THE PROTECTION OF CONFIDENTIAL DATA STORED IN A SEQUENTIAL ACCESS STATISTICAL DATABASE

E. Noren and S. Keller-McNulty
Win Noren, Department of Statistics, Kansas State University, Manhattan KS 66506

Key Words: Tracker Attacks, Compromise, Disclosure Risk

Abstract
Methods of protecting sequentially accessed statistical databases from unauthorized use have been investigated by both statisticians and computer scientists. Proposed methods of control include minimum query set size restrictions, uniform rounding, slicing, partitioning, random sample queries, and data imputation. The effectiveness of these control methods against database attacks such as tracker attacks; i.e. the manipulation of legitimate marginal data, is evaluated.

1. Introduction

Agencies and individuals gather information in many ways. Private companies conduct telephone polls, news organizations conduct surveys, the United States Bureau of the Census conducts surveys and censuses. All of this information is stored in a variety of ways. Many companies form databases to store the information. This information is used by the companies for internal purposes as well as by outside researchers. In some instances detailed microdata may need to be released in some form.

This paper briefly looks at the federal legislation that governs the collection and release of information. For a detailed review of the legislation see Cecil (1993). Next methods of data storage and release are described and a definition for data disclosure is given. Finally a selected method of protecting data is described and evaluated. An expanded version of the material in this paper can be found in Noren (1993).

2. Privacy

Individuals, as well as organizations such as businesses, have a right to privacy. A general interpretation of the right to privacy, is the right to control information about oneself. This also implies a right to limit the intrusiveness of data gathering. A balance must be made between society's need to gather information and the individuals right to privacy (ASA 1977). This can be seen in the various legislation that controls federal agencies. Three important acts of legislation are briefly discussed.

2.1. Title 13

Title 13 of the U.S. Code is legislation that governs the U.S. Bureau of the Census. It contains guidelines for the disclosure of census records and it requires everyone to respond to the decennial census questionnaires. Section 9 (a) of Title 13 requires:

(1) the use of Census Bureau information for statistical purposes only;
(2) that information be published only in a way that prevents the identification of an individual;
(3) employees to take an oath to uphold Title 13 before examining individual reports.

Title 13 permits "statistical compilations," such as tables, to be published, as long as individual information is not released (§8 b).

2.2 Freedom of Information Act

The Freedom of Information Act (Public Law 89-487 and Public Law 93-502) is concerned with public access to government information. In general, the government cannot withhold information that is requested. There are several restrictions relevant to statistical information:

(1) Information specifically exempted from disclosure by statue, such as Title 13;
(2) Trade secrets and commercial or financial information;
(3) Personal and medical files.

2.3. Privacy Act

The Privacy Act of 1974 is an attempt to afford comprehensive protection to an individual's right to privacy by governing the collection, management, and disclosure of personal information maintained by government agencies (5 U.S.C. § 552 a). While recognizing that there are

legitimate uses for identifying information, the Privacy Act gives individuals some measure of control over that information. The Privacy Act requires federal agencies to:

(1) allow individuals access to their identifiable records;
(2) confirm that information is correct and timely, and restrict the gathering of unnecessary data;
(3) restrict the release of identifiable information to third parties (Cecil 1993).

Anonymous information is not regulated by the Privacy Act, and may be released to anyone. As long as the record is "in a form which is not individually identifiable" (§ 552 a (b) (5)), it may be released. While the legislation described does not place restrictions on non-Federal organizations, the impact of this legislation affects businesses and individuals alike.

3. Data Release and Disclosure

In dealing with Federal agencies, compromise of privacy is governed by legislation. Private companies do not have those restrictions, but are controlled by the media and public pressure. When data is released, a wealth of information is available for research. Some individuals desire more information than is directly released either out of curiosity or for malicious purposes and may seek to manipulate the data in order to determine previously unknown or unreleased and frequently sensitive information.

3.1. Data Storage and Release

A ***database*** consists of the organization of data records having several fields. The fields could contain information about attributes such as gender, salary, or occupation. Identifying information such as name, social security number, and address, may be included in the database, but may not be accessible to all users. This, however, is not necessarily enough control to prevent someone from directly identifying confidential information. A statistical database has been proposed to provided more protection. A ***statistical database*** is a database which enables users to access information only through aggregate statistics.

An ***intruder*** is an individual who exploits the vulnerability of the database to learn sensitive information, also called an ***attacker*** or ***data spy***. Frequently, the intruder is a legitimate user with various degrees of access to the database (Keller-McNulty and Unger 1993).

A user of a statistical database accesses the information through ***queries*** to the database. These queries may take on several forms. If a database contains information about gender, rank, department, and salary, a user may query the database for the sum of the salaries of all the women in the statistics department. To query a database, a ***characteristic formula***, is used. This consists of arbitrary logical formulas using category-values connected by the operators AND($\cap$), OR($\cup$), and NOT($^-$). For example Sum(Male $\cup$ Professor) would yield the sum of everyone who is either male or is a professor contained in the database (Denning, Denning, and Schwartz 1979). The sum query is denoted by

$$q(G) = \sum_{i \in G} y_i ,$$

where $G = \{1,2,\ldots,n_G\}$ is an index set for the records in the database defined by the characteristic formula G.

3.2 Data Disclosure

Keller-McNulty and Unger (1993) assert that two separate conditions are required for a disclosure to occur. First a specific entity (e.g. an individual, household, company, or industry) must be linked to one or more records or fields in the database. The second requirement is that confidential or sensitive information about the entity be learned by the intruder. This has also been called attribute disclosure. Duncan and Lambert (1989) point out that an attribute disclosure can occur with or without re-identification. It is possible to determine sensitive information about an individual even if that individual is not represented in the database.

4. Methods of Attack

It is understood that there is some risk in releasing information. This risk is related to the method the intruder uses to attack the database.

In particular, if a query set in the database or a cell in a cross tabular table contains too few entities, the value of the cell may need to be protected. Methods of attack can easily overcome simple suppression of sets or cells with small frequencies. These methods can be characterized as a solution to a system of linear equations and hence are called linear system attacks.

A special characteristic formula that allows the intruder to compromise the data with a linear systems attack is called a ***tracker***. The general principle behind the tracker, is that the intruder can compromise a database by posing a few answerable queries to a database that requires queries to be of some minimum size. It has been shown by Denning, Denning, and Schwartz (1979) that any characteristic formula T whose query set size is in the range [2k, N-2k], where k is the size of the minimum set that will not be suppressed and N is the size of the database, can be used as a tracker. This implies that the query T would be answerable by the database since the query set is well within the nonsuppressable range. The choice of the tracker T is somewhat arbitrary. To find a tracker, the intruder must find a formula T such that $2k \leq n_T \leq N\text{-}2k$. For example, if gender is a field in the database, it is likely that it can be used as a tracker. The value of any unanswerable query q(C) can be computed using a tracker T as

$$q(C) = q(C \bigcup T) + q(C \bigcup \bar{T}) - [q(T) + q(T)].$$

5. Methods of Control

Many methods of control that have been proposed to protect the confidentiality of the information stored in a statistical database. Little formal work has been done to evaluate the effectiveness of control techniques against an attack on the database using a tracker. In this section, a statistical framework for the evaluation of these control techniques is developed and applied to one control method.

Recall a sum query can be characterized as

$$q(G) = \sum_{i \in G} y_i$$

where the $i \in G$ denotes the records in the database that belong to the set characterized by G.

The tracker is found by asking the data base four queries and then adding and subtracting the results. The tracker equation partitions the database into four sets as diagramed below.

	T	$\bar{T}$
C	u	v
$\bar{C}$	w	x

$$q(C) = q(C \bigcup T) + q(C \bigcup \bar{T}) - q(T) - q(\bar{T})$$
$$= \sum_{i \in (u+v+w)} y_i + \sum_{i \in (u+v+x)} y_i - \sum_{i \in (u+w)} y_i - \sum_{i \in (v+x)} y_i$$

Depending on the control techniques, an individual query as well as q(C) may or may not return the true sum over the corresponding set.

A control method proposed by Venulapalli and Unger (1991), reports the results for a query by first perturbing each record, y_i, in the query set. The perturbation is defined as

$$Y_i^* = y_i + X_i H_i \bar{y},$$

where

$$X_i = \begin{cases} 1 & \text{with probability } p_1 \\ -1 & \text{with probability } p_2 \\ 0 & \text{with probability } 1 - p_1 - p_2 \end{cases}$$

$$H_i \sim U(L,U) \text{ for } 0 \leq L \leq U \leq 1,$$

and X_i and H_i are independent. The expected value and variance for a sum query are

$$E[q(G)] = E\left[\sum_{i \in G} Y_i^*\right] = n_G \bar{y}_G + n_G \bar{y}_G \left[\frac{U+L}{2}\right](p_2 - p_1)$$

and

$$Var[q(G)] = \frac{n_G \bar{y}_G^2}{12}[(p_1 + p_2)(4U^2 + 4UL + 4L^2) - 3(p_2 - p_1)^2(U + L)^2].$$

Notice that this technique results in a biased perturbation to the query set.

Assuming only one record in C, say in $(C \cap T)$, the cumulative effect of this perturbation

technique on the four queries required for a tracker attack results in

$$E[q(C)] = y_u + \frac{U+L}{2}[(p_2 - p_1)[(\bar{y}_{u+w} + \bar{y}_{u+x}) + n_x(\bar{y}_{u+x} - \bar{y}_x)]]$$
$$\approx y_u + \frac{U+L}{2}[(p_1 - p_2)(\bar{y}_{u+v} + \bar{y}_{u+x})]$$

and

$$Var[q(C)] = [n_{u+x}\bar{y}^2_{u+x} + 2n_{u+w}\bar{y}^2_{u+w} + n_x\bar{y}^2_x] \times \frac{1}{12}[(p_1 + p_2)(4U^2 + 4UL + 4L^2) - 3(p_2 - p_1)^2(U+L)^2],$$

The details of these computations as well as expected values and variance for arbitrary set sizes defined by C are given in Noren (1993).

This method yields a biased estimator of q(C) with a non-zero variance, thus it offers some level of protection to the information stored in the database. The bias depends on the mean of the data, as well as the upper and lower limits, U and L, and p_1 and p_2.

To gain an understanding of the behavior of E[q(C)] and Var[q(C)], the perturbation technique was applied to a generic college student database. The database consisted of seven characteristics: sex, marital status, class, citizenship, ethnic origin, curriculum, and high school city. The database contained information on 6,270 students. Grade point average (GPA) was the attribute attacked.

It was of interest to determine if certain ranges of GPA were more likely to be at risk of disclosure than others. Figure 5.1 displays the distribution of grade point averages. There were 3,965 "trackable" individuals in the database. The second histogram displays the distribution of GPA for these trackable individuals. The distribution of the GPA of the trackable individuals is remarkable similar to the distribution of the GPA for the entire database. The last graph is the distribution of GPA for students that can be characterized by different values of the characteristic equation C=(class, ethnic origin, curriculum).

Three parameter settings were used to perturb the data. The data perturbation occurred every time a query was issued. The parameter settings

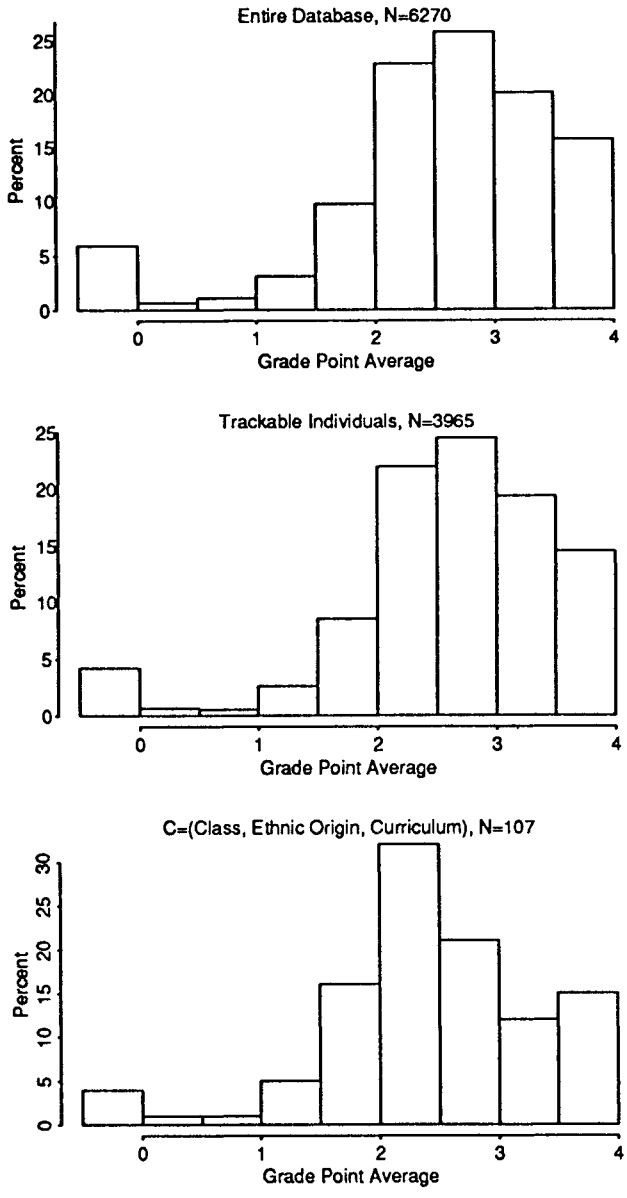

Figure 5.1 Distribution of Grade Point Average

were (L=0.02, U=0.08, $p_1=p_2=0.4$), (L=0.02, U=0.08, $p_1=0.4$, $p_2=0.3$), and (L=0.02, U=0.08, $p_1=0.05$, $p_2=0.1$). For a ***legitimate*** single query over all the men in the database, Table 5.1 gives the query results across the three perturbations settings for the sum of their GPA. The true value of GPA over all men is 7,247.499. Since there are 2,933 men, the mean GPA is 2.471. It is interesting that the values for the mean grade point average for the men in the

database are close to the true value. This indicates that for this legitimate single query the results may be acceptable.

Table 5.2 displays the same results for an instance of tracking. In this table the cumulative effect of the perturbation acting on the four queries needed to isolate a single GPA can be seen. The variance in these examples are approximately four times larger than in the single query case, as might be expected. The intervals in Table 5.2 also indicate that by combining the four queries a user may obtain negative values or values greater than 4.0 which are out of the range for actual grade point averages.

Table 5.1 Legitimate Query Results

Method	Sum of Male GPA	Std Dev of Sum	Mean	Interval *
$p_1=.4$ $p_2=.4$	7247	6.33	2.472	7203.619 (2.46) 7291.379 (2.49)
$p_1=.4$ $p_2=.3$	7609	5.92	2.595	7569.094 (2.58) 7650.654 (2.61)
$p_1=.05$ $p_2=.1$	7428	2.74	2.533	7409.826 (2.53) 7447.546 (2.54)

* The interval is Sum $\pm$ 2 [Standard Deviation of Sum]. The values in parenthesis are for the Interval/2933, i.e. an interval for the mean GPA based on the true, unperturbed count for the query.

The empirical work in this section was of great benefit in understanding the expected values and variances of the method described here as well as several other methods in Noren (1993). It appears that the results for legitimate queries may be acceptable when computing simple descriptive statistics although work needs to be done to understand the usefulness of perturbed data for other statistical applications. When four legitimate queries are combined in a tracker equation the results are no longer valid. This result is what is desired in order to protect the confidentiality of individuals in a statistical database.

Table 5.2 Expected Values, Standard Deviations and Two Standard Deviation Intervals

	GPA = 2.469		
Method	Expected	Standard Deviation	Interval
$p_1=.4$ $p_2=.4$	2.463	13.7228	-24.9826 to 29.9086
$p_1=.4$ $p_2=.3$	2.4432	12.7544	-23.0656 to 27.9520
$p_1=.05$ $p_2=.1$	2.482	5.8778	-9.2736 to 14.2376

6. Conclusion

Federal legislation governs the release of information by Federal agencies. While non-Federal organizations are not restricted by legislation, they are controlled by the media and public pressure. Information that is released, even in a controlled method, are at risk of being compromised. This risk must be weighed against the benefits the release of the data can give to researchers.

The use of a tracker to attack a database has been well studied, but the usefulness of a tracker against control techniques has not been statistically examined. The results presented in this paper show that while the results for legitimate single queries are close to the true value, the results for an instance of tracking are often outside the range of a valid response. This is the desired result of using control techniques to protect statistical databases. Although it appears that acceptable results are given for legitimate quires, more work needs to be done to understand the usefulness of perturbed data for statistical applications.

Acknowledgements

Funded in part by the U.S. Bureau of the Census. Computer implementation was done by David Rodgers, Computer and Information Sciences, Kansas State University.

References

American Statistical Association. (1977), "Report of Ad Hoc Committee on Privacy and Confidentiality." *The American Statistician*, 31(2):59-78.

Cecil, J.S. (1993), "Confidentiality Legislation and the Federal Statistical System," *Journal of Official Statistics*, 9(2):519-536.

Denning, D.E, P.J. Denning, and M.D. Schwartz. (1979), "The Tracker: A Threat to Statistical Database Security," *ACM Transactions on Database Systems*, 4(1):76-96.

Keller-McNulty, S. and E. Unger. (1993), "Database Systems: Inferential Security," *Journal of Official Statistics*, 9(2):475-500.

Keller-McNulty, S. and E. Unger. (1989), "The Protection of Confidential Data," *Computer Science and Statistical Proceedings on the Interface: Computer Science and Statistics*. 215-219.

Noren, Ewing E. (1993). "The Protection of Confidential Data Stored in a Sequential Access Statistical Database," MS Report, Kansas State University, Department of Statistics.

Venulapalli, K.C. and E.A. Unger (1991). "Output Perturbation Techniques for the Security of Statistical Databases," Technical Report 91-10, Department of Computer and Information Sciences, Kansas State University.

13 U.S. Code.

5 U.S. Code § 552a.