ON SECURITY OF STATISTICAL DATABASES

R. AHLSWEDE AND H. AYDINIAN
DEPARTMENT OF MATHEMATICS
UNIVERSITY OF BIELEFELD
POB 100131, D-33501 BIELEFELD, GERMANY
EMAIL: AHLSWEDE@MATH.UNI-BIELEFELD.DE
AYD@MATH.UNI-BIELEFELD.DE

Abstract.
A statistical database (SDB) is a database that is used to return statistical information derived from the records to user queries for statistical data analysis. Sometimes, by correlating enough statistics, confidential data (stored in a SDB) about an individual can be inferred. Examples of confidential information stored in a SDB might be salaries or data concerning the medical history of individuals. An important problem is to provide security to SDB against the disclosure of confidential information. A statistical database is said to be secure if no protected data can be inferred from the available queries. One of the security-control methods suggested in the literature consists of query restriction: the security problem is to limit the use of the SDB, introducing a control mechanism, such that no protected data can be obtained from the available queries. Chin and Ozsoyoglu [7] introduced a control mechanism, called Audit Expert, where only SUM queries, that is only certain sums of individual records, are available for the users. This SUM query model leads to several challenging optimization problems. Assume there are n numeric records $\{z_1, \ldots, z_n\}$ stored in a database. A natural problem is to maximize the number of answerable SUM queries, that is the number of subset sums of $\{z_1, \ldots, z_n\}$ (possibly with some additional constraints) that can be returned, such that none of numbers z_i (or sums of subsets with the size not exceeding a specified number) can be inferred from these queries. In this paper we give tight bounds for this number under constraints on size and dimension of query subsets.

Key words. Database security, Subset sums, Dimension constraints

AMS subject classifications. 68R05, 68P15, 05D05

1. Introduction. The problems of statistical database security have been of growing concern in recent years [6],[8], [12-16]. A statistical database (SDB) is a database that is used to return statistical information derived from the records to user queries for statistical data analysis. Sometimes, by correlating enough statistics, confidential data (stored in a SDB) about an individual can be inferred. Examples of confidential information stored in a SDB might be salaries or data concerning the medical history of individuals. An important problem is to provide security to SDB against the disclosure of confidential information. A statistical database is said to be *secure* if no protected data can be inferred from the available queries. When users are able to infer protected information in the SDB from responses to queries, the SDB is said to be *compromised*.

Security-control methods suggested in the literature (see [9], [10]) are classified into four general approaches: *conceptual, query restriction, data perturbation, and output perturbation*.

We are interested in query restrictions where the security problem is to limit the use of the SDB, introducing a control mechanism, such that no protected individual data can be obtained from the available queries. Such a control mechanism for query restriction, called AUDIT EXPERT, was proposed in Chin and Ozsoyoglu [7], where only SUM queries, that is only certain sums of individual records, are available for the users. This model of security leads to several optimization problems which arise in a natural way.

As an example consider a company N with n employees. Suppose that for each

member of N is recorded the sex, age, rank, length of her/his employment with N, salary etc. The salaries $\{z_1, \ldots, z_n\}$ of the individual employees are confidential. Suppose that only SUM queries are allowed, i.e. the sum of the salaries of the specified people is returned. For example one might pose the query: What is the sum of salaries for males above 50, working with N during the last 10 years?
How large can be the number of SUM queries, preventing compromise (i.e. no individual salary z_i can be inferred using the outcomes from the list of allowed SUM queries)?

More generally, let $z_1, \ldots, z_n$ be nonzero real numbers which are n confidential records stored in a database. A possible SUM query for users is $S_A := \sum_{i \in A} z_i$ for some $A \subset [n] := \{1, \ldots, n\}$ with $|A| > 1$.

A natural problem is to maximize the number of SUM queries, possibly with some other side constraints, without compromise. This problem was originally stated in Chin and Ozsoyoglu [7] (see also [14]) and is studied (in different settings) in [6], [8], [12-16].

In particular, consider the problem (without constraints): Maximize the umber M of subsets (answerable query sets) $A_1, \ldots, A_M \subset [n]$ (or the ratio $M/2^n$ called the *usability* of SDB) such that the knowledge of the corresponding sums $S_{A_1}, \ldots, S_{A_M}$ does not enable one to determine any of records z_i.

The problem can be reduced to the following optimization problem. For a subset $A \subset [n]$, its characteristic vector is defined by $\chi(A) = (x_1 \ldots, x_n)$, where $x_i = 1$ if $i \in A$ and $x_i = 0$, if $i \notin A$. Thus each SUM query S_A can be represented by the characteristic vector of A. Let $X \subset \{0,1\}^n \subset \mathbb{R}^n$ be the set of characteristic vectors corresponding to a family of query sets $A_1, \ldots, A_M$ avoiding compromise. It is clear that $\mathrm{span}(X) \neq \mathbb{R}^n$, that is $\dim(\mathrm{span}(X)) \leq n-1$ (otherwise the SDB is compromised). This means that X lives in an $(n-1)$-dimensional subspace $U \subset \mathbb{R}^n$. Suppose now U is defined by $U = \{(x_1, \ldots, x_n) \in \mathbb{R}^n : a_1x_1 + \ldots + a_nx_n = 0\}$ where $a_1, \ldots, a_n \in \mathbb{R}$. Note then that $a_i \neq 0$ $(i = 1, \ldots, n)$, otherwise there exists a unit vector in $\mathrm{span}(X)$, (that is the SDB is compromised).
Thus, we come to the following problem:
Given nonzero real numbers $a_1, \ldots, a_n$, determine the maximum possible number of subsets with a zero sum, that is determine the maximum number of (0,1)-solutions of an equation

$$a_1x_1 + \ldots + a_nx_n = 0 \quad (a_i \in \mathbb{R} \setminus \{0\}). \tag{1.1}$$

Miller et al. [14] solved this problem reducing it to a combinatorial extremal problem and using symmetric chain decomposition of the Boolean lattice (see [5] or [11]).

Theorem 1 ([14]). *(i) The maximum number of answerable SUM queries without compromise, from a database of n real entries z_i is $\binom{n}{\lfloor n/2 \rfloor}$.*

(ii) (Griggs [12]) *The maximum is achieved iff the set of entries is partitioned into two parts, of sizes $\lfloor \frac{n}{2} \rfloor$ and $\lceil \frac{n}{2} \rceil$, and all query sets have an equal number of elements from each part. Equivalently, the maximum number of (0,1)-solutions of (1,1), assumed for $a_i = -a_{i+1}$ $(i = 1, \ldots, n-1)$, is unique up to permutations of the*

coordinates.

In a series of papers [6], [13-16] Miller et all. introduced and studied other models of compromise. Among them so called *relative compromise* where either some record z_i or some difference $z_i - z_j$ $(i \neq j)$ can be inferred from available queries. This model leads to the famous Erdös–Moser problem (determine the largest possible number of subsets of a set of nonzero real numbers $\{a_1, \ldots, a_n\}$ having a common sum of elements) and its generalizations. In an excellent survey paper by Griggs [12], further fundamental models of database compromise, *group-security, internal-security etc.*, were proposed. It was shown that they lead to challenging combinatorial, number theoretic, and geometric problems.

All these problems can be formulated in terms of (0,1)–solutions of some linear equations (over real numbers) with certain restrictions.

In the model called *group–security model* (see [12]), not only individual data but also subset sums of subsets $I \subset [n]$ with small size, say $0 < |I| \leq g$, must be protected. By the observation above, this problem is equivalent to the following one.

Problem 1. Determine the maximum number $G(n, g)$ of (0,1)–solutions of equation (1.1) provided there are no nonzero solutions of Hamming weight less than $g+1$. In other words $G(n, g)$ is the maximum number of (0,1)–vectors of an $(n-1)$–dimensional subspase (of $\mathbb{R}^n$) not containing a nonzero (0,1)–vector of weight less than $g+1$.

Problem 2 (with a size restriction on inquired subsets). Assume that the number of elements in the SUM queries are restricted by the size constraint: only sums of m (or at most m) elements are considered. This is a natural restriction since in the applications the size of the data stored in a SDB is usually huge, while the number of operations could be limited. The problem for $g = 1$ was considered and solved for $n \gg m$ in [8]. An equivalent formulation of the problem is the finding of maximal number of (0,1)-solutions of weight m (or **Problem** 2^*: weight not exceeding m) of equation (1.1), provided there are no (nonzero) solutions of weight less than $g+1$. We denote this quantity by $G(n, m, g)$ (resp. $G(n, \leq m, g)$).

Theorem 2 (Demetrovich et al. [8]). *For integers* $1 < m \leq n$ *and* $t := \lfloor n/m \rfloor$, *holds*

(i) $G(n, m, 1) = t\binom{n-t}{m-1}$, *if* $n \gg m$.

(ii) $G(n, \leq m, 1) = t\binom{n-t}{m-1}(1 + o(1))$, *as* $n \to \infty$.

In [8] it was also shown that the bound in (i) is tight if the query sets are from two consecutive levels $m, m-1$. In fact, as we see below, the equality in (i) holds for all integers $1 < m < n$.

Let us consider the following more general problem, which clearly makes sense theoretically and hopefully also practically.

Problem 3. Under similar restrictions as in Problems 1,2 determine or estimate the maximal number of (0,1)–solutions of a linear equation

$$B(x_1, \ldots, x_n)^T = 0, \tag{1.2}$$

where B is a real $r \times n$ matrix of rank r.

In particular, for integers $1 \leq g, k \leq n$ let $G_k(n, g)$ denote the maximum number of (0,1)–solutions of equation (1.2) such that $rank(B) = n - k$, provided there are no

solutions of the weight g or less. Clearly the set of SUM queries (with the characteristic vectors) corresponding to these solutions does not lead to g-group compromise.

This problem was also addressed in [12] as an extention of the group security problem to higher dimension. Note that $G(n,g) := G_{n-1}(n,g)$. As a motivation for study of Problem 3 let us also mention the notion of *internal security* introduced in [12]. Let $\{z_1,\dots,z_n\}$ be confidential records (say salaries in company $N=\{1,\dots,n\}$). Suppose a coalition $K \subset N$ of $h-1$ members of the company can produce a linear combination $\sum_{i\in I}\alpha_i z_i$ $(\alpha_i \neq 0)$, $K \subset I \subset N$ with $|I| = h$ using allowable SUM queries $A_1,\dots,A_M \subset N$. Then they can infer the record (salary) z_j where $\{j\} = I - K$. The database is called then *h–inside compromised*. Griggs [12] observed that the maximum number of SUM queries avoiding h–inside compromise equals the maximum number of $(0,1)$–solutions of equation (1.2) with $h = r$ and every h columns of B are linearly independent. Note that in the case when B is a matrix (of rank r) without zero columns then a coalition of $h-1$ members, $1 \leq h \leq r-1$, can infer at most $n-r+h$ protected records.

In this paper we study the group security problems stated above. We give all exact solutions to Problem 2, thus we determine also $G(n,m,g)$ for all parameters. Surprisingly the answer is the same as for 1-security, that is $G(n,m,g) = G(n,m,1)$. We solve Problem 2* for $n \geq m^2$ showing that $G(n,\leq m,g) = G(n,m,1)$. We also determine $G(n,g)$ (as well as $G_k(n,g)$), within a constant factor less than $1/2$, thus answering the question raised in Griggs [12] about the usability of AUDIT EXPERT for the g-group security model (also for the higher dimensional case). For this case it turns out that for all $1 < g < \frac{n}{2}$ the number of answerable queries (without g-group compromise) decreases less than two times as compared with 1-security, that is $G(n,g) > \frac{1}{2}G(n,1)$. Our main results are stated and proved in Section 2. In Section 3 we discuss the results and some open problems.

2. Main results. We need some notation and definitions. Throughout the paper we use the abbreviation $[m,n]$ for the interval of integers $\{m,m+1,\dots,n\}$ and $[n] := [1,n]$. We also use the notation: $2^{[n]} = \{A : A \subset [n]\}$, $\binom{[n]}{k} = \{A \subset 2^{[n]} : |A| = k\}$, $\binom{[n]}{\leq k} := \{A \subset 2^{[n]} : |A| \leq k\}$, $E^n = \{0,1\}^n \subset \mathbb{R}^n$, and $E^n_k = \{x \in E^n : x \text{ has } k \text{ ones}\}$ for $n,k \in \mathbb{N}$ $(k \leq n)$.

A family $\mathcal{A} = \{A_1,\dots,A_m\} \subset 2^{[n]}$ is called a chain of size m if $A_1 \subset \cdots \subset A_m$. If $m = n+1$ then $\mathcal{A}$ is called a maximal chain. $\mathcal{A} \subset 2^{[n]}$ is called an antichain if $A_i \not\subset A_j$ holds for all distinct $A_i, A_j \in \mathcal{A}$.

Let us also recall two classical results concerning antichains in $2^{[n]}$ (see textbooks [5], [11])

Sperner's Theorem. *Let $\mathcal{A} \subset 2^{[n]}$ be an antichain, then $|\mathcal{A}| \leq \binom{n}{\lfloor \frac{n}{2} \rfloor}$ and the maximum is achieved only for $\mathcal{A} = \binom{[n]}{\lfloor \frac{n}{2} \rfloor}$ or $\binom{[n]}{\lceil \frac{n}{2} \rceil}$.*

LYM inequality (Lubell-Yamamoto-Meshalkin). *Let $\mathcal{A} \subset 2^{[n]}$ be an antichain, then*

$$\sum_{A\in\mathcal{A}} \frac{1}{\binom{n}{|A|}} \leq 1.$$

Let $\mathcal{A} \subset 2^{[n]}$ be a maximal family of query sets avoiding compromise for a database of n records. As we note d above, there exist nonzero real numbers $a_1,\dots,a_n$ such that for each set $A \in \mathcal{A} \subset 2^{[n]}$ the corresponding subset sum $\sum_{i\in A} a_i = 0$. Let $X := \chi(\mathcal{A})$ be the set of characteristic vestors corresponding to the members of $\mathcal{A}$,

that is X is the set of (0,1)-solutions of equation (1.1). Without loss of generality we may write the equation (1.1) in the following form

$$a_1x_1+\ldots+a_\ell x_\ell-a_{\ell+1}x_{\ell+1}-\ldots-a_nx_n=0, \quad \text{where all } a_i>0 \text{ and } \ell\in[n-1]. \tag{2.1}$$

Let the ground set $[n]$ be partitioned into two parts $[n]=[\ell]\cup[\ell+1,n]$. Observe that $\mathcal{A}$ satisfies the following property (P):

(P) For all $A,B\in\mathcal{A}$

(i) $A\cap[\ell]=\emptyset$ or $A\cap[\ell+1,n]=\emptyset$ iff $A=\emptyset$

(ii) $(A\cap[\ell])\subseteq(B\cap[\ell])$ implies $(A\cap[\ell+1,n])\not\supseteq(B\cap[\ell+1,n])$.

Indeed, (i) is obvious; assuming the opposite in (ii) we get $\sum_{i\in B}a_i-\sum_{i\in A}a_i>0$, a contradiction. Later on we assume, without loss of generality, that $1\le\ell\le\frac{n}{2}$.

Our first result sharpens Theorem 2 and generalizes it to g-group security fo r arbitrary g. The result is easily derived from a result in [2]. Given $n,m,w\in\mathbb{N}$ let $F(n,m,w)$ denote the maximum number of (0,1)–vectors $X\subset\mathbb{R}^n$ of weight m such that the $span(X)$ does not contain (0,1)-vectors of weight w. Similarly is defined the function $F(n,w)$ where again vectors of weight w are forbidden but we have no restriction on the weights of (0,1)–vectors corresponding to the query sets (the unrestricted case).

In [2] $F(n,m,w)$ is determined for all parameters $1\le w<m<n$. Results for $F(n,w)$ are presented in [3].

Theorem 3 [2]. *For integers $1\le w<m<n$ and $t:=\lfloor\frac{n}{m}\rfloor$ we have*

$$F(n,m,w)=\binom{t}{1}\binom{n-t}{m-1}. \tag{2.2}$$

It is clear that $F(n,m,1)=G(n,m,1)$ and $F(n,1)=G(n,1)$.

However. note that the query (sets corresponding to the (0,1)-vectors) satisfying the restriction for $F(n,m,g)$ (as well as for $F(n,g)$) with $g\ge2$ does not necessarily avoid $(g-1)$-compromise. Thus, clearly we have $F(n,m,g)\ge G(n,m,g)$ (and $F(n,g)\ge G(n,g)$). Surprisingly, one has also the following.

Lemma 1. *For integers $1\le g<m\le n$ we have $F(n,m,1)=F(n,m,g)=G(n,m,g)$.*

Proof. Theorem 3 shows that the quantity $F(n,m,w)$ does not depend on w (for $1\le w<m$) thus $F(n,m,1)=F(n,m,w)$.

Note now that the equality in (2.1) is achieved for the set

$Y=\{(x_1,\ldots,x_n)\in E(n,m):(m-1)x_1+\ldots+(m-1)x_t-x_{t+1}-\ldots-x_n=0\}$.

Note also that $X:=\text{span}(Y)\cap E(n)=\{(x_1,\ldots,x_n)\in E(n):(m-1)x_1+\ldots+(m-1)x_t-x_{t+1}-\ldots-x_n=0\}$ consists only of vectors of weight 0 modulo m. This implies that $G(n,m,g)\ge F(n,m,g)$ concluding the result. □

Theorem 4. (i) *For integers $1\le g<m\le n$ let $t\in\{\lfloor n/m\rfloor,\lfloor(n+1)/m\rfloor\}$. Then we have*

$$G(n,m,g)=t\binom{n-t}{m-1}. \tag{2.3}$$

(ii) An optimal set of SUM queries corresponds to the set of (0,1)-solutions of weight m of equation $(m-1)x_1+\ldots+(m-1)x_t-x_{t+1}-\ldots-x_n=0$ and is unique (up to the permutations of the elements) if $\lfloor n/m\rfloor=\lfloor(n+1)/m\rfloor$. If $\lfloor n/m\rfloor\neq\lfloor(n+1)/m\rfloor$, then there are two optimal configurations with $t=\lfloor n/m\rfloor$ or $t=\lfloor n/m\rfloor+1$.

Proof. Easy calculation shows that in case $\lfloor n/m\rfloor\neq\lfloor(n+1)/m\rfloor$, that is for $n=(t_1+1)m-1$, where $t_1:=\lfloor n/m\rfloor$, we have

$$t_1\binom{n-t_1}{m-1}=(t_1+1)\binom{n-t_1-1}{m-1}. \tag{2.4}$$

Thus, we have only to prove the second part of the theorem. Given a partition $[n]=[\ell]\cup[\ell+1,n]$, let us represent the elements of $2^{[n]}$ by pairs $(A_1,A_2):=A_1\cup A_2$ where $A_1\subseteq[\ell]$ and $A_2\subseteq[\ell+1,n]$.

Let $\mathcal{A}\subset 2^{[n]}$ be an antichain satisfying property (P). That is for every $(A_1,A_2),(B_1,B_2)\in\mathcal{A}$ either A_1 and B_1 or A_2 and B_2 form an antichain.

Then one has the following generalization of the LYM inequality.

Lemma 2 [2].

$$\sum_{(A_1,A_2)\in\mathcal{A}}\frac{1}{\binom{\ell}{|A_1|}\binom{n-\ell}{|A_2|}}\leq 1. \tag{2.5}$$

The proof exploits Lubell's argument (for the LYM inequality). For completeness we present it here.

Proof. Consider the set of all direct products $\sigma:=\{\mathcal{C}_1\times\mathcal{C}_2\}$, where $\mathcal{C}_1$ and $\mathcal{C}_2$ are maximal chains in $[\ell]$ and $[\ell+1,n]$ respectively. Note that $|\sigma|=\ell!(n-\ell)!$. Given $A\in 2^{[n]}$ there are $|A_1|!(\ell-|A_1|)!|A_2|!(n-|A_2|)!$ members of σ containing A and each member of σ contains at most one element from $\mathcal{A}$. Hence the probability that a randomly chosen member of σ meets an element from our family $\mathcal{A}$ is $\sum_{(A_1,A_2)\in\mathcal{A}}|A_1|!(\ell-|A_1|)!|A_2|!(n-|A_2|)!/\ell!(n-\ell)!\leq 1$, which implies (2.5). □

Clearly (2.5) implies

$$\frac{|\mathcal{A}|}{\max_{(A_1,A_2)\in\mathcal{A}}\binom{\ell}{|A_1|}\binom{n-\ell}{|A_2|}}\leq\sum_{(A_1,A_2)\mathcal{A}}\frac{1}{\binom{\ell}{|A_1|}\binom{n-\ell}{|A_2|}}\leq 1. \tag{2.6}$$

For $\mathcal{A}\subset\binom{[n]}{m}$ we can rewrite (2.5) as

$$\sum_{(A_1,A_2)\in\mathcal{A}}\frac{1}{\binom{\ell}{|A_1|}\binom{n-\ell}{m-|A_1|}}\leq 1. \tag{2.7}$$

Hence we also have

$$\frac{|\mathcal{A}|}{\max\limits_{\substack{1\leq\ell<n\\1\leq i\leq m-1}}\binom{\ell}{i}\binom{n-\ell}{m-i}}\leq\frac{|\mathcal{A}|}{\max_{1\leq i\leq m-1}\binom{\ell}{i}\binom{n-\ell}{m-i}}\leq\sum_{(A_1,A_2)\in\mathcal{A}}\frac{1}{\binom{\ell}{|A_1|}\binom{n-\ell}{m-|A_1|}}\leq 1. \tag{2.8}$$

For ℓ fixed (2.8) implies that $|\mathcal{A}| \le \max_{1\le i<m} \binom{\ell}{i}\binom{n-\ell}{m-i}$. Moreover it is easy to show (see [2]) that given integers $0 < \ell, m < n$ we have

$$\max_{1\le i<m} \binom{\ell}{i}\binom{n-\ell}{m-i} = \binom{\ell}{s}\binom{n-\ell}{m-s}, \quad \text{where } s := \lceil \frac{\ell m}{n+1} \rceil \tag{2.9}$$

Thus $|\mathcal{A}|$ and hence $G(n,m,g)$ is upper bounded by $\max \binom{\ell}{i}\binom{n-\ell}{m-i}$ taken over all $1 \le \ell < n$ and $1 \le i \le m-1$. It is proved in [2] that

$$\max_{\substack{1\le \ell<n \\ 1\le i\le m-1}} \binom{\ell}{i}\binom{n-\ell}{m-i} = \binom{t}{1}\binom{n-t}{m-1}. \tag{2.10}$$

For the proof of (2.10) it was shown that for $1 \le i \le m-1$ one has

$$\max_{1\le \ell<n} \binom{\ell}{i}\binom{n-\ell}{m-i} = \binom{\ell_i}{i}\binom{n-\ell_i}{m-i}, \quad \text{where } \ell_i := \lfloor \frac{(n+1)i}{m} \rfloor, \tag{2.11}$$

and for $2 \le i \le m-2$ holds

$$\binom{t}{1}\binom{n-t}{m-1} > \binom{\ell_i}{i}\binom{n-\ell_i}{m-i}. \tag{2.12}$$

Remark 1. In [2] is also remarked that using the same approach as for the proof of (2.10) one can show that for $n \ge \frac{m^2}{2}$ holds

$$\max_{\substack{1\le \ell<n \\ i\le j\le m-j}} \binom{\ell}{j}\binom{n-\ell}{m-j} = \binom{\ell_i}{i}\binom{n-\ell_i}{m-i}, \quad \text{where } \ell_i := \lfloor \frac{(n+1)i}{m} \rfloor. \tag{2.13}$$

From (2.11) and (2.12) we conclude that the maximum in (2.10) is achieved if and only if $i = 1$ and $t \in \{\lfloor n/m \rfloor, \lfloor (n+1)/m \rfloor\}$. If now $|\mathcal{A}| = G(n,m,g)$ then all inequalities in (2.8) must hold with equalities. This together with the previous observation implies that $|A \cap [\ell]| = 1$ for all $A \in \mathcal{A}$ and $\ell = t$. $\square$

Clearly Lemma 1 together with Theorem 2(ii) implies the following.

Corollary 1. Given integers $1 \le g < m$ we have

$$G(n, \le m, g) = t\binom{n-t}{m-1}(1+o(1)) \quad as\ n \to \infty. \tag{2.14}$$

Note that (2.14) follows from the fact that (for m fixed) the order of magnitude of $G(n,m,g)$ is n^m while $G(n, \le m, g) - G(n,m,g) < G(n,m-1,g) + \ldots + G(n,g+1,g)$ has the order of magnitude n^{m-1}.

Remark 2. Suppose a query family $\mathcal{B} \subset 2^{[n]}$ consists of two consecutive levels, that is $\mathcal{B} \subset \binom{[n]}{m} \cup \binom{[n]}{m-1}$. Observe then that in view of property (P) $\mathcal{B}$ is an antichain. Therefore we can repeat all arguments above and extend Theorem 4 to two consecutive levels. Moreover, we are able to describe all optimal query sets.

Theorem 4*. (i) *For integers* $1 \le g < m-1 < n$ *let* $t_1 := \lfloor \frac{n}{m} \rfloor$ *and* $t_2 := \lfloor \frac{n}{m-1} \rfloor$. *Then*

$$G(n,\{m-1,m\},g)=\max\left\{t_1\binom{n-t_1}{m-1},t_2\binom{n-t_2}{m-2}\right\}. \tag{2.15}$$

(ii) Let $\mathcal{A}\subset(\binom{[n]}{m}\cup\binom{[n]}{m-1})$ *be an optimal query family, that is* $|\mathcal{A}|=G(n,\{m-1,m\},g)$. *If* $\mathcal{A}\subset\binom{[n]}{m}$ *or* $\mathcal{A}\subset\binom{[n]}{m-1}$ *then it is determined by Theorem 4. All optimal families are described below.*

(a) If $n\geq 2m$ *then* $\mathcal{A}\subset\binom{[n]}{m}$.

(b) If $n\leq 2m-2$ *then* $\mathcal{A}\subset\binom{[n]}{m-1}$ *(and is unique).*

(c) If $n=2m-1$ *then there are exactly four optimal query sets:*

$\ell=1$*:* $\mathcal{A}\subset\binom{[n]}{m}$ *(and is unique),*

$\ell=2$*:* $\mathcal{A}\subset\binom{[n]}{m}$ *(and is unique),*

$\ell=2$*:* $\mathcal{A}\subset\binom{[n]}{m-1}$ *(and is unique),*

$\ell=2$*:* $\mathcal{A}=\mathcal{A}_1\cup\mathcal{A}_2$ *with* $\mathcal{A}_1=\{A\in\binom{[n]}{m}: A\cap[2]=\{1\}\}$, $\mathcal{A}_2=\{A\in\binom{[n]}{m-1}: A\cap[2]=\{2\}\}$.

In other words $\mathcal{A}$ *corresponds to the set of solutions* $X\subset(E(n,m)\cup(E(n,m-1))$ *of equation* $(m-1)x_1+(m-2)x_2-x_3-\ldots-x_n=0$.

Proof. Part (i) directly follows from (2.5).

Suppose $\mathcal{A}\subset(\binom{[n]}{m}\cup\binom{[n]}{m-1})$ is an optimal query family. The cases (a) and (b) follow by easy calculations for the maximum in (2.11). Suppose now $n=2m-1$. By Theorem 4 we have $|\mathcal{A}|=\binom{2m-2}{m-1}=2\binom{2m-3}{m-1}$ and $\ell=1$ or 2. Moreover, in view of (2.6), for every $A,B\in\mathcal{A}$ we must have $\binom{\ell}{|A\cap[\ell]|}\binom{n-\ell}{m-|A\cap[\ell]|}=\binom{\ell}{|B\cap[\ell]|}\binom{n-\ell}{m-|B\cap[\ell]|}=\binom{2m-2}{m-1}$, otherwise $\mathcal{A}$ is not optimal. If $\ell=1$ then $1\in A$ and $|A\cap[2,n]|=m-1$ for each $A\in\mathcal{A}$, thus $\mathcal{A}\subset\binom{[n]}{m}$. If $\ell=2$ then by the observation above $|A\cap[2]|=1$ for every $A\in\mathcal{A}$. Define $\mathcal{A}_i=\{A\in\mathcal{A}: i\in(A\cap[2])\}$ so that $\mathcal{A}=\mathcal{A}_1\cup\mathcal{A}_2$. Note that (in view of optimality of $\mathcal{A}$) $|\mathcal{A}_1|=|\mathcal{A}_2|=\binom{2m-3}{m-1}$. Now Sperner's Theorem implies that for every $A,B\in\mathcal{A}_i$ holds $|A\cap[2,n]|=|B\cap[2,n]|$. which means that $\mathcal{A}_i\subset\binom{n}{m-1}$ or $\binom{n}{m}$ (i=1,2). This completes the proof of part (ii). □

Remark 3. We already mentioned that $F(n,g)\geq G(n,g)$, note however that $F(n,g)\neq G(n,g)$. For example, observe that $F(n,n-1)\geq|\{0,1\}^{n-2}\times\{(0,0)\}|=2^{n-2}$, while clearly $G(n,n-1)=1$.

Our next goal is a sharpenning of Corollary 1.

Let $[n]=[\ell]\cup[\ell+1,n]$ and let $\mathcal{F}\subset 2^{[n]}$ satisfy property (P). Let us represent the elements of $2^{[n]}$ by pairs (F_1,F_2) where $F_1\subseteq[\ell],F_2\subseteq[\ell+1,n]$.

We say that $\mathcal{F}$ is a *homogeneous* family if $(F_1,F_2)\in\mathcal{F}$ implies $\{(E_1,E_2)\in 2^{[n]}: |E_1|=|F_1|,|E_2|=|F_2|\}\subseteq\mathcal{F}$.

Lemma 3. Given integers $1\leq\ell\leq n-1$ there exists an optimal homogeneous family $\mathcal{F}\subset 2^{[n]}$ satisfying property (P).

Proof. Let $\mathcal{A}\subset 2^{[n]}$ be an optimal family satisfying property (P). Let now σ be the set of maximal product chains defined in the proof of Lemma 2. For each $\mathcal{C}\in\sigma$ and $A=(A_1,A_2)\in\mathcal{A}$ define the family $\mathcal{F}_{\mathcal{C},A}=\{(F_1,F_2)\in 2^{[n]}: |F_1|=|A_1|,|E_2|=|A_2|\}$, if $A\in\mathcal{C}$ and $\mathcal{F}_{\mathcal{C},A}=\emptyset$ if $A\notin\mathcal{C}$. It is clear that $\mathcal{F}_{\mathcal{C},A}\cap\mathcal{F}_{\mathcal{C},B}=\emptyset$ for every distinct $A,B\in\mathcal{A}$ and $|\mathcal{F}_{\mathcal{C},A}|=\binom{\ell}{|A_1|}\binom{n-\ell}{|A_2|}$ for each nonempty family $\mathcal{F}_{\mathcal{C},A}$. Define also the family $\mathcal{F}_{\mathcal{C},\mathcal{A}}=\bigcup_{A\in\mathcal{A}}\mathcal{F}_{\mathcal{C},A}$. Observe that each $\mathcal{F}_{\mathcal{C},\mathcal{A}}$ is a homogeneous family satisfying

property (P). Indeed, if for some $A, B \in \mathcal{A}$ the family $\mathcal{F}_{\mathcal{C},A} \cup \mathcal{F}_{\mathcal{C},B}$ does not satisfy (P) then clearly property (P) is violated for A and B. Since for each $A \in \mathcal{A}$

there are exactly $|A_1|!(\ell - |A_1|)!|A_2|!(n - \ell - |A_2|)!$ members of σ containing A, it follows that $|\bigcup_{\mathcal{C}\in\sigma} \mathcal{F}_{\mathcal{C},A}| = \ell!(n-\ell)!$. Recall also that $|\sigma| = \ell!(n-\ell)!$. Therefore the average number of elements contained in a family $\mathcal{F}_{\mathcal{C},\mathcal{A}}$ is

$$\frac{1}{\ell!(n-\ell)!}\sum_{\mathcal{C}\in\sigma}|\mathcal{F}_{\mathcal{C},\mathcal{A}}| = \frac{1}{\ell!(n-\ell)!}\sum_{A\in\mathcal{A}}|\bigcup_{\mathcal{C}\in\sigma}\mathcal{F}_{\mathcal{C},A}| = |\mathcal{A}|.$$

Thus, there exists a homogeneous family $\mathcal{F}$ satisfying property (P) with $|\mathcal{F}| \geq |\mathcal{A}|$. $\square$

A homogeneous family $\mathcal{F}$ can be described by the set of pairs $I(\mathcal{F}) = \{(i,j) \in [0,\ell] \times [0, n-\ell] : \exists\ (F_1, F_2) \in \mathcal{F}$ such that $|F_1| = i, |F_2| = j\}$. Let us introduce a partial ordering: for $(i_1, j_1), (i_2, j_2) \in [0,\ell] \times [0, n-\ell]$ we write $(i_1, j_1) < (i_2, j_2)$ if $i_1 < j_1$ and $i_2 < j_2$.

Let $\mathcal{F} \subset 2^{[n]}$ satisfy property (P). For convenience of description later on we assume (w.l.o.g.) that $\emptyset \notin \mathcal{F}$. Let now $\mathcal{F} \subset \binom{n}{\leq m}$ be a homogeneous family satisfying property (P), so that $I(\mathcal{F}) \subset [\ell] \times [n-\ell]$. Then clearly every two pairs in $I(\mathcal{F})$ are comparabl e. That is $I(\mathcal{F}) = \{(i_1, j_1), \ldots, (i_k, j_k)\}$ for some $k \leq \frac{m}{2}$ where $(i_1, j_1) < \ldots < (i_k, j_k)$ and $i_k + j_k \leq m$.

Suppose now $\mathcal{F}$ is optimal. Note then that all pairs in $I(\mathcal{F})$ are consecutive, that is $(i_{r+1}, j_{r+1}) = (i_r + 1, j_r + 1), r = 1, \ldots, k-1$. Indeed, if, say, $i_{r+1} \geq i_r + 2$, then the replacement of (i_{r+1}, j_{r+1}) by $(i_{r+1} - 1, j_{r+1})$ or (i_r, j_r) by $(i_r + 1, j_r)$ in $I(\mathcal{F})$ gives us a larger family. Therefore we have the following

Corollary 2. For an optimal family $\mathcal{F} \subset \binom{[n]}{\leq m}$ satisfying property (P) we have

$$|\mathcal{F}| = \max \sum_{i=0}^{\min\{k-1, s-1\}} \binom{\ell}{k-i}\binom{n-\ell}{s-i}, \tag{2.16}$$

where the maximum is taken over all integers $1 \leq \ell \leq \frac{n}{2}$ and $k, s \geq 1$ with $k + s \leq m$.

Lemma 4. Given integers m, ℓ, n; $1 \leq \ell, m < \frac{n}{3}$ and $r := \min\{\lfloor \frac{m}{2} \rfloor, \lfloor \frac{\ell}{2} \rfloor\}$, let $\mathcal{F} \subset \binom{[n]}{\leq m}$ be an optimal family satisfying property (P). Then

$$|\mathcal{F}| = \max_{1\leq k\leq r} \sum_{i=0}^{k-1} \binom{\ell}{k-i}\binom{n-\ell}{m-k-i}. \tag{2.17}$$

Proof. By Lemma 3 we may assume that $\mathcal{F}$ is homogeneous and (in view of Corollary 2) $|\mathcal{F}| = \sum_{i=1}^{\min\{k-1,s-1\}} \binom{\ell}{k-i}\binom{n-\ell}{s-i}$ for some k and s. Note first that $k+s = m$, otherwise we can replace $(k-i, s-i)$ by $(k-i, s-i+1)$ or $(k-i, s-i)$ by $(k-i+1, s-i)$ in $I(\mathcal{F})$ getting a larger family. Thus, we have only to show that $k \leq \min\{\lfloor \frac{m}{2} \rfloor, \lfloor \frac{\ell}{2} \rfloor\}$. Suppose first that $k > \frac{\ell}{2}$. Then we replace $(k, m-k)$ and $(k-1, m-k-1)$ in $I(\mathcal{F})$ by $(k-1, m-k+1)$ obtaining a new set $I' \subset [\ell] \times [n-\ell]$. Note now that the new family $\mathcal{F}'$ associated with I' satisfies property (P). Moreover, we claim that $|\mathcal{F}'| > |\mathcal{F}|$, or equivalently

$$\binom{\ell}{k}\binom{n-\ell}{m-k} + \binom{\ell}{k-1}\binom{n-\ell}{m-k-1} < \binom{\ell}{k-1}\binom{n-\ell}{m-k+1}. \tag{2.18}$$

The LHS of (2.18) is not greater than $\binom{\ell}{k-1}(\binom{n-\ell}{m-k} + \binom{n-\ell}{m-k-1}) = \binom{\ell}{k-1}\binom{n-\ell+1}{m-k}$ and simple calculations show that $\binom{n-\ell+1}{m-k} < \binom{n-\ell}{m-k+1}$ for $n \geq 3m$. This contradiction implies that $k \leq \frac{\ell}{2}$.

Suppose now that $k > \frac{m}{2}$. Then we replace $(k-i, m-k-i)$ in $I(\mathcal{F})$ by $(m-k-i, k-i)$ $(i = 0, \dots, m-k-1)$ getting a larger family, a contradiction to the optimality of $\mathcal{F}$. $\square$

Let $P(n, \le m)$ denote the maximum size of a family $\mathcal{F} \subset \binom{[n]}{\le m}$ satisfying property (P). Clearly Lemma 4 implies that for $n \ge 3m$ $(1 \le \ell \le \frac{n}{2})$

$$G(n, \le m, g) \le P(n, \le m) = \max_{\substack{1 \le k \le \frac{m}{2} \\ 2k \le \ell \le \frac{n}{2}}} \sum_{i=0}^{k-1} \binom{\ell}{k-i} \binom{n-\ell}{m-k-i}. \tag{2.19}$$

Theorem 5. For integers $m \ge 8, n \ge m^2$, $t := \lfloor \frac{n}{m} \rfloor$ one has

$$G(n, \le m, g) = t \binom{n-t}{m-1}. \tag{2.20}$$

Proof. We need two simple lemmas.

Lemma 5.

$$\frac{\binom{t}{1}\binom{(m-1)t+r}{m-1}}{\max\limits_{\substack{1 \le \ell < n \\ 2 \le j \le m-2}} \binom{\ell}{j}\binom{n-\ell}{m-j}} > \frac{1}{2}\left(1 + \frac{m-2}{(m-1)(t-1)}\right)^{t-1}. \tag{2.21}$$

Proof. Let $n = tm+r$, $0 \le r \le m-1$, and let $\alpha(j) := \lfloor j(r+1)/m \rfloor$, $1 \le j \le m-1$. In view of (2.13)

$$\max_{\substack{1 \le \ell < n \\ 2 \le j \le m-2}} \binom{\ell}{j}\binom{n-\ell}{m-j} = \binom{2t+\alpha(2)}{2}\binom{(m-2)t+r-\alpha(2)}{m-2}, \tag{2.22}$$

where $\alpha(2) = \lfloor 2(r+1)/m \rfloor \in \{0, 1, 2\}$. Further we have
$\frac{\binom{t}{1}\binom{(m-1)t+r}{m-1}}{\binom{2t+\alpha(2)}{2}\binom{(m-2)t+r-\alpha(2)}{m-2}} = \frac{2t}{2t+\alpha-1} \cdot \frac{(m-1)t+r}{(m-1)(2t+\alpha(2))} \cdot \frac{(m-1)t+r-1}{(m-1)(t-1)+r} \cdot \frac{(m-1)t+r-2}{(m-1)(t-1)+r-1} \cdot \dots \cdot \frac{(m-2)t+r-\alpha(2)+1}{(k-2)(t-1)+r-\alpha(2)+1}$.
Simple calculation shows that for $\alpha(2) = 1, 2$ the product of the first three factors (in the RHS of the last equation) is not less than $1/2$. The same holds for the first two factors when $\alpha(2) = 0$. Therefore,
$\frac{\binom{t}{1}\binom{(m-1)t+r}{m-1}}{\binom{2t+\alpha(2)}{2}\binom{(m-2)t+r-\alpha(2)}{m-2}} > \frac{1}{2}(\frac{(m-1)t+r-1}{(m-1)(t-1)+r})^{t-1} = \frac{1}{2}(1 + \frac{m-2+r}{(m-1)(t-1)+r})^{t-1} \ge \frac{1}{2}(1 + \frac{m-2}{(m-1)(t-1)})^{t-1}$. $\square$

Corollary 3. Let us denote the RHS of (2.21) by $\varphi(t)$. Calculations show that $\varphi(t) > 1$ for $n \ge m^2$ (with $m \ge 6$). In particular for $m \ge 8$ we have $\varphi(t) > 1.22$.

Lemma 6. For integers $1 \le \ell \le n/2$, $1 \le k \le \frac{\ell}{2}$, $m \le n^{\frac{1}{2}}$ holds

$$\sum_{i=0}^{k-1} \binom{\ell}{k-i}\binom{n-\ell}{m-k-i} < \frac{n}{n-m}\binom{\ell}{k}\binom{n-\ell}{m-k}. \tag{2.23}$$

Proof. Let us show first that for $i = 0, \ldots, k-1$ holds

$$\frac{\binom{\ell}{k-i}\binom{n-\ell}{m-k-i}}{\binom{\ell}{k-i-1}\binom{n-\ell}{m-k-i-1}} > \frac{n}{m}. \tag{2.24}$$

One can easily verify that
$\frac{\binom{\ell}{k-i}\binom{n-\ell}{m-k-i}}{\binom{\ell}{k-i-1}\binom{n-\ell}{m-k-i-1}} = \frac{(\ell-k+i+1)(n-m-\ell+k+i+1)}{(k-i)(m-k-i)} \geq \frac{(k+1)(n-k-m+1)}{k(m-k)} > \frac{n-k-m+1}{m-k} > \frac{n}{m}$.
Now in view of (2.24) we infer

$$\sum_{i=0}^{k-1} \binom{\ell}{k-i}\binom{n-\ell}{m-k-i} < \binom{\ell}{k}\binom{n-\ell}{m-k}\left(1+\frac{m}{n}+\ldots+\frac{m^{k-1}}{n^{k-1}}\right) < \frac{n}{n-m}\binom{\ell}{k}\binom{n-\ell}{m-k}.$$

□

Let now $\mathcal{F} \subset \binom{[n]}{\leq m}$, $n \geq m^2$ be an optimal family satisfying property (P). In view of Lemmas 4 and 6, for some $1 \leq k \leq \frac{m}{2}$ we have

$$|\mathcal{F}| = \sum_{i=0}^{k-1} \binom{\ell}{k-i}\binom{n-\ell}{m-k-i} < \frac{n}{n-m}\binom{\ell}{k}\binom{n-\ell}{m-k}. \tag{2.25}$$

On the other hand for $k \geq 2$ Lemma 5 together with Corollary 3 implies

$$t\binom{n-t}{m-t} > 1.2\binom{\ell}{k}\binom{n-\ell}{m-k} > \frac{n}{n-m}\binom{\ell}{k}\binom{n-\ell}{m-k}. \tag{2.26}$$

This completes the proof of Theorem 5. □

Remark 4. We notice that Theorem 5 holds also for $4 \leq m \leq 7$ when $n \geq cm^2$ for some constant c (this follows directly from the proof of Theorem 5). Moreover by direct calculations one can show that it holds also for $n \geq m^2$.

We consider now the group security model without restriction on the size of a query set. We determine $G(n, g)$ within a constant factor less than $1/2$ for arbitrary parameters n and g.

Theorem 6. (i) *For $2 \leq g < n/2$ we have*

$$\frac{n+1}{n-1}\binom{n-1}{\frac{n-3}{2}} \leq G(n,g) < 2\binom{n-1}{\frac{n-3}{2}}, \quad \text{if } 2 \nmid n, \tag{2.27}$$

$$\frac{n+2}{2n-2}\binom{n}{\frac{n-2}{2}} \leq G(n,g) < \binom{n}{\frac{n-2}{2}}, \quad \text{if } 2 \mid n. \tag{2.28}$$

(ii) For $n/2 \leq g < n$ we have

$$\frac{g+1}{n}\binom{n}{g+1} < G(n,g) \leq \binom{n}{g+1}. \tag{2.29}$$

Proof. (i): Since $G(n,g) \geq G(n,m,g)$, Theorem 4 implies that for $m = \lfloor n/2 \rfloor$ and $g \leq m-1$ we have $G(n,g) \geq G(n,m,g) = 2\binom{n-2}{\lfloor \frac{n-2}{2} \rfloor}$. The latter equals $\frac{n+1}{n-1}\binom{n-1}{\frac{n-3}{2}}$, if $2 \nmid n$ and $\frac{n+2}{2n-2}\binom{n}{\frac{n-2}{2}}$, if $2 \mid n$, thus obtaining the lower bound.

For the upper bound we use the following result from [1].

Lemma 7 [1]. *Let $a_1, \ldots, a_n \in \mathbb{R} \setminus \{0\}$, $b \in \mathbb{R}$ and $|a_i| \neq |a_j|$ for some $i, j \in [1,n]$.*

Let Y be the (0,1)–solutions of the equation $a_1x_1 + \ldots + a_nx_n = b$.

Then

$$|Y| \leq \begin{cases} 2\binom{n-1}{\frac{n-3}{2}}, & \text{if } 2 \nmid n \\ \binom{n}{\frac{n-2}{2}}, & \text{if } 2 \mid n. \end{cases} \tag{2.30}$$

Let now $X \subset E(n)$ be the set of (0,1)-solutions of equation (2.1). Note then that for $g \geq 2$ there exist $i \in [\ell]$ and $j \in [\ell+1, n]$ such that $a_i \neq a_j$, for otherwise X contains a vector $(x_1, \ldots, x_n)$ of weight 2 with $x_i = x_j = 1$. This together with Lemma 2 completes the proof of case (i).

(ii): Let the ground set $[n]$ be partitioned into $[\ell] \cup [\ell+1, n]$ and let $\mathcal{A} \subseteq \binom{[n]}{\geq g+1}$, with $n/2 \leq g+1 \leq n$, be a family satisfying property (P). Now using the same argument as for Corollary 2 we get the following equality

$$|\mathcal{A}| = \max \sum_{i=0} \binom{\ell}{k+i}\binom{n-\ell}{g+1-k+i}, \tag{2.31}$$

where the maximum is taken over all $1 \leq \ell \leq \frac{n}{2}, 1 \leq k \leq g$.

Clearly the RHS of (2.31) is less than $\binom{n}{g+1}$. Since $G(n,g) \leq |\mathcal{A}|$ we get the simple upper bound $G(n,g) \leq \binom{n}{g+1}$.

For the lower bound in (2.29) note that $G(n,g) \geq G(n,g+1,g)$. Since $g+1 > \frac{n}{2}$ Theorem 4 implies that $G(n,g+1,g) = \binom{n-1}{g} = \frac{g+1}{n}\binom{n}{g+1}$. $\square$

We turn now to Problem 3. Let $\mathcal{A} \subset 2^{[n]}$ be a family of query sets avoiding g-group compromise under restrictions like in Problems 1,2. The only difference we have now is that $\dim(\text{span}\chi(\mathcal{A})) \leq k$ for given integers $1 \leq k \leq n-1$. Note that w.l.o.g. we may assume that $\dim(\text{span}\chi(\mathcal{A})) = k$ since every subspace $U \subset \mathbb{R}^n$ of dimension less than k can be embedded in a k-dimensional subspace V such that $U \cap E(n) = V \cap E(n)$.

It is not hard to see that $|\mathcal{A}|$ is upper bounded by 2^k (even if $\mathcal{A}$ is compromised). The following statement is based on that simple fact.

Proposition. *For integers $1 \leq k, g < n$ holds $G_k(n,g) = 2^k$ if and only if $n \geq k(g+1)$.*

Proof. Let us denote $X = \chi(\mathcal{A})$ and let $P = \{b_1, \ldots, b_k\}$ be a basis of a k-dimensional space $V \supseteq span(X)$. We assume that P is represented as row vectors of a $k \times n$ matrix. W.l.o.g. we may also assume that P has the echelon form $(I_k|M)$ (where I_k is the $k \times k$ identity matrix). It is clear that all linear combinations of P giving (0,1)-vectors must have (0,1)-coefficients which implies that $|X| \leq 2^k$. If now $|X| = 2^k$, then clearly each column-vector of M is either a unit vector or an all-zero vector. Note also that each row of P contains at least $g+1$ ones, otherwise $\mathcal{A}$ is g-compromised. This clearly implies that $n \geq k(g+1)$. $\square$

The next result is a generalization of Theorem 1.

Theorem 7. *(i) For integers* $\frac{n}{2} \le k < n$ *we have*

$$G_k(n,1) = \binom{2k-n+2}{\lfloor \frac{2k-n+2}{2} \rfloor} 2^{n-k-1} \tag{2.32}$$

(ii) An optimal set of SUM queries corresponds to the following set of vectors $X = X_1 \times X_2 \subset \{0,1\}^n$ *with* $X_1 := \{(x_1, \ldots, x_{2k-n+2}) : x_1 + \ldots + x_{\lfloor \frac{s}{2} \rfloor} - x_{\lfloor \frac{s}{2} \rfloor + 1} - \ldots - x_s = 0\}$ *and* $X_2 := \{00, 11\}^{n-k-1}$, *where* $s := 2n - k + 2$.

The optimal construction is unique, up to the permutations of the coordinates, if $2|n$. *If* $2 \nmid n$ *there is another optimal configuration with* $s = 2n - k + 3$ *and* $X_2 := \{00, 11\}^{n-k-2} \times \{0\}$.

Proof. Observe that $|X| = \binom{2k-n}{\lceil \frac{2k-n}{2} \rceil} 2^{n-k}$ and $\dim \operatorname{span}(X) = k$. Moreover $\operatorname{span}(X)$ has no vectors of weight less than 2. The upper bound directly follows from a result in [4], where it is proved that the maximum number of (0,1)-solutions of equation (1.2), where $r = n - k$ and B does not contain zero columns, is upper bounded by the RHS of (2.32). The description of all optimal constructions is also easily derived from that result. □

Note that in case $k = n - 1$ we have $G(n,1) = \binom{n}{\lfloor n/2 \rfloor}$ (Theorem 1).

Theorem 8. For $2g < n < (g+1)k$ we have

$$\frac{1}{2} G_k(n,1) < G_k(n,g) \le G_k(n,1). \tag{2.33}$$

Proof. Consider first the case $n \ge 2k$. Define the set $X = \{01, 10\}^{k-1} \times \{1^{n-2k+2}\}$. Clearly $|X| = 2^{k-1}$ and all vectors of X have weight $n - k + 1 > \frac{n}{2} > g$. Note also that $\dim \operatorname{span}(X) = k$. Finally observe that $\operatorname{span}(X) \cap E^n = X \cup \{0^n\}$, that is $\operatorname{span}(X)$ contains no other nonzero $(0,1)$-vectors besides those that are in X. Thus, the set of queries $\mathcal{A} \subset 2^{[n]}$ corresponding to vectors $X \cup \{0^n\}$ is not g-group compromised. Moreover $|\mathcal{A}| = 2^{k-1} + 1 > \frac{1}{2} G_k(n,1)$. Let now $n \le 2k$. Define the set $X = E_t^{2k-n} \times \{01, 10\}^{n-k}$, where $t = \lceil (2k-n)/2 \rceil$. Note that $|X| = \binom{2k-n}{\lceil \frac{2k-n}{2} \rceil} 2^{n-k}$ and $\dim \operatorname{span}(X) = k$. Note also that all vectors of X have weight $\lceil \frac{n}{2} \rceil$. Moreover, $\operatorname{span}(X) \cap E^n$ contains only vectors of weight 0 modulo $\lceil \frac{n}{2} \rceil$. This (together with Theorem 7) implies $G_k(n,g) \ge |X| = \binom{2k-n}{\lceil \frac{2k-n}{2} \rceil} 2^{n-k} > \frac{1}{2} \binom{2k-n+2}{\lfloor \frac{2k-n+2}{2} \rfloor} 2^{n-k-1} = \frac{1}{2} G_k(n,1)$. □

3. Concluding remarks. We considered combinatorial problems in connection with a security control mechanism in statistical databases under SUM query restrictions. We gave tight bounds for the maximum number of answerable queries without g-group compromise.

One of our objectives in this paper was to demonstrate how useful for applications is the subject called *Extremal Problems under Dimension Constraints* which was introduced in [1] and studied in a series of papers mentioned in [1]. Quite surprisingly, the results presented above are either direct consequences of results in those papers or can be easily derived using methods and tools developed in them.

It should be mentioned that the fact $\frac{1}{2} G(n,1) < G(n,g) < G(n,1)$ (a weaker form of Theorem 6) seems to be somewhat surprising. It shows that providing g–group security costs almost nothing as compared with the simplest case $g = 1$ (that

is just prevention of compromise)! The same we have for arbitrary dimension as is shown in Theorem 7.

Finding exactly $G(n,g)$ seems to be more difficult. The first open case is $g=2$. A good candidate for SUM query sets is the family $\mathcal{A}$ corresponding to the set of (0,1)–solutions of equation (1.1) where $a_1=\ldots=a_t=2$, $a_{t+1}=\ldots=a_n=-1$ with $t:=\lfloor n/3\rfloor$ (which can be shown to be superior to the lower bound in Theorem 6). Note that the (0,1)–solutions of this equation consist only of vectors of weight 0 modulo 3 and $|\mathcal{A}|=\sum_{i=0}^{t}\binom{t}{i}\binom{n-t}{2i}$. A similar construction seems to be "good" for $g=3$ (with $t:=\lfloor n/4\rfloor$, $a_1=\ldots=a_t=3$, $a_{t+1}=\ldots=a_n=-1$). We believe that these two constructions are optimal. In particular we have

Conjecture. *For an integer $n\geq 3$ and $t:=\lfloor n/3\rfloor$ holds*

$$G(n,2)=\sum_{i=0}^{t}\binom{t}{i}\binom{n-t}{2i}. \tag{3.1}$$

Another question is to clarify how sharp is the restriction $n\geq m^2$ in Theorem 5.

Problem 1 (as well as Problems 2,3) can be viewed as a coding problem: we seek for a largest binary code $C\subset E^n\subset\mathbb{R}^n$ of length n such that $(\mathrm{span}(C))\cap E^n$ does not contain (nonzero) vectors of Hamming weight less than $g+1$. Note however that unlike the classical error correcting codes (we assume w.l.o.g. that the code contains the zero vector), the minimum weight here does not equal the minimum Hamming distance, unless $g=1$. The restriction on minimum distance may lead to other interesting problems for further research.

REFERENCES

[1] R. Ahlswede, H. Aydinian, and L.H. Khachatrian, Extremal problems under dimension constraints, Discrete Mathematics, Special issue: EuroComb'01, Edited by J. Nesetril, M. Noy and O. Serra, vol. 273, no. 1-3, 9-21, 2003.

[2] R. Ahlswede, H. Aydinian, and L.H. Khachatrian, Forbidden (0,1)-vectors in hyperplanes of $\mathbb{R}^n$: the restricted case, Designs, Codes and Cryptogr., vol. 29, 17-28, 2003.

[3] R. Ahlswede, H. Aydinian, and L.H. Khachatrian, Forbidden (0,1)-vectors in hyperplanes of $\mathbb{R}^n$: the unrestricted case, Des., Codes and Cryptogr., vol. 37, 151-167, 2005.

[4] R. Ahlswede, H. Aydinian, and L.H. Khachatrian, Maximum number of constant weight vertices of the unit n–cube contained in a k–dimensional subspace, Paul Erdös and his mathematics, Combinatorica, vol. 23 (1), 5–22, 2003.

[5] I. Anderson, Combinatorics of Finite Sets, Clarendon Press, Oxford, 1987.

[6] L. Branković, P. Horak, and M. Miller, An optimization problem in statistical databases, SIAM J. Discrete Math., vol. 13, 346-353, 2000.

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

[8] J. Demetrovich, G.O.H. Katona, and D. Miklos, On the security of individual data, Lecture Notes in Comp. Sci., vol. 2942, 49-58, Springer 2004.

[9] D.E.R. Denning, Cryptology and Data Security, Addison-Wesley, Reading, MA, 1982.

[10] J. Domingo-Ferrer, Inference Control in Statistical Databases, Springer, Berlin, 2002.

[11] K. Engel, Sperner Theory, Cambridge University Press, 1997.

[12] J. R. Griggs, Database security and the distribution of subset sums in R^m. Graph theory and combinatorial biology (Balatonlelle, 1996), 223-252, Janos Bolyai Math. Soc., 1999.

[13] P. Horak, L. Branković, and M. Miller, A combinatorial problem in database security, Discrete Appl. Math., vol. 91, no. 1-3, 119-126, 1999.

[14] M. Miller, I. Roberts, and I. Simpson, Application of symmetric chains to an optimization problem in the security of statistical databases, Bull. ICA 2, 47-58, 1991.

[15] M. Miller, I. Roberts, and I. Simpson, Prevention of relative compromise in statistical databases using audit expert, Bull. ICA 10, 51-62, 1994.

[16] M. Miller and J. Seberry, Relative compromise of statistical databases, Austral. Computer J., vol. 21(2), 51-62, 1994.