vakgroep informatica R.U. Utrecht

√

COMPROMISING STATISTICAL DATA-BASES WITH A FEW KNOWN ELEMENTS IN A COMBINATORIAL MODEL

Jan van Leeuwen

RUU-CS-78-2

January 1978

Rijksuniversiteit Utrecht

Vakgroep Informatica

Budapestlaan 6
Utrecht 2506
Telefoon 030-531454

ge

7901/4-1-11

COMPROMISING STATISTICAL DATA-BASES WITH A FEW
KNOWN ELEMENTS IN A COMBINATORIAL MODEL

Jan van Leeuwen

Technical Report RUU-CS-78-2

January 1978

Department of Computer Science
University of Utrecht
P.O.Box 80.012
3508 TA Utrecht, the Netherlands

All correspondence to:

Dr. Jan van Leeuwen
Department of Computer Science
University of Utrecht
P.O.Box 80.012
3508 TA Utrecht, the Netherlands

COMPROMISING STATISTICAL DATA-BASIS WITH A FEW KNOWN ELEMENTS IN A COMBINATORIAL MODEL

Jan van Leeuwen
Department of Computer Science
University of Utrecht
3508 TA Utrecht, the Netherlands

Abstract. For a few simple models Dobkin, Jones and Lipton proved that a database may be compromised when statistical querying is permitted. In particular, for a database of n items let $S(n,p,q,r)$ be the minimum number of averages of samples of a fixed size p needed to deduce at least one new item, assuming that q items of the database are known already and any two distinct samples may overlap for at most r items. Reiss showed that $S(n,p,q,r) \geq \frac{2p-(q+1)}{r}$, but little is known about the quality of this bound. For r=1 we improve Reiss' bound slightly to $S(n,p,q,1) \geq 2p-q$ when $q \geq 2$, obtaining the interesting conclusion that knowing 2 items of the database has no advantage over knowing 1 item (which in itself does have an advantage of 1 query over knowing no items). We show that bounds of the form $2p-\Omega(\sqrt{q})$ are achievable.

1. Introduction

Consider a database of numeric items $d_1,\ldots,d_n$ and let $d_1,\ldots,d_q$ be known. Assuming that the items d_j should remain protected for $j>q$, a serious threat to the security can occur 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 user inference Dobkin, Jones and Lipton [1] discussed the complexity of actually compromising a database, for a few types of querying which users typically request. Subsequently, the work was substantially extended by Reiss [5] and in Dobkin, Lipton and Reiss [2]. In this note we shall consider some interesting questions concerning the security problem when a user can request the average of any fixed-size sample 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 $d_1,\ldots,d_q$ are known and any two distinct samples queried may not overlap for more than r items. Reiss [5] proved

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

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

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 $r=1$). This admittedly restrictive case seems to be of interest, as even here only a few results are known. In particular, it follows from Dobkin, Jones and Lipton [1] and from Reiss [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 compared to when no item is known prior to querying at all, we shall prove in section 2 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 argument can be extended for a few more "small" values of q, and can be derived from the following slight improvement of Reiss' bound.

Proposition A. $S(n,p,q,1) \geq 2p-q$ for $q \geq 2$

From examples one might suspect that knowing any number of items $q \geq 2$ is of no help, but this is not true. We shall prove (and specify it more precisely in section 3) that

Proposition B. Under suitable conditions for p and q we can have $S(n,p,q,1) \leq 2p-\Omega(\sqrt{q})$.

The Ω-notation should be customary, and is taken from Knuth [4] (see also Weide [6]).

2. Constructions for proposition A

Determining or precisely estimating $S(n,p,q,1)$ does not just give information for one case of overlap. Only slightly extending a similar result of Dobkin, Jones and Lipton [1] we can show

Proposition 2.1. 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 with the given constraints to infer d_{qt+t}. Construct a "new" database $e_1,e_2,\dots$ by taking $e_j=d_{(j-1)t+1}+\dots+d_{jt}$. It follows that precisely $e_1,\dots,e_q$ are known and d_{qt+t} can easily be deduced once the database is compromised for e_{q+1}. Any $S(n,p,q,1)$ - method for doing so easily translates to 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: $Q_j=d_{j1}+\ldots+d_{jp}$. 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

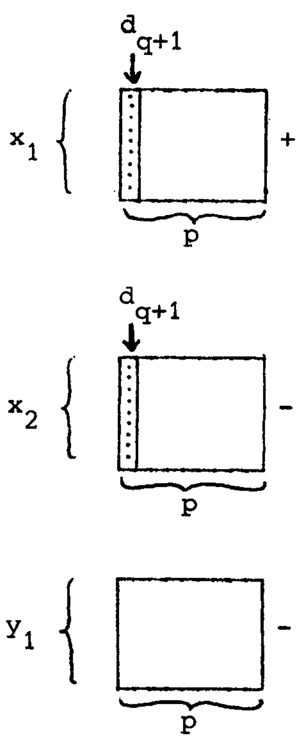
$$d_{q+1}=\sum_{i=1}^{n}\Big(\sum_{j=1}^{s}\delta_{ij}\alpha_j\Big)d_i + \langle\text{lin. combin. of } d_1,\ldots,d_q\rangle \tag{2.2}$$

It follows that $\sum_{j=1}^{s}\delta_{ij}\alpha_j = 0$ 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 queries (with p elements each) which merely satisfy the overlap restriction and (2.3), then the minimum number of queries possible in any set of this sort is certainly a lower-bound for $S(n,p,q,r)$. For r=1 we can make a rather precise picture of the combinatorial structure of such sets:

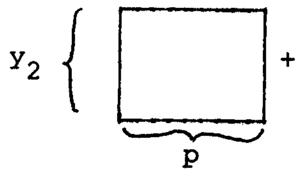


A: the positive queries containing d_{q+1}

B: the negative queries containing d_{q+1}

(2.4)

C: the negative queries not containing d_{q+1}



D: the positive queries not containing d_{q+1}

with the following conditions satisfied

(i) each line (query) intersects any other line in at most one point

(ii) each element d_i (<u>with</u> $i > q+1$) on a positive line must occur on a negative line also, and vice versa

(iii) $x_1 + x_2 > 0$

Thus, we have a structure not unlike a block-design (see e.g. Hall [3]) which deserves further study within the scope of combinatories. From (1.2) we know that such designs exist for any pair of p,q-values. Let $R(p,q)$ be the minimum number of lines in any such design.

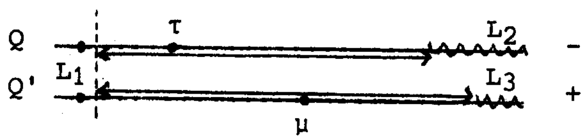
<u>Lemma</u>. $R(p,q) \geq 2p-q$ for $q \geq 2$ (and $p > q+1$)

<u>Proof</u>

We should require $p > q+1$ as the analysis would be meaningless otherwise. The argument is for a considerable part merely a refinement of Reiss' proof [5] (which in turn was a refinement of the proof in Dobkin, Jones and Lipton [1]). We distinguish two cases

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

Consider any two C and D lines



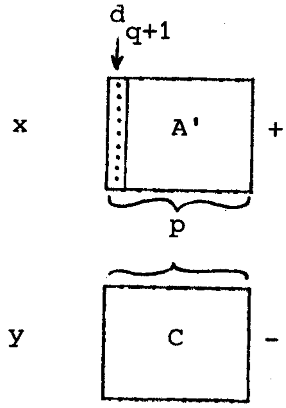
(2.5)

with an overlap of $L_1 \leq 1$ and containing L_2 and L_3 items d_i with $i \leq q$. Note that d_{q+1} does not occur on either line. Each point τ must occur also on some positive line (necessarily different from Q'), and each point μ must occur also on some negative line (likewise necessarily different from Q). The number of points τ is $p-L_1-L_2$, of points μ it is $p-L_1-L_3$. Thus

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

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

By symmetry we may assume that $y_2 = 0$. It follows that $x_2 > 0$! For, let us assume otherwise. Any point d_i with $i > q+1$ on a line in B (as $p > q+1$ such points exist) must also occur on a line in A, because it is the only possibility to be on a positive line. The A and B line would intersect in 2 points (d and d_{q+1}), which is a contradiction. The combinatorial structure has become at least easier to display:



(2.6)

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:

<u>Case</u> I. $x \geq p$

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

$$x(p-1) - q \leq y.p \quad \Rightarrow \quad y \geq (1 - \frac{1}{p})x - \frac{q}{p}$$

and for the total number of lines we obtain

$$y + x \geq (2 - \frac{1}{p})x - \frac{q}{p} \geq 2p - 1 - \frac{q}{p} \geq 2p - 1 - \frac{q}{q+2} > 2p - 2 \qquad (2.7)$$

<u>Case</u> II. $x \leq p-1$.

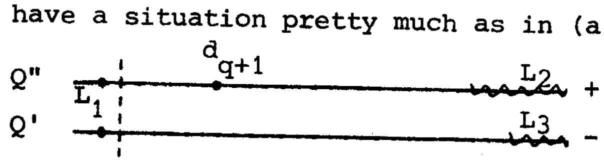
Observe first that A' and C must contain the same d_i with $i > q+1$, by (2.3). Because a C-line can contain at most one point from each A-line (thus having at most x A-points in all), each C-line must contain at least $p-x \geq 1$ points d_i with $i \leq q$. Consider any C-line Q', containing some $q'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

(i) Q" and Q' intersect in a d_i with $i \leq q$

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

(iii) Q" and Q' do not intersect.

We have a situation pretty much as in (a)



and we get

$$y + x \geq 2 + (p-2-L_2) + (p-1-L_3) = 2p - 1 - (L_2+L_3) \geq 2p - q \qquad (2.8)$$

for (i) and (ii), and

$$y + x \geq 2 + (p-1-L_2) + (p-1-L_3) = 2p - (L_2+L_3) \geq 2p - q \qquad (2.9)$$

for (iii).

This leaves very few possibilities open. 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 $i \le q$ on any C-line Q' and A-line Q" sums to q. It means that for any pair Q", Q' the set of points d_i with $i \le q$ contained in Q" is precisely the complement of the similar set contained in Q' in the collection $\{d_1,\ldots,d_q\}$. If there was another A-line Q"' , then it would contain the same points d_i with $i \le q$ as does Q". As A-lines can only intersect at d_{q+1} we have two possibilities remaining:

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

It follows that y=1 (as $q \ge 2$ the intersection contraint would be violated otherwise), and there are precisely p-q elements d_i with $i > q+1$. Considering an arbitrary A-line we see that $\ge (p-1) - (p-q) = q-1 \ge 1$ elements cannot possibly occur on a negative line, contradicting (2.3).

(v) there is one A-line Q" (x=1), and it contains $q'' \ge 1$ elements d_i with $i \le q$.

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

We conclude that also in case II $(x \le p-1)$ the desires inequality holds.

□

Clearly $S(n,p,q,1) \ge R(p,q)$, and proposition A follows from the lemma. An interesting conclusion is obtained for q=2.

<u>Corollary</u>. $S(n,p,2,1) = S(n,p,1,1) = 2p-2$

<u>Proof</u>

By (1.3) and proposition A we have

$$2p-2 \le S(n,p,2,1) \le S(n,p,1,1) = 2p-2$$

□

This shows 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 (for the case that averages of fixed-size samples can be asked).

3. Constructions for proposition B.

The proof that $S(n,p,q,1) \geq 2p-q$ $(q \geq 2,\ p > q+1)$ holds no clue as to whether the lower-bound can be achieved or not. Reiss [5] (sect 6) noted for his bound that it is not likely achieved everywhere, and the precise value of $S(n,p,q,1)$ will vary depending on some purely number-theoretic connections for p and q. Trying for small values of q, one might tend to a feeling that the 2p-2 upperbound is hard to beat. We present a general argument that one can beat it, and show that bounds of the form $2p - \Omega(\sqrt{q})$ can be achieved for a wide range of p,q values. We shall be using some considerations from the study of (v,k,λ)-designs as presented in e.g. Hall [3].

Let D be a "master" $(v,k,1)$-design, for parameters v and k which we shall fix in terms of p and q later. The blocks $B_1,\ldots,B_b$ of D will appear to be of great value in designing a set of averages which overlap in at most one sample-element. The number of blocks (b) 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. Marginal further savings are possible in the distribution of distinct elements over queries, but these will not affect the range of the result in any major way.

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

Proof

By assumption there is an integer $\alpha \geq 1$ such that

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

Design the following queries, shown in a diagram which patterns samples as indicated in (2.4). Elements a_i and c_{ij} denote distinct and "general" elements, elements b_i and D_{ij} are to be chosen from among the q known $d_1,\ldots,d_q$. The element d_{q+1} will be denoted merely as d.

$$+\quad \boxed{d \;\Big|\; \overbrace{a_1 \;\cdot\;\cdot\;\cdot\; a_{p-k}}^{p-k} \;\Big|\; \overbrace{b_1 \;\cdot\;\cdot\;\cdot\; b_{k-1}}^{k-1}} \qquad A$$

$$
-\ p-k\left\{\begin{array}{|c|ccc|c|}
\multicolumn{1}{c}{} & \multicolumn{3}{c}{\overbrace{\qquad\qquad\qquad\qquad\qquad}^{p-2}} & \multicolumn{1}{c}{} \\
\hline
a_1 & c_{11} & \cdots & c_{1t} & b_k \\
a_2 & c_{21} & \cdots & c_{2t} & b_k \\
\vdots & \vdots & & \vdots & \vdots \\
a_{p-k} & c_{s1} & \cdots & c_{st} & b_k \\
\hline
\end{array}\right.
$$

B (with s=p-k, t=p-2)

(3.3)

$$
+\ p-2\left\{\begin{array}{|ccc|c|}
\multicolumn{3}{c}{} & \multicolumn{1}{c}{\overbrace{\qquad\qquad}^{k}} \\
\hline
c_{11} & \cdots & c_{s1} & D_1\text{-blocks} \\
\cline{4-4}
c_{12} & \cdots & c_{s2} & D_2\text{-blocks} \\
\cline{4-4}
\vdots & & \vdots & \vdots \\
c_{1t} & \cdots & c_{st} & \\
\hline
\end{array}\right.
\begin{array}{l} \\ b \\ b \\ \\ \\ \end{array}
$$

C

with each full set of b D_i-blocks composed of elements D_{ij} $(j=1,\ldots,v)$ which are instances of the D-points.

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

$$d = (\text{sum of A}) - (\text{sum of B}) + (\text{sum of C}) \tag{3.4}$$

in which the "unknowns" cancel and only "known" elements remain. The database is compromised for d in $1+(p-k)+(p-2) = 2p-k-1$ queries, as was to be shown.

All we need to verify is that sufficiently many "known" elements" are at hand to choose distinct $b_1,\ldots,b_k$ and D_{ij} from among them. (Note that points $b_1,b_2,\ldots$ could be allowed to occur among the D_{ij}, but we shall not explore this.) By (3.2) we know that at most α sets of D_i-blocks are needed to fill the right part of the C-table in (3.3). Assuming the worst, let's see how many "known" elements we need to build the queries: $(k-1)+1+\alpha.v = \alpha v + k$. Using (3.2) we see that

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

and the construction works.

□

The lemma has reduced the question of determining a "small" set of queries to compromise the database to the question of finding a (v,k,1)-design D such that

$$\lceil \frac{p-2}{b} \rceil \leq \lfloor \frac{q-k}{v} \rfloor \qquad (3.5)$$

Given p and q, can we find a design D with the right parameters for satisfying (3.5). As $p > q+1$, we must find designs with b "large" compared to v. The reason why this strategy does not work in general clearly is the fact that in block-designs the value of b cannot be "arbitrarily large" in terms of v, see (3.1). If we write

$$p = \beta . b + \gamma + 2 \qquad (0 < \gamma \leq b) \qquad (3.6)$$

for some integers β and γ, then (3.5) is satisfied precisely when

$$q \geq (\beta+1)\, v + k \qquad (3.7)$$

If we give q its smallest possible value, then the usual constraint that $p \geq q+2$ leads to the following reformulation of (3.5):

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

$$\Rightarrow \beta(b-v) + (\gamma-v) \geq k \qquad (3.8)$$

By Fisher's inequality we know that $b \geq v$ (Hall [3], 10.2.3). As (3.8) can impossibly be satisfied for $b = v$, we conclude that necessarily $b > v$. Fix γ to the more tighter range $v \leq \gamma \leq b$, and let $\beta \geq \lceil \frac{k}{b-v} \rceil$. As (3.8) is satisfied under these assumptions we obtain

$$2p - q \leq S(n,p,\underbrace{(\beta+1)v+k}_{q},1) \leq 2p - k \qquad (3.9)$$

Now observe that for $\beta = \lceil \frac{k}{b-v} \rceil$: $q = v + \theta(k)$. To get a tight bound, we must be able to choose a design D in which v remains strongly bounded in terms of k (while $b > v$). From (3.1) one can easily derive that v is at best quadratic in k.

<u>Proposition</u> B. For an infinite range of p,q values we have $2p - q \leq S(n,p,q,1) \leq 2p - \Omega(\sqrt{q})$.

<u>Proof</u>

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 $q = \theta(v^2)$.

□

An interesting problem is to prove or disprove that for fixed $q \geq 2$: $\lim_{p\to\infty}\{S(n,p,q,1) - 2p\} = 2$.

4. References.

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

[2] Dobkin, D., R.J. Lipton and S.P. Reiss, Aspects of the database security problem, Proc. of a Conference on Theoretical Computer Science, University of Waterloo, August 1977, pp. 262-274

[3] Hall Jr., M., Combinatorial theory, Blaisdell Publ. Company, Waltham, Mass. (1967)

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

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

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