ON COMPROMISING STATISTICAL DATA-BASES WITH A FEW KNOWN ELEMENTS

Jan VAN LEEUWEN
Department of Computer Science, University of Utrecht, P.O. Box 80.012, 3508 TA Utrecht, Netherlands

Received 22 March 1978; revised version received 12 September 1978

Database, statistical querying, security

1. Introduction

Consider a database of numeric items $d_1, ..., d_n$ and let $d_1, ..., d_q$ be known. Assuming that the items d_j should remain protected for $j > q$, a serious threat to security occurs when a user is permitted to ask statistical information of fixed or variable-sized samples of the database. In a first study of the possible protection against used inference Dobkin, Jones and Lipton [1] discussed the complexity of actually compromising a database for a few types of queries which can arise in practice. The work was substantially extended by Reiss [5] and in Dobkin, Lipton and Reiss [2]. In this note we shall consider some interesting further questions concerning the security problem when a user can request the average of fixed-size samples of items.

Let $S(n, p, q, r)$ be the minimum number of averages of samples of a fixed size p needed to infer d_{q+1}, assuming that $d_1, ..., d_q$ are known and any two distinct samples queried may not overlap by more than r items. We shall assume throughout this paper that $p > q + 1$, to exclude trivial cases of the model. Reiss [5] proved under this assumption that

$$S(n, p, q, r) \geq \frac{2p - (q+1)}{r} \tag{1.1}$$

but little seems known about the quality of this bound. Reiss [5] presented several results for $q = 0$, and proved that the bound of (1.1) is achievable to within 1 query if we extend the model and permit the querying of samples of arbitrary sizes $\geq p$.

Keeping the sample-size fixed at p, we shall study the complexity of inferring d_{q+1} when samples are allowed to overlap by at most 1 element (thus $r = 1$) and, as ever, $p > q + 1$. This admittedly restrictive case is of interest, because even here very little is known. It follows from [1] and from [5] that

$$S(n, p, 0, 1) = 2p - 1 \tag{1.2}$$

$$S(n, p, 1, 1) = 2p - 2 \tag{1.3}$$

and also that $S(n, p, q, 1) \geq 2p - q - 1$. Whereas (1.2) and (1.3) show that it is strictly easier to compromise the database when one item is known prior to querying compared to when no item is known at all, we shall prove that there is no such advantage when 2 items are known. Thus, knowing 2 items makes it no easier to compromise the database than knowing 1 item does. The result follows from an improvement of Reiss' bound [5], which we derive in Section 2:

Proposition 1.

$$S(n, p, q, 1) \geq 2p - q \quad \text{for } q \geq 2$$

One might now suspect that knowing any number of items $q \geq 2$ is of no help, but this is not true. We shall prove in Section 3:

Proposition 2. *Under suitable conditions for p and q one can have*

$$S(n, p, q, 1) \leq 2p - \Omega(\sqrt{q}).$$

The Ω-notation is taken from [4] (see also [6]).

2. Constructions for Proposition 1

Estimating $S(n, p, q, 1)$ does not just give us information for one case of overlap only. Slightly extending a similar result of [1] we can show:

Proposition 3. *For any* $t \geq 1$,

$$S(n, pt, qt + t - 1, t) \leq S(n, p, q, 1).$$

Proof. We consider $S(n, pt, qt + t - 1, t)$, which is the number of averages required to infer d_{qt+t} (under the usual constraints). Construct a 'new' database $e_1, e_2, \ldots$ by taking $e_j = d_{(j-1)t+1} + \cdots + d_{jt}$. It follows that $e_1, \ldots, e_q$ are 'known', and d_{qt+t} can be deduced once the 'new' database is compromised for e_{q+1}. Any $S(n, p, q, 1)$-method for doing so easily translates into an $S(n, pt, qt + t - 1, t)$ algorithm for finding d_{qt+t}.

Consider $S(n, p, q, r)$. Queries $Q_1, \ldots, Q_s$ are averages, but we may just as well take them as sums: $Q_j = d_{j1} + \cdots + d_{jp}$. If we can infer d_{q+1}, then there must be coefficients $\alpha_1, \ldots, \alpha_s$ such that

$$d_{q+1} = \sum_{j=1}^{s} \alpha_j Q_j + \langle \text{lin. combination of } d_1, \ldots, d_q \rangle \tag{2.1}$$

Defining $\delta_{ij} = 1(0)$ if d_i is (is not) in Q_j, one can easily rearrange (2.1) to obtain

$$d_{q+1} = \sum_{i=1}^{n} \left(\sum_{j=1}^{s} \delta_{ij} \alpha_j \right) d_i + \langle \text{lin. combin. of } d_1, \ldots, d_q \rangle . \tag{2.2}$$

It follows that necessarily $\Sigma_{j=1}^{s} \delta_{ij} \alpha_j = 0$ for $j > q + 1$, and, as in Dobkin, Jones and Lipton [1] or Reiss [5], we conclude that this can only be when

for $i > q + 1$, each d_i occurs at least once in a query with positive α and at least once in a query with negative α. (2.3)

From now on it should be clear what we mean by a 'positive' and a 'negative' query.

If we consider sets of positive and negative queires which merely satisfy the overlap constraints and (2.3), then the minimum number of q ieries possible in any set of this sort is certainly a lower-bound for $S(n, p, q, r)$. For $r = 1$ the typical combinatorial

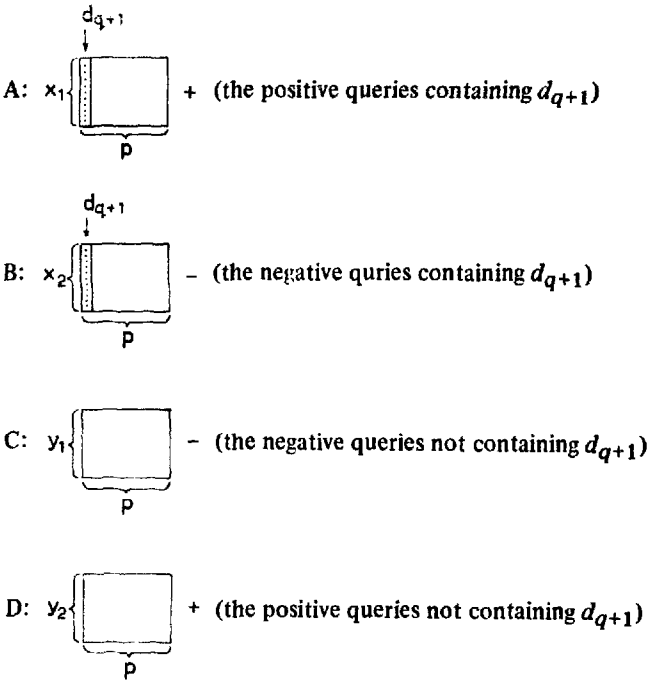

Fig. 1.

structure of such a set is shown in Fig. 1. The following conditions must be satisfied:

(i) each line (row or query) can intersect another line in at most one point,

(ii) each element d_i $(i > q + 1)$ on a positive line must occur somewhere on a negative line, and vice versa, and

(iii) $x_1 + x_2 > 0$.

Thus, we have a structure not unlike a block-design (see e.g. [3]). Let $R(p, q)$ be the minimum number of lines in such a structure.

Lemma. $R(p, q) \geq 2p - q$ *for* $q \geq 2$ *(and* $p > q + 1$*)*

Proof. Recall that $p > q + 1$ by assumption. The argument is for a considerable part merely a refinement of Reiss' proof [5] (which in turn was a refinement of a proof in [1]). We distinguish two main cases:

(a) $y_1 > 0$ and $y_2 > 0$.

Consider any two lines α and β from C and D,

Fig. 2.

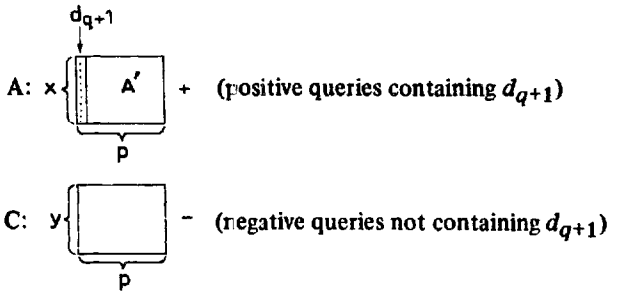

Fig. 3.

with an overlap of $L_1 \leqslant 1$ and containing L_2 and L_3 items d_i with $i \leqslant q$ (respectively) in the nonoverlapping part, as shown in Fig. 2.

Note that d_{q+1} does not occur on either line. Each point τ in the region shown on α must also occur on some positive line (necessarily different from β), and each point μ in the similar region shown on β must also occur on some negative line (likewise necessarily different from α). The number of points τ is $p - L_1 - L_2$, of points μ is $p - L_1 - L_3$. Thus

$$R(p, q) \geqslant 2 + (p - L_1 - L_2) + (p - L_1 - L_3)$$
$$\geqslant 2p - (L_2 + L_3) \geqslant 2p - q$$

(b) $y_1 = 0$ or $y_2 = 0$.

By symmetry we may assume that $y_2 = 0$. It follows that $x_2 = 0$. Otherwise any point d_i with $i > q + 1$ on a line in B (as $p > q + 1$ such points exist) would also occur on a line in A, because it is the only possibility to be on a positive line. These A and B lines would then intersect in 2 points (d_i and d_{q+1}), a contradiction! The combinatorial structure simplifies to Fig. 3 with $x > 0$ and $y > 0$, the elements in A' all distinct (as otherwise the overlap restriction would be violated) and C not containing d_{q+1}. Clearly, the design conditions remain in effect. We distinguish two further cases:

Case 1, $x \geqslant p$. A' contains at least $x(p - 1) - q$ elements d_i with $i > q + 1$ which all have to occur in C at least once in order to be on a negative line. Thus

$$x(p - 1) - q \leqslant y \cdot p \Rightarrow y \geqslant \left(1 - \frac{1}{p}\right)x - \frac{q}{p}.$$

The total number of lines can be estimated by

$$x + y \geqslant \left(2 - \frac{1}{p}\right)x - \frac{q}{p} \geqslant 2p - 1 - \frac{q}{p}$$
$$\geqslant 2p - 1 - \frac{q}{q+2} > 2p - 2 \qquad (2.4)$$

Case 2, $x \leqslant p - 1$. Observe that A' and C must contain the same d_i with $i > q + 1$, by (2.3). Because a C-line can contain at most one point from each A-line (thus having at most x A-points in all), each C-line must contain at least $p - x \geqslant 1$ points d_i with $i \leqslant q$. Consider any C-line α, containing some $q' \geqslant 1$ points d_i with $i \leqslant q$. Let there be an A-line β containing some q'' points d_i with $i \leqslant q$ ($q'' \geqslant 0$) such that one of the following conditions holds:

(i) α and β intersect in a d_i with $i \leqslant q$

(ii) α and β intersect in a d_i with $i > q + 1$, but $q' + q'' < q$

(iii) α and β do not intersect.

The situation can be sketched (using conventions as in (a)) as in Fig. 4., and we get

$$x + y \geqslant 2 + (p - 1 - L_2) + (p - 2 - L_3)$$
$$= 2p - 1 - (L_2 + L_3) \geqslant 2p - q \qquad (2.5)$$

for (i) and (ii), and

$$x + y \geqslant 2 + (p - 1 - L_2) + (p - 1 - L_3)$$
$$= 2p - (L_2 + L_3) \geqslant 2p - q \qquad (2.6)$$

for (iii).

The only situation left to consider is where each C-line intersects each A-line in a point d_i with $i > q + 1$, and the total number of (necessarily distinct) points d_i with $i \leqslant q$ on any C-line α and A-line β sums to q. It means that for any pair α, β the set of points d_i with $i \leqslant q$ contained in α is precisely the complement of the similar set contained in β in the collection $\{d_1, \ldots, d_q\}$. Hence, any other A-line γ will contain the same points d_i with $i \leqslant q$ as does β. As A-lines can only intersect at d_{q+1} two possibilities remain:

(iv) A-lines contain no points d_i with $i \leqslant q$, but each C-line contains all q of them.

It follows that $y = 1$, as the intersection constraint would be violated otherwise (note that $q \geqslant 2$). Consequently there are precisely $p - q$ elements d_i with $i > q + 1$. Considering an arbitrary A-line we see that

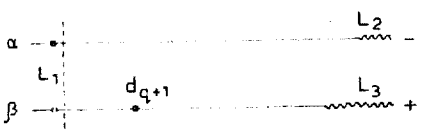

Fig. 4.

at least $(p-1)-(p-q)=q-1\geqslant 1$ elements cannot possibly occur on a negative line, contradicting (2.3).

(v) there is only one A-line β ($x=1$), and it contains $q''\geqslant 1$ elements d_i with $i\leqslant q$.

By the same argument as above we conclude that each C-line must contain the same set of $q-q''$ elements d_i with $i\leqslant q$. Because of the overlap constraint on the one hand ($q-q''\leqslant 1$) and the observation that each C-line contains at least one such element ($q-q''\geqslant 1$), we obtain $q''=q-1$. Thus, the C-lines contain precisely one (identical) d_i with $i\leqslant q$. Observe that this time at least $(p-1)-(p-1-(q-1))=q-1\geqslant 1$ elements d_i with $i>q+1$ in C find no compensation in A, contradicting (2.3).

Hence the desired inequality holds also in Case 2 ($x\leqslant p-1$).

As $S(n,p,q,1)\geqslant R(p,q)$, Proposition 1 easily follows from the Lemma. An interesting conclusion is obtained for $q=2$.

Corollary. $S(n,p,2,1)=S(n,p,1,1)=2p-2$

Proof. By (1.3) and Proposition 1 we have: $2p-2\leqslant S(n,p,2,1)\leqslant S(n,p,1,1)=2p-2$.

This proves the interesting phenomenon discussed in Section 1, that knowing 2 elements of the database does not make it easier to compromise the data than knowing just 1 element does (for the case that averages of fixed-size samples can be asked).

3. Constructions for Proposition 4

The proof that $2p-2\geqslant S(n,p,q,1)\geqslant 2p-q$ ($q\geqslant 2, p>q+1$) holds no clue as to whether the lower-bound can be achieved or not. Reiss [5, Section 6] noted for his bound that it is not likely to be achieved everywhere, and the precise value of $S(n,p,q,1)$ will vary depending on purely number-theoretic connections between p and q. Trying for small values of q, one might tend to believe that the $2p-2$ upperbound is hard to beat. We present a general method to do better, achieving bounds of the form $2p-\Omega(\sqrt{q})$ for a wide range of p, q values. We use some elementary facts about (v,k,λ)-designs, taken from [3].

A (v,k,λ)-design is a structure of v points and b blocks of k points each such that

(i) all points occur in the same number of blocks, and

(ii) each pair of distinct points occurs in exactly λ distinct bloakcs. Let D be a 'master' $(v,k,1)$-design, with parameters v and k to be fixed later. The blocks $B_1,\ldots,B_b$ of D will be of great value in designing a set of averages which overlap in at most one sample-element. The number of blocks in D is completely determined by v and k (see Hall [3, p. 101]):

$$b=\frac{v(v-1)}{k(k-1)} \tag{3.1}$$

Let $D_1, D_2, \ldots$ be copies of D.

The following Lemma shows a strategy for compromising a database of sufficiently many elements in the $S(n,p,q,1)$-sense.

Lemma. *If* $\lceil (p-2)/b\rceil\leqslant\lfloor (q-k)/v\rfloor$, *then one can compromise a database with q known elements for* d_{q+1} *by asking the average of at most* $2p-k-1$ *samples of size p which overlap in at most one point.*

Proof. By assumption there is an integer $\alpha\geqslant 1$ such that

$$\frac{p-2}{b}\leqslant\alpha\leqslant\frac{q-k}{v} \tag{3.2}$$

Design the set of queries shown in Fig. 5. Elements a_i and b_{ij} denote unknown items, elements c_i and D_{ij} are to be chosen from among the q known items $d_1,\ldots,d_q$. The element d_{q+1} is denoted merely as d.

If there are sufficiently many 'known' elements (we will check it in a moment), then a design as in Fig. 5 does exist and satisfies all criteria for size and permissible overlap. Element d follows by noting that

$$d=(\text{sum of A})-(\text{sum of B})+(\text{sum of C}) \tag{3.3}$$

in which the 'unknowns' cancel and only 'known' elements on the right-hand side remain. The database is compromised for d in $1+(p-k)+(p-2)=2p-k-1$ queries, as was to be shown. We only need to verify that sufficiently many 'known' elements are at hand to choose distinct $c_1,\ldots,c_k$ and D_{ij} from among them. By (3.2) we know that at most α full sets of D_i-blocks are needed to fill the right part

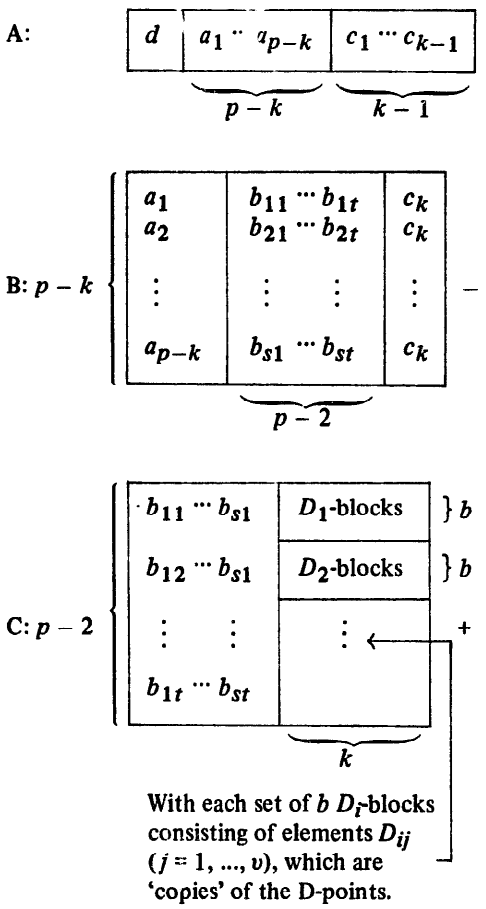

Fig. 5. ($s = p - k$, $t = p - 2$).

in C in Fig. 5. Assuming the worst, we need $(k-1) + 1 + \alpha v = \alpha v + k$ elements to build the queries. By (3.2)

$$\alpha v + k \leqslant \frac{q-k}{v} v + k = q$$

and we see that sufficiently many 'known' elements are available to make the construction work.

The Lemma has reduced the question of how to compromise the database efficiently to the question of finding a $(v, k, 1)$-design D such that

$$\lceil (p-2)/b \rceil \leqslant \lfloor (q-k)/v \rfloor \qquad (3.4)$$

If we write

$$p = \beta b + \gamma + 2 \qquad 0 > \gamma \leqslant b$$

for certain integers β and γ, then (3.4) is satisfied precisely when

$$q \geqslant (\beta + 1) v + k .$$

If we give q its smallest possible value, then $p \geqslant q + 2$ leads to the condition:

$$\beta b + \gamma + 2 \geqslant (\beta + 1) v + k + 2 ,$$

or (3.5)

$$\beta(b - v) + (\gamma - v) \geqslant k$$

By Fisher's inequality [3; 10.2.3] $b \geqslant v$. Fix γ to the range $v \leqslant \gamma \leqslant b$, and let $\beta \geqslant \lceil k/(b-v) \rceil$. As (3.5) is satisfied under these assumptions, we obtain

$$2p - q \leqslant S(n, p, \underbrace{(\beta + 1) v + k}_{q}, 1) \leqslant 2p - k - 1 \qquad (3.6)$$

To get a tight bound, we must choose designs D in which v remains as small as possible in terms of k (while $b > v$). From (3.1) one can easily derive that v is at best quadratic in k.

Proposition 4. *For an infinite range of p, q values (unbounded in both p and q) we have* $2p - q \leqslant S(n, p, q, 1) \leqslant 2p - \Omega(\sqrt{q})$.

Proof. Let k be a prime-power, and let D be the design of lines in the 2-dimensional affine space over $GF(k)$. D has $b = k(k+1)$ and $v = k^2$ (thus $b > v$), and $q = \theta(v^2)$. Substitute these values in (3.6).

An interesting problem would seem to prove or disprove that for fixed $q \geqslant 2$:

$$\lim_{p \to \infty} \{S(n, p, q, 1) - 2p\} = 2 .$$

References

[1] D. Dobkin, A.K. Jones and R.J. Lipton, Secure databases: protection against user inference, Research Rep. 65, Department of Computer Science, Yale University (1976).

[2] D. Dobkin, R.J. Lipton and S.P. Reiss, Aspects of the database security problem, Proc. Conf. Theoretical Comput. Sci., University of Waterloo (August 1977) 262–274.

[3] M. Hall Jr., Combinatorial Theory, (Blaisdell Waltham, MA 1967).

[4] D.E. Knuth, Big omicron and big omega and big theta, SIGACT News 8 (1976) 18–24.

[5] S.P. Reiss, Security in data bases: a combinatorial study, Research Rep. 77, Department of Computer Science, Yale University (1976).

[6] B. Weide, A survey of analysis techniques for discrete algorithms, ACM Comput. Surveys 9 (1977) 291–313.