CUCS-36-82

STATISTICAL SECURITY OF A STATISTICAL DATA BASE

J.F. Traub
Department of Computer Science
Columbia University
New York, New York 10027

Y. Yemini
Department of Computer Science
Columbia University
New York, New York 10027

H. Woźniakowski

Institute of Informatics
University of Warsaw
Warsaw, Poland

and

Department of Computer Science
Columbia University
New York, New York 10027

September 1981
Revised December 1982

This research was supported in part by the National Science Foundation under Grants MCS-782376 and MCS-8110319. J.F. Traub and Y. Yemini wish to acknowledge the support of the Advanced Research Projects Agency under contract N00039-82-C-0427.

ABSTRACT

This paper proposes a statistical perturbation scheme to protect a statistical database against compromise. The proposed scheme can handle the security of numerical as well as non-numerical sensitive fields or a combination of fields. Furthermore, knowledge of some records in a database does not help to compromise unknown records. We use Chebychevs inequality to analyze the tradeoffs between the magnitude of the perturbations, the error incurred by statistical queries and the size of the query set to which they apply. We show that if the statistician is given absolute error guarantees, then a compromise is possible but the cost is made exponential in the size of the database.

1. INTRODUCTION

The problem of security of a statistical data base involves three hypothetical individuals: the statistician, whose interest is to obtain aggregate data (means, medians, frequency) from the data base, the owner of the data who wishes to secure individual records, and the data base administrator, who needs to satisfy both. The interested reader may consult [8, 1, 2, 5, 6, 7] for a survey of the problem and for further references.

One possible approach to a solution of the statistical data base security problem is to restrict the size and overlap between query sets[1] available to the statistician. However, the major result of a series of papers is that "Compromise is straightforward and cheap", to quote the conclusions of [1]. Accordingly, it is suggested that "The requirement of complete secrecy... is not consistent with the requirement of producing exact statistical measures... At least one of these requirements must be relaxed...".

In this paper we study alternative solutions to the problem of securing statistical data bases, based on random perturbations of the data base. We quantify and analyze the tradeoffs between "security" of the database and precision of the data extracted through statistical queries, using Chebyshev's inequality.

Securing data through statistical perturbations has been previously considered; see for example [8]. The idea is quite simple, numerical fields in the data base are randomly perturbed; a statistical query provides an estimator of the required quantity. If a query set is sufficiently large, the law of large numbers causes the error in the query to be significantly less than the perturbations of individual records. The dangers involved in a statistical perturbation scheme are:

1. that the data base may be compromised if the statistician is allowed sufficiently many independent estimates;
2. the statistician is not guaranteed error bounds on query responses.

---

[1] A query set is a set of records whose aggregate statistics is subject to a query.

Recently, L. Beck [7] suggested a statistical perturbation scheme which parametrizes the variance of the perturbations. He demonstrates there that a proper parametrization guarantees the data base against a statistical compromise using linear queries. There are two limitations to this method: First, protection is only proved for restricted forms of attack (e.g., linear queries). Second, the statistician may only be provided with statistical error guarantees (i.e., bounds on the variance of the estimator used). It is possible that the result of any specific query grossly deviates from the actual value of the answer. These two limitations are removed by the scheme proposed in this paper.

To introduce the proposed scheme consider, for example, a data base containing employee data where one of the fields indicates salary. Let $\underline{d}=(d_1, d_2, \ldots, d_n)$ describe the salary field of the n records in the data base. Let $C \subset (1,2,\ldots,n)$ be a query set and let $|C|$ be the number of elements in C. A linear statistical query over C is the function:

$$q_C(\underline{d}) = \sum_{i \in C} d_i$$

The values of individual d_i can be compromised in spite of query set size and overlap restrictions; see[1,2].

Let us assume that rather than storing the actual vector $\underline{d}$ in the data base, a vector $\underline{d}'$ is stored, where $\underline{e} = \underline{d}' - \underline{d}$ is a random perturbation vector, whose components are independent random variables whose mean is 0. Note that unlike the scheme proposed by Beck [7], which allows the statistician to obtain unlimited number of sample estimates to any given query, the perturbed data is fixed for the full database and the perturbation is not changed from query to query.

The statistician may now make any query of his choice, he may even acquire the vector $\underline{d}'$. However, by properly selecting the perturbation, knowledge of $\underline{d}'$ does not constitute a danger of compromising the actual records $\underline{d}$. On the other hand, the statistics of $\underline{d}'$ may be used to provide a fairly accurate description of the statistics of $\underline{d}$. Note again that since the perturbation is fixed, it is impossible for the statistician to improve the estimate of any given query result by repeating the query. Therefore the concept of a statistical compromise in

the sense of [7], measuring the variance of the estimator of individual records, is not relevant here. The only estimator available for any given record is its perturbed value d'_i .

To fix these ideas, assume that the perturbations e_i are identically distributed with variance σ^2 and consider a linear query q_C. Chebyshev's inequality yields:

$$\text{Prob}\{|q_C(\underline{d}') - q_C(\underline{d})| \geq \varepsilon|C|\} \leq \frac{\sigma^2}{|C|\varepsilon^2} \quad .$$

The left hand side represents the probability that the error in the query exceeds some bound. The right hand side bounds this probability in terms of the variance of individual perturbations, the error bound sought and the size of the query set. The statistician would like to keep the probability of large errors in the result of a query small, while the data base owner would like to keep the perturbations (i.e., σ) in individual records large so as to protect individual records. The data base administrator can choose σ to satisfy both.

To illustrate this solution, let us consider a numerical example. Suppose the average salary in the data base is \$20,000 and the data base owner would like the standard deviation to be at least 20% of this average. The statistician, on the other hand, would like the error in queries not to exceed \$1,000. The data base administrator can set σ=4000 and ε=1000 and thus guarantee the statistician that the error $\frac{1}{|C|}|q_C(\underline{d}') - q_C(\underline{d})|$ exceeds \$1,000 with probability smaller than $16/|C|$. Therefore the data base owner can be fully satisfied that individual records are adequately perturbed while the statistician gets his queries answered within the required accuracy with probability which depends on the size of his query sets. The larger the query set, the smaller the probability of error.

Put another way, the error bounds agreeable to the statistician, ε, and the probability, $1 - \delta$, that this error bound holds, depend on the size of the query set $|C|$ and the standard deviation of the perturbation (as determined by the desires of the data base owner) through the formula:

$$\varepsilon^2\delta|C| \leq \sigma^2.$$

Once the data base administrator determines the right hand side, the tradeoffs between query size, error bound and error probability are given by the left hand side. This inequality gives an uncertainty principle among the error in queries, the probability of error, and the size of the query set. Also note that the magnitude of the perturbations in individual records, measured by σ, may usually be substantially larger than the magnitude of the error in the queries. The statistician takes advantage of the law of large numbers to increase the precision of the results of queries by applying queries to large query sets.

The above idea generalizes to an arbitrary numerical field $\underline{d}$ in a trivial manner.

2. Variations on the Basic Idea

A number of objections might be raised to the above protection scheme. We list four of them and show how our basic protection scheme might be modified to meet these objections.

(i) A first objection might be that the error guarantees provided to the statistician are probabilistic. Given a perturbed data base $\underline{d}'$, a query $q_C(\underline{d}')$ may result in a large error relative to $q_C(\underline{d})$. This may happen even when the query set C is very large. Chebyshev's inequality provides an assurance that such an error is unlikely for a given query set and a random perturbation. However, once $\underline{d}'$ is selected, clearly a large error is possible for some queries.

If the statistician is not content to live with occasional large errors, there is a variation of the scheme, to be described below, which will permit the data base to be statistically compromised but at such a high cost that this compromise is not feasible.

The variation is to monitor the error in the query $|q_C(\underline{d}') - q_C(\underline{d})|$ and to check whether it exceeds the error bound requested by the statistician. If the error bounds are exceeded and if the size of the query set $|C|$ exceeds a certain threshold size n_0, set by the data base administrator, then a correction mechanism is invoked. The correction mechanism simply perturbs $q_C(\underline{d})$ by adding some random variable (with 0-mean and σ^2_1-variance) until adequate error bounds are met. That is, we add per(C) to $q_C(\underline{d})$ where $|per(C)|$ does not exceed the error bound requested by the statistician.

While this scheme may keep the statistician happy, the data base is no longer secure (in a statistical sense). An attack on the data base would first identify query sets C, of size greater than n_0, that result in an excessive error and thus cause the correction mechanism to be invoked. We call such sets compromising sets. Once compromising sets are identified, a statistical compromise of the data base is easy.

We illustrate this by an example. Suppose that the data base is perturbed such that $\underline{d}' = \underline{d} + \underline{e}$ where $\underline{e} = (e_1, \ldots, e_n)$ with $e_i = \pm\sigma$ and + and − are selected with probability .5. Without loss of generality, assume that n is even, $n = 2p$, and the number of + and − is p. Let every query be equiprobable. Then the average query length is

$$2^{-n} \sum_{k=0}^{n} k\binom{n}{k} = \frac{n}{2} .$$

Suppose that the error agreeable to the statistician is

$$|q_C(\underline{d}') - q_C(\underline{d})| \leq \varepsilon \frac{n}{2}$$

for some (presumably small) positive ε. Define

$$I = \{i: e_i = +\sigma\}.$$

Then $|I| = p$. If $C \subset I$ and

$$|C| > a = \max(n_0, \left|\frac{\varepsilon n}{2\sigma}\right|)$$

then C is a compromising set. Note that $a << n$ for n_0, ε and σ of practical interest. Let

$$k = \left\lfloor \frac{p-1}{a} \right\rfloor = \left\lfloor \frac{n-2}{2a} \right\rfloor .$$

Let $I = \{i_1, i_2, \ldots i_p\}$. Define the query set C_j as

$$C_j = \{i_{(j-1)a+1}, i_{(j-1)a+2}, \ldots, i_{(j-1)a+a}, i_p\}, \quad j = 1, 2, \ldots, k$$

$$C_{k+1} = \{i_1, i_2, \ldots, i_{p-1}\}.$$

Then $|C_j| = s+1$, $j = 1,2,\ldots,k$, $|C_{k+1}| = p-1$ and $C_j \subset I$. Thus all the sets C_j are compromising sets.

Since C_j is a compromising set we get

$$x(C_j) = q_{C_j}(\underline{d}) + per(C_j), \quad j = 1,2,\ldots, k+1.$$

Knowing $x(C_j)$ we compute

$$x = \frac{1}{k}(-x(C_{k+1}) + \sum_{j=1}^{k} x(C_j)).$$

Let $m = i_p$. It is easy to check that

$$x = d_m + \frac{1}{k}(- per(C_{k+1}) + \sum_{j=1}^{k} per(C_j)).$$

Chebyshev's inequality yields

$$Prob\{ \frac{1}{k} |-per(C_{k+1}) + \sum_{j=1}^{k} x(C_j)| \geq \delta\} \leq \frac{k+1 \quad \sigma_1^2}{k^2 \delta^2}.$$

Since k is large, $x - d_m$ is small with probability close to one. Thus x is a good approximation of the individual record d_m. This proves a statistical compromise of the data base.

Therefore the problem of compromise boils down to that of identifying compromising sets. Let us first note that if the error bound ε is sufficiently small then every perturbation scheme has compromising sets. Indeed, let $C = \{i: e_i > 0\}$. Then

$q_c(\underline{d}') - q_c(d) = \sum_{i \in C} e_i$. If $\varepsilon \frac{n}{2} < \sum_{i \in C} e_i$ then the absolute error exceeds $\varepsilon \frac{n}{2}$ and C is a compromising set.

We now derive the complexity of identifying a compromising set for the example presented

above. For simplicity we set the threshold size $n_0 = 0$. We first note that the number of query sets for which the absolute error is s, i.e., $|q_C(\underline{d}') - q_C(\underline{d})| = |\sum_{i \in C} e_i| = s$, is equal to

$$\sum_{i=0}^{p} \binom{p}{i} \left[\binom{p}{i-s} + \binom{p}{i+s} \right] = 2\binom{2p}{p-s},$$

due to [4]. The number of query sets for which the correction mechanism is invoked, i.e., $s > \frac{en}{2}$, is equal to

$$N = 2\sum_{s=s^*}^{p} \binom{2p}{p-s} = 2\sum_{i=0}^{p-s^*} \binom{2p}{i}, \quad s^* = \left\lfloor \frac{en}{2} \right\rfloor + 1.$$

It is known, see for instance [3], that for large n, $N \leq 2^{n+1}\, q(\varepsilon)^n$ where

$$q(\varepsilon) = \frac{1}{(1-\varepsilon)^{\frac{1-\varepsilon/2}{}} (1+\varepsilon)^{\frac{1+\varepsilon/2}{}}}$$

and $\frac{1}{2} < q(\varepsilon) < 1$.

From this we conclude that the probability of the query sets for which the correction mechanism is invoked satisfies the inequality

$$\text{Prob}(\{C: |q_C(\underline{d}') - q_C(\underline{d})| > \varepsilon \tfrac{n}{2}\}) \leq 2q(\varepsilon)^n.$$

Thus, it decreases exponentially to zero with n. This proves that almost all queries are answered within the acceptable error bound and the correction mechanism is invoked extremely rarely.

How hard is it to find a compromising set? More precisely, what is the minimal number of queries to find a compromising set with probability $1 - \delta$? If we have k queries, $k \ll 2^n$, the probability that all of them are answered within the acceptable error bound is approximately at least $(1-2q(\varepsilon)^n)^k$. Thus $1 - (1-2q(\varepsilon)^n)^k$ is an upper bound on the probability of finding a compromising set. Hence k satisfies the inequality

$$1 - (1-2q(\varepsilon)^n)^k \geq 1 - \delta$$

which yields

$$k \geq \frac{\ln \delta}{\ln (1-2q(\varepsilon)^n)} \cong \frac{\ln \delta^{-1}}{2} \left(\frac{1}{q(\varepsilon)}\right)^n.$$

This proves that one has to have an exponential number or queries to find a compromising set. This makes compromising the data base infeasible.

(ii) A second objection that might be raised to the proposed statistical security scheme is that the perturbed vector $\underline{d}'$ requires additional storage. This is easily circumvented by perturbing the elements of $\underline{d}$ every time they are accessed. To guarantee that a given query always receives the same answer, the perturbation of d_i may be obtained using a pseudo random generator with a fixed seed which is given by some function of i. Therefore, one has a choice of either storing the perturbed values or computing them every time they are needed.

(iii) A third objection might be that the perturbations of individual records should not be identically distributed. For instance, in the salaries example above, a \$4,000 perturbation may be suitable to hide information about salaries that do not exceed the average salary (\$20,000) too far. However, if one of the salaries is \$175,000, a perturbation of \$4,000 does not adequately protect the respective record. The discussion above assumed that a record is compromised only if its real value becomes available to the statistician. This example suggests that an alternative notion of compromisability is required. Given a constant c, let us define a perturbation scheme to be non-compromisable with respect to c if $cd_i < \sigma(e_i)$ for all records. That is, if the standard deviation of each perturbation exceeds a given fraction (c) of the magnitude of the respective item.

This problem may be solved using a multiplicative rather than an additive perturbation. That is, the perturbed record d'_i is generated by selecting a random factor a_i and multiplying by d_i. The perturbations caused by this process are given by $e_i = d'_i - d_i = d_i(a_i-1)$ and are thus proportional to the value of d_i. Let A indicate the diagonal matrix formed from the

multipliers a_i. The perturbed data is related to the original data through $\underline{d}' = A\underline{d}$. If the random variables $a_i - 1$ are independent and identically distributed with zero mean and variance σ^2 then the perturbation elements e_i are independent, have zero mean and a variance $d_i^2\sigma^2$, respectively. Clearly the scheme is non-compromisable (in the sense defined above) with respect to any constant that does not exceed σ. Furthermore, Chebyshev's inequality may be applied to obtain:

$$\text{Prob}\{|q_C(\underline{d}') - q_C(\underline{d})| > \varepsilon|C|\} \leq \frac{\sigma^2\sum d_i^2}{|C|^2\varepsilon^2} \leq \frac{\sigma^2 d^2}{|C|\varepsilon^2}$$

where d denotes the maximal element among the d_i

The tradeoffs analysis performed in the previous section thus trivially generalizes for this multiplicative perturbation scheme. One may generalize this last scheme to allow perturbations by arbitrary random matrices. It has been shown by S.L. Warner [9] that this general "regression" perturbation of data might be applied to perturb non-numerical statistical data, as well as combinations of sensitive fields in the database. A detailed discussion of perturbation through matrix multiplication is beyond the scope of this paper; such details are available in [9]. Again, the tradeoff analysis easily generalizes to the case of arbitrary random matrix perturbation.

(iv) A fourth objection might be that the above perturbation scheme only applies to linear queries. This objection is valid since some queries (e.g., what is the maximal salary?) cannot benefit from the effects of the law of large numbers when one uses a statistical perturbation scheme. From a security point of view the perturbation schemes discussed above offer no risk when it comes to an arbitrary query: the statistician may acquire a complete knowledge of the perturbed values $\underline{d}'$ without compromising the real data. However, the statistician may be unhappy with the answers if they deviate grossly from the real value. A possible solution is to generalize the process used to answer objection (i) above. Given an arbitrary query, define its query set as the set of records which are essential to the answer (a record is essential to a query if its omission from the database will change the answer). The modified mechanism to handle arbitrary queries is invoked only when the size of the respective query set exceeds some preset threshold. It proceeds by computing both the answer to

the query based upon the perturbed data and an answer based upon the real data. If the two answers deviate by more than $\varepsilon|C|$, where ε is the error guarantee provided and C is the respective query set, then a new perturbation of the real answer is sampled until the error bounds are met. An arbitrary query may be represented as a function $Q(\underline{d})$ applied to the database. If the function Q satisfies some smoothness conditions (e.g., $|Q(\underline{d}) - Q(\underline{d}')| \leq K|\sum e_i|$ where K is a constant), it is possible to generalize the results of (i) to prove that compromise is exponentially hard. A general proof is beyond the scope of this paper.

3. ACKNOWLEDGMENTS

The authors gratefully acknowledge the helpful and constructive comments of the referees.

REFERENCES

[1] Denning, D.E., Denning, P.J., and Schwartz, M.D.
The Tracker: A Threat to Statistical Database Security.
ACM Trans. Database Syst. Vol. 4 :76-96, 1979.

[2] Dobkin, D., Jones, A.K., and Lipton, R.
Secure Databases: Protection Against User Influence.
ACM Trans. Database Syst. 4 :97-106, 1979 .

[3] Hamming, R.W.
Coding and Information Theory.
Prentice-Hall, 1980.

[4] Knuth, D.E.
The Art of Computer Programming, Vol. 1.
Addison-Wesley , 1973.

[5] Schlorer, J.
Disclosure From Statistical Databases: Quantitative Aspects of Trackers. .
ACM Trans. Database Syst. 5 :467-492, 1980.

[6] Ullman, J.D. .
Principles of Database Systems.
Computer Science Press, 1980.

[7] Beck, L.
A Security Mechanism for Statistical Data Bases.
ACM Transactions of Databases, 5,1 , 1980.

[8] Conway, R. and Strip, D.
Selective Partial Access to a Data Base.
In Proc. ACM Nat'l. Conf., pages 85-89. ACM, Oct. 1976 .

[9] Warner, S.L.
The Linear Randomized Response Model.
J. Amer. Stat. Assoc. 66 :884-888, 1971.