Security in Data Bases: A Combinatorial Study

Steven P. Reiss

Yale Research Report #77

October 1976

This research was supported by a National Science Foundation graduate fellowship.

ABSTRACT

Security is considered in the context of the abstract model of a data base proposed by Dobkin, Jones and Lipton. Limiting the number of types of queries that are asked of the data base is investigated as a means for providing statistical access while insuring that no data is compromised. The resultant questions are primarily combinatoric in nature and are interesting of themselves.

1. Introduction

The study of data bases has become an important area of computer science. In the light of public and private concern over the possible misuse of computerized data bases the need for security has received considerable attention. This need is filled by insuring that the information in the data base is only available to those who should have access to it. This in turn means insuring first that people who should not have access to any part of the data base do not and second that people who should only have access to part of the data base cannot access the remainder. These are two distinct problems. The former is related to the security of operating systems and computer systems as a whole and will not be considered in this paper. The latter problem is usually handled by restricting certain parts of the data base so that the nonprivileged user cannot access them at all. This approach is effective and fairly easy to implement but it can be too severe for some applications. In particular there are occasions where each individual datum is privileged but statistical information concerning a larger sample of the data is not privileged and should be accessible to all users. For example, in a university data base individual students' grades would be privileged, but the average grade of a class or in a given field should be considered general information. This paper will be concerned with a method of allowing such statistical access.

2. An Example

A data base generally consists of a number of records each of which contains several fields of information. As an example consider the data base of figure 1 that depicts nine fictional students. Each student is represented by a record in the data base corresponding to a line of the figure. This record is composed of six fields, the student's name, his class, his geographical origin, his major area, his family income and his grade point average. Suppose that the grade point average of any individual student is privileged information but statistical information such as the mean or median grade point average of a group of students is not privileged. This is a reasonable assumption since it allows useful information to be obtained from the data base without necessarily violating the privacy of any individual student. The set of legal requests or queries of the data base would include such questions as "What is the mean grade point average of all members of the class of '78?" or "What is the mean grade point average of all students whose major area is in the humanities?" Twelve such queries that can be asked of the data base in figure 1 are summarized in figure 2. Our goal is to establish when a data base is secure with respect to a set of queries. To show that the data base of figure 1 is secure with respect to the twelve queries of figure 2 we must show that no individual student's grade point average can be obtained from the given information.

In a slightly more complex system other queries such as "What is the mean grade point average of all social science majors from the west?" can be asked. This query refers to a single student and hence if an answer to it is given the data base is compromised. To prevent this from occurring some restriction must be placed on the set of queries that can be asked. The simplest such restriction is to insure that all legal queries refer to at least some minimum number of records. For instance in the example given we could restrict queries so that each must refer to at least three students. Then the twelve queries of figure 2 would all be legal as would more complicated ones that still involve at least three students.

Using sophisticated queries it is possible to ask for the grade point average of any set of three or more students. Since we are only interested in the security of individual grade point averages, this obviates the need for the other fields of the data base and allows us to consider the data base shown in figure 3 consisting solely of each student's name and his grade point average. A query here consists of naming a set of students and returning some statistical function of their grade point averages. In particular, the queries we are considering require a set of three or more students and request the mean of their grade point averages.

A data base is secure if the user cannot determine any information which he should not know. Our simplified data base would be secure if a user could not determine the grade point average of any individual student by asking queries regarding the mean grade point average of sets of three or more students. However, if the user can ask enough such queries he can determine all of the individual grade point averages by solving a system of linear equations in nine unknowns. It can also be shown that any individual student's grade point average can be

Student	Grade Point Average
A	4
B	3
C	2
D	4
E	1
F	4
G	4
H	2
I	3

Figure 3. Simplified Sample Data Base

Query 1: Mean Grade Point Average of $\{A, B, C\} = Q_1$

Query 2: Mean Grade Point Average of $\{B, C, D, E\} = Q_1$

Quary 3: Mean Grade Point Average of $\{A, D, E\} = Q_3$

$$\text{Grade Point Average of Student A} = \frac{1}{2}[3*Q_1 + 3*Q_3 - 4*Q_2]$$

Figure 4: Compromising the Data Base with 3 Queries.

if it is in this set.

We can now formally define when a data base is compromised.

Definition: An allowed sequence of queries $q_1, \ldots, q_m$ *compromises* a data base $D = \{d_1, \ldots, d_n\}$ if and only if there is some i such that for any data base $D' = \{d'_1, \ldots, d'_n\}$ with the same response to $q_1, \ldots, q_m$ as D, $d_i = d'_i$.

This definition simply says that a data base is compromised by some set of queries if the value of some element of the data base is uniquely determined by the responses to the queries.

We are now ready to define the concepts of security problem and security measure which are the focus of this paper. Recall that there are three factors we will consider that influence a security measure. The first two of these, the number of records in a query and the amount of overlap between two queries, are incorporated in the notions of allowed query and allowed query sequence respectively. The third factor relates to a portion of the data base being known in advance. We incorporate this factor in our definition of security problem.

Definition: A *security problem* is a 3-tuple $<D,D_0,\overline{Q}>$ where D is the particular data base, D_0 is the subset of D known to the user before any queries are asked, and $\overline{Q}$ is the set of allowed query sequences.

A security measure is defined in terms of a specific security problem. In particular, the *security measure* of a security problem $<D,D_0,\overline{Q}>$ is the minimum length of an allowed query sequence that compromises the data base D provided D_0 is known.

We begin our study of this security problem by establishing a good lower bound for its security measure. Using an extension of the proof technique of [2] where

$$S(n,k,r,\ell) \geq 1 + \frac{k-(\ell+1)}{r}$$

was shown, we prove

Theorem 1: $S(n,k,r,\ell) \geq \frac{2k-(\ell+1)}{r}$.

Proof: Suppose that after t queries we can determine the value of $d_{\ell+1}$ and that $d_1, \ldots, d_\ell$ are the elements of the data base known in advance. Let the queries be represented as

$$q_i = \sum_{j=1}^{k} d_{m_{ij}} \text{ , } i = 1, \ldots, t$$

where $1 \leq m_{ij} < \ldots < m_{ik} \leq n$ for $1 \leq i \leq t$ and where $\{m_{i1}, \ldots, m_{ik}\} \cap \{m_{j1}, \ldots, m_{jk}\}$ contains at most r elements for $i \neq j$. Then $q_1 \ldots q_t$ is an allowed query sequence. We can represent the fact that $d_{\ell+1}$ can be determined from these queries by

$$\sum_{i=1}^{t} \alpha_i q_i = \sum_{j=1}^{\ell+1} \beta_j d_j \text{ , } \beta_{\ell+1} \neq 0, \quad \alpha_i \neq 0 \quad 1 \leq i \leq t. \tag{1}$$

The left-hand side can be rewritten as

$$\sum_{i=1}^{t} \alpha_i q_i = \sum_{i=1}^{t} \alpha_i \sum_{j=1}^{k} d_{m_{ij}} = \sum_{\sigma=1}^{n} \left(\sum_{i=1}^{t} \alpha_i \delta_{i\sigma} \right) d_\sigma \tag{2}$$

where $\delta_{i\sigma} = 1$ if d_σ is used in query q_i and $\delta_{i\sigma} = 0$ otherwise. From (1) and (2) at most $\ell + 1$ of the terms $\sum_{i=1}^{t} \alpha_i \delta_{i\sigma}$, $1 \leq \sigma \leq n$ are non-zero.

appear in both queries. Then because of the first query we need

$$\frac{k - (L1 + L3 + 1 + L5)}{r}$$

additional queries with a negative α_i, and because of the second query we need

$$\frac{k - (L2 + L3 + L4 + L5)}{r}$$

additional queries with a positive α_i. Thus the total number of queries needed is:

$$2 + \frac{k - (L1 + L3 + 1 + L5)}{r} + \frac{k - (L2 + L3 + L4 + L5)}{r}$$

$$= 2 + \frac{2k - (L1 + L2 + L3 + 1) - (L3 + L4 + L5) - L5}{r}$$

But $L1 + L2 + L3 \le \ell$ and $L5 \le L3 + L4 + L5 \le r$ and hence

$$\ge 2 + \frac{2k - (\ell + 1) - r - r}{r}$$

$$= \frac{2k - (\ell + 1)}{r}. \qquad \square$$

This theorem can be used to show that the upper bounds achieved in [2] for this security measure are optimal for sufficiently large data bases. These upper bounds were given as:

(a) $S(n, k, 1, 0) \le 2k - 1$ for $n \ge k^2 - k + 1$

(b) $S(n, k, 1, 1) \le 2k - 2$ for $n \ge (k - 1)^2 + 2$

(c) $S(n, kp+d, p, 2d-1) \le 2k$ for $n \ge k^2p + 2d$

(d) $S(n', kr, r, r-1) \le S(n,k,1,0)$ for $n' \ge rk^2$.

5. A Special Case

In this section we will consider a special case of the security problem presented in section 4 in which no portion of the data base is known in advance. This case is interesting because of the contrast between the behavior of our security measure when no data is known in advance and its behavior when data is known in advance. From corollary 2c) we know that for any fixed overlap we can achieve the lower bound of theorem 1 for infinitly many k when at least one element of the data base is known. The best result that can be obtained with the methods of [2] for the case $S(n,k,r,0)$ where nothing is known in advance is $2k - 1$ queries regardless of the value of r. Ideally when we allow an overlap of about $r = \frac{k}{c}$ elements for some constant $c \geq 1$, a constant number,

$$\frac{2k}{r} \approx \frac{2k}{k/c} = 2c,$$

of queries should suffice. While this doesn't seem possible here, we will prove a significant improvement over previous results by showing an $O(\sqrt{k})$ upper bound on the number of queries needed.

We first consider the case where we allow an arbitrary amount of overlap.

<u>Theorem 3</u>: $S(n,k^2 + 1, k^2, 0) \leq 2k + 2$ for $n \geq 2k^2 + k + 1$.

<u>Proof</u>: Let the data base be $D = \{d_0, \ldots, d_{n-1}\}$ and let $y_i = \sum_{j=1}^{k} d_{ik+j-k}$ for $1 \leq i \leq 2k + 1$. Then let the queries be

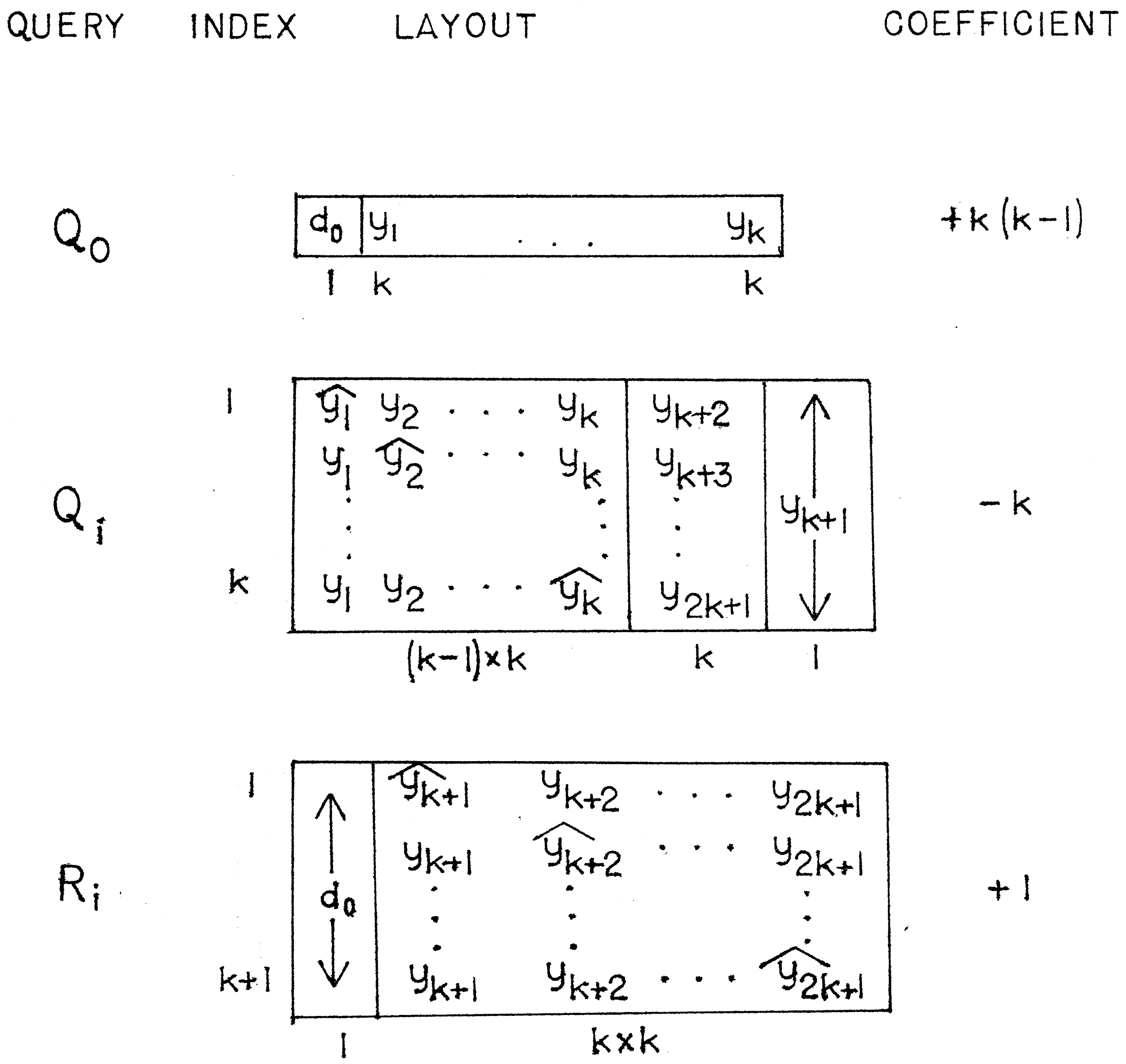
QUERY
INDEX
LAYOUT
COEFFICIENT
Q_0
d_0
y_1 . . . y_k
1
k
k
$+k(k-1)$
Q_i
1
k
$\hat{y}_1$ y_2 . . . y_k
y_1 $\hat{y}_2$ . . . y_k
y_1 y_2 . . . $\hat{y}_k$
y_{k+2}
y_{k+3}
y_{2k+1}
y_{k+1}
$(k-1)\times k$
k
1
$-k$
R_i
1
k+1
d_0
$\hat{y}_{k+1}$ y_{k+2} . . . y_{2k+1}
y_{k+1} $\hat{y}_{k+2}$ . . . y_{2k+1}
y_{k+1} y_{k+2} . . . $\hat{y}_{2k+1}$
1
$k\times k$
$+1$

Figure 5: Queries for Theorem 3

Then let the queries be

$$Q_0 = d_0 + \sum_{i=0}^{\ell-1} \sum_{j=1}^{pq/\ell} d_{ij}$$

$$Q_i = \sum_{j=1}^{pq/\ell} d_{(i \bmod \ell)j} + \sum_{j=1}^{p} \sum_{k=1}^{q-\frac{q}{\ell}} e_{jik} + f_i, \qquad 1 \le i \le q$$

$$R_i = d_0 + \sum_{j=1}^{p} \sum_{k=1}^{q} e_{jki}, \qquad 1 \le i \le q - {}^{q}/\ell$$

$$S_i = d_0 + \sum_{j=1}^{q} f_i + \sum_{j=1}^{q} \sum_{k=1}^{p-1} g_{ijk}, \qquad 1 \le i \le p$$

$$T_i = d_0 + \sum_{j=1}^{p} \sum_{k=1}^{q} g_{jki}, \qquad 1 \le i \le p - 1$$

Then d_0 can be computed as:

$$d_0 = \frac{1}{pq+1} \left[\frac{pq}{\ell} Q_0 - p \sum_{j=1}^{q} Q_j + p \sum_{j=1}^{q-\frac{q}{r}} R_j + \sum_{j=1}^{p} S_j - \sum_{j=1}^{p-1} T_j \right]$$

The number of queries used is $1 + q + (q - {}^{q}/\ell) + p + (p - 1) = 2p + 2q - {}^{q}/\ell$.

The layout of the queries can be seen in figure 6. □

In particular, for sufficiently large n and appropriate p and q, we have

$$S(n,\ pq + 1,\ {}^{pq}/2,\ 0) \le 2p + {}^{3}/2\ q.$$

$$S(n,\ pq + 1,\ {}^{pq}/3,\ 0) \le 2p + {}^{5}/3\ q.$$

$$S(n,\ pq + 1,\ q + 1,\ 0) \le 2p + 2q - 1.$$

Setting p = q shows that these are all about $4\sqrt{K}$ where K is the query size.

and

$$Q'^{k}_{\text{mean}} = \{ \sum_{\ell \in S} d_\ell \mid S \subseteq D \text{ and } |S| \geq k\}.$$

For this problem we define the security measure $S'(n,k,r,\ell)$ where $n = |D|$, $\ell = |D_0|$ and k and r come from $\overline{Q}'^{k,r}_{\text{mean}}$.

We first note that the lower bound established in theorem 1 for $S(n,k,r,\ell)$ also holds for $S'(n,k,r,\ell)$. This follows since the proof of the theorem relies only on limited overlap and the fact that at least k elements must appear in each test. Hence we have

$$S'(n,k,r,\ell) \geq \frac{2k-(\ell+1)}{r} .$$

With our original model we found that this bound is actually achievable for infinite families of n, k, r and ℓ, but that for some cases there does not seem to be a way to even approach it. In contrast to this, the extended model can almost always achieve an upper bound within one query of the lower bound.

Theorem 5:

$$S'(n,k,r,\ell) \leq \left\lceil \frac{2k-(\ell+1)}{r} \right\rceil + 1 \text{ for } \ell \leq 2r - 1 \text{ and } n \geq r\left\lceil\frac{k}{r}\right\rceil^2 + \ell + 1.$$

Proof: Choose ℓ_1 and ℓ_2 such that

$$0 \leq \ell_1 \leq r - 1$$

$$0 \leq \ell_2 \leq r$$

$$\ell_1 + \ell_2 = \ell$$

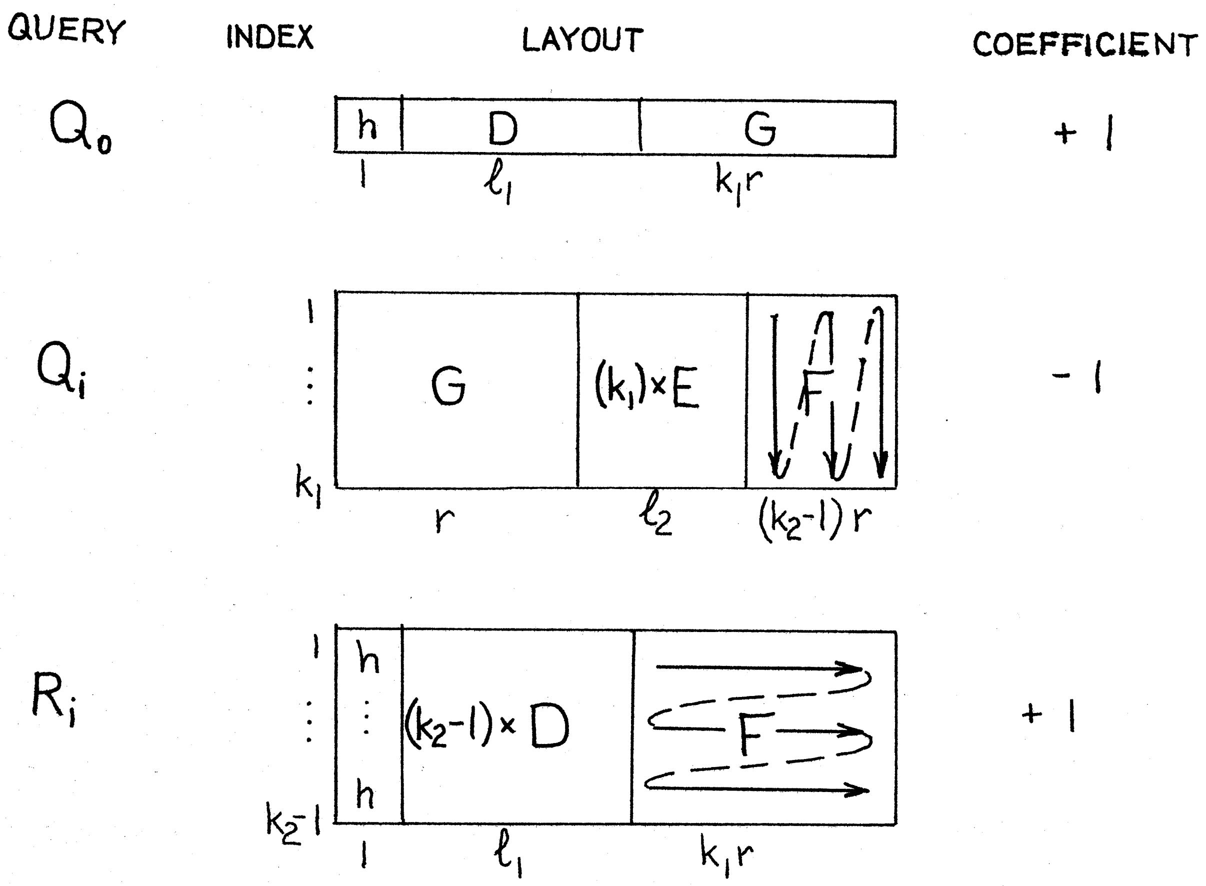

Figure 7: Queries for Theorem 5

7. Medians

So far we have considered only security problems in which a query requests the mean of a set of elements of the data base. In this section we will consider a security problem in which we are interested in the median. In particular we consider the problem $<D, \emptyset, \bar{Q}^{k,r}_{median}>$ where D is the data base, nothing is known in advance, and $\bar{Q}^{k,r}_{median}$ is the set of allowed query sequences. We define

$$\bar{Q}^{k,r}_{median} = \{q_1 \ldots q_m \mid q_i \in Q^k_{median} \text{ and for } 1 \leq i < j \leq m, |S_i \cap S_j| \leq r\}$$

where

$$Q^k_{median} = \{\underset{\ell \in S}{\text{median}} \{d_\ell\} \mid S \subseteq D \text{ and } |S| = k \text{ and } k \text{ is odd}\}$$

and where S_i is the set S used in query q_i.

As mentioned in [2], there are two different security measures to be considered for this security problem. A set of queries involving the median can give information with one sequence of answers and yet the same queries with different responses may give no information. To see this, consider the example in figure 8. If the three queries yield the results shown for trial 1, then it is fairly simple to show $d_0 = 10$ (see theorem 7). On the other hand, if the queries all yield the same responses as in Trial 2 then no specific information is forthcoming. Ideally a security measure for this model would take into account the actual data values. However, this is neither appropriate nor possible in an abstract model and hence two distinct security measures are applicable. The first considers worst case behavior, that is, given the worst data for guaranteeing security, how many queries are sufficient to compromise the data base. The second considers best case behavior, that is, given arbitrary data how many queries are required to insure that the data base is compromised. Of these two measures the first is the more practical. Unlike sorting and other computational problems where it is important to know what the longest computation can be, the interest here lies in the minimum number of queries necessary to compromise the data base. Moreover, if our model were to be used to insure the integrity of a data base, the first measure would provide an enforcable security scheme while the second would only alert the system when the data base must have been compromised. More formally, for $n = |D|$ and k and r from $\overline{Q}^{k,r}_{median}$, the two security measures will be denoted as $M(n,k,r)$ and $M'(n,k,r)$ respectively. Let $D_0, D_1, D_2, \ldots$ be all possible data bases such that $|D_i| = n$ and let m_i be the minimum size of an allowed query sequence that compromises D_i. Then we can define the measures as

$$M(n,k,r) = \min_i m_i$$

and

$$M'(n,k,r) = \max_i m_i .$$

In [2] and [3] the security measure $M'(n,k,r)$ is considered but only with respect to data bases with the additional property that all elements are unique. Let the two security measures restricted to such data bases be $M_{\neq}(n,k,r)$ and $M'_{\neq}(n,k,r)$ respectively. Then the following were shown:

a) $M'_{\neq}(n,k,k-1) \le \frac{3}{2}(k+1) + 1$ for $n \ge k + 2$

b) $M'_{\neq}(n,k,1) \le \begin{cases} k^2 + 1 & \text{if} \quad \text{is a prime power} \\ 4(k^2 + 1) & \text{otherwise} . \end{cases}$

We next show how to compromise a data base with three unlimited overlap queries. Let the data base be $D = \{d_1, d_2, \ldots, d_{k+2}\}$ and let the queries be

$$q_0 = \text{median } \{d_1, d_2 \ldots d_k\} = \alpha$$

$$q_1 = \text{median } \{d_{k+1}, d_2 \ldots d_k\} = \beta < \alpha$$

$$q_2 = \text{median } \{d_{k+2}, d_2 \ldots d_k\} = \gamma > \alpha.$$

From queries q_0 and q_1 we can easily deduce that $d_{k+1} < \alpha \leq d_1$ since this is the only conceivable explanation. Similarly, from queries q_0 and q_2 we can deduce that $d_{k+2} > \alpha \geq d_1$. But then $\alpha \geq d_1 \geq \alpha$ and hence $d_1 = \alpha$. □

<u>Theorem 8</u>:

a) $M(n,k,1) \geq \frac{3}{4}(k + 1)$ for $k \geq 3$

b) $M(n,k,1) \leq 3k - 5$ for $n \geq k^2 - 2k + 4$.

<u>Proof</u>: a) We first claim that in any query either at least $(k+1)/2$ of the elements must be referred to in another test or that the element being determined is in this query and at least $(k-1)/2$ elements are referred to in another test. Suppose not. Then there are $(k+1)/2$ elements in the query whose values are never needed. These are enough to insure that any desired median is achieved regardless of the values of the other elements and hence the query yields no useful information.

We know that at least one query is needed. This query must contain at least $\frac{k-1}{2}$ elements which appear in other queries as well as the element being determined. As the maximum overlap between any pair of queries is one, there must be $\frac{k-1}{2}$ queries to accomodate the $\frac{k-1}{2}$ elements. Each of these queries involves only one known element and hence must also include at least $\frac{k-1}{2}$ new elements. (The element being determined cannot appear

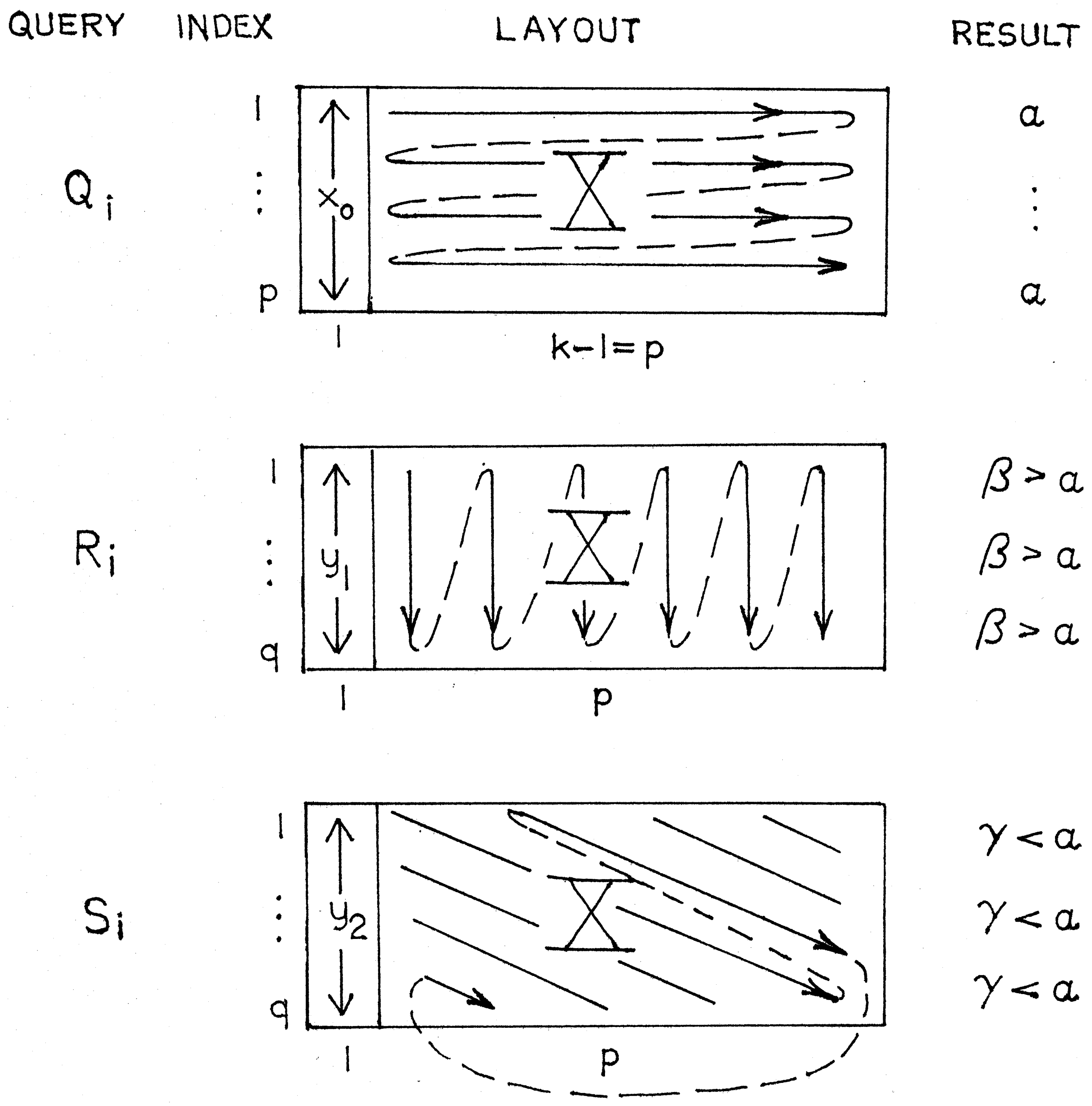

Figure 9: Queries for Theorem 8b.

8. Summary

In this paper we studied some security issues of computerized data bases using a limited abstract model. We considered problems that involved limiting the number and type of queries that are asked of a data base in order to insure that the data base is not compromised. The questions that these problems brought up are primarily combinatoric in nature and, although this study relates them to data bases and security, they are interesting of themselves.

In section 4 and 5 we considered the security measure $S(n,k,r,\ell)$ based on a data base where we only allow queries asking for the mean of exactly k elements. We proved a lower bound which can actually be achieved for a large class of such problems. However, for the problem where no elements are known in advance the best we could do did not even asymtotically approximate this lower bound. The question remains whether this bound is achievable in this case and, if not, if our upper bound of $O(\sqrt{k})$ is optimal. The contrast between our upper and lower bounds brought to life the open question of whether $S(n,k,r,\ell) \geq S(n,k',r,\ell)$ for $k \geq k'$. In section 6 the problems with this model are seen to be an outgrowth of the requirement that all rules involve exactly k elements of the data base. We showed that if this requirement is relaxed to allow queries that involve k or more elements then we can almost always demonstrate an upper bound within one query of the lower bound. Whether this result is optimal remains an open question. Finally in section 7 we considered a security problem where we allowed queries that ask for the median of k elements of the data base. While two differing security measures were considered here, we only presented results

Acknowledgements: I would like to thank Professors David Dobkin and Richard Lipton for their comments, my wife Loretta for her assistance, and Stephanie Lukes for her typing.