draft date: August 17, 1998

Database Security and the Distribution of Subset Sums in $\mathbf{R}^m$

Jerrold R. Griggs[1]
Department of Mathematics
University of South Carolina
Columbia, SC 29208 USA
email: griggs@math.sc.edu

Abstract

We discuss applications of combinatorial arguments to database security: maximizing the "usability" of a statistical database under the control of the mechanism Audit Expert of Chin and Ozsoyoglu. As modelled by Mirka Miller *et al.*, the goal is to maximize the number of SUM queries from a database of real numbers without compromising it. Via linear algebra, direct connections emerge between such database query models and problems that concern maximimizing, over all choices of n nonzero elements $a_1, \ldots, a_n$ in $\mathbf{R}^m$, the number of the 2^n subset sums $\sum_{i \in I} a_i$, over all index sets I, belonging to some specified target set T. Well-known problems of this kind that arise, including the Littlewood-Offord and Erdős-Moser problems in number theory, have been successfully attacked using results and methods for partially ordered sets. We survey these connections and their extensions to higher dimensions. Fascinating new challenges have emerged with two new models of compromise that we introduce here.

Running head: Database Security and Subset Sums

[1] Research supported in part by grants NSA/MSP MDA904–95H1024 and NSF DMS–9701211.

Section 1. Introduction

A series of papers by Mirka Miller and a changing cast of co-authors describes a remarkable connection between maximizing the usability of a certain database mechanism and fundamental problems about concentrating subset sums for collections of numbers or vectors. The two principal models of database compromise in past studies are solved by surprising applications of the combinatorial theory of finite posets.

In this survey we offer streamlined proofs that simplify earlier approaches while giving more insight into the original database security problems. We also describe natural extensions of the associated subset sum problems, which are interesting for their own sake, to higher dimensions or to more general target sets. We introduce two new models of database compromise that lead to new subset sum problems meriting future study.

The origin of the topic is the analysis of the statistical database control mechanism called the AUDIT EXPERT devised by Chin and Ozsoyoglu [6]. Using it, one can ask for numerical information from a database by making a series of queries. An example that is particularly helpful is to imagine a mathematics department with n members such that for each member is recorded the name, date of birth, sex, rank, salary, and so on. Let us suppose that the salaries are confidential, but that one may ask for information about the salaries of selected groups of people, such as the maximum or average salary for an identifiable subgroup within the department. For instance, we can determine which people are over 60 years old and ask for the maximum salary of any of these people. The papers of M. Miller *et al.* present a more detailed example of this sort. What leads to particularly nice problems is to restrict attention to the case that the user can only ask SUM queries, in which the sum of the salaries of the specified people is returned.

We imagine then a database with n confidential records, which are real numbers $x_1, \ldots, x_n$. Let $[n] := \{1, \ldots, n\}$ and $2^{[n]} := \{I \subseteq [n]\}$. The user may specify any subset J of the index set $[n]$ and request the subset sum $\sum_{j \in J} x_j$. AUDIT EXPERT will keep track of which queries it has previously answered and decline to answer the next query if it would, together with the previous answers, lead to a compromise of the database. We shall consider several models of "compromise". Using only SUM queries, our analysis will not depend in any way on the actual values of the x_j's.

We generally consider AUDIT EXPERT running in the *static mode* in which it is decided in advance of the session which queries are answerable. In order to be as informative as possible, the goal is to maximize the number of answerable queries. That is, we seek families of subsets $\mathcal{S} \subseteq 2^{[n]}$ such that answering all queries $\sum_{j \in J} x_j, J \in \mathcal{S}$ does not compromise the database, with $|\mathcal{S}|$ as large as possible.

The most-studied model of compromise is this: We say that the database is *absolutely compromised* if some x_i can be determined. Of course, x_i need not be revealed

directly. For instance, $\mathcal{S}$ is absolutely compromised if $\{1,2\},\{1,3\},\{2,3\} \in \mathcal{S}$ since

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

Let B_n denote the Boolean lattice, which is the poset consisting of all 2^n subsets of the set $[n]$ ordered by inclusion. Miller, Roberts, and Simpson [23] converted the maximum usability problem here to a matrix problem, which in turn exposed a subset sum problem for real numbers, which they could solve using symmetric chain decompositions of the Boolean lattice B_n. We observed [12] that there is a close correspondence between their subset problem and the famous Littlewood-Offord problem of combinatorial number theory.

In Sections 2 and 3, we provide a careful description of the reduction of the absolute compromise problem to a matrix problem and then to a subset sum problem. An appropriate reduction permits a simple new solution to all of these problems via Sperner's Theorem. A new consequence of this approach is the determination of all extremal solutions for the absolute compromise model, which was announced independently by Branković, M. Miller, and Širáň [4].

Section 4 examines natural exensions of the subset sum results in Section 3, although their potential applicability to database security is not yet evident. In particular, we consider how to maximize the number of subset sums, for a set of n nonzero vectors in $\mathbf{R}^m$, that hit a target set of k elements in $\mathbf{R}^m$.

In Section 5, we survey results about using the AUDIT EXPERT in other modes, where the answerable queries are not decided in advance, but rather in response to the user's queries.

The second major model of compromise is analyzed in Section 6. We say the database is *relatively compromised* if either some record x_i or some difference of two records $x_i - x_j, i \neq j$ can be determined. Again, via a matrix description, we end up with a problem concerning the maximum concentration of subset sums $\sum_{i \in I} a_i$ at a target point, where now the nonzero numbers a_i must be distinct. This is the famous Erdős-Moser problem, which was solved, in large part, by using Sperner theory of ranked posets. We present a derivation of this solution for the database problem, based largely on its discovery by M. Miller *et al.* [24]. Unfortunately, we still cannot describe all optimal query sets $\mathcal{S}$. However, we shall see in Section 7 that one can describe asymptotically as $n \to \infty$ the maximum number of answerable queries, up to a constant factor, thanks to work on the Erdős-Moser problem. We also survey extensions of the Erdős-Moser problem to higher dimensions, including important results of Halász.

Two fundamental new models of compromise are put forth in the remainder of the paper. We first describe *g-group-compromise*, which means that some sum $\sum_{j \in J} x_j$ with $0 < |J| \leq g$ can be deduced from $\mathcal{S}$. This problem is translated via a matrix description to an apparently new problem on subset sums for a set of real numbers, a refinement of the Littlewood-Offord problem. Further work is needed on this topic.

Second, in Section 9 we describe a model of "internal security." We propose *h-inside-compromise*, which means that some linear combination $\sum_{j \in I} \alpha_j x_j$ with $0 <$

$|I| \leq h$ is revealed, where the α_j's are nonzero reals. The h-inside security problem is translated into a striking fundamental geometric question about concentrating subset sums for a collection of vectors in $\mathbf{R}^h$ in general position. Indeed, the security and general position subset sum problems turn out to have the same optimal values.

In Section 10, we describe two constructions of sets of vectors in two dimensions, which we learn in the following section are asymptotically optimal to within a constant factor. Upper bounds derived from Halász's analytical studies are presented. For general h, we provide an asymptotic upper bound on the maximum number of queries.

The survey concludes by reviewing the main directions for future study.

The close relationship between these natural (even practical) database security problems about query sums and fundamental mathematical problems is striking. What is also exciting is to see how, due to the combined insights of many researchers, methods from linear algebra, number theory, geometry, Sperner theory, probability, and analysis can be brought to bear on these problems.

Section 2. Basis Matrix

Throughout the paper $[n]$ represents the set $\{1, \ldots, n\}$. A SUM query from our database has the form $\sum_{j \in S} x_j$, where the index set $S \subseteq [n]$. Given a collection of distinct SUM queries $\mathcal{S} := \{S_1, \ldots, S_m\}$, where each $S_i \subseteq [n]$, we can determine any *weighted* query they generate by taking linear combinations of these sums. Now for each SUM query S_i there corresponds naturally the $0-1$ row vector $[s_{i1}, \ldots, s_{in}] \in \mathbf{R}^n$. Then from the collection of SUM results for $\mathcal{S}$, we can determine all sums of the form $\sum_{j=1}^n a_j x_j$, $a_j \in \mathbf{R}$ for all j, where the row vector of coefficients, $[a_1, \ldots, a_n] \in \mathbf{R}^n$, is any linear combination of the SUM query coefficient vectors,

$$[a_1, \ldots, a_n] = \sum_{i=1}^{m} \lambda_i [s_{i1}, \ldots, s_{in}],$$

where the λ_i's are arbitrary real numbers.

The set of row vectors $[a_1, \ldots, a_n]$ for which we are certain to know the weighted sum is then described as the row space $R(\mathcal{S})$ of the $m \times n$ matrix $A(\mathcal{S}) = [s_{ij}]$. Since SUM queries correspond to $0-1$ vectors in $\mathbf{R}^n$, we seek to maximize the number of $0-1$ vectors in the space $R(\mathcal{S})$. Avoiding absolute compromise means that none of the standard basis vectors $e_j = [0, \ldots, 1, 0, \ldots, 0]$, with 1 in position j, is in $R(\mathcal{S})$.

Let d denote the dimension of $R(\mathcal{S})$ as a subspace of $\mathbf{R}^n$. Clearly, $d < n$ or there would be compromise. Let $B = B(\mathcal{S})$ be a basis matrix for $R(\mathcal{S})$ obtained by elementary row operations (Gaussian elimination) on $A(\mathcal{S})$. Then $B(\mathcal{S})$ is a $d \times n$ matrix whose rows form a basis for $R(\mathcal{S})$. Also, d of its columns are the columns of the identity matrix I_d. If any row of B is a standard basis vector (all zero except the

single one), then $\mathcal{S}$ produced an absolute compromise. Otherwise, it is clear that no e_j belongs to $R(\mathcal{S})$, and there is no compromise.

Let us continue in the case that there is no compromise.

Claim. *If $d < n-1$, then we may add $n-1-d$ independent rows to B while still avoiding absolute compromise. Equivalently, there is an $(n-1) \times n$ matrix with independent rows (rank $n-1$ over $\mathbf{R}$) which has a row space that contains all query vectors $[s_{i1}, \ldots, s_{in}]$ but none of the standard basis vectors e_j, which are forbidden.*

Proof. Here is one argument to prove this claim. It suffices to show how to add one row to $B = B(\mathcal{S})$; the operation can be repeated until there are $n-1$ rows: Let us assume that the first d columns of B are are linearly independent, after reordering the columns, if necessary. Let's add the row given by the vector $e_{d+1} + \alpha e_n$. For each row i of the matrix, only one possible value of α must be avoided to keep e_i out of the row space. At most $d+1$ values of α must be avoided altogether, and any other value is okay. We may select α to be rational, or even integer. Continuing in this way, we can go until the matrix has $n-1$ independent rows. When we started the procedure, the matrix $B(\mathcal{S})$, obtained by Gaussian elimination from a $0-1$ matrix, had rational entries. So when we conclude the procedure, with $n-1$ rows, we may assume all the entries are rational.

A second approach to verifying the claim that we can add rows to the basis matrix B works as follows: Let $r_1, \ldots, r_d$ be the rows of $B(\mathcal{S})$. We claim there is another row vector $v \notin R(\mathcal{S})$ such that, for all j,

$$e_j \notin \langle r_1, \ldots, r_d, v \rangle,$$

where $\langle w, x, \ldots \rangle$ denotes the span of vectors $w, x, \ldots$ (over $\mathbf{R}$). Such v avoids creating one of the forbidden standard basis vectors that would compromise the database. Equivalently,

$$v \notin \langle r_1, \ldots, r_d, e_j \rangle, 1 \leq j \leq n,$$

so that v avoids the finite union of hyperplanes (or smaller subspaces when $d < n-2$). Such v clearly exists, even with all rational entries. ∎

This second method of reasoning works well with other models of compromise we present later in the paper.

After further row reduction, we see that for any set $\mathcal{S}$ of queries avoiding compromise, we can produce (after possibly permuting columns) an $(n-1) \times n$ *basis matrix*

$$M = M(\mathcal{S}) = \begin{pmatrix} 1 & 0 & 0 & \ldots & 0 & a_1 \\ 0 & 1 & 0 & \ldots & 0 & a_2 \\ 0 & 0 & 1 & \ldots & 0 & a_3 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & 1 & a_{n-1} \end{pmatrix}. \tag{1}$$

The row space of $M(\mathcal{S})$ includes the vectors for all queries $S_j \in \mathcal{S}$, but excludes the standard basis vectors e_j. In particular, the entries $a_1, \ldots, a_{n-1}$ in the last column are all *nonzero* reals.

It follows that any set of $n-1$ columns of $M(\mathcal{S})$ is linearly independent. So regardless of what order the columns were originally, we can reach a basis matrix of the form (1) by suitable row operations.

While we may construct such a matrix for given $\mathcal{S}$ in which all a_i are rational, one can construct examples of query collections $\mathcal{S}$ such that no matter how the columns are ordered the basis matrix $M(\mathcal{S})$ will contain some non-integer entries.

To maximize the number of sum queries we can answer without absolute compromise, we see that it is enough to maximize the number of $0-1$ vectors in the row space of the basis matrix, which can be any matrix of form (1). Given nonzero reals a_i, such vectors are obtained as the sum of some subset of the set of rows of $M(\mathcal{S})$ in which the sum of corresponding entries in column n is 0 or 1. The empty query, for $S = \emptyset$, is one such query. Our problem is to choose M to maximize the number of such sums.

Section 3. Absolute Compromise and the Littlewood-Offord Problem

The problem of maximizing the number of answerable queries without absolute compromise has now been translated into one about concentrating the subset sums for a collection of real numbers: We seek to choose nonzero reals $a_1, \ldots, a_{n-1}$ such that we maximize the number of the 2^{n-1} subset sums

$$\sum_{i \in I} a_i = 0 \quad \text{or} \quad 1,$$

as I ranges over all subsets of $[n-1]$.

Using symmetric chain decompositions, Miller, Roberts, and Simpson [23] determined that this maximum is $\binom{n}{\lfloor \frac{n}{2} \rfloor}$. It can be achieved by selecting the a_i, $1 \le i \le n-1$, to be

$$1, -1, 1, -1, 1, \ldots \qquad \text{or} \qquad 1, 1, -1, 1, -1, 1, -1, \ldots.$$

For odd n, these coincide. Subsequently, Kevin Miller and Sarvate [22] proved the uniqueness of these maximum solutions to this real number problem, provided the a_i's are assumed to be integers. We shall see that the uniqueness extends to all reals a_i.

In [12] we discussed the close connection between this real number problem and a famous problem of Littlewood and Offord [21] concerning the concentration of subset sums for complex numbers.

The Littlewood-Offord Problem. *How does one select numbers $a_1, \ldots, a_n \in \mathbf{C}$, not necessarily distinct, with $|a_i| \ge 1$ for all i, and an open unit diameter ball B, so as to maximize the number of the 2^n sums $\sum_{i \in I} a_i$, $I \subseteq [n]$, lying inside B?*

A good strategy is to take all $a_i = 1$ and centering the ball B at $\lfloor n/2 \rfloor$ to get $\binom{n}{\lfloor \frac{n}{2} \rfloor}$ sums in B. This turns out to be optimal.

Erdős [7] solved the Littlewood-Offord problem under the restriction that the a_i's be real. He did this by exploiting a nice connection he found to the theory of ordered sets via a simple trick that applies as well to our problem: Given any collection of nonzero numbers $a_1, \ldots, a_n \in \mathbf{R}$, not necessarily distinct, observe that replacing any one of them, say a_1, by its negative, $-a_1$, merely translates the complete constellation of 2^n subset sums $\sum_{i \in I} a_i$, $I \subseteq [n]$, but does not in any way change their relative position. (To see this, notice that the sums $\sum_{i \in I} a_i$ with $1 \in I$ translated by $-a_1$ are the sums $\sum_{i \in I \setminus \{1\}} a_i$, while the sums $\sum_{i \in I} a_i$ with $1 \notin I$ translated by $-a_1$ are the sums $\sum_{i \in I \cup \{1\}} a_i$ except a_1 is replaced by $-a_1$.) Thus, to maximize the number of subset sums inside any unit diameter ball B, it suffices to consider the maximum over all *positive* reals a_i, which means all $a_i \geq 1$, and all balls B.

For any such a_i's, and for any index sets I, J with $I \subset J \subseteq [n]$, we have

$$\left| \sum_{i \in J} a_i - \sum_{i \in I} a_i \right| = \sum_{i \in J \setminus I} a_i \geq |J \setminus I| \geq 1.$$

Thus, not both $\sum_{i \in J} a_i$ and $\sum_{i \in I} a_i$ belong to any ball B.

Hence, the collection of subsets

$$\{I \subseteq [n] : \sum_{i \in I} a_i \in B\}$$

is an antichain in the Boolean lattice B_n of all subsets of $[n]$, ordered by inclusion. (An *antichain* in a poset $P = (P, \leq)$ is a subset A of P such that no two of its elements $p \neq q \in A$ satisfy $p \leq q$.) As Erdős observed, Sperner's Theorem [30] applies. It says that $\binom{n}{\lfloor \frac{n}{2} \rfloor}$ is the largest size of any antichain in B_n.

For our real number problem, one can introduce an nth number, denote it by a_n, which we shall set to -1. Then the number of subsets $I \subseteq [n-1]$ with $\sum_{i \in I} a_i = 0$ or 1 is the same as the number of subsets $J \subseteq [n]$ such that $\sum_{i \in J} a_i = 0$. We then consider the broader problem of maximizing, over all choices of nonzero reals $a_1, \ldots, a_n$ and all target sums t, the number of subsets $I \subseteq [n]$ such that $\sum_{i \in I} a_i = t$. We no longer have the lower bound $|a_i| \geq 1$ from the Littlewood-Offord problem, since we only demand that $a_i \neq 0$. However, since our target set is just a point now, instead of a ball, the same arguments apply, and we get that the maximum is again $\binom{n}{\lfloor \frac{n}{2} \rfloor}$. This can be achieved with $t = 0$ and $a_n = -1$ when we pick $a_i = (-1)^i$ for $1 \leq i \leq n-1$.

Further, for our real number problem, our approach yields all maximum solutions. Sperner described all maximum-sized antichains in B_n: For all n, one can take the collection of all subsets of size $\lfloor \frac{n}{2} \rfloor$, or for odd n, another choice is the collection of

all subsets of size $\lceil \frac{n}{2} \rceil$. These solutions force all a_i to be equal up to sign in our real number problem. For a target value $t = 0$, the only choices of $a_1, \ldots, a_{n-1}$ to go with $a_n = -1$ that achieve the maximum are, up to reordering the a_i's, the one (resp., two) choices for odd (resp., even) n described earlier.

Determining all maximum solutions for the basis matrix problem solves a problem of Mirka Miller, Roberts, and Simpson [23, p.57]. A paper by Kevin Miller and Sarvate [22] proved that these are the only choices for the a_i's in $M(\mathcal{S})$ allowing the maximum number of queries, but only in the case that the a_i's are restricted to being integer. Further, their arguments, using the symmetric chain decomposition, are more elaborate than ours. However, their success in the integer case provided the inspiration for us to continue the project.

We originally solved the maximum solution problem for the basis matrix (or the equivalent real number version) in our earlier paper [12]. Our new proof above here, which uses the trick of introducing $a_n = -1$ (an idea also used in [4]), is actually shorter and relies only on Sperner's original theorem, whereas the previous proof needed the essential uniqueness of extremal *k-families* in B_n. On the other hand, the earlier proof extended to more general problems.

Different proofs of this matrix result were developed independently by Branković and Miller, using symmetric chain decompositions in one instance and Lubell-type inequalities for the other [2, 22], but they are not as simple as the proof here.

Looking at what the matrix interpretation for the maximum solution(s) tells us about the original Absolute Compromise Problem, we find that there is a simple description of all maximum-sized sets of SUM queries.

Theorem 3.1. *Let $\mathcal{S} = \{S_1, \ldots, S_m\} \subseteq 2^{[n]}$ be a collection of index sets of distinct SUM queries that can be answered without absolute compromise from a database of n real entries x_i. Then $m \leq \binom{n}{\lfloor \frac{n}{2} \rfloor}$. Equality is achieved for all n precisely when the set of entries is partitioned into two parts, of sizes $\lfloor \frac{n}{2} \rfloor$ and $\lceil \frac{n}{2} \rceil$, and $\mathcal{S}$ consists of all queries with an equal number of elements from each part.*

Proof. The discussion above reduced the problem of maximizing the number of SUM queries avoiding absolute compromise to that of considering the entries $a_1, \ldots, a_{n-1}$ in the last column of a basis matrix $M(\mathcal{S})$. For an optimal solution of this problem, let I consist of indices i such that $a_i = 1$, and let $\bar{I} = [n] \backslash I$. The allowable queries for the solution correspond naturally to the $0-1$ characteristic vectors that indicate which entries x_i are to be added. Each is in the row space of M. We observe that each vector in the row space has half of its entries indexed in I and half indexed in $\bar{I}$: If the vector has last component 0, then it sums as many 1's as -1's (but leaves out x_n), while if it is 1, then it sums one more term in I than in $[n-1] \backslash I$, but the 1 indicates that x_n is also to be included, and $n \in \bar{I}$. ∎

To better understand how the collection $\mathcal{S}$ in the theorem works, consider the example where the database entries are the salaries of the members of a particular mathematics department. Suppose the department is evenly split between two factions,

good guys and bad guys. To maximize the number of answerable queries, only those sums will be given that involve the salaries of exactly as many good guys as bad guys. That this does not compromise the database can be seen by observing that if all bad guys are given a raise in salary of D dollars, while all good guys are given a cut of D dollars, there is no change to any of the answerable queries, and no change is detected. Note that this works even when n is odd, and there is an "extra" person in the department, good or bad.

Independently of our working out the theorem above, an equivalent result has been announced by M. Miller *et al.* [5].

M. Miller *et al.* define the *usability* U of a database to be the ratio of the number of answerable queries avoiding compromise to the total number of possible queries, 2^n for a database of n records. It is important to consider the maximum usability asymptotically as $n \to \infty$, which we can do with Stirling's formula.

Corollary 3.2. [23] *The maximum usability of a database of n records that avoids absolute compromise is $U = \binom{n}{\lfloor \frac{n}{2} \rfloor}/2^n = \Theta(n^{-1/2})$ as $n \to \infty$.* ∎

Section 4. Extensions to Higher Dimensions

The original Littlewood-Offord problem allowed complex numbers a_i, and Erdős's solution for real a_i could not be extended to two dimensions, though asymptotic results suggested that $\binom{n}{\lfloor \frac{n}{2} \rfloor}$ is still the answer in the complex case. It was not until 20 years later that Katona and Kleitman independently proved it [16,17]. Then in 1970 Kleitman [18] showed that in general dimension $\mathbf{R}^m$ the maximum number of subset sums $\sum_{i \in I} a_i$ inside an open unit diameter ball remains $\binom{n}{\lfloor \frac{n}{2} \rfloor}$, over all $a_1, \ldots, a_n \in \mathbf{R}^m$ with $|a_i| \geq 1$ and over all such balls.

The methods devised to obtain these results–all related to Sperner-theoretic arguments about the maximum size of a family of subsets of $[n]$ satisfying some condition–can be adapted to the concentration of subset sums in $\mathbf{R}^m$ at a target set of one or several points.

In our situation, we can even describe all extremal solutions, which is not really possible when the target is a ball instead of a finite collection of points. Obtaining these extremal solutions depends on our using the appropriate tools, since although Kleitman's proof in $\mathbf{R}^m$ gives the correct bound, it does not reveal much information about the extremal solutions themselves.

Theorem 4.1. [12] *Let $a_1, \ldots, a_n \in \mathbf{R}^m \backslash \{0\}$. Let $T = \{x_1, \ldots, x_k\} \subseteq \mathbf{R}^m$, $k \leq n+1$. The number of sums $\sum_{i \in I} a_i \in T$ is maximum, the sum of the k middle binomial coefficients in n, if and only if each $a_i = a_1$ or $-a_1$, and, letting $\lambda = |\{i : a_i = a_1\}|$, the set T contains k middle points in the sequence*

$$\{(\lambda - n)a_1, (\lambda - n + 1)a_1, \ldots, \lambda a_1\}.$$

The proof relies on the essential uniqueness of the maximum-sized k-families (union of k antichains) in the Boolean lattice B_n.

Section 5. Complexity of Dynamic Mode Use of Audit Expert

Thus far, we have only considered Audit Expert used in the non-adaptive "static mode" in which it is decided in advance which queries are answerable. We determined how to do this so as to maximize the number of answerable queries.

An adaptive set-up, called "dynamic mode", answers queries successively one-by-one in a session, but refuses to answer any query that when combined with previously answered ones leads to absolute compromise. At any given time in the session, a basis matrix can be stored describing what is already known (but the columns may not be in order as in (1)). When another query is made, the corresponding row vector is attached to the matrix, Gaussian elimination is applied, and we can check whether

(a) the new row is dependent on the old ones (then answer the query but drop the new row from the basis matrix),

(b) it is independent of the others and leads to a compromise (then drop the row, but don't answer the query), or

(c) it is independent but does not lead to compromise (answer it and retain the row).

This can be done efficiently.

But compromise under "dynamic mode" can happen quickly: any collection of n linearly independent vectors span $\mathbf{R}^n$, so compromise the database. In fact, M. Miller *et al.* [23] observed that as few as n queries could be answered: If $a_i = 1$ for all i in $M(\mathcal{S})$ this happens, *i.e.*, if one answers the queries $x_i + x_n$, $1 \leq i \leq n-1$, then no further query can be answered, other than the empty one, without compromising the database! The usability in this instance is merely $n/2^n$.

Branković, Miller, and Širáň [4] proposed a "hybrid mode" as a way to react dynamically while maintaining high usability. The user's question will be answered whenever the answered queries belong to some set of answerable queries of maximum usability. In view of Theorem 3.1, we must check whether there is a 2-coloring of the index set $[n]$, which is *equitable* in the sense that the color classes are used the same number of times within one, such that every answered query involves an equal number of indices of each color.

The natural decision problem suggested now is this:

Problem AE

Instance: A set $[n]$ and a collection $\mathcal{S} \subseteq 2^{[n]}$.

Question: Is there an equitable 2-coloring of $[n]$ into sets R, B such that for every $S \in \mathcal{S}$, $|S \cap R| = |S \cap B|$?

Theorem 5.1. [4] *Problem AE is NP-complete.*

Branković *et al.* prove this result by a transformation from the problem 3SAT. Our description of the problem is simpler, though equivalent; in [5] they have essentially arrived (independently) at the same equitable 2-coloring idea.

Of course, if $\mathcal{S}$ contains any subset (query) of odd size, there is no such 2-coloring, so there is no way to achieve maximum usability.

In [4] the authors go on to consider the restriction of this equitable 2-coloring problem to the case that every subset (query) in $\mathcal{S}$ has size 2. They show the problem is polynomially solvable in this case. We give the following self-contained explanation: The subsets can be viewed as edges of a graph with vertex set $[n]$. We are then asking whether there is an equitable 2-coloring of the vertices where each edge meets both color classes. Of course, this is impossible when there is an odd cycle, since no 2-coloring at all is possible.

If there is no odd cycle, then the coloring is forced except in every component one can interchange the two color classes. Let us say that there are c components in this graph, and the absolute value of the difference in the sizes of the two color classes in the ith component is d_i, $1 \leq i \leq c$. The problem now is whether there exists $I \subseteq [c]$ such that $\sum_{i \in I} d_i = \sum_{i \notin I} d_i$. Starting at $j = 1$ we keep track of all possible values of $\sum_{i=1}^{j} \varepsilon_i d_i$, where each ε_i is 1 or -1. An update to $j+1$ simply requires working out $s + d_{j+1}$ and $s - d_{j+1}$ for each sum s achieved with d_1 through d_j. Since we started with n vertices, the possible values of s are the integers in $[-n, n]$. At the end, when we treat d_c, we must only check whether 0 occurs as one of the sums, to decide whether there is a coloring as desired. Further, we can obtain such a coloring explicitly if we store a sign for each s at each stage $j+1$ to indicate how we got there from the previous value j.

Section 6. Relative Compromise and the Erdős-Moser Problem

There is a second model of compromise that further restricts the allowable sets of queries and again has interesting connections to number theory and to ordered sets. This was described in another paper by M. Miller *et al.* [24, cf. 25]. We hope that discussing it here, building on our approach to the absolute compromise problem, will lead to a better understanding of it.

In terms of the salary model, we now want to prevent not only the revelation of some individual's salary, but also it cannot be determined how much more someone makes than someone else. To be precise, in terms of the database, we say there is a *relative compromise* if our queries determine either some entry x_i or some difference of entries $x_i - x_j, i \neq j$.

The arguments leading to the basis matrix for absolute compromise adapt easily to relative compromise. Again we are avoiding just a finite set of vectors in the row space $R(\mathcal{S})$, as it cannot contain any standard basis vector e_i or any difference $e_i - e_j, i \neq j$. Every compromise-avoiding collection $\mathcal{S}$ is spanned by a $(n-1) \times n$ basis matrix

$M(\mathcal{S})$ of form (1) as before, with rational entries. The conditions on the last column are now more restrictive: Each $a_i \neq 0$ (to avoid e_i), $a_i \neq -1$ (to avoid $e_i - e_n$), and $a_i \neq a_j, i < j < n$ (to avoid $e_i - e_j$). Again, any $n-1$ columns are linearly independent, so we are not giving up anything by putting the identity matrix I_{n-1} in the first $n-1$ columns. We now face a number theory problem: We seek to select a_i's subject to these conditions to maximize the number of subset sums $\sum_{i \in I} a_i, I \subseteq [n-1]$ that equal 0 or 1.

A trick similar to the absolute compromise problem simplifies matters here. Artificially introduce a term $a_n = -1$ and look only for subset sums equal to 0. That is, it is equivalent to ask for distinct nonzero reals $a_1, \ldots, a_n$ to maximize the number of subset sums $\sum_{i \in I} a_i, I \subseteq [n]$ equal to 0. Note that for any solution to this last problem, we need only divide all a_i by the same quantity, $-a_n$, to get another solution in which $a_n = -1$ that we can use to give us $a_1, \ldots, a_{n-1}$ for our matrix problem.

What changed from the case of absolute compromise is that we now require *distinct* numbers a_i. In fact, Erdős and Moser [8] posed a general version of this natural question over 30 years ago.

The Erdős-Moser Problem. *How does one select distinct nonzero reals $a_1, \ldots, a_n$ and a target sum t to maximize the number of subset sums $= t$?*

A wise choice, for it turns out to be an optimal one, is to select n integers closest to 0 and target $t = 0$. It is fortunate that optimality can be achieved with $t = 0$, since our problem required this. Before describing the surprising link between this number theory problem and ordered sets, we record what this means for the database problem.

Theorem 6.1. [24] *Let $\mathcal{S}$ be the index set of SUM queries from a database of n entries that avoids relative compromise. Then $|\mathcal{S}|$ is maximized if it corresponds to the row space of the the matrix $M(\mathcal{S})$ with values in the last column given by*

$$1, 2, -2, 3, -3, 4, \ldots.$$

The pattern for the last column begins irregularly: It omits -1 since that had to be reserved for a_n. We cannot think of as nice a description of $\mathcal{S}$ as we did for the absolute compromise problem in Theorem 3.1, so we are forced to rely on the description in terms of $M(\mathcal{S})$. We also do not believe anyone has been able to describe all solutions to the number theory problem; we are able only to describe some of them with this theorem. (We discuss the asymptotics in the following section.)

To see the connection to ordered sets, it is best to consider first the variant of the Erdős-Moser problem that asks for n distinct *positive* real numbers a_i and a target t to maximize the number of subset sums $\sum_{i \in I} a_i = t$. For instance, a wise choice is to take $a_i = i$ and $t = \lfloor (1 + 2 + \cdots + n)/2 \rfloor$.

Lindström [20] observed that for given positive a_i's, say $0 < a_1 < \cdots < a_n$, and target t, the collection

$$\{I \subseteq [n] : \sum_{i \in I} a_i = t\}$$

contains no two sets I, J where J is obtained from I by some combination of inserting elements and/or replacing elements by larger ones. That is, this collection of subsets is an antichain, not merely in the Boolean lattice, but in the more fully ordered poset, we denote by $M(n) = (2^{[n]}, \leq)$, on the subsets of $[n]$ ordered as follows: We have $I \leq J$ whenever we insert elements into I or we replace elements by larger ones. For instance, in $M(5)$ we have $\{2,4\} \leq \{1,3,5\}$. This poset $M(n)$ is ranked, since an element $\{i_1, \ldots, i_r\}$ has rank $\sum_j i_j$. The rank of the whole poset is $(1 + 2 + \cdots + n)$. It is a self-dual poset.

For the "wise choice" $a_i = i$ above, setting $t = \lfloor (1 + 2 + \cdots + n)/2 \rfloor$ gives all elements I in the middle rank t of $M(n)$ as the sums $\sum_{i \in I} a_i = t$. Lindström [20] observed that this choice must be optimal, provided one can prove that the middle rank set in $M(n)$ is an antichain of maximum size.

Stanley proved this is indeed true (and much more) [31, cf. 32] when he developed machinery to construct nice decompositions of various ranked posets into chains. Specifically, he showed that an assortment of posets have collections of symmetric chains as follows: Letting n denote the rank of such a poset P, there exists for each $i < n/2$ a collection of disjoint chains that meet each of the ranks from i through $n - i$ and which cover ranks i and $n - i$. In particular, the chains each meet the middle rank(s). It can be deduced that the middle rank of such posets P is an antichain of maximum size. The poset $M(n)$ turned out to be one of the posets Stanley came up with, and Harper noted the connection to Lindström's work.

The property that no antichain is larger than the largest rank is called the *Sperner property*, a term motivated by Sperner's Theorem, which implies that the Boolean lattice B_n has this property. The author [11] noted that Stanley's chain partition implies stronger Sperner-type properties for his posets. What has come to be known as the *Peck property* holds if a ranked poset P has a symmetric, unimodal sequence of rank sizes, and for all k, selecting the union of k middle ranks in P gives a maximum-sized union of k antichains. Stanley's chain partition shows in particular that $M(n)$ is Peck.

This is important for solving the Erdős-Moser problem. For let us consider a general set of distinct nonzero reals a_i, with, say, l positive ones and $n - l$ negative ones, which we may denote by

$$b_l > \cdots > b_1 > 0 \qquad \text{and} \qquad 0 > c_1 > \cdots > c_{n-l}.$$

For any index sets $J \subseteq [l]$ and $K \subseteq [n - l]$, the corresponding subset sum

$$\sum_{j \in J} b_j + \sum_{k \in K} c_k$$

increases if J goes up in order $M(l)$ or K goes down in order $M(n-l)$. For any target sum t, the collection of index sets

$$\{I \subseteq [n] : \sum_{i \in I} a_i = t\},$$

which corresponds to

$$\{(J,K) : J \subseteq [l], K \subseteq [n-l], \qquad \sum_{j \in J} b_j + \sum_{k \in K} c_k = t\},$$

forms an antichain in the product poset $M(l) \times M(n-l)^D$, where P^D denotes the order dual of poset P. As Stanley observed [31], this poset must also be Peck (by the Peck product theorem, for instance), hence also Sperner, and its middle level is maximum-sized. So one can do no better than to select $b_j = j, c_k = -k$ for all j, k.

But what value of l is the best for the Erdős-Moser problem? It is easily checked that the rank-generating function for $M(l)$ is

$$(1+q)(1+q^2)\cdots(1+q^l),$$

where the coefficient of q^i in the expansion is the number of elements of rank i. The product of the functions for $M(l)$ and $M(n-l)$ is what we need, and it can be shown that the maximum middle coefficient is achieved by taking $l = \lfloor n/2 \rfloor$. In this way, the Erdős-Moser problem is solved, and we obtain Theorem 6.1. Peck [27] describes a proof of Erdős-Moser from the positive number version.

We note that Stanley's original chain construction arguments relied on results from algebraic geometry. For the case of $M(n)$, Proctor showed that linear algebra arguments are sufficient to derive the Peck property, although some choices in his proof are motivated by insights from Lie algebras [28]. We are not aware of any purely elementary proofs that $M(n)$ is Peck.

Section 7. Asymptotics and Extensions for Relative Compromise

Now we wish to determine the asymptotic behavior of the maximum number of queries avoiding relative compromise. We now know how to exactly achieve the maximum number of queries for a database with n records without a relative compromise. But what is this maximum asymptotically, in comparison to the $C2^n n^{-1/2}$ behavior for absolute compromise? For the Erdős-Moser problem, Sárközy and Szemerédi [29] already worked out the asymptotics, up to a constant factor, some ten years before it was actually solved. Applied to relative compromise we obtain the following bounds.

Theorem 7.1. *Let $r(n)$ denote the maximum size of a set $\mathcal{S}$ of queries from a database of n records that avoids relative compromise. There exist constants $C, C' > 0$ such that for all n,*

$$C2^n n^{-3/2} < r(n) < C'2^n n^{-3/2}.$$

Although Sárközy and Szemerédi only considered sets of n distinct *positive* real numbers a_i, it is simple to extend their result (up to a constant factor) to the Erdős-Moser problem with distinct nonzero reals. The clever indirect proof of the Sárközy-Szemerédi upper bound relies in part on Sperner's Theorem, incidentally. For their

lower bound, they only say that it is "*leicht zu sehen*" (easy to see). It will be instructive for us to sketch a proof here, to illustrate a method that we will use again later on.

Observe that for any n reals a_i, the distribution of subset sums $\sum_{i\in I} a_i$ is the same, after multiplication by 2^{-n}, as that of the distribution of the random variable $X := \sum_{i=1}^{n} a_i X_i$, where the X_i are i.i.d. $0-1$ variables. We examine this random variable next.

Lemma 7.2. *Let $a_1, \ldots, a_n \in \mathbf{R}$ and let $X_1, \ldots, X_n$ be independent random variables, where each X_i is 0 or 1 with probability $1/2$ each. Define*

$$X := \sum_{i=1}^{n} a_i X_i.$$

Then

$$E(X) = \sum_{i=1}^{n} a_i/2 \qquad \text{and} \qquad Var(X) = \sum_{i=1}^{n} a_i^2/4. \ \blacksquare$$

In the solution to the Erdős-Moser problem, the a_i are n integers closest to zero. Applying the Lemma, our variable X has mean μ either zero (even n) or $\pm(n+1)/4$ (odd n) and standard deviation $\sigma \sim (n^3/48)^{1/2}$. Now by Chebyshev's inequality,

$$\Pr(|X-\mu| \le 2\sigma) \ge 1 - (1/2)^2 = 3/4,$$

so that at least $3/4$ of the 2^n subset sums lie in an interval of just $\sim Cn^{3/2}$ consecutive integers. Thus, some value occurs at least $C2^n n^{-3/2}$ times as a subset sum, which gives the stated lower bound.

Extensions of the Erdős-Moser problem were given by the imaginary mathematician, G. W. Peck [27]. Specifically, the same maximum number of subset sums equal to a target point applies in arbitrary dimension, where $a_1, \ldots, a_n$ are distinct nonzero vectors in $\mathbf{R}^m$ and the target $t \in \mathbf{R}^m$. More generally, suppose we do not insist on distinct nonzero vectors a_i, but we still restrict the number of times that any vector is used.

Theorem 7.3. [27] *Fix integers k and $b_1 \ge b_2 \ge \cdots \ge b_s > 0$, $\sum_{i=1}^{s} b_i = n$. Let $a_1, \ldots, a_n \in \mathbf{R}^m$ such that no u vectors in $\mathbf{R}^m$ occur altogether more than $\sum_{i=1}^{u} b_i$ times. Then for any set T of k target points in $\mathbf{R}^m$, the number of subset sums $\sum_{i\in I} a_i \in T$ is at most the sum of the k middle coefficients of the polynomial*

$$2^{b_1} \prod_{j=1}^{\lfloor s/2 \rfloor} \left(1+q^j\right)^{b_{2j}+b_{2j+1}}.$$

The bound is achieved by taking the a_i's to consist of b_1 0's, b_2 1's, b_3 -1's, and so on.

Halász considers results of both Littlewood-Offord and Erdős-Moser types in dimensions $m \geq 2$. He investigates what happens when we add the restriction that not all n vectors a_i can be confined to the neighborhood of a lower-dimensional space. This leads to more restrictive bounds on the number of subset sums all the same (or inside a unit ball):

Theorem 7.4. [14] *Let B be an open ball of unit diameter in $\mathbf{R}^m$, m arbitrary. Let $0 < \delta < 1$. Suppose $a_1, \ldots, a_n \in \mathbf{R}^m$ are such that for any unit vector $e \in \mathbf{R}^m$, the inner products*

$$|(a_i, e)| \geq 1$$

for at least δn vectors a_i. Then there exists a constant $C = C(m, \delta)$ such that the number of the 2^n subset sums $\sum_{i \in I} a_i \in B$ is at most $C 2^n n^{-m/2}$. Moreover, there exists $C' = C'(m, \delta)$ such that if, in addition to the above conditions,

$$|a_i - a_j| \geq 1 \qquad \text{for all} \quad i \neq j,$$

then the number of subset sums $\sum_{i \in I} a_i \in B$ is at most $C' 2^n n^{-1-m/2}$.

Section 8. Group Security and Avoiding Zero Sums for Small Subsets

Other models of compromise arise naturally in the context of the database SUM query set-up. Besides their intrinsic interest and potential applicability, they may give rise to interesting new problems about concentrating subset sums.

One such model, suggested by my current student, Éva Czabarka, we call the *group security model*. Fix a group size g, which is an integer ≥ 1. We say that a collection $\mathcal{S}$ of SUM queries $\sum_{j \in S_i} x_j$ from our database, $S_i \in \mathcal{S}$, produces a *g-group-compromise* if some nontrivial sum of at most g entries (*i.e.*, $\sum_{j \in J} x_j$ for some index set $J \subseteq [n]$, $1 \leq |J| \leq g$) is determined. Of course, 1-group-compromise is the same as absolute compromise.

For example, taking the salary model with $g = 5$, such a compromise would mean that the salary sum–and, hence, the average salary–would be determined for some nonempty group of at most 5 people in the department.

For a collection $\mathcal{S}$ of distinct queries, $\mathcal{S}$ avoids g-group-compromise whenever the row space $R(\mathcal{S})$ does not contain any of the vectors $\sum_{j \in J} e_j$, $\emptyset \neq J \subseteq [n]$ with $|J| \leq g$. We still forbid just a finite number of vectors, so as with absolute compromise, we can check via Gaussian elimination whether $\mathcal{S}$ produces g-group-compromise. If not, we can construct a basis matrix of form (1) that avoids compromise and generates all vectors for $\mathcal{S}$. Again, we may assume all a_i are rational.

Again, we may artifically introduce $a_n = -1$ to simplify our problem. The maximum number of subset sums $= 0$ or 1 that we want for the matrix will be the same as for this real number problem:

The Problem of Avoiding Zero Sums for Small Subsets. *Given $n \geq g \geq 1$, we seek $a_1, \ldots, a_n \in \mathbf{R}$ such that the number of subset sums $\sum_{i \in I} a_i = 0$, $I \subseteq [n]$, is maximized subject to the restriction that no sum $\sum_{i \in I} a_i = 0$ with $0 < |I| \leq g$.*

This is an interesting, if rather specialized, extension of the real number problem (a variation of the Littlewood-Offord problem) we encountered in connection with avoiding absolute compromise. It deserves study for its own sake. However, so far we have found no connection to posets similar to that for absolute or relative compromise. Naturally, it would be interesting to consider the extension of this problem to higher dimensions.

Section 9. Internal Security and Sums of Vectors in General Position

Motivated particularly by the salary model, we introduce another new model of database compromise. It leads to an interesting new subset sum problem. Let the data records x_i represent salaries of department members. Suppose the ith person in the department, Gyuszi, determines some linear combination $\alpha x_i + \beta x_j$ of his own salary and that of the jth department member, Zsuzsi. Here, $j \neq i$ and α, β are any real numbers with $\beta \neq 0$. Then Gyuszi can solve for Zsuzsi's salary, x_j, since he already knows x_i, something our absolute and relative compromise models do not allow for.

To prevent compromise by queries from an "insider" like Gyuszi, we see that we must prevent the determination of any such expression $\alpha x_i + \beta x_j$ with $i \neq j$ except when $\alpha = \beta = 0$. More generally, for fixed integer $h \geq 1$, we say a collection of queries $\mathcal{S}$ produces *h-inside-compromise* if one can determine some linear combination $\sum_{j \in I} \alpha_j x_j$ where all $\alpha_j \neq 0$ and $0 < |J| \leq h$. So absolute compromise is equivalent to 1-inside-compromise, while 2-inside-compromise is the situation in the last paragraph that we want to avoid. Relative compromise implies 2-inside-compromise, but not necessarily the other way around.

We seek the maximum number of queries $|\mathcal{S}|$ avoiding h-inside-compromise. Let us denote this maximum by $q_h(n)$. For h-inside-compromise, it is difficult to come up with a candidate optimal solution, or even one that merely seems very good, even for $h = 2$. We must be satisfied with asymptotic bounds at this time. However, we see that our query problem is actually equivalent to a natural, and evidently new, problem about maximally concentrating sums of vectors in "general position" in $\mathbf{R}^h$. Let us now derive this result.

We consider the matrix interpretation of maximizing the number of queries. Forming a $d \times n$ basis matrix $B(\mathcal{S})$ as in Section 2, we require that the row space $R(\mathcal{S})$ contains no nonzero vector of weight $\leq h$, where the weight of $v = (v_1, \ldots, v_n) \in \mathbf{R}^n$ is the number of components $\neq 0$.

Next we discuss adding rows to the matrix $B(\mathcal{S})$. Let us denote the rows of $B(\mathcal{S})$ by $b_1, \ldots, b_d \in \mathbf{R}^n$. We want that their span $\langle b_1, \ldots, b_d \rangle = R(\mathcal{S})$ contains no nonzero

vector of weight $\leq h$. Equivalently, for any h standard basis vectors $e_{i_1}, \ldots, e_{i_h} \in \mathbf{R}^n$, $i_1 < \cdots < i_h$, the set $\{b_1, \ldots, b_d, e_{i_1}, \ldots, e_{i_h}\}$ is linearly independent. For $d < n - h$ we claim we can add another row to $B(\mathcal{S})$: Select any

$$b_{d+1} \notin \cup_{1 \leq i_1 < \cdots < i_h \leq n} \langle b_1, \ldots, b_d, e_{i_1}, \ldots, e_{i_h} \rangle,$$

so that we merely have to avoid the union of $\binom{n}{h}$ hyperplanes (or smaller subspaces) in $\mathbf{R}^n$. In fact, one may assume that all entries of b_{d+1} are rational, or even integer.

We may continue to add independent rows to $B(\mathcal{S})$ until we have $n - h$ of them. Performing Gaussian elimination, we obtain, after permuting the columns if necessary, an $(n - h) \times n$ basis matrix

$$M = M(\mathcal{S}) = \begin{pmatrix} 1 & 0 & 0 & \ldots & 0 & a_{11} & \ldots & a_{1h} \\ 0 & 1 & 0 & \ldots & 0 & a_{21} & \ldots & a_{2h} \\ 0 & 0 & 1 & \ldots & 0 & a_{31} & \ldots & a_{3h} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & a_{n-h,1} & \ldots & a_{n-h,h} \end{pmatrix}.$$

This matrix consists of the identity I_{n-h} on the left together with rows $a_1, \ldots, a_{n-h} \in \mathbf{R}^h$ on the right. It turns out that avoiding h-inside-compromise means that any set of h vectors from

$$N(\mathcal{S}) := \{a_1, \ldots, a_{n-h}, e_1, \ldots, e_h\},$$

where the e_i are the standard basis vectors in $\mathbf{R}^h$, is linearly independent, *i.e.*, a basis. Our goal is to choose such a_i so as to maximize the number of subset sums $\sum_{i \in I} a_i$, $I \subseteq [n-h]$, that belong to the target set $\{0, 1\}^h$. Each such sum in $\{0, 1\}^h$ corresponds to a query vector in $\mathbf{R}^n$.

A question that arises naturally in connection with this is to estimate how many subset sums, at most, equal the same target vector for a set of vectors in $\mathbf{R}^h$, any h of which are a basis.

General Position Subset Sum Problem. *Given positive integers n, h, how can one select vectors $a_1, \ldots, a_n \in \mathbf{R}^h$ and a target $t \in \mathbf{R}^h$ to achieve the maximum number $f_h(n)$ of the 2^n subset sums $\sum_{i \in I} a_i$, where $I \subseteq [n]$, equal to t, provided that every h of the vectors a_i are linearly independent?*

Notice that for $h = 1$ dimension, this is just our real number problem from Section 3, the point target analogue of the real version of the Littlewood-Offord problem. For general h, it is a higher-dimensional analogue of the Erdős-Moser problem, in which "distinctness" of real numbers a_i is replaced by the "general position" condition on the vectors a_i. Indeed, this subset sum problem is in fact *equivalent* to our h-inside-compromise problem.

Theorem 9.1. *For all $n, h \geq 1$, the maximum number $q_h(n)$ of queries from a set of n records that avoids h-inside-compromise equals the maximum number $f_h(n)$ of subset sums equal to the same target vector t over all sets of n vectors in $\mathbf{R}^h$ in general position and all targets $t \in \mathbf{R}^h$.*

Proof. For $n \leq h$ we immediately get $q_h(n) = f_h(n) = 1$ from the definitions. Henceforth, assume $n > h$.

As described above, our database problem has been transformed to a subset sum concentration problem in $\mathbf{R}^h$. It suits our purposes, given vectors $a_1, \ldots, a_{n-h} \in \mathbf{R}^h$ that satisfy the condition on $N(\mathcal{S})$ above, to artificially introduce vectors

$$a_{n-h+i} = -e_i, \qquad 1 \leq i \leq h.$$

Defining $N'(\mathcal{S}) = \{a_1, \ldots, a_n\}$, the subset sums over $a_1, \ldots, a_{n-h}$ that belong to $\{0,1\}^h$ correspond naturally to subset sums over $N'(\mathcal{S})$ that equal the zero vector. By our conditions, any h vectors in $N'(\mathcal{S})$ form a basis. So we have an instance of the general position problem, where the target sum $t = 0$ and where the last h vectors a_i form the standard basis.

Notice that setting $t = 0$ is no restriction since for any instance of the general position problem, we may apply Erdős's sign reversal trick: Take any I such that $\sum_{i \in I} a_i = t$, and replace each $a_i, i \in I$ by $-a_i$, which will then give just as many sums $= 0$ as were originally $= t$. Further, including the standard basis vectors is no restriction in the general position problem! For suppose we have vectors $a_1, \ldots, a_n$ as in the general position problem, where any h of them are linearly independent, and target $t = 0$. Let us define a linear transformation generated by the basis correspondence $a_{n-h+i} \mapsto -e_i$, $1 \leq i \leq h$. This produces a new set of n vectors in $\mathbf{R}^h$, such that any h of them form a basis, with just as many subset sums equal to 0. This is an instance of the problem we got in reducing the query problem. Therefore, both problems have the same maximum, and the theorem is proved. ∎

Section 10. Constructions for the General Position Problem in Two Dimensions

We now concentrate on the case $h = 2$ of the general position problem. We have seen that it is of particular interest in connection with the salary model for the database. Unlike the Littlewood-Offord problem (and its real analogue), it is not so obvious what the optimal configuration of a_i's is likely to be. We shall describe two approaches that turn out to be asymptotically optimal to within a constant factor. Both can be analyzed by the elementary probabilistic methods we used earlier.

Our first construction is to adapt the one-dimensional Erdős-Moser solution by taking the a_i's to be

$$(1,0), (1,1), (1,-1), (1,2), (1,-2), \ldots,$$

which are the n lattice points closest to the origin on the line $x = 1$. These vectors are clearly pairwise independent. Arguing as with the Erdős-Moser solution, we find that over half of the 2^n subset sums hit lattice points in a rectangle centered at $(n/2, 0)$ with width $\sim Cn^{1/2}$ in the x-direction and width $\sim C'n^{3/2}$ in the y-direction, so that some sum occurs at least on the order of $2^n n^{-2}$ times.

Another construction, which effectively spreads out the vectors in a different way, was suggested by Füredi at the Balatonlelle conference (July, 1996). Fix an integer $k > 0$ and consider the set of vectors

$$S_k := \{(i, j) \in \mathbf{Z}^2 : 1 \le i, j \le k\},$$

which consists of the lattice points in a square in the first quadrant. Define the set F_k to consist of those vectors $(i, j) \in S_k$ such that i and j are relatively prime. This set is closely related to the Farey series in number theory. A well-known result of Dirichlet [15, Thm. 331; *cf.* 19, p. 337] informs us that over half the vectors in the square S_k are included, specifically,

$$|F_k| \sim \frac{6}{\pi^2} k^2.$$

Thus, we get n vectors in F_k by taking $k \sim (\pi/6^{1/2})n^{1/2}$, denote them by $a_1, \ldots, a_n \in \mathbf{R}^2$ and write $a_i = (u_i, v_i)$. Our vectors are pairwise independent, as we have selected the shortest vector in every direction from S_k, eliminating all longer multiples.

For this Farey construction, we apply the probabilistic method separately to each coordinate, with $X = \sum_{i=1}^{n} u_i X_i$, to show that at least half of the 2^n subset sums for the a_i's hit lattice points in a square centered at the point (m, m) with

$$m = E(X) = \Theta(n^{3/2}),$$

with the length of each side being 4σ, where

$$\sigma = \left(\sum u_i^2 / 4 \right)^{1/2} = \Theta(k^2) = \Theta(n^1).$$

So our square encloses just $\Theta(n^2)$ lattice points, and some sum must occur at least on the order of $2^n n^{-2}$ times as $n \to \infty$.

It follows from either of the two constructions that $f_2(n)$ grows with n at least as fast as some constant times $2^n n^{-2}$. (For more details, please refer to [13].)

Theorem 10.1. *There exists a constant $C > 0$ such that for all $n \ge n_o$,*

$$f_2(n) = q_2(n) > C 2^n n^{-2}. \quad \blacksquare$$

We described this problem and these constructions to Paul Erdős, in our last meeting with him in July, 1996, at the Balatonlelle conference. This was before we had worked out the asymptotic lower bounds above. He expressed his firm belief that the second (Farey) construction must be essentially the right one for the two-dimensional subset sum concentration function, $f_2(n)$. Indeed, we propose the

Two-Dimensional Conjecture. *We expect that there exists a constant C such that as $n \to \infty$,*

$$f_2(n) = q_2(n) \sim C2^n n^{-2}.$$

In the next section, we shall complete the proof that this is true to within a constant factor. To honor Erdős's legacy, those of us who survive him must complete the job on this problem and extend the solution to higher dimensions, which is not yet as well understood.

Section 11. Upper Bounds for the Internal Security Problem

In Balatonlelle (July, 1996) I asked for a general upper bound on $q_h(n)$, a bound that would tighten as h grows. Noga Alon suggested a sphere-packing argument that gives such a bound.

Theorem 11.1. *The maximum number of distinct queries of a set of n records that avoids h-inside-compromise satisfies*

$$\begin{aligned} q_h(n) &\le 2^n / \sum_{i=0}^{\lfloor h/2 \rfloor} \binom{n}{i} \\ &= O(2^n n^{-\lfloor h/2 \rfloor}), \qquad \text{as} \quad n \to \infty. \end{aligned}$$

Proof. A subset of $[n]$ corresponding to a SUM query in turn corresponds to a $0-1$ vector of length n, its characteristic vector. Define the ball $B(I) \subseteq 2^{[n]}$ for $I \subseteq [n]$ to consist of all subsets $J \subseteq [n]$ within Hamming distance $\lfloor h/2 \rfloor$ of I (*i.e.*, $|J \Delta I| \le \lfloor h/2 \rfloor$). Then, for any two sets $I \neq I' \in \mathcal{S}$, where $\mathcal{S}$ avoids h-inside-compromise, the Hamming distance between I and I' is more than h, so the balls $B(I)$ and $B(I')$ are disjoint. The bound follows. ∎

For $h = 2$ dimensions, this gives us

$$f_2(n) = q_2(n) = O(2^n n^{-1}),$$

but we can do better. For any vectors $a_1, \ldots, a_n \in \mathbf{R}^2$ in general position (meaning here that no vector a_i is a multiple of some a_j, $j \neq i$), take any vector $v \in \mathbf{R}^2$ such that the projections $a_1 \cdot v, \ldots, a_n \cdot v$ are distinct and nonzero. Then $\sum_{i \in I} a_i = t$ implies $\sum_{i \in I} a_i \cdot v = t \cdot v$. Thus, the maximum number of subset sums equal to the same target, $q_2(n)$, is at most the number of solutions to the Erdős-Moser problem with n real numbers, which by Sárközy-Szemerédi is $\sim C2^n n^{-3/2}$. This improves our bound by a factor of $n^{-1/2}$.

Just as we were finishing this survey, the paper of Halász mentioned earlier [14] came to our attention, and it leads to a further improvement of the upper bounds above in all dimensions $h \ge 2$. Its impact on our study of internal compromise is fundamental.

Theorem 11.2. *Let $h \geq 2$. Then as $n \to \infty$,*

$$f_h(n) = q_h(n) = O(2^n n^{-1-h/2}).$$

Proof. We seek to use Halász's Theorem 7.4. Suppose we have vectors $a_1, \ldots, a_n \in \mathbf{R}^h$, $h \geq 2$, that achieve the maximum concentration $f_h(n) = q_h(n)$ at a target t in the general position subset sum problem. Since the vectors are distinct, there exists $\epsilon_1 > 0$ so that

$$|a_i - a_j| \geq \epsilon_1 \qquad \text{for all} \quad i \neq j.$$

For any unit vector $e \in \mathbf{R}^h$, at most $h-1$ of the vectors a_i are orthogonal to e, due to the general position condition. Let us define $s(e)$ to be the hth smallest of the n values $|(a_i, e)|$. By compactness, there exists $\epsilon_2 > 0$ so that

$$s(e) \geq \epsilon_2 \qquad \text{for all unit vectors} \quad e.$$

"Blowing up" the vectors a_i and the target t by scalar multiplication by the same real number $\lambda > 0$ preserves the maximum concentration at a single target point. For any fixed $\delta < 1$, blowing up by sufficiently large λ ensures that the conditions of the second part of Theorem 7.4 are satisfied for n sufficiently large (for $\delta = 1/2$, say, $n \geq 2h$ will do). We then obtain the stated bound on the number of sums hitting the same target point. ∎

This result, for $h = 2$ dimensions, combined with Theorem 10.1, yields the solution to the Two-Dimensional Conjecture, to within a constant factor.

Corollary 11.3. *As $n \to \infty$,*

$$f_2(n) = q_2(n) = \Theta(2^n n^{-2}). \quad ∎$$

Halász notes two constructions in connection with the second part of Theorem 7.4 that relate directly to our situation. One method Halász describes is to take an extremal configuration in dimension $h-1$ and then translate it orthogonally by a fixed large vector to get a configuration in dimension h. For instance, if we take the solution to the Erdős-Moser problem, an orthogonal translate by one unit in two dimensions is essentially our first construction in Section 10. (However, going up to three dimensions, taking an orthogonal translate of this two-dimensional construction, does not appear to satisfy the first condition of 7.4; we're not sure what Halász meant.)

The second Halász construction does the following for general dimension h. The method is to take all integer component vectors that fit inside a ball centered at the origin, where the radius is taken to be just large enough that we have as many vectors as we need. In $h = 2$ dimensions, we easily extract a large set of vectors in general position from this set by tossing out multiples of vectors, including the zero vector and negative multiples. This is essentially the "rounded version" of the Farey construction

for $\mathbf{R}^2$ that we described earlier, provided we take $\sim n/2$ vectors in the first quadrant and then add as many vectors in the second quadrant by changing the sign of the first coordinates. Unfortunately, in $h > 2$ dimensions, it is not enough to take the second Halász construction and just toss out multiples of vectors, since we generally still have collections of three vectors that are dependent. So it is not yet clear how to obtain a collection of vectors in general position that achieves the upper bound in Theorem 11.2.

What if instead of requiring any h of the vectors to be a basis, we only require that any l of them be independent, for some $l < h$? Can we do any better than the general position problem in $\mathbf{R}^l$? The answer is no.

Theorem 11.4. *For all $n, h \geq l \geq 1$, the maximum number of subset sums equal to the same target vector, over all sets of n vectors in $\mathbf{R}^h$, any l of which are independent, and over all targets $t \in \mathbf{R}^h$, is $f_l(n) = q_l(n)$.*

Proof. There is nothing to prove unless $h > l$. Let $a_1, \ldots, a_n \in \mathbf{R}^h$ and target $t \in \mathbf{R}^h$ achieve the stated maximum. Select any vector v such that

$$v \notin \cup_{i_1 < \cdots < i_l} \langle a_{i_1}, \ldots, a_{i_l} \rangle.$$

Then project the vectors a_i and the target t onto the hyperplane

$$H_v = \{w \in \mathbf{R}^h : (w, v) = 0\}.$$

At least as many subset sums equal the projection of the target as equalled the target originally, and any l of the projections of the a_i's are independent. So we have reduced the dimension to $h - 1$. We can repeatedly project down until dimension l is reached, where the maximum is $f_l(n)$. ∎

Section 12. Future Study

A construction that achieves the bound in Theorem 11.2 for general h would be especially nice to see. Perhaps the general position condition, which is forced on us by the original internal security problem, causes $f_h(n)$ to be of lower order than the bound in Theorem 11.2? Continuing studies are needed to resolve this issue.

We would like to see the Two-Dimensional Conjecture proved; we wonder which construction is asymptotically optimal. The analogue of this conjecture for higher dimensions would then be worthy of consideration.

We are anxious to see progress made on questions motivated by the group security model in Section 8. There may be other pertinent new models for the database security problem worth investigating.

Finally, we point out that Halász's proofs employ sophisticated analytical arguments involving measure theory, Fourier transforms, and inequalities of Schwartz, Esséen, and Wiener. It would be nice to derive by more elementary methods the upper bound in Theorem 11.2, or at least Corollary 11.3 (in two dimensions), in order to have an accessible, self-contained proof.

Acknowledgements

We are grateful to Mirka Miller and Ljiljana Branković for sharing their results and enthusiasm for this project. We thank Noga Alon, Éva Czabarka, and Zoltán Füredi for suggesting crucial ideas that furthered our progress. Daphne Liu, László Székely, and Chih-Chang Ho also provided valuable assistance.

References

1. I. Anderson, *Combinatorics of Finite Sets*, Clarendon Press (1987).

2. L. Branković and M. Miller, An application of combinatorics to the security of statistical databases, *Austral. Math. Soc. Gazette* **22** (1995), 173–177.

3. L. Branković and M. Miller, An application of antichains to an optimisation problem in data security, preprint(1995).

4. L. Branković, M. Miller, and J. Širáň, Towards a practical auditing method for the prevention of statistical database compromise, *Proc. 7th Australasian Database Conf., Austral. Comp. Sci. Commun.* **18(no. 2)** (1996), 177–184.

5. L. Branković, M. Miller, and J. Širáň, Graphs, 0-1 matrices, and usability of statistical databases, *Congr. Numer.* **120** (1996), 169–182.

6. F. Y. Chin and G. Ozsoyoglu, Auditing and inference control in statistical databases, *IEEE Trans. Software Engin.* **SE-8** (1982), 574–582.

7. P. Erdős, On a lemma of Littlewood and Offord, *Bull. Amer. Math. Soc. (2nd ser.)* **51** (1945), 898–902.

8. P. Erdős, Extremal problems in number theory, in *Theory of Numbers* (A. L. Whiteman, ed.), Amer. Math. Soc., Providence (1965) 181–189.

9. P. Frankl and Z. Füredi, The Littlewood-Offord problem in higher dimensions, *Annals Math.* **128** (1988), 259–270.

10. C. Greene and D. Kleitman, Proof techniques in the theory of finite sets, in *Studies in Combinatorics* (G.-C. Rota, ed.), Math. Assn. America (1978) 22–79.

11. J. R. Griggs, On chains and Sperner k-families in ranked posets, *J. Combin. Th. (ser. A)* **28** (1980), 156–168.

12. J. R. Griggs, Concentrating subset sums at k points, *Bull. Inst. Comb. Applns.* **20** (1997), 65–74.

13. J. R. Griggs and G. Rote, On the distribution of sums of vectors in general position, *Proceedings of the DIMATIA/DIMACS Conference on The Future of Discrete Mathematics, Stiřín (1997), Amer. Math. Soc.,* to appear.

14. G. Halász, Estimates for the concentration function of combinatorial number theory and probability, *Periodica Math. Hungar.* **8 (3-4)** (1977), 197–211.

15. G. H. Hardy and E. M. Wright, *An Introduction to the Theory of Numbers, 5th ed.*, Clarendon Press, Oxford (1979).

16. G. O. H. Katona, On a conjecture of Erdős and a stronger form of Sperner's theorem, *Studia Sci. Math. Hungar.* **1** (1966), 59–63.

17. D. J. Kleitman, On a lemma of Littlewood and Offord on the distribution of certain sums, *Math. Z.* **90** (1965), 251–259.

18. D. J. Kleitman, On a lemma of Littlewood and Offord on the distributions of linear combinations of vectors, *Advances in Math.* **5** (1970), 155–157.

19. D. E. Knuth, *The Art of Computer Programming, v. 2, Seminumerical Algorithms*, 2nd ed., Addison-Wesley Publ., Reading, Mass. (1981).

20. B. Lindström, Conjecture on a theorem similar to Sperner's, in *Combinatorial Structures and Their Applications* (R. Guy *et al.*, eds.), Gordon and Breach, New York (1970) 241.

21. J. Littlewood and C. Offord, On the number of real roots of a random algebraic equation III, *Mat. Sbornik* **12** (1943), 277–285.

22. K. Miller and D. Sarvate, Application of symmetric chains to a statistical database compromise prevention problem, *Bull. ICA* **13** (1995), 57–64.

23. M. Miller, I. Roberts, and J. Simpson, Application of symmetric chains to an optimization problem in the security of statistical databases, *Bull. ICA* **2** (1991), 47–58.

24. M. Miller, I. Roberts, and J. Simpson, Prevention of relative compromise in statistical databases using audit expert, *Bull. ICA* **10** (1994), 51–62.

25. M. Miller and J. Seberry, Relative compromise of statistical databases, *Austral. Computer J.* **21(2)** (1989), 56–61.

26. M. Miller and J. Seberry, Audit expert and statistical database security, in *Databases in the 1990's* (B. Srinivasan and J. Zeleznikow, eds.), (1990) 149–174.

27. G. W. Peck, Erdős conjecture on sums of distinct numbers, *Studs. Appl. Math.* **63** (1980), 87–92.

28. R. A. Proctor, Solution of two difficult combinatorial problems with linear algebra, *Amer. Math. Monthly* **89** (1982), 721–734.

29. A. Sárközy and E. Szemerédi, Über ein Problem von Erdős und Moser, *Acta Arith.* **11** (1965), 205–208.

30. E. Sperner, Ein Satz über Untermengen einer endlichen Menge, *Math. Z.* **27** (1928), 544–548.

31. R. P. Stanley, Weyl groups, the hard Lefschetz theorem, and the Sperner property, *SIAM J. Alg. Discr. Meths.* **1** (1980), 168–184.

32. R. P. Stanley, Quotients of Peck posets, *Order* **1** (1984), 29–34.