum possible number of allowed queries is $n-1$ dimensional. Take a basis $\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_{n-1}$ and the matrix with these rows. Using the standard Gauss-elimination the matrix can be brought in the following form:

$$\begin{pmatrix} 1 & 0 & \ldots & 0 & a_1 \\ 0 & 1 & \ldots & 0 & a_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & 1 & a_{n-1} \end{pmatrix}.$$

Of course, these rows form another basis. Therefore none of the a_i's are equal to 0 due to the fact that the unit vectors are not in $\mathbf{V}$.

We have to maximize the number of 0,1 vectors (satisfying an additional condition) in $\mathbf{V}$, that is among the linear combinations of the rows of the matrix above. 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 and 1. That is it is sufficient to look at the subsums of these rows. Take a subset $A \subseteq [n-1]$ of the rows and add these rows. The first $n-1$ coordinates will trivially be 0's and 1's. However the last coordinate, $\sum_{i \in A} a_i$ can be different from 0 and 1. Therefore, we have to maximize the number of sums $\sum_{i \in A} a_i$ which are either 0's or 1's, where As are an arbitrary subsets of $[n-1]$. 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 $[n] = \{1, 2, \ldots, n\}$. There is a natural one-to-one correspondence between

these two sets of sums. ($B = A \cup \{n\}$ if the sum for A is 1, and $B = A$ otherwise.) Our original question

Problem 1. 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. 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 $[n]$.

Let us remark that when we allow an arbitrary a_n, not only -1, we weaken our assumptions, that is, we might have enlarged the searched maximum (however, it was shown not to be the case, the upper bound proved on the question of Problem 2 can be reached by a construction in case of Problem 1 as well).

Assuming additional condition(s) concerning the SUM queries require careful investigation, since the condition(s) might be changed by the transformation leading to Problem 2 from Problem 1. In our present case, however, we are lucky: The number of 1's in the resulting 0,1 vector is equal to $|B|$. Therefore if the additional condition is a size constraint then the transformation from Problem 1 to Problem 2 is still working.

Let us formulate it for our case. The main question of the present paper,

Problem 3. Determine the maximum possible number of SUM queries containg at most k records from a set of n records without compromising the database.

is reduced to the following one:

Problem 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 $B \subseteq [n], |B| \leqslant k$.

Let us remark that with the step when we allowed an arbitrary a_n, not only -1, we weakened our assumptions once more, that is, we might have enlarged the searched maximum.

Denote the family of B's in question by $\mathcal{B} = \{B_1, B_2, \ldots, B_m\}$ where $\sum_{i \in B_j} a_i = 0$. Separate the a_i's according to their signs. Define $X_1 = \{i\colon\ a_i > 0\}$ and $X_2 = \{i\colon\ a_i < 0\}$. Since none of the a_i's is zero, $[n] = X_1 \cup X_2$ is a partition. Observe that the sets B_i have the property

$$B_i \cap X_1 \neq \emptyset, \quad B_i \cap X_2 \neq \emptyset, \tag{i}$$

since the sum of all negative or all positive numbers cannot be equal to zero. (Let us not here that in theory it would be possible that both $B_i \cap X_1$ and $B_i \cap X_2$ are empty, when B_i is empty. This corresponds to the case of the empty sum which is always 'disclosed.' For technical reasons we exclude it from $\mathcal{B}$, decreasing the number of possible sums by one.)

If we consider two members B_1 and B_2 of $\mathcal{B}$, then $\sum_{i\in B_1} a_i = \sum_{i\in B_2} a_i = 0$ implies $\sum_{i\in B_1 - B_2} a_i = -\sum_{i\in B_1\cap B_2} a_i = \sum_{i\in B_2 - B_1} a_i$. Therefore

$$B_1 - B_2 \subset X_1 \text{ and } B_2 - B_1 \subset X_2 \text{ cannot simultaneously hold.} \qquad \text{(ii)}$$

Definition 2.3. Let $[n] = X_1 \cup X_2$ be a partition. We say that a family $\mathcal{B} \subseteq 2^{[n]}$ is non-difference-separated, or shortly nodifsep with respect to the partition $X_1 \cup X_2$ if its members satisfy (i) and every pair of its members satisfies (ii).

In this way we obtained a further weakened variant of our Problem 3.

Problem 5. Determine $M(n,k) = \max|\mathcal{B}|$ under the conditions $\mathcal{B} \subseteq 2^{[n]}$, $B \in \mathcal{B}$ implies $|B| \leqslant k$, and $\mathcal{B}$ is nodifsep with respect to a non-trivial partition of $[n]$.

This purely combinatorial problem is formulated for families of subsets. The goal of the next section is to solve this problem for the case $n(k) \leqslant n$, that is when n is large relative to k. In case of combinatorial problems $n(k)$ usually denotes some (calculable) constant which depends only on k and in many times the obtained results (for a parameter n) is valid only if $n > n(k)$. The evaluation of the exact value of $n(k)$ is usually omitted, since theoretically we only are interested in the fact that the result is valid for big enough n's (that is, almost always).

3. The combinatorial theorems

We introduce the notation

$$\binom{[n]}{\leqslant k} = \{F : F \in 2^{[n]}, |F| \leqslant k\}.$$

Let $M(n_1, n_2, k)$ be the maximum analogous to $M(n,k)$ for the case when the partition sizes are fixed. That is, fix the partition $X_1 \cup X_2 = [n]$ with sizes $|X_1| = n_1, |X_2| = n_2$. Suppose that $k \leqslant n_1 + n_2$. Then $M(n_1, n_2, k)$ is the maximum size of a family $\mathcal{F}$ such that $\mathcal{F} \subseteq \binom{[n]}{\leqslant k}$, and $\mathcal{F}$ is nodifsep with respect to the partition $X_1 \cup X_2 = [n]$.

Our first theorem determines this maximum up to a certain extent.

Theorem 3.1. Let $2 \leqslant k \leqslant n_1 + n_2$. Then

$$M(n_1, n_2, k) = \max \sum_{\ell=1}^{r} \binom{n_1}{i_\ell} \binom{n_2}{j_\ell} \tag{3.1}$$

where the maximum is taken for all integers $1 \leqslant r, 0 < i_1 < \ldots < i_r \leqslant n_1, 0 < j_1 < \ldots < j_r \leqslant n_2$ satisfying $i_r + j_r \leqslant k$.

The proof is based on an easy lemma. Fix a permutation π_i of the elements of $X_i (i = 1, 2)$. A set $G_i \subseteq X_i$ is called an *initial segment* if it consists of the first $|G_i|$ elements of X_i with respect to π_i. The set $G \subseteq [n]$ is called a *combined initial segment* if both $G \cap X_1$ and $G \cap X_2$ are initial segments. Let $A(n_1, n_2, k)$ denote the right hand of (3.1).

Lemma 3.2. Let the nodifsep family (with respect to the partition $X_1 \cup X_2 = [n]$) $\mathcal{G} \subseteq \binom{[n]}{\leqslant k}$ consist of combined initial segments with respect to some fixed permutations π_1, π_2. Then

$$\sum_{G \in \mathcal{G}} \binom{n_1}{|G \cap X_1|} \binom{n_2}{|G \cap X_2|} \leqslant A(n_1, n_2, k) \tag{3.2}$$

holds.

Proof. Let G, H be distinct members of $\mathcal{G}$. Then the nodifsep property implies that the inequalities $|G \cap X_1| \leqslant |H \cap X_1|$ and $|G \cap X_2| \geqslant |H \cap X_2|$ cannot both simultaneously hold, since they would imply $H - G \subseteq X_1$ and $G - H \subseteq X_2$. The same can be said about the pair of inequalities $|G \cap X_1| \geqslant |H \cap X_1|$ and $|G \cap X_2| \leqslant |H \cap X_2|$. Consequently, either

$$|G \cap X_1| < |H \cap X_1| \quad \text{and} \quad |G \cap X_2| < |H \cap X_2| \tag{3.3}$$

or

$$|G \cap X_1| > |H \cap X_1| \quad \text{and} \quad |G \cap X_2| > |H \cap X_2| \tag{3.4}$$

must hold.

Let $|\mathcal{G}| = r$ and suppose $1 \leqslant r$. Choose a member $G \in \mathcal{G}$. Since G is a combined initial segment, the sizes $i = |G \cap X_1|, j = |G \cap X_2|$ uniquely determine it, that is the members G can be determined by the pairs (i, j). If (i_1, j_1) and (i_2, j_2) are two such pairs then either $i_1 < i_2, j_1 < j_2$ or $i_1 > i_2, j_1 > j_2$ must hold by (3.3) and (3.4). Index the members of $\mathcal{G}$ in such a way that $0 < i_1 < \ldots < i_r, 0 < j_1 < \ldots < j_r$ holds. Here the inequality $0 <$ comes from property (i). $i_r \leqslant n_1, j_r \leqslant n_2$ are trivial. Since the sets G

are not larger than k, $i_r + j_r \leqslant k$ also holds, the conditions on the sum on the left hand side of (3.2) coincide with the conditions in the definition of $A(n_1, n_2, k)$, proving the statement. □

Proof of Theorem 3.1. Suppose that $\mathcal{F}$ satisfies the conditions of the theorem (that is, $\mathcal{F} \subseteq \binom{[n]}{\leqslant k}$, and $\mathcal{F}$ is nodifsep with respect to the partition $X_1 \cup X_2 = [n]$) and consider the sum

$$\sum_{\pi_1,\pi_2,F} \binom{n_1}{|F \cap X_1|}\binom{n_2}{|F \cap X_2|} \tag{3.5}$$

where π_1, π_2 run over all permutations of X_1, X_2, respectively, and F is a combined initial segment with respect to these permutations. This sum will be calculated in two different ways.

Calculate first

$$\sum_{\pi_1,\pi_2} \binom{n_1}{|F \cap X_1|}\binom{n_2}{|F \cap X_2|} \tag{3.6}$$

for a fixed F. The number of permutations π_1 where $F \cap X_1$ is an initial segment in X_1 is $|F \cap X_1|!(n_1 - |F \cap X_1|)!$. Similarly, the number of permutations π_2 where $F \cap X_2$ is an initial segment in X_2 is $|F \cap X_2|!(n_2 - |F \cap X_2|)!$. Therefore the number of pairs π_1, π_2 where F is a combined initial segment is $|F \cap X_1|!(n_1 - |F \cap X_1|)!|F \cap X_2|!(n_2 - |F \cap X_2|)!$. For fixed F the summands in (3.6) are constants, therefore (3.6) is equal to

$$|F \cap X_1|!(n_1 - |F \cap X_1|)!|F \cap X_2|!(n_2 - |F \cap X_2|)!\binom{n_1}{|F \cap X_1|}\binom{n_2}{|F \cap X_2|} = n_1!n_2!.$$

Hence (3.5), calculating in this order, is

$$\sum_F \sum_{\pi_1,\pi_2} \binom{n_1}{|F \cap X_1|}\binom{n_2}{|F \cap X_2|} = |\mathcal{F}|n_1!n_2!. \tag{3.7}$$

Fix now a pair of permutations. The members of $\mathcal{F}$ which are combined initial segments with respect to these permutations satisfy the conditions of Lemma 3.2, therefore

$$\sum_F \binom{n_1}{|F \cap X_1|}\binom{n_2}{|F \cap X_2|}$$

can be upper-bounded by $A(n_1, n_2, k)$ and since the number of pairs of permutations is $n_1!n_2!$, we obtain

$$\sum_{\pi_1,\pi_2} \sum_F \binom{n_1}{|F \cap X_1|}\binom{n_2}{|F \cap X_2|} \leqslant n_1!n_2!A(n_1, n_2, k). \tag{3.8}$$

The comparison of (3.7) and (3.8) results in the inequality

$$|\mathcal{F}|n_1!n_2! \leqslant n_1!n_2!A(n_1,n_2,k),$$

proving the difficult part of (3.1).

It remains to show that there is an $\mathcal{F}$ of size $A(n_1,n_2,k)$, satisfying the conditions of the theorem. Take the integers $1 \leqslant r, 0 < i_1 < \ldots < i_r \leqslant n_1, 0 < j_1 < \ldots < j_r \leqslant n_2, (i_r + j_r \leqslant k)$ providing the maximum in $A(n_1,n_2,k)$. The family consisting of all sets F satisfying $|F \cap X_1| = i_\ell, |F \cap X_2| = j_\ell$ for some $1 \leqslant \ell \leqslant r$. It obviously meets all the requirements of the theorem and its size is really $A(n_1,n_2,k)$. □

Now we will show that $i_r + j_r = k$ can be supposed in the definition of $A(n_1,n_2,k)$ if $k < \frac{n_1+n_2}{2}$, that is, there is always a parameter set with $i_r + j_r = k$ and $\sum_{\ell=1}^{r} \binom{n_1}{i_\ell}\binom{n_2}{j_\ell} = A(n_1,n_2,k)$. It will be shown by increasing $\binom{n_1}{i_r}\binom{n_2}{j_r}$ if $i_r + j_r < k$. Indeed, $\frac{n}{2} = \frac{n_1+n_2}{2} \geqslant k > i_r + j_r$ implies $i_r \leqslant \frac{n_1-1}{2}$ or $j_r \leqslant \frac{n_2-1}{2}$ and thus at least one of i_r and j_r can be increased by one without decreasing $\sum_{\ell=1}^{r} \binom{n_1}{i_\ell}\binom{n_2}{j_\ell} = A(n_1,n_2,k)$.

Let us sketch the rest of the content of the present section. In the sequel we will assume that – as in Theorem 3.7 – n is large enough compared to k. In particular, $n \geqslant 2k$ and so, by the previous comment, $i_r + j_r = k$ may be assumed, and so, the term $\binom{n_1}{i_r}\binom{n_2}{j_r}$ in the sum $A(n_1,n_2,k)$ can be chosen to be equal to constant times n^k. On the other hand, all other terms are $O(n^{k-2})$. That is, the dominating term is of the form $\binom{n_1}{i_r}\binom{n_2}{k-i_r}$. We will maximize this quantity as a function of n_1 and i_r for large n. It will turn out that one of i_r and $j_r = k - i_r$ is 1, that is there is no other term in $A(n_1,n_2,k)$.

The proof is broken into lemmas.

Lemma 3.3. Suppose $1 \leqslant i < \ell, i \leqslant n_1, \ell - i \leqslant n - n_1$. Then

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

Proof. Compare two consecutive expressions:

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

After carrying out the possible cancellations

$$(n-n_1)(n_1-i+1) \leqslant (n_1+1)(n-n_1-\ell+i)$$

is obtained what is equivalent to

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

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 3.4. Suppose $0 < \frac{\ell}{2} \leqslant i \leqslant \ell - 1$. Then there is a constant $n_0(\ell)$ depending on $\ell < n$ only, such that if $n_0(\ell) \leqslant n$ then

$$\binom{\left\lfloor \frac{(n+1)i}{\ell} \right\rfloor}{i}\binom{n - \left\lfloor \frac{(n+1)i}{\ell} \right\rfloor}{\ell - i} < \binom{\left\lfloor \frac{(n+1)(i+1)}{\ell} \right\rfloor}{(i+1)}\binom{n - \left\lfloor \frac{(n+1)(i+1)}{\ell} \right\rfloor}{\ell - i - 1}. \tag{3.9}$$

where the last binomial coefficient is understood to be 1 when $i = \ell - 1$.

Proof. Since $\lfloor x \rfloor$ differs from x by at most one, $\frac{1}{n}\left\lfloor \frac{(n+1)i}{\ell} \right\rfloor$ tends to $\frac{i}{\ell}$ when n tends to infinity. The same is true for $\left\lfloor \frac{(n+1)i}{\ell} \right\rfloor - 1, \left\lfloor \frac{(n+1)i}{\ell} \right\rfloor - 2, \ldots$. Therefore

$$\lim_{n\to\infty} \frac{1}{n^i}\binom{\left\lfloor \frac{(n+1)i}{\ell} \right\rfloor}{i} = \left(\frac{i}{\ell}\right)^i \frac{1}{i!}.$$

We can obtain the following limits in the same way.

$$\lim_{n\to\infty} \frac{1}{n^{i+1}}\binom{\left\lfloor \frac{(n+1)(i+1)}{\ell} \right\rfloor}{i+1} = \left(\frac{i+1}{\ell}\right)^{i+1} \frac{1}{(i+1)!},$$

$$\lim_{n\to\infty} \frac{1}{n^{\ell-i}}\binom{n - \left\lfloor \frac{(n+1)i}{\ell} \right\rfloor}{\ell - i} = \left(\frac{\ell-i}{\ell}\right)^{\ell-i} \frac{1}{(\ell-i)!},$$

$$\lim_{n\to\infty} \frac{1}{n^{\ell-i-1}}\binom{n - \left\lfloor \frac{(n+1)(i+1)}{\ell} \right\rfloor}{\ell - i - 1} = \left(\frac{\ell-i-1}{\ell}\right)^{\ell-i-1} \frac{1}{(\ell-i-1)!}.$$

To prove the lemma we have to see that the limit of the left hand side of (3.9) divided by $\frac{1}{n^\ell}$ is less than the same for the right hand side of (3.9):

$$\left(\frac{i}{\ell}\right)^i \frac{1}{i!}\left(\frac{\ell-i}{\ell}\right)^{\ell-i}\frac{1}{(\ell-i)!} <$$

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

An equivalent inequality is

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

Here the assumptions of the lemma imply $i > \ell - i - 1$. Therefore $\frac{i}{i+1} < \frac{\ell-i-1}{\ell-1}$ holds and the inequality above contains a higher power of the smaller quantity. □

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

The information what is really needed from Lemmas 3.3 and 3.4 is collected in the following lemma.

Lemma 3.5. Suppose that $2 \leqslant i \leqslant k-2, i \leqslant n_1, k-i \leqslant n-n_1$. Then there is a constant $n_0(k)$ depending on k only, such that if $n_0(k) \leqslant n$ then

$$\binom{n_1}{i}\binom{n-n_1}{k-i} \leqslant \binom{\left\lfloor\frac{(n+1)(k-2)}{k}\right\rfloor}{k-2}\binom{n-\left\lfloor\frac{(n+1)(k-2)}{k}\right\rfloor}{2}. \tag{3.10}$$

On the other hand, if $k-1 \leqslant n_1, 1 < n-n_1$ then

$$\binom{n_1}{k-1}(n-n_1) \leqslant \binom{\left\lfloor\frac{(n+1)(k-1)}{k}\right\rfloor}{k-1}\left(n-\left\lfloor\frac{(n+1)(k-1)}{k}\right\rfloor\right). \tag{3.11}$$

Finally,

$$\binom{\left\lfloor\frac{(n+1)(k-2)}{k}\right\rfloor}{k-2}\binom{n-\left\lfloor\frac{(n+1)(k-2)}{k}\right\rfloor}{2} \leqslant$$

$$\leqslant \binom{\left\lfloor\frac{(n+1)(k-1)}{k}\right\rfloor}{k-1}\left(n-\left\lfloor\frac{(n+1)(k-1)}{k}\right\rfloor\right).$$

We need somewhat more than the last inequality, namely that the largest one of these products of binomial coefficients really stands out of the other ones, that is the last inequality is strict, the two sides are apart from each other by a difference equal to a positive constant times n^k. This is expressed by the following lemma, it can be proved by elementary calculus.

Lemma 3.6.

$$\binom{\left\lfloor\frac{(n+1)(k-1)}{k}\right\rfloor}{k-1}\left(n-\left\lfloor\frac{(n+1)(k-1)}{k}\right\rfloor\right)$$

is asymptotically equal to

$$c_1(k)n^k=\frac{1}{(k-1)!}\frac{(k-1)^{k-1}}{k^k}n^k,$$

$$\binom{\left\lfloor\frac{(n+1)(k-2)}{k}\right\rfloor}{k-2}\binom{n-\left\lfloor\frac{(n+1)(k-2)}{k}\right\rfloor}{2}$$

is asymptotically equal to

$$c_2(k)n^k=\frac{2}{(k-2)!}\frac{(k-2)^{k-2}}{k^k}n^k$$

where $c_2(k)<c_1(k)$ holds for $4\leqslant k$.

Note that Lemma 3.6 defines two constants $c_1(k)>c_2(k)$ for $4\leqslant k$ which will be used below. Now we are ready to formulate the main statement of this section which answers Problem 5 for large n.

Theorem 3.7. Suppose $4\leqslant k$. Then there is a constant $n_1(k)$ depending on k only such that if $n_1(k)\leqslant n$, then

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

Proof. Observe that

$$M(n,k)=\max_{1\leqslant n_1<n}M(n_1,n-n_1,k).$$

Consider a choice of n_1 giving the maximum here and then a choice of the parameters $r,i_1,\ldots,i_r,j_1,\ldots,j_r$ giving the maximum in Theorem 3.1. ($i_r+j_r=k$ is supposed.) Two cases will be distinguished.

Case 1: One of i_r and j_r, say the latter one is 1. Then $r = 1$, the sum in Theorem 3.1 is consists of only one term. This only term is upper-bounded by (3.11), proving the theorem in this case.

Case 2: $2 \leqslant i_r, j_r \leqslant k-2$. In this case (3.10) gives an upper estimate on

$$\binom{n_1}{i_r}\binom{n-n_1}{k-i_r}.$$

This upper bound is asymptotically equal to $c_2(k)n^k$. Since $i_1 + j_1 < i_2 + j_2 < \ldots < i_{r-1} + j_{r-1} \leqslant k-2$, the sum of all other terms in Theorem 3.1 are $O(n^{k-2})$, consequently the total sum is $c_2(k)n^k$, asymptotically. For large n this is smaller than

$$\binom{\left\lfloor \frac{(n+1)(k-1)}{k} \right\rfloor}{k-1}\left(n - \left\lfloor \frac{(n+1)(k-1)}{k} \right\rfloor\right)$$

which is asymptotically $c_1(k)n^k$. □

4. Answer to the database question

Theorem 4.1. Let $4 \leqslant k$ and n be integers, where n is large relative to k : $n_1(k) \leqslant n$. The maximum number of SUM queries involving at most k data of the n numerical data is

$$\binom{\left\lfloor \frac{(n+1)(k-1)}{k} \right\rfloor}{k-1}\left(n - \left\lfloor \frac{(n+1)(k-1)}{k} \right\rfloor\right). \tag{4.1}$$

Proof. Problem 3 in Section 2 raises the question of the present theorem. Problem 3 is there reduced to Problem 5 with a weaker condition. Therefore the upper bound of Theorem 3.7 answering Problem 5 is an upper bound for Problem 3, too. We only have to show that the bound is sharp here, too. That is, one can construct a family of SUM queries (over properly chosen set of numbers) of this size with the given properties.

Let us consider n equal real numbers and divide them 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 these numbers of k elements such that $k-1$ are chosen from set B_1 and 1 from set B_2. The sets really have size k, their number is equal to (4.1). We only have to show that knowing the answer to these queries it does not determine any of the individual data.

Now increase all of the elements of B_1 by 1 and decrease all of the elements of B_2 by $k-1$. 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. □

5. Concluding remarks

The restriction to the validity of Theorem 4.1 that n is relatively large is not essential from the practical point of view. In the real situations the total number of numerical data is huge, the restriction on the size of a query makes sense only when it is really limited, that is, k is much smaller than n, as we assume.

It is not true however from the mathematical point of view, the ratio of n and k plays important role. Further work is needed in this direction. Some easy calculations, completing our proofs, can give the value of $n_1(k)$ is in Theorem 3.7. It is, however, very far from the real necessary bound. Still, if n and k are close to each other, the statement of Theorem 4.1 is not true. It is known [7] that

$$M(n,n) = \binom{n}{\lfloor \frac{n}{2} \rfloor} - 1 = \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} \binom{\lceil \frac{n}{2} \rceil}{i} \binom{\lfloor \frac{n}{2} \rfloor}{\lfloor \frac{n}{2} \rfloor - i} = \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} \binom{\lceil \frac{n}{2} \rceil}{i} \binom{\lfloor \frac{n}{2} \rfloor}{i},$$

where the right hand side equation is given by the following construction (showing, at least, that $M(n,n) \geqslant \binom{n}{\lfloor \frac{n}{2} \rfloor} - 1$): Consider again n equal numbers, divided into two groups, B_1 and B_2 of sizes $\lfloor \frac{n}{2} \rfloor$ and $\lceil \frac{n}{2} \rceil$, respectively. Take all subsums of these numbers of $2i$ elements such that i are chosen from both of these sets. Knowing the answer to these queries does not determine any of the individual data, since by increasing all of the elements of B_1 by 1 and decreasing all of the elements of B_2 by 1, the answers to these queries will remain the same.

It is expected that this construction will remain the best if n is not much larger than k (suppose for convenience that k is even):

$$M(n,k) = \sum_{i=1}^{\frac{k}{2}} \binom{\lceil \frac{n}{2} \rceil}{i} \binom{\lfloor \frac{n}{2} \rfloor}{i}.$$

For example, we have the following examples: $M(20,6) = \binom{17}{5} \cdot 3$, but $M(12,6) = \binom{6}{3}\binom{6}{3} + \binom{6}{2}\binom{6}{2} + \binom{6}{1}\binom{6}{1} = 661 > 504 = \binom{10}{5}\binom{2}{1}$.

Let us mention that in practical situations the restriction on the possible sums does not necessarily come in the form of a 'size $\leqslant k$.' The restriction comes in some other form (like only at most one datum may be asked for a large group data), what may imply the restriction on the size. Then not all of the sums containing at most k data can be disclosed, our results give only an upper bound which is not tight in general.

References

[1] F.Y. Chin, G. Ozsoyoglu, Auditing and inference control in statistical databases, IEEE Transactions on Software Engineering SE-8 (1982) 574–582.

[2] D.E. Denning, *Cryptography and Data Security* (Addison-Wesley, Sydney, 1982).

[3] D.E. Denning and J. Schlorer, Inference controls for statistical databases, Computer (1983), 69–82.
[4] J.R. Griggs, Concentrating subset sums at k points, Bulletin of the Institute of Combinatorics and its Applications 20 (1997) 65–74.
[5] J.R. Griggs, Database security and the distribution of subset sums in $\mathbf{R}^m$, in: "Graph Theory and Combinatorial Biology," BSMS 7, eds. L. Lovász et al. (Bolyai Society, Budapest, 1999).
[6] J. Littlewood and C. Offord, On the number of real roots of a random algebraic equation III, Mat. Sbornik 12 (1943) 277–285.
[7] M. Miller, I. Roberts and I. Simpson, Application of symnmetric chains to an optimization problem in the security of statistical databases, Bulletin ICA 2 (1991) 47–58.
[8] M. Miller, I. Roberts, I. Simpson, Preventation of relative compromise in statistical databases using Audit Expert, Bulletin ICA 10 (1994) 51–62.