

**OPTIMAL DISCLOSURE LIMITATION STRATEGY IN STATISTICAL DATABASES:
DETECTING TRACKER ATTACKS THROUGH ADDITIVE NOISE**

George T. Duncan¹ and Sumitra Mukherjee²

¹Heinz School of Public Policy and Management
Carnegie Mellon University
Pittsburgh, PA 15213
Phone/FAX: (412) 268-2172/7036
e-mail: gd17+@andrew.cmu.edu

²School of Computer and Information Sciences
Nova Southeastern University
Fort Lauderdale, FL 33315
Phone/FAX: (954) 262-2079/3915
Email: sumitra@scis.nova.edu

June 16, 1998

ABSTRACT

Disclosure limitation methods transform statistical databases to protect confidentiality. A statistical database responds to queries with aggregate statistics. The database administrator should maximize legitimate data access while keeping the risk of disclosure below an acceptable level. Legitimate users seek statistical information, generally in aggregate form; malicious users—the data snoopers—attempt to infer confidential information about an individual data subject. Tracker attacks are of special concern for databases accessed online. This article derives optimal disclosure limitation strategies under tracker attacks for the important case of data masking through additive noise. Operational measures of the utility of data access and of disclosure risk are developed. The utility of data access is expressed so that tradeoffs can be made between the quantity and the quality of data to be released.

The article shows that an attack by a data snooper is better thwarted by a combination of query restriction and data masking than by either disclosure limitation method separately. Data masking by independent noise addition and data perturbation are considered as extreme cases in the continuum of data masking using positively correlated additive noise. Optimal strategies are established for the data snooper. Circumstances are determined under which adding autocorrelated noise is preferable to using existing methods of either independent noise addition or data perturbation. Both moving average and autoregressive noise addition is considered.

KEYWORDS: Confidentiality, Privacy, Data Access, Computer Databases, Disclosure Limitation, Data Perturbation, Autocorrelation

George T. Duncan is Professor of Statistics, Heinz School of Public Policy and Management and Department of Statistics, Carnegie Mellon University, Pittsburgh, PA, 15213. Sumitra Mukherjee is Associate Professor, School of Computer and Information Sciences, Nova Southeastern University, Fort Lauderdale, FL 33315. The authors thank the Associate Editor and the anonymous reviewers for helpful comments and the National Science Foundation for grants IRI-9312143 and SES 91-10512.

Duncan and Mukherjee, Optimal Disclosure Limitation Strategy

1. INTRODUCTION

Demand for on-line data access has been pushed by the rapid growth of information technology and the information requirements of digital libraries, electronic commerce, and, notably, the health care and financial industries. Many of the data attributes—such as medical diagnoses, salaries, and academic transcripts—are confidential. Government, as a major collector of data, has become increasingly sensitive to its stewardship responsibilities to disseminate information while maintaining confidentiality (Duncan, Jabine, and de Wolf 1993). Protecting databases against the attack of a data snooper is of personal concern to data providers, of practical concern to database administrators, and of methodological concern to statisticians and computer scientists (Duncan and Pearson 1991). Legitimate users seek statistical information, generally in aggregate form; malicious users—the data snoopers—attempt to infer confidential information about an individual data subject (Chin and Ozsoyoglu 1981).

For legal, ethical, and practical reasons, statistical users—in contrast to authorized administrative users—are typically not entitled to obtain identifiable information on individual data subjects (Bethlehem, Keller, and Pannekoek 1990). A statistical database responds to queries with aggregate statistics. The individual records—microdata—contained in the database are not to be linked with the data subjects (Duncan and Lambert 1989). Preventing disclosure is complicated; simple anonymization of records is inadequate because of the possibility of reidentification (inferential disclosure). Disclosure limitation methods must ensure that the information provided is statistically useful while protecting confidentiality.

Disclosure limitation research has focused on responsible release of public use microdata files and dissemination of tabular data (Cassel 1976, Cox 1980, Cox 1995, Dalenius 1982, Duncan and Lambert 1986, Frank 1983, Keller-McNulty et al 1989, Paass 1988, Spruill and Gastwirth 1982). Increasingly, however, information is stored and accessed online (Rainwater and Smeeding 1988) through database management systems (DBMS). The fact that a DBMS can be sequentially queried raises dynamic disclosure control concerns that are not present with more traditional modes of data dissemination, such as tables and public use microdata files. Disclosure risk issues are highlighted in Ahituv, Lapid, and Neumann (1988), Keller-McNulty and Unger (1993), and Lambert (1993). As a contribution to this relatively new literature, this article presents optimal disclosure limitation procedures for sequentially queried databases. The key to the optimization formulation is to maximize legitimate data access subject to appropriate constraints on disclosure risk.

Duncan and Mukherjee, Optimal Disclosure Limitation Strategy

Several methods have been developed to limit disclosure in statistical databases (see Adam and Wortmann (1989), Adam, Gangopadhyay, and Holowczak (1998), and Jabine (1993) for surveys of methods). Existing methods of disclosure limitation in sequentially accessed statistical databases may be broadly classified into one of two classes—query restriction methods and response modification methods.

An important query restriction technique is *query set size* (QSR) control: a query is disallowed if the number of records satisfying the query's conditions is too small (by inference from the complementary query, too large). The motivation behind this inference control scheme is that if a query was to yield a unique record, then this record may be identifiable and sensitive information obtained from the record. To illustrate, consider a company's employee database. If a query for all records of employees earning over \$300,000 yields just one record, this may be enough to have identified the CEO's record and so link sensitive information on the record to the CEO. Further, just banning uniqueness is not sufficient: a query with just a few valid records may be enough to make identification (Fellegi 1972, Friedman and Hoffman 1980). Commonly, aggregate information based on three or fewer data subjects is withheld because two data subjects may collude to compromise confidentiality of the third subject's information (Frank 1983). Nullifying the promise of the query restriction approach is the finding that through certain sequences of seemingly innocuous, and hence unrestricted, queries called trackers, the answers to restricted queries can always be deduced (Schlörer 1975). Trackers are considered to be one of the most significant threats to security in on-line statistical databases (Denning, Denning, and Schwartz 1978). Hence this article focuses on protection against tracker attacks. One attempt to protect against tracker attacks is through *query size overlap* control (Dobkin, Jones and Lipton 1979). This form of response modification restricts the number of overlapping records in successive queries. It has the two drawbacks of not dealing with collusion of data snoopers and precluding subset queries.

By releasing other than the true response to a query, the *response modification approach* introduces uncertainty in the response. This method fuzzes linkage to key values known by the user and reduces the precision with which sensitive values may be inferred, thereby reducing the risk of disclosure. The challenge is to do this fuzzing in a way that still permits valid statistical analysis. This article treats a common implementation of response modification in which data are masked using zero mean additive noise (Dalenius 1988, Kim 1986, Tendick 1991). Under many such additive noise schemes the sequence of modified responses to repeated queries are stochastically independent. Consequently—and compromising to confidentiality—

Duncan and Mukherjee, Optimal Disclosure Limitation Strategy

sensitive values may be estimated with increasing precision by making inferences based on responses to repeated queries. Faced with the possibility of repeated query attack by a data snooper, a database administrator has a simple parry: provide the same modified response every time a query is made. The data snooper would, in effect, be making queries of a once-and-for-all-modified database and so gain no additional information by repeating queries. This technique of disclosure limitation is termed the data perturbation method (Adam and Wortmann 1989). The data perturbation method has the drawback that it may lead to substantial selection bias in the set of records returned in response to a query (Matloff 1986).

To provide a broader range of response modification methods for sequential queries, we observe that independent noise addition and data perturbation are the two extreme cases in a continuum of adding positively correlated noise. We demonstrate that there are practical circumstances when intermediate cases are preferable to either extreme.

Section 2 reviews how tracker attacks may be used to compromise query size restriction control in a statistical database. Section 3 presents a mathematical programming formulation of the problem of maximizing data access while ensuring that disclosure risks are acceptably low. Section 4 establishes the risk of disclosure under various disclosure limitation schemes. Section 5 uses these results to obtain optimal disclosure limitation strategies. Section 6 presents conclusions and discussion.

2. TRACKER ATTACKS AGAINST QUERY SIZE RESTRICTION METHOD

We adopt the model for a sequentially queried statistical database of Denning and Schlörer (1980). The database is a collection of N records, each record corresponding to a data subject. A record X_i about subject i contains K attributes, some of which may be sensitive. Identifiers such as names and identity numbers are not included among the available fields—the database has been anonymized. Further, we assume that the database values do not change during the relevant period. A statistical query is expressed using a characteristic formula C . This formula is a logical expression involving conditions on values of the attributes. It identifies a subset of records R_C that satisfy the conditions. Responses may involve statistics such as sums, averages, moments, and

Duncan and Mukherjee, Optimal Disclosure Limitation Strategy

regression coefficient estimates. To demonstrate the utility of our approach we give explicit results for the sum

$$R_C = \sum_{i \in C} X_i$$
 of a univariate sensitive attribute X.

Under Query Size Restriction (QSR) control, a response R_C is denied when the cardinality $|C|$ of the query set C (or the cardinality of the complement of C) is less than or equal to a specified integer k , $0 < k < N/2$. A query that is not answered under QSR is a restricted query. In spite of its intuitive appeal, QSR control is compromised because the answer to a restricted query may be computed based on a finite sequence of legitimate queries. Schlörer (1975, 1980) demonstrated this computation using a tracker formula:

$$R_C = R_{C \cup T} + R_{C \cup \bar{T}} - R_T - R_{\bar{T}} \quad \text{for } |C| \leq k, \quad (1)$$

where T is a subset of data subjects chosen so that $2k < |T| < N - 2k$. Each of the four terms used to compute the restricted value results from queries that are not restricted. While this article considers the sum query, trackers are quite general and can be used with other statistical queries. Further, while we use a single sensitive attribute to demonstrate our approach, the tracker is equally applicable to vectors of attributes. It has been shown that in any practical database a tracker can always be found (Denning 1978). Hence trackers are a major security threat in statistical databases.

Given this vulnerability to tracker attacks, a common strategy for disclosure limitation is to mask the released data. Duncan and Mukherjee (1991) show that a combination of query size restriction control and additive noise data masking significantly decreases the risk of disclosure through tracker attacks. For this case we cast the database administrator's task in selecting disclosure limitation strategy as a constrained optimization problem.

3. A CONSTRAINED OPTIMIZATION FORMULATION

Disclosure limitation involves minimizing restrictions on data access while keeping the risk of disclosure acceptably low. In our context, restrictions on access are imposed by the QSR method, which leaves some queries unanswered, as well as by data masking, which adds variability to the data.

Duncan and Mukherjee, Optimal Disclosure Limitation Strategy

We first propose a measure of restriction on data access when both QSR control and additive noise masking are employed. Under QSR control alone, the proportion of queries that are not answered may be used as a measure of restriction on access (Michalewicz, Li, and Chen 1990). Taking the query size Q as a random variable, its probability distribution can be estimated for a particular database application. Since under QSR control the query size must be greater than k and less than or equal to $N-k$ for a query to be answered, the proportion of restricted queries is given by $P[Q \leq k] + P[Q > N-k]$. Note that $k=0$ represents the special case when QSR control is not used. While the proportion of queries not answered serves as a measure of the *quantity* of data inaccessible to the legitimate user, the *quality* of the available data is also affected by data masking.

With noise addition, the larger is the variance of noise, the worse is the quality of data available to the legitimate user for statistical purposes. Hence the variance σ^2 of the additive noise may be used as a measure of restrictions on data access. Several other measures have been suggested in the literature to capture the effects of data masking (Dalenius 1982, Frank 1976). These include increased entropy, and the bias and the increased variance in the estimators of parameters such as regression coefficients. While these measures are appropriate for specific applications, the reliability of the statistical inference process under data masking invariably decreases with increasing variance of additive noise.

In the general case when both QSR control and data masking schemes are in place, we take a weighted sum of the proportion of restricted queries and the variance of the noise added as an overall measure of restrictions on data access. The weighted sum captures the adverse consequences of the disclosure limitation methods for the legitimate user both in terms of the quantity of data available and the quality of the released data. This may be formalized as the objective function:

$$L = (1-\alpha)(P[Q \leq k] + P[Q > N-k]) + \alpha\sigma^2, \quad (2)$$

where α is an appropriately selected weight, ranging from 0 to 1. This parameter α reflects the relative importance of the two component measures of restrictions on data access. The database administrator assigns a value to α appropriate to the data dissemination problem. Lower values of α indicate that it is more important to provide responses to queries even though this may require the addition of higher levels of noise. Higher values of α reflect greater concerns about the quality of data. Thus specification of α allows an explicit

Duncan and Mukherjee, Optimal Disclosure Limitation Strategy

tradeoff between data quantity and data quality. Subject to constraints on disclosure risk, the database administrator seeks to minimize L , thereby maximizing the utility of data access. We now consider measures for disclosure risk.

4. MEASURES OF DISCLOSURE RISK

The larger the variance of an estimator of a sensitive value R_C , the lower we take the risk of disclosure to be. While other measures have been suggested (Lambert 1993 and Frank 1978), the primary threat is the ability of the snooper to infer attribute values for data subjects. This motivates our choice of the variance of estimators of sensitive values as a measure of disclosure risk. That disclosure risk must be lower than some acceptable threshold may be operationalized as constraints of the form

$$V_i \geq I_i \text{ for } i = 1, 2, \dots, N, \quad (3)$$

where V_i is the variance of an estimator of a response R_C with query size $|C| = i$ and λ_i is the minimum acceptable threshold for the variance. The variance V_i in (3) depends both on the method of disclosure limitation as well as on the estimator used by the data snooper. We now discuss how the threshold parameters λ_i may be selected by the database administrator.

The smaller the query size i , the greater must be the disclosure protection. This is because small groups of data subjects may be atypical and hence more easily identifiable. In the most sensitive case—for a query involving a single data subject—the snooper gains no additional information about the target subject if the variance of the estimator V_1 is greater than the variance of the sensitive attribute in the population. Hence, selecting λ_1 equal to the variance of the sensitive attribute in the population offers acceptable protection. For a query about two data subjects, consider the case when the query is posed by one of the data subjects included in the query set. Since this user knows the attribute value for his own record, the variance of the estimator is the variance associated with the attribute value of the other data subject. It is then conservative and so sufficient to ensure that the variance of the estimator involving two data subjects is equal to the variance of an estimator of a single data subject's sensitive value, that is, to set $\lambda_2 = \lambda_1$. Queries about larger groups of data subjects require less protection. Hence we take λ_i to be non-increasing. Expressions for disclosure risk are obtained next.

Under QSR control, an answer to a restricted query (one involving fewer than k data subjects) can be inferred using the tracker formula (1). Responses to queries that are not restricted may of course be obtained directly. However, under additive noise masking these responses are fuzzed and true values must be estimated using masked observations. Existing implementations of additive noise data masking either use independent noise or follow the data perturbation scheme (in which the same masked response is provided upon repeated queries). Recognizing that these schemes are the two extreme cases in the continuum of data masking using positively correlated noise, we analyze the general case.

Positively correlated stationary noise may be added to mask responses to repeated queries. We propose a masking scheme (Fuller1993) under which a masked value M_{it} for the t^{th} query involving the value X_i for data subject i is generated as $M_{it} = X_i + \mathbf{e}_{it}$ for $i = 1, \dots, N; t = 1, 2, \dots$, where \mathbf{e}_{it} is zero mean additive noise with variance \mathbf{s}^2 . The masked response released to the user is hence

$$M_{Ct} = R_C + \sum_{i \in C} \mathbf{e}_{it} \quad \text{for } t = 1, 2, \dots, \text{ where } R_C \text{ is the true response to the query.}$$

Depending on the implementation, the noise component may be generated only for those values that are queried or for all values in the database. For the purpose of our analysis we consider the former case since it is computationally less demanding to implement. Further, we do not distinguish between queries posed by different users. This practice is conservative since it provides protection against collusion among all users. Also, in implementation this practice is more practical since it does not require tracking the activities of individual users.

The noise process can be described by the covariance matrix Σ of the noise vector. Stationary noise has the property that the (s, t) -element of the covariance matrix Σ is given by

$$\Sigma_{st} \equiv E[\mathbf{e}_{it} \mathbf{e}_{is}] = \rho_{|t-s|} \mathbf{s}^2 \quad \text{for } i = 1, \dots, N, \text{ where } \rho_{|t-s|} \text{ is the correlation coefficient.}$$

Duncan and Mukherjee, Optimal Disclosure Limitation Strategy

For an unrestricted response R_C (so $k < |C| \leq N-k$), we consider the case where a snooper uses the mean of r repeated responses as an estimator \hat{R}_C . The variance of \hat{R}_C is given by

$$V[\hat{R}_C] = \frac{|C|}{r^2} \mathbf{1}' \Sigma \mathbf{1}, \quad (4)$$

where $\mathbf{1}$ is a vector with each of its elements 1. In the special case of independent noise addition this variance is $|C|\sigma^2/r$. Under data perturbation (since repeated queries garner no new information) the variance is $|C|\sigma^2$.

For restricted queries (with $|C| \leq k$ or $|C| > N-k$) a data snooper may use masked observations in the tracker formula to estimate restricted values. Further, repeated masked observations may be used to increase the precision with which sensitive values may be estimated. A snooper can adopt one of two alternative approaches to estimate restricted values:

- (i) Repeated observations can be obtained for each component of the tracker formula. The mean for each component can be calculated and used in the tracker formula to estimate a restricted value.
- (ii) A set of masked observations for each of the four components of the tracker formula can be obtained in sequence and used to compute a tracker result. This process can be repeated and the mean of a sequence of tracker results used as an estimator of the restricted sensitive value.

Under the assumption that the correlation between noise components of repeated responses is non-increasing with time, the latter approach results in lower variance for the estimator and is hence the preferred strategy of the snooper. This notion can be formalized in the following results, which have proofs in the appendix:

Result 1: When repeated masked observations are available, the optimal strategy for a snooper is to compute tracker results based on the four components of a tracker formula obtained in sequence, and use a mean of a sequence $(\hat{R}_{C_1}, \dots, \hat{R}_{C_r})$ of such tracker results as an estimator of a restricted value R_C .

Further, the variance of the estimator of a tracker result depends on the order in which the tracker components are obtained. It can be shown that:

Result 2: When masked values are used in a tracker formula to estimate a restricted value R_C , the optimal strategy for a snooper is to obtain components of the tracker formula in the order $(R_{C \cup T}, R_T, R_{-T}, R_{C \cup -T})$. No other sequence of tracker components result in a lower variance for the estimator.

Duncan and Mukherjee, Optimal Disclosure Limitation Strategy

Results (1) and (2) yield the optimal strategy for a data snooper attempting to estimate restricted values when both QSR control and data masking have been imposed. In our analysis we consider the case when the data snooper adopts this strategy and uses the mean of a sequence of results obtained from the application of the tracker formula r times to estimate a restricted value R_C . The variance for this estimator \hat{R}_C is given by

$$V[\hat{R}_C] = \frac{1}{r^2} \mathbf{1}' \Psi \mathbf{1}, \quad (5)$$

where $\mathbf{1}$ is a vector with each of its elements 1 and Ψ is the covariance matrix of a sequence of tracker results $(\hat{R}_{C1}, \dots, \hat{R}_{Cr})$. It can be shown that the diagonal elements $\Psi_{ii} \equiv \Psi_0$ for $i = 1, \dots, r$, are given by

$$V[\hat{R}_C] = \mathbf{s}^2 \left[(2N+|C|) - 2(N \mathbf{r}_1 + |C|(\mathbf{r}_1 - \mathbf{r}_2)) \right], \text{ and the off-diagonal elements } \Psi_{ij} \equiv \Psi_t, \text{ where } t = |i-j|, \text{ are given by}$$

$$\mathbf{y}_t = \mathbf{s}^2 \left[|C|(\mathbf{r}_{3t-2} - 2\mathbf{r}_{3t-1} + 3\mathbf{r}_{3t} - 2\mathbf{r}_{3t+1} + \mathbf{r}_{3t+2}) + (N-|C|)(-\mathbf{r}_{2t-1} + 2\mathbf{r}_{2t} - \mathbf{r}_{2t+1}) \right].$$

In the special case of independent noise addition, the variance of the estimator of a restricted query is

$$V[\hat{R}_C] = \frac{\mathbf{s}^2}{r} (2N+|C|). \quad (6)$$

Under data perturbation,

$$V[\hat{R}_C] = \mathbf{s}^2/|C|. \quad (7)$$

Notice that under data perturbation the expression for the variance of the estimator of R_C is the same for both restricted and unrestricted queries. This is because the static noise components of the various terms in the tracker formula cancel each other out. This result may be formalized as follows:

Result 3: Under data perturbation the variance of an estimator of a sensitive value R_C computed using the tracker formula is the same as in the case when access to R_C is not restricted under QSR control and can be obtained directly. Hence restricting access under QSR control when data perturbation is used provides no additional protection.

In the general case when positively correlated noise is added to mask the data, a data snooper may choose to use the average of a sequence of tracker results to estimate a restricted value R_C . The variance of the estimator is given by Equation (5). For unrestricted queries the variance of the estimator may be computed using

Duncan and Mukherjee, Optimal Disclosure Limitation Strategy

Equation (4). We take the variance of an estimator as a measure of disclosure risk: the greater the variance, the lower the risk. In the next section we use these results to present solutions to the constrained optimization problem formulated in Section 3.

5. OPTIMAL SOLUTIONS UNDER DISCLOSURE LIMITATION METHODS

In our framework, the database administrator seeks to select the size k of the restricted set and the variance of the additive masking noise σ^2 so as to minimize the objective function (2) subject to the constraints specified by (3). For each specified disclosure limitation scheme, the minimum value of the objective function L may be determined. The scheme with the lowest minimum L is the optimal method. Optimal solutions to this problem under various disclosure limitation methods are compared next.

5.1. Data Perturbation

Restricting access under QSR control provides no additional protection when data perturbation is used (Result 3). Hence the optimal choice for the restricted set size k is 0 when data perturbation is in place. Further, since λ_i is non-increasing in i and V_i is increasing in i , ensuring that $V_1 \geq \lambda_1$ guarantees that $V_i \geq \lambda_i$ for all i . The variance V_1 under data perturbation is σ^2 . Hence for the additive noise the minimum acceptable variance is $\sigma^2 = \lambda_1$. Substituting $k=0$ and $\sigma^2 = \lambda_1$ we obtain the optimal value for the objective function (2) under data perturbation as

$$L[k_0, \mathbf{s}_0^2] = \mathbf{a} \mathbf{1}_1, \quad (8)$$

If the objective function value is lower than the minimum L achieved under alternate disclosure limitation methods, then data perturbation should be used.

5.2. Combination of QSR Control and Independent Noise Addition

Under independent noise addition the variance of an estimator of a restricted query (so $i \leq k$ or $i > N-k$) is given by $V_i = \frac{\mathbf{s}^2}{r}(2N + i)$. Since V_i is increasing in i and λ_i is non-increasing in i , ensuring that $V_1 \geq \lambda_1$ guarantees

Duncan and Mukherjee, Optimal Disclosure Limitation Strategy

that $V_i \geq \lambda_i$ for all other restricted queries (i.e. for $1 < i \leq k$ or $i > N - k$). Hence $\frac{\mathbf{s}^2}{r}(2N + 1) \geq \mathbf{I}_1$ is the only binding constraint for restricted queries.

For unrestricted queries (so $k < i \leq N - k$), the variance of an estimator is given by $i\sigma^2/r$. Since V_i is increasing in i and λ_i is non-increasing in i , ensuring that $V_{k+1} \geq \lambda_{k+1}$ guarantees that $V_i \geq \lambda_i$ for other unrestricted queries (i.e. for $k+1 < i \leq N - k$). Hence $(k+1)\sigma^2/r \geq \lambda_{k+1}$ is the only binding constraint for unrestricted queries.

Consider the only two binding constraints in the problem: $\frac{\mathbf{s}^2}{r}(2N + 1) \geq \mathbf{I}_1$ and $\frac{\mathbf{s}^2}{r}(k + 1) \geq \mathbf{I}_{k+1}$.

Both constraints are satisfied for a choice of $\mathbf{s}^2 = \frac{\mathbf{I}_1 r}{2N + 1}$ when $k \geq (2N + 1) \frac{\mathbf{I}_{k+1}}{\mathbf{I}_1} - 1$. Since k is to be

minimized, the optimal choice for k is $k_1 = \text{maximum integer } k \text{ such that } k < (2N + 1) \frac{\mathbf{I}_{k+1}}{\mathbf{I}_1}$.

For this choice of parameters, the value for the objective function is hence

$$L[k_1, \mathbf{s}_1^2] = (1 - \mathbf{a})(P[Q \leq k_1] + P[Q > N - k_1]) + \mathbf{a} \frac{\mathbf{I}_1 r}{2N + 1} . \quad (9)$$

For any choice of $k < k_1$, both binding constraints may be satisfied by selecting $\mathbf{s}^2 = \frac{\mathbf{I}_{k+1} r}{k + 1}$.

Hence for a choice of $k < k_1$, the minimum value for the objective function is given by

$$L[k_2, \mathbf{s}_2^2] = (1 - \mathbf{a})(P[Q \leq k_2] + P[Q > N - k_2]) + \mathbf{a} \frac{\mathbf{I}_{k_2+1} r}{k_2 + 1} . \quad (10)$$

The values of L as obtained in Equations (9) and (10) can be evaluated to select the optimal parameters under independent noise addition. Because the size of the restricted set k is typically small, the optimal parameters can be easily computed using this method.

5.3. Combination of QSR Control and Autocorrelated Noise Addition

Duncan and Mukherjee, Optimal Disclosure Limitation Strategy

Under autocorrelated noise addition, the variance for an estimator of a restricted value is given by Equation (5) and the variance for an estimator of an unrestricted query is specified by Equation (4). Using the same arguments as in the case of independent noise addition, it can be shown that the optimization problem involves only two binding constraints— $V_1 \geq \lambda_1$ for restricted queries and $V_{k+1} \geq \lambda_{k+1}$ for unrestricted queries. Since the parameter σ^2 is to be determined as a solution to the constrained optimization problem, we use the normalized matrices $\Sigma_z = \Sigma/\sigma^2$ and $\Psi_z = \Psi/\sigma^2$. Solutions to the problem may be obtained as follows:

Case 1. For a choice of $k_3 = \text{maximum integer } k \text{ such that } k < \frac{\mathbf{1}_{k+1}' \Psi_z \mathbf{1}}{\mathbf{1}' \Sigma_z \mathbf{1}}$ both binding constraints may be satisfied

with $\mathbf{s}^2 = \frac{\mathbf{1}_1 r^2}{\mathbf{1}' \Psi_z \mathbf{1}}$. The minimum value for the objective function is hence

$$L[k_3, \mathbf{s}_3^2] = (1 - \mathbf{a})(P[Q \leq k_3] + P[Q > N - k_3]) + A \frac{\mathbf{1}_1 r^2}{\mathbf{1}' \Psi_z \mathbf{1}}. \quad (11)$$

Case 2. For $k < k_3$, both constraints can be satisfied for a choice of $\mathbf{s}^2 = \frac{\mathbf{1}_{k+1} r^2}{(k+1) \mathbf{1}' \Sigma_z \mathbf{1}}$.

For these parameters the optimal value for the objective function is given by:

$$L[k_4, \mathbf{s}_4^2] = (1 - \mathbf{a})(P[Q \leq k_4] + P[Q > N - k_4]) + \mathbf{a} \frac{\mathbf{1}_{k_4+1} r^2}{(k_4 + 1) \mathbf{1}' \Sigma_z \mathbf{1}}. \quad (12)$$

The objective functions (11) and (12) may be evaluated to select the optimal parameters for autocorrelated noise addition. Again, the solutions are not computationally demanding since k is typically small.

Having determined the optimal solution under each disclosure limitation method, the database administrator's best decision is to select the method with the lowest minimum objective function value. The analysis presented in this section is adequate to determine an optimal disclosure strategy for fixed values of the weight parameter α and the maximum number r of repeated queries. The database administrator may, however, want to know how sensitive the optimal strategy is to the choice of α and r . This issue is investigated in the next section.

5.4. Optimal Disclosure Limitation Strategy as a Function of the Weight Parameter

Duncan and Mukherjee, Optimal Disclosure Limitation Strategy

The objective function values L depend on the weight parameter α and the number r of repeated queries. We use the results obtained in the previous section to determine regions in the α - r plane over which various disclosure limitation strategies are optimal. The database administrator could use such an α - r diagram to determine the sensitivity of a selected method to the choice of the weight parameter α .

5.4.1. Data Perturbation versus Independent Noise Addition

Comparing Equations (8) and (9) we find that data perturbation is preferable to independent noise addition with $k=k_1$ as long as

$$r \geq (2N + 1) \left(1 - \frac{(1 - \mathbf{a})}{\mathbf{a}} \frac{P[Q \leq k_1] + P[Q > N - k_1]}{I_1} \right). \quad (13)$$

Comparing Equations (8) and (10) we find that data perturbation results in a lower minimum value of the objective function than independent noise addition with $k=k_2$ for

$$r \geq (k_2 + 1) \left(\frac{I_1}{I_{k_2+1}} - \frac{(1 - \mathbf{a})}{\mathbf{a}} \frac{P[Q \leq k_2] + P[Q > N - k_2]}{I_{k_2+1}} \right). \quad (14)$$

Moreover, k_2 can assume any value less than k_1 . Equation (10) shows that a choice of $k_2=i$ is preferable to selecting $k_2=i-1$ for

$$r \geq \frac{(1 - \mathbf{a})}{\mathbf{a}} \frac{P[Q = i] + P[Q = N - i + 1]}{I_i / i - I_{i+1} / (i + 1)}. \quad (15)$$

Figure 1 summarizes the results of (13), (14), and (15). It identifies regions in the α - r plane for which each disclosure limitation strategy is optimal when data perturbation and independent noise addition are the available data masking schemes. Data perturbation is preferred in the region above the envelope formed by segments of the curves demarcating data perturbation from independent noise addition with different values of k (as specified by (13) and (14)). Below this envelope (when independent noise is preferable), optimal values of k are differentiated by segments of the hyperbolas specified by (15), with higher values of k being preferred as α increases.

Under circumstances where the number of repeated queries cannot be restricted to less than $2N+1$, data perturbation should be the only disclosure limitation method imposed. When the number of repeated queries

can be restricted to a lower value, data perturbation should be used only for lower values of α —reflecting circumstances when the level of masking noise is relatively less important. As the concern about keeping the level of noise low increases (reflected by increasing values of α) a combination of QSR control and independent noise addition is preferred, with increasing restricted set size k and decreasing noise level σ^2 .

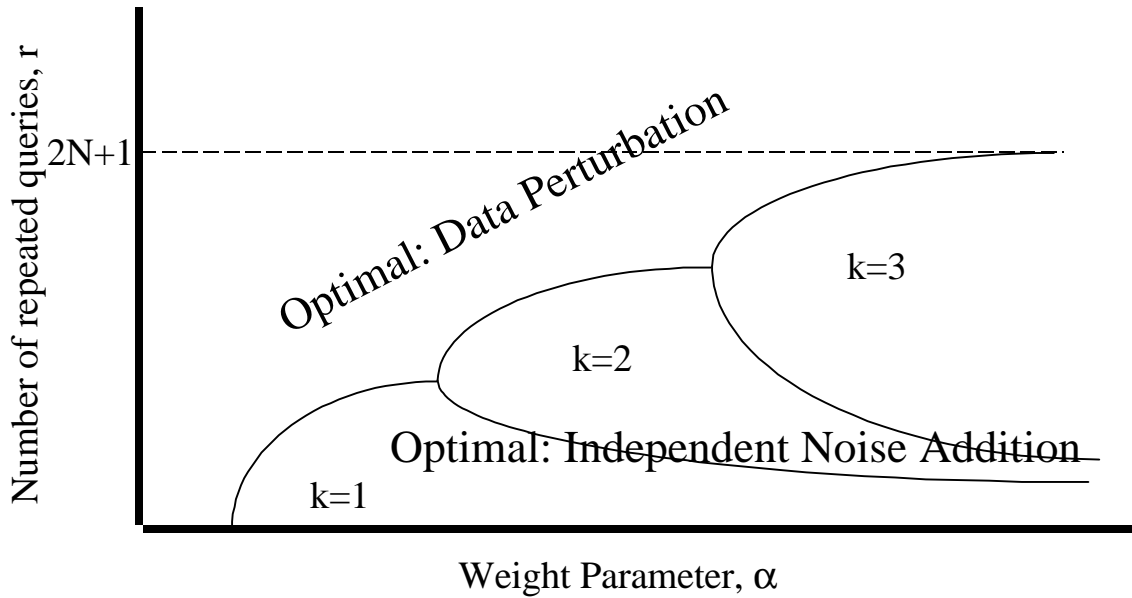


Figure 1. Optimal choice of disclosure limitation scheme when data perturbation and independent noise addition are the only available data masking methods.

5.4.2. Optimal Strategy when Autocorrelated Noise is Permissible

For the same choice of k and σ^2 the variance of an estimator of a restricted value is greater under independent noise addition than can be achieved when autocorrelated noise is used. This is because under the latter scheme the noise components in the tracker terms cancel each other out to some extent. Hence, independent noise addition with $k=k_1$ is always preferable to autocorrelated noise addition with $k=k_3$.

However, for unrestricted queries the variance of an estimator based on repeated queries is always greater when autocorrelated noise is used than when the additive noise is independent. It follows that for $k < k_1$, autocorrelated noise is always preferable to adding independent noise. Using Equations (8) and (12) a family of curves may be generated. The envelope of these curves, for any value of α , determines the number of

Duncan and Mukherjee, Optimal Disclosure Limitation Strategy

repeated queries above which the data perturbation method is preferable to autocorrelated noise with $1 \leq k < k_1$. Equation (12) can also be used to generate a series of hyperbolas demarcating the regions for which different values of k_4 are optimal.

Figure 2 partitions the α - r plane based on the optimal disclosure limitation strategy when autocorrelated noise addition is a possible option. When a data snooper can make more than $2N+1$ repeated queries, the only disclosure limitation method imposed should be data perturbation. When the number of repeated queries can be further restricted, data perturbation is optimal only when the level of masking noise is relatively less important, as reflected by lower values of α . With increasing concern about the level of additive noise (reflected by increasing values of α) a combination of QSR control and autocorrelated noise addition is optimal. As the weight factor α increases, the restricted set size k increases, accompanied by a decreasing noise level σ^2 . Above a certain weight factor α , a combination of independent noise addition (with $k=k_1$) and QSR control becomes optimal.

The database administrator may use the diagram to determine how sensitive a selected disclosure limitation strategy is to the subjectively determined parameter α . Sensitivity analysis may suggest, for example, that computational costs make data perturbation preferable to a more complex scheme based on addition of autocorrelated noise, even though autocorrelated noise may be marginally superior in L .

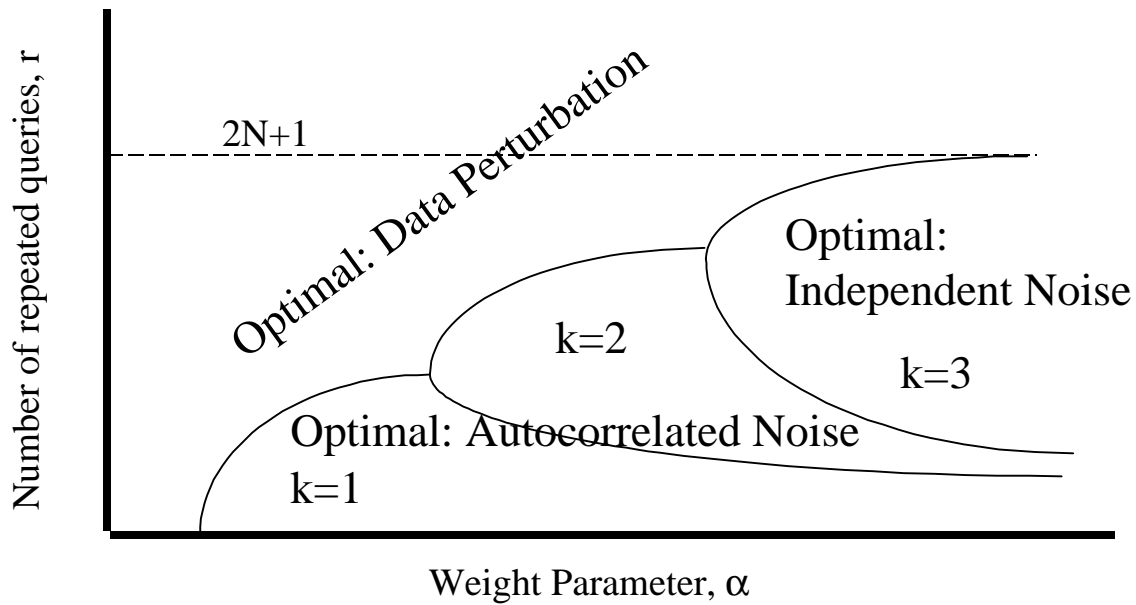


Figure 2. Optimal choice of disclosure limitation scheme when data perturbation, independent noise addition, and autocorrelated noise addition are available for data masking.

In the next section we use an example to demonstrate how our approach can be used to determine an optimal disclosure limitation strategy. The example illustrates that under certain conditions autocorrelated noise addition may be the preferred strategy.

5.5. Optimal Disclosure Limitation Strategy: an Example

Consider a database with $N=100$ records maintaining an audit trail (see Javitz and Valdes (1991)) which prevents data users from making more than $r=20$ repeated queries. The probability distribution of the query size Q has been determined and some of the probabilities are specified in the following table:

q	1	2	3	4	5	...	96	97	98	99	100
$P[Q=q]$.005	.005	.01	.01	.0201	.01	.02	.02	.1

Duncan and Mukherjee, Optimal Disclosure Limitation Strategy

The relatively high probability associated with $q=100$ is due to the fact that in a statistical database a large proportion of the queries may involve the entire set of records available.

Concerns about disclosure risk might suggest the following threshold parameters:

$$I_1 = I_2 = 10, \quad I_3 = 2, \quad I_4 = 0.5, \quad I_i = 0 \quad \text{for } 5 \leq i \leq N$$

The choice of the threshold parameter $\lambda_1 = 10$ may be appropriate when the variance of the sensitive attribute in the population is 10. Under these circumstances a snooper can infer the sensitive attribute value for a target data subject with no more precision than is possible from legitimately available population statistics. The choice of $\lambda_2 = \lambda_1$ is motivated by the concern that one of two data subjects included in a query set (of size 2) may be a snooper attempting to estimate the other subject's sensitive value. Decreasing, but non-zero values for λ_3 and λ_4 reflect the diminishing concern about collusion among three or even four data snoopers. Queries involving five or more data subjects are considered legitimate statistical queries (as is often the policy adopted by statistical agencies). Hence there are no restrictions imposed on the variance of estimators for query sets of size greater than four.

The weight parameter α is chosen to reflect the concern about degradation of data quality due to noise addition relative to the concern about query restriction. For illustration here, we take α to be 0.8. The sensitivity of the solution to the choice of the weight parameter will be investigated.

Under these conditions, if data perturbation and independent noise addition were the only available masking schemes, the optimal method may be determined to be a combination of QSR control and independent noise addition with parameters $k=3$ and $\sigma^2=1$. The solutions are obtained by evaluating and comparing the objective functions specified by (8), (9), and (10) for the given database parameters. Note that the optimal case is specified by (10). Further, as conditions (14) and (15) allow, these results are valid for a wide range (0.16 to 0.98) of the weight parameter α .

To broaden the class of disclosure limitation options, consider adding *autocorrelated* noise. For the sake of simplicity, we analyze first-order stationary noise processes. First, we adopt a scheme proposed by Beck (1980) which uses moving average additive noise, and consider a noise generation method following a first

Duncan and Mukherjee, Optimal Disclosure Limitation Strategy

order moving average (MA[1]) process. Next, we analyze a first order autoregressive (AR[1]) process. We allow the database administrator control over the correlation parameter of the stationary processes.

Moving Average Noise:

When noise follows an MA[1] process with correlation coefficient ρ , constraint (3) for an unrestricted query,

(when $i > k$) is given by $V_i = \frac{i\mathbf{S}^2}{r^2}[r + 2(r-1)\mathbf{r}] \geq I_i$.

Comparison of the objective function values for our example indicates that when moving average noise is added, the optimal choice of k is 3, and for a choice of $\rho = 0.2$, it is sufficient to add noise with variance $\sigma^2 = 0.72$. This is an improvement over the noise variance of 1 required under independent noise. This result holds over a wide range of correlation coefficient values, indicating that autocorrelated noise addition provides better quality data access while ensuring the same level of protection as independent noise addition.

Autoregressive Noise:

When noise follows an AR[1] process with correlation coefficient ρ , constraint (3) for an unrestricted query

(when $i > k$), is given by $V_i = \frac{i\mathbf{S}^2}{r^2}[r + 2(r-1)\mathbf{r}] \geq I_i$.

Under autoregressive noise, comparison of the objective function values for our example indicates that the optimal choice of k is 3, and for a choice of $\rho = 0.2$, it is sufficient to add noise with variance $\sigma^2 = 0.68$. Again, over a range of correlation coefficient values, autocorrelated noise can offer the same level of protection as independent noise while using a lower noise level.

This example demonstrates how an optimal disclosure limitation strategy is determined in our framework. Further it shows that adding positively correlated noise is, under some conditions, preferable to using either data perturbation or independent noise addition.

6. CONCLUSION

We have investigated disclosure limitation methods for statistical databases. In doing so we have focused on protection against tracker attacks, a problem of special importance when the databases may be accessed online

Duncan and Mukherjee, Optimal Disclosure Limitation Strategy

through database management systems. The tension between data access and disclosure is modeled as an optimization problem: The goal of the database administrator is to minimize restrictions on legitimate access while ensuring that the risk of disclosure is below an acceptable threshold. Operational measures of restrictions on data access and disclosure risk are developed. The utility of data access is expressed so that tradeoffs can be made between the quantity and the quality of data to be released. Optimal disclosure limitation strategies are derived as solutions to the problem.

Our results indicate that a combination of existing disclosure control methods based on query restriction and data masking provides better protection than when these methods are separately used. While the vulnerability of data masking methods using independent additive noise to estimators based on repeated queries is well understood, we demonstrate that data perturbation provides no additional protection against tracker attacks and hence should not be used in combination with query size restriction control.

Recognizing that independent noise addition and data perturbation are extreme cases in the continuum of data masking using positively correlated additive noise, we analyze the general case. Optimal strategies are obtained for the data snooper attempting to estimate sensitive values. Expressions are derived for the variance of these estimators as functions of the covariance matrix of the noise process. Solutions to the optimization problem indicate that under certain conditions adding autocorrelated noise is preferable to using either independent noise addition or data perturbation for data masking.

The approach adopted in this study provides database administrators with practical guidance in selecting appropriate disclosure limitation measures. Optimal disclosure limitation strategies are specified as functions of parameters that reflect concerns about restrictions on legitimate access and disclosure risk. This specification allows the database administrator to consider an explicit tradeoff between data quantity and data quality. We provide some guidance on the choice of these parameters.

This study raises a research question. How can our approach be extended to multivariate settings? To keep our analysis tractable while demonstrating the utility of our framework, we restricted our focus to sum queries on a single sensitive attribute. The problem addressed is, however, a much more general one. Tendick and Matloff (1994) address bias issues in the multivariate data perturbation case. Tracker attacks are equally applicable to queries such as averages and moments involving a vector of attributes. Interesting aspects of this question

Duncan and Mukherjee, Optimal Disclosure Limitation Strategy

include the choice of appropriate measures for disclosure risk and restrictions on access. Further, problems may arise when concerns about two or more attributes lead to conflicting objectives (Mukherjee 1998).

APPENDIX : PROOFS

Result 1: Assume that a user adopts the optimal sequence ($R_{C \cup T}$, R_T , R_{-T} , $R_{C \cup -T}$) of tracker terms (as specified by Result 2), and the noise components of only the queried records are updated upon each query.

Then the covariance matrix $\underline{\Omega}$ of the tracker terms has the form

$$\mathbf{s}^2 \begin{bmatrix} |C \cup T| & \mathbf{r}_1|T| & \mathbf{r}_2|C \cap -T| & \mathbf{r}_3|C| \\ \mathbf{r}_1|T| & |T| & 0 & \mathbf{r}_2|C \cap T| \\ \mathbf{r}_2|C \cap -T| & 0 & |-T| & \mathbf{r}_1|-T| \\ \mathbf{r}_3|C| & \mathbf{r}_2|C \cap T| & \mathbf{r}_1|-T| & |C \cup -T| \end{bmatrix}.$$

Hence, the variance of the estimator of a restricted value R_C , based on a single tracker application is

$$\begin{aligned} V[\tilde{R}_C] &= (1 \ -1 \ -1 \ 1) \underline{\Omega} (1 \ -1 \ -1 \ 1)' \\ &= \mathbf{s}^2 \left[(2N+|C|) - 2(N \mathbf{r}_1 + |C|(\mathbf{r}_2 - \mathbf{r}_3)) \right] \end{aligned}$$

The above result indicates that the variance of a tracker estimator decreases with increasing covariance between the terms in the tracker formula. Furthermore, the variance of the mean of a set of observations decreases with decreasing correlation between the observations. A snooper attempting to minimize the variance of an estimator may satisfy both these goals by applying the tracker formula repeatedly, and using the mean of a sequence ($\hat{R}_{C1}, \dots, \hat{R}_{Cr}$) of tracker results as an estimator of R_C .

Result 2: When the masked values ($R_{C \cup T}$, R_T , R_{-T} , $R_{C \cup -T}$) are used in the tracker formula (1) to estimate a restricted value R_C under QSR control, the variance of the estimator $\underline{\tilde{R}}_C$ based on a single application of the tracker formula may be expressed as

Duncan and Mukherjee, Optimal Disclosure Limitation Strategy

$$V[\tilde{R}_C] = V[R_{C \cup T}] + V[R_{C \cup \neg T}] + V[R_T] + V[R_{\neg T}] + \\ 2(\text{Cov}[R_{C \cup T}, R_{C \cup \neg T}] - \text{Cov}[R_{C \cup T}, R_T] - \text{Cov}[R_{C \cup T}, R_{\neg T}] \\ - \text{Cov}[R_{C \cup \neg T}, R_T] - \text{Cov}[R_{C \cup \neg T}, R_{\neg T}] + \text{Cov}[R_T, R_{\neg T}]).$$

A data snooper, in attempting to minimize the variance, seeks to keep the covariance term making a positive contribution small and those making a negative contribution large.

The covariance between $\underline{R_T}$ and $\underline{R_{\neg T}}$ is 0. This is because the noise added to X_i is independent of noise added to X_j if $i \neq j$, and since T and $\neg T$ have no elements in common.

Under a noise addition scheme in which the covariances between responses are non-increasing in time (so, $\rho_s \leq \rho_t$ for $s > t$), a data snooper would obtain $\hat{R}_{C \cup T}$ first and $\hat{R}_{C \cup \neg T}$ last to minimize the positive contribution of this positive covariance term. All the other covariance terms make a negative contribution and hence the snooper aims to maximize them.

The greater the numbers of records two query sets have in common, the higher is their covariance. The size of the overlap between the sets $C \cup T$ and T is greater than the overlap between $C \cup T$ and $\neg T$. This is because, $|(C \cup T) \cap \neg T| \leq |C| \leq k < |T| = |(C \cup T) \cap T|$.

It follows that the covariance between $\hat{R}_{C \cup T}$ and \hat{R}_T makes a more significant contribution towards the variance than does the covariance between $\hat{R}_{C \cup T}$ and $\hat{R}_{\neg T}$. By a similar argument, the covariance between $\hat{R}_{C \cup \neg T}$ and $\hat{R}_{\neg T}$ makes a more significant contribution towards the variance than does the covariance between $\hat{R}_{C \cup \neg T}$ and \hat{R}_T . Hence the optimal strategy for the snooper while employing the tracker is to sequence the queries to get responses in the order $(R_{C \cup T}, R_T, R_{\neg T}, R_{C \cup \neg T})$.

REFERENCES

- Adam, N. R., Gangopadhyay, A., and Holowczak, R. (1998), "A Survey on Research on Database Protection," Proceedings of Statistical Data Protection '98, Eurostat, Lisbon, Portugal, March 25-27.
- Adam, N. R. and Wortmann, J.C. (1989), "Security-Control Methods for Statistical Databases: A Comparative Study," ACM Computing Surveys, 21, 4, 515-556.

Duncan and Mukherjee, Optimal Disclosure Limitation Strategy

- Ahituv, N., Lapid, Y. and Neumann, S. (1988), "Protecting Statistical Databases Against Retrieval of Private Information," *Computers & Security*, 7, 59-63.
- Beck, L.L. (1980), "A Security Mechanism for Statistical Databases," *ACM Trans. Database Systems*, 5, 3, 316-338.
- Bethlehem, J.G., Keller, W.J. and Pannekoek, J. (1990), "Disclosure Control of Microdata," *Journal of American Statistical Association*, 85, 409, 38-45.
- Cassel, C.M. (1976), "Probability Based Disclosures," in *Personal Integrity and the Need for Data in the Social Sciences*, eds T. Dalenius and A. Klevmarck, Stockholm: Swedish Council for Social Science Research, 189-193.
- Chin, F. Y. and Ozsoyoglu, G. (1981) "Statistical Database Design," *ACM Transactions on Database Systems*, 6, 113-139.
- Cox, L.H. (1980), "Suppression Methodology and Statistical Disclosure Control," *Journal of American Statistical Association*, 75, 377-385.
- Cox, L.H. (1995), "Network Models for Complementary Cell Suppression," *Journal of American Statistical Association*, 90, 1453-1462.
- Dalenius, T. and Reiss, S. P. (1982), "Data Swapping: A Technique for Disclosure Control," *Journal of Statistical Planning and Inference*, 6, 73-85
- Dalenius T. (1988), "Controlling Invasion of Privacy in Statistical Research Unit," *Statistics Sweden*, Stockholm.
- Denning, D. E., Denning, P. J. and Schwartz, M. D. (1978), "The Tracker: A Threat to Statistical Database Security," *ACM Trans. Database Systems*, 4, 1, 7-18.
- Denning, D. E., and Schlörer, J. (1980), "A Fast Procedure for Finding a Tracker in a Statistical Database," *ACM Transactions on Database Systems*, 5, 1, 88-102.
- Dobkin D., A.K. Jones, and R.J. Lipton, (1979) *Secure Databases: Protection against User Influence*, *ACM Trans. on Database Systems*, 4, 1, 97-106.
- Duncan, G. T., Jabine, T. B., and de Wolf, V. A. (1993) *Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics*. Washington, D.C.: National Academy Press.
- Duncan, G.T. and Lambert, D. (1986), "Disclosure - Limited Data Dissemination," *Journal of American Statistical Association*, 81, 10-28 (with discussion by L. Cox, O. Frank, J. Gastwirth, and H. Roberts)
- Duncan, G.T. and Lambert, D. (1989), "The Risk of Disclosure of Microdata," *Journal of Business and Economic Statistics*, 7, 207-217.
- Duncan, G.T. and Mukherjee, S. (1991), "Microdata Disclosure Limitation in Statistical Databases: Query Size Restriction and Random Sample Query Control," *Proceedings of the 1991 IEEE Computer Society Symposium on Research in Security and Privacy*, 278-287.
- Duncan, G.T. and Mukherjee, S. (1992), "Disclosure Limitation in Statistical Databases Using Autoregressive Noise," *Database Security VI: Status and Prospects*, (ed. C. Landwehr and B. Thuraisingham), *IFIP Transactions A-21*, North Holland, pp 211-224.
- Duncan, G.T. and Pearson, R.W. (1991), "Enhancing Access to Data while Protecting Confidentiality: Prospects for the Future," *Statistical Science*, 6, 3, 219-239.

Duncan and Mukherjee, Optimal Disclosure Limitation Strategy

- Fellegi, I. P. (1972), "On the Question of Statistical Confidentiality," *Journal of the American Statistical Association*, 67, 7-18.
- Frank, O. (1976), "Individual Disclosures From Frequency Tables," in *Personal Integrity and the Need for Data in the Social Sciences*, eds T. Dalenius and A. Klevmarcken, Stockholm: Swedish Council for Social Science Research, 175-187.
- Frank, O. (1978), "An Application of Information Theory to the Problem of Statistical Disclosure," *Journal of Statistical Planning and Inference*, 2, 143-152.
- Frank, O. (1983), "Statistical Disclosure Control," *Statistical Review*, 5, 173-178.
- Friedman, A. D. and Hoffman, L. J. (1980) "Towards a fail-safe approach to statistical databases," *Proceedings of IEEE Symposium on Security and Privacy*.
- Fuller, W. A. (1993), "Masking Procedures for Microdata Disclosure Limitation," *Journal of Official Statistics*, 9, 2, 383-406.
- Jabine, T. B. (1993) "Statistical Disclosure Limitation Practices of the United State Statistical Agencies," *Journal of Official Statistics*, 9, 2, 427-454.
- Javitz, H.S. and Valdes, A. (1991), "The SRI IDES Statistical Anomaly Detector," *Proceedings of the IEEE Symposium on Research in Security and Privacy*.
- Keller-McNulty, S., McNulty, M. S., and Unger, E.A. (1989), "The Protection of Confidential Data," *Proceedings of the 21st Symposium on Interface*, American Statistical Association, Alexandria, VA, pp. 215-219.
- Keller-McNulty, S. and Unger, E.A. (1993), "Database Systems: Inferential Security," *Journal of Official Statistics*, 9, 2, 475-500.
- Kim, J. J. (1986) "A Method for Limiting Disclosure in Microdata Based on Random Noise and Transformation," *American Statistical Association, Proceedings of the Section on Survey Research Methods*, 456-461.
- Lambert, D. (1993) "Measures of Disclosure Risk and Harm," *Journal of Official Statistics*, 9, 2, 313-331.
- Matloff, N. S. (1986), "Another Look at Noise Addition for Database Security," *Proceedings of the IEEE Computer Society Symposium on Research in Security and Privacy*, 173-180.
- Michalewicz, Z., Li, J. and, Chen, K. (1990), "Ranges and Trackers in Statistical Databases," *Proceedings of the 5th International Conference on Statistical and Scientific Databases*, Springer-Verlag, Lecture Notes on Computer Science, No. 420, 65-79.
- Mukherjee, S. (1998), "Should Non-Sensitive Attributes be Masked? Data Quality Implications of Data Perturbation in Regression Analysis," *Proceedings of the Thirty-First Annual Hawaii International Conference on System Sciences*, 6, 223-231.
- Paass, G. (1988), "Disclosure Risk and Disclosure Avoidance for Microdata," *Journal of Business and Economic Statistics*, 6, 487-500.
- Rainwater, L. and Smeeding, T. M. (1988), "The Luxembourg Income Study: The Use of International Telecommunications in Comparative Social Research," *ANNALS, AAPSS*, 495, 95-105.
- Schlörer, J. (1975), "Identification and Retrieval of Personal Records from a Statistical Databank," *Methods of Inform. in Medicine.*, 14, 1, 7-13.

Duncan and Mukherjee, Optimal Disclosure Limitation Strategy

- Schlörer J. (1980), "Disclosure from Statistical Databases: Quantitative Aspects of Trackers," *ACM Transactions on Database Systems*, 5, 4, 467-492.
- Spruill, N.L. and Gastwirth, J.L. (1982), "On the Estimation of the Correlation Coefficient from Grouped Data," *Journal of the American Statistical Association*, 77, 614-620.
- Tendick, P. (1991), "Optimal Noise Addition for the Preservation of Confidentiality in Multivariate Data," *Journal of Statistical Planning and Inference*, 27, 342-353.
- Tendick, P. and Matloff, N. (1994) "A Modified Random Perturbation Method for Database Security." *ACM Transactions on Database Systems*, 19, 47-63.