The Security of Confidential Numerical Data in Databases

Rathindra Sarathy • Krishnamurty Muralidhar
Department of Management, Oklahoma State University, Stillwater, Oklahoma 74078-4011
School of Management, Gatton College of Business & Economics, University of Kentucky, Lexington, Kentucky 40506-0034
sarathy@okstate.edu • kmurald@pop.uky.edu

Organizations are storing large amounts of data in databases for data mining and other types of analysis. Some of this data is considered confidential and has to be protected from disclosure. When access to individual values of confidential numerical data in the database is prevented, disclosure may occur when a snooper uses linear models to predict individual values of confidential attributes using nonconfidential numerical and categorical attributes. Hence, it is important for the database administrator to have the ability to evaluate security for snoopers using linear models. In this study we provide a methodology based on Canonical Correlation Analysis that is both appropriate and adequate for evaluating security. The methodology can also be used to evaluate the security provided by different security mechanisms such as query restrictions and data perturbation. In situations where the level of security is inadequate, the methodology provided in this study can also be used to select appropriate inference control mechanisms. The application of the methodology is illustrated using a simulated database.

(Confidentiality; Data Perturbation; Database Security; Inferential Disclosure; Inferential Security)

1. Introduction

Organizations increasingly face the problem of protecting confidential information contained in their databases. In many cases, legal requirements dictate that sensitive data regarding individuals and organizations are protected from disclosure (Goldstein 1992). Protecting databases from unauthorized users (hackers) has received considerable attention, while less attention has been paid to preventing disclosure of confidential data to "snoopers." The term snooper refers to a person who is authorized to access the database and misuses such access to gain unauthorized information (Adam and Wortmann 1989, Hoffer and Straub 1989). The security threat posed by snoopers generally takes the form of undesired inferences about confidential data using other data available either within or outside the database. The rapidly growing array of tools for data mining and other legitimate sophisticated analysis, coupled with the proliferation of online databases, also increases this threat.

The context in this study is the evaluation of security of confidential numerical data residing in organizational databases. It is assumed that individual values of confidential numerical attributes in the database are not available to users. The issue of disclosure is evaluated from the perspective of a snooper who may use legitimate queries to infer information regarding confidential numerical attributes (inferential value disclosure). Government agencies such as the Census Bureau, for reasons of anonymity, consider disclosure to have occurred even when a snooper is able to identify that a specific individual is a part of a given database.
(Duncan and Pearson 1991). By contrast, the knowledge that an individual (or entity) is a part of the database generally does not constitute disclosure in an organizational context. Disclosure is said to occur if the snooper is able to explain a higher proportion of variability in the confidential numerical attributes than intended by the database administrator (partial value disclosure), even if the exact value is not disclosed (Adam and Wortmann 1989, Palley and Simonoff 1987). As we show later, when access to individual values of the confidential attributes are prevented, the database is susceptible to partial value disclosure when a snooper uses linear models.

A key element in preserving confidentiality of sensitive data is the ability to evaluate the extent of disclosure for such data. Such an evaluation will allow the database administrator (DBA) to select the most appropriate security control mechanism for the database. Conversely, an inappropriate evaluation of security could lead to disclosure of confidential data that is greater than intended by the organization, resulting in decreased confidence among users of the database as well as potential legal action. Palley and Simonoff (1987) showed that even when security mechanisms are in place, a snooper would be able to gain accurate estimates of confidential attributes using simple queries and linear models. Therefore, it is important for the database administrator (DBA) to be able to evaluate the security of the database under such conditions. Existing methods of security evaluation do not provide the DBA with this ability. The objective of this study is to develop a methodology for evaluating inferential value disclosure of either individual or linear combinations of confidential numerical attributes by snoopers using linear models.

The remainder of this paper is organized as follows. The next section discusses existing methods for evaluating security. The third section provides a theoretical basis for using canonical correlation analysis for evaluating security. The fourth section illustrates security evaluation in the context of an organizational database. The fifth section discusses the impact of various inference control mechanisms on security. The sixth section discusses the computational issues and the final section provides the conclusions of the study.

2. Existing Methods for Evaluating Security of Confidential Numerical Data

Consider a database consisting of a set of $N$ records with $K$ numerical, confidential attributes $X$ with mean vector $\mu_X$ and covariance matrix $\Sigma_{XX}$. Assume that the database also consists of a set of $L$ nonconfidential attributes $S$ with mean vector $\mu_S$ and covariance matrix $\Sigma_{SS}$. Let $\Sigma_{XS}$ represent the covariance between $X$ and $S$. The objective of a snooper is to predict $X$ using $S$. Further, users are allowed access only to aggregate information regarding $X$, and access to individual values is prevented. Other than mechanisms used to prevent access to individual values of confidential attributes, no other security mechanisms are assumed. The impact of security mechanisms is discussed in a later section. Also, no distributional assumptions regarding $X$ and/or $S$ are made and the results shown in this study are applicable irrespective of the underlying distribution of the database.

The use of linear models to compromise confidential numerical information in databases was illustrated by Palley and Simonoff (1987). They showed that a snooper intending to estimate a confidential attribute $X_i \in X$, using some linear combination of $S$, could create a synthetic database that captures the inherent structure (the statistical and numerical characteristics of individual attributes and the relationships among the attributes) of the database. This structure can be created using only typical queries such as COUNT, MEAN, STANDARD DEVIATION, CORRELATION, etc. Once the structure is identified, the snooper can then apply linear regression models to this synthetic database to predict the (unobservable) values of the confidential attribute $X_i$ using the (observable) values of the nonconfidential attributes $S$.

Note that the snooper need not have access to any individual values of confidential attributes $X_i$ when estimating them using linear models because analytical expressions can be derived for estimating the prediction equation using only the means and covariance structure of the attributes. Such analytical expressions for the prediction equation cannot be derived for non-linear models, and it is necessary to use numerical search procedures that would require the snooper to
have access to individual values of confidential attributes (Neter et al. 1990). Hence, it would not be possible for snoopers to use only the structure of the database to estimate nonlinear models. In some situations, snoopers may be able to estimate higher-order moments of the distribution of the confidential attributes conditioned on the nonconfidential attributes. This may provide the snooper with a predictive ability that is higher than that resulting from a linear model. In this study, we focus only on the more widely applicable case where a snoop employs linear models for estimating confidential information.

Palley and Simonoff (1987) showed that the snoop is able to perform accurate linear estimation of a confidential attribute even when most existing inference controls are employed. In measuring the predictive ability of the snoop, Palley and Simonoff used $R^2$ (the coefficient of determination obtained by a regression of $S$ on $X$). Concerning $R^2$ they indicate:

This statistic measures the proportion of variability in the confidential attribute that is accounted for by the regression. $R$-squared is also intimately connected with the gain in predictive accuracy when using regression. Specifically, the proportional reduction in the length of the confidence interval for the prediction of the confidential attribute by using regression is roughly $1 - (1 - R^2)^{1/2}$ (Palley and Simonoff 1987, p. 598)

Thus, when a snoop is able to predict $R^2$ proportion of the variability in an individual confidential attribute, the level of security provided is the proportion of unexplained variability, $1 - R^2$.

When a database has multiple confidential attributes, the threat of disclosure can be magnified further. Tendick (1991) showed that even if the level of security provided for a single confidential attribute is adequate, the level of security provided for linear combinations of confidential attributes could be very low. For example, consider a situation where a snoop is attempting to estimate the confidential attribute PROFIT based on nonconfidential attributes. The snoop could use Palley and Simonoff's approach to estimate REVENUE and EXPENSES individually, and estimate PROFIT as (ESTIMATED REVENUE - ESTIMATED EXPENSES). However, if the snoop also estimated linear combinations of REVENUE and EXPENSES, it is likely (though not necessary) that the prediction of the linear combination (PROFIT = REVENUE - EXPENSES) has a higher level of accuracy (Tendick 1991). It is also possible that some other linear combination of REVENUE and EXPENSES can be predicted with an even higher level of accuracy than (REVENUE - EXPENSES). Therefore, it is important that any evaluation of inferential security take into account the potential gain in accuracy that a snoop might obtain by estimating linear combinations of confidential attributes. In §4, we illustrate this further using a simulated database.

The evaluation of security provided for linear combinations of confidential attributes can be formalized. Consider one possible linear combination involving $X$, 

$$ Z = c^TX = c_1X_1 + c_2X_2 + \ldots + c_KX_K. $$

(1)

Tendick provided a measure to determine the proportion of variability in $Z$ that can be explained using a linear combination of the *perturbed* values of $X$ (Tendick 1991, p. 345, Equation 3.8). As with $X$, the snoop can estimate the values of $Z$ by using the relationships between $Z$ and the nonconfidential attributes in the database, even if access to individual values of $Z$ are prevented. We can modify the measure proposed by Tendick (1991) to account for the proportion of variability in $Z$ that is explained using a linear combination of the nonconfidential attributes $S$ as (the coefficient of determination of $Z$ and $S$):

$$ R_{Z|S}^2 = \frac{c^T \Sigma_{ZS} \Sigma_{SS}^{-1} \Sigma_{SY}c}{c^T \Sigma_{XX}c}, $$

(2)

where the covariance matrix between $(S^T, Z)^T$ is partitioned as (Tendick 1991):

$$ \begin{pmatrix} \Sigma_{SS} & \Sigma_{SY}c \\ c^T \Sigma_{XS} & c^T \Sigma_{XX}c \end{pmatrix}. $$

When a snoop attempts to estimate a single confidential attribute $X_j$ directly (such as PROFIT), $c_i = 1$ and $c_j = 0$, for all $j \in K, j \neq i$ in Equation (2). In this case, $R_{Z|S}^2$ is the same as the $R^2$ measure indicated in Palley and Simonoff (1987). Thus, $R_{Z|S}^2$ can be considered a generalization of the $R^2$ measure for linear combinations of confidential attributes.

Tendick's study highlights an important area of concern for the DBA in assessing inferential value disclosure. A DBA could conclude (based on $R^2$) that adequate security has been provided for an individual
attribute. However, a snooper using nonconfidential attributes to estimate a linear combination of confidential attributes could actually explain more than $R^2$ (i.e., $R^2_{ZIS}$ may be greater than $R^2$). The resulting lower security would be given by $(1 - R^2_{ZIS})$. Note that this represents the security provided for a single linear combination $Z$. There is no guarantee that other linear combinations do not result in a lower security than the specific linear combination considered by the DBA. Most organizational databases typically contain numerous attributes that could lend themselves to potentially thousands of linear combinations. In many instances, linear combinations of different attributes are not stored separately in the database because they can be computed from other attributes. Thus, the necessity to evaluate a very large number of linear combinations limits the usefulness of Equation (2) as a general measure for evaluating security. For comparing security provided by perturbation methods, Muralidhar et al. (1999) overcame this problem by using canonical correlation analysis. In the following section we show that canonical correlation analysis can be employed in the context of an organizational database, regardless of whether perturbation is employed, to evaluate inferential value disclosure.

3. Canonical Correlation Analysis as a Security Evaluation Approach

Canonical Correlation Analysis (CCA) is a statistical procedure that is used to identify and quantify the relationship between two sets of variables. In the context of security evaluation in organizational databases, individual values of one set of variables (nonconfidential attributes) is observable to a snooper, while the individual values of the second set of variables (confidential attributes) are unobservable (or masked). CCA identifies a linear combination of variables in one set that have the highest correlation with a linear combination of variables in another set (Johnson and Wichern 1992). Hence, canonical correlation analysis is well suited for evaluating the level of security when estimating linear combinations of the (unobservable) confidential attributes using the (observable) nonconfidential attributes.

3.1. Canonical Correlation as a General Measure of Security

Consider two linear combinations $c^TX$ and $d^TS$. The correlation between $c^TX$ and $d^TS$ can be determined as (see Johnson and Wichern 1992, p. 462):

$$\text{Corr}(c^TX, d^TS) = \frac{c^T \Sigma_{XS} d}{\sqrt{c^T \Sigma_{XX} c} \sqrt{d^T \Sigma_{SS} d}}$$

(3)

Let $a = \Sigma_{XX}^{1/2} c$ and let $b = \Sigma_{SS}^{1/2} d$, then

$$\text{Corr}(c^TX, d^TS) = \frac{a^T \Sigma_{XX}^{1/2} \Sigma_{XS} \Sigma_{SS}^{-1/2} b}{\sqrt{a^T a} \sqrt{b^T b}}.$$ 

(4)

Using the Cauchy-Schwarz inequality, we can show that:

$$\sqrt{(a^T \Sigma_{XX}^{-1/2} \Sigma_{XS} \Sigma_{SS}^{-1/2} \Sigma_{SS}^{-1/2} d) b^T b} \geq a^T \Sigma_{XX}^{-1/2} \Sigma_{XS} \Sigma_{SS}^{-1/2} b.$$ 

(5)

Since $\Sigma_{XX}^{-1/2} \Sigma_{XS} \Sigma_{SS}^{-1/2} \Sigma_{SS}^{-1/2}$ is symmetric (Johnson and Wichern 1992, result (2.51)), it can also be shown that:

$$a^T \Sigma_{XX}^{-1/2} \Sigma_{XS} \Sigma_{SS}^{-1/2} \Sigma_{SS}^{-1/2} a = \lambda a^T a,$$

(6)

where $\lambda$ is the largest eigenvalue of $\Sigma_{XX}^{-1/2} \Sigma_{XS} \Sigma_{SS}^{-1/2} \Sigma_{SS}^{-1/2}$. This implies:

$$\sqrt{a^T a} \sqrt{b^T b} \geq \sqrt{(a^T \Sigma_{XX}^{-1/2} \Sigma_{XS} \Sigma_{SS}^{-1/2} \Sigma_{SS}^{-1/2} d) b^T b} \geq a^T \Sigma_{XX}^{-1/2} \Sigma_{XS} \Sigma_{SS}^{-1/2}.$$ 

(7)

Therefore,

$$\text{Corr}(c^TX, d^TS) \leq \frac{a^T \Sigma_{XX}^{-1/2} \Sigma_{XS} \Sigma_{SS}^{-1/2} b}{\sqrt{a^T a} \sqrt{b^T b}} \leq \sqrt{\lambda},$$

(8)

where $\sqrt{\lambda}$ is the largest canonical correlation between $X$ and $S$. Also, substituting $c = \Sigma_{XX}^{1/2} \Sigma_{SS}^{-1/2} a$ back in Equation (6) and reorganizing yields,

$$\frac{c^T \Sigma_{XS} \Sigma_{SS}^{-1/2} \Sigma_{SS}^{-1/2} c}{c^T \Sigma_{XX} c} \leq \lambda.$$

(9)

The left-hand side of the equation in (9) is exactly the same as the expression in (2). Hence,

$$R^2_{ZIS} = \frac{c^T \Sigma_{XS} \Sigma_{SS}^{-1/2} \Sigma_{SS}^{-1/2} c}{c^T \Sigma_{XX} c} \leq \lambda.$$ 

(10)

Equation (10) shows that, irrespective of the specific linear combination $c^TX$ being considered, the proportion of variability explained using a linear combination of $S$ is
less than or equal to \( \lambda \), in the worst case. It also highlights the advantage of canonical correlation over previous measures. The canonical correlation analysis eliminates the need for the DBA to evaluate a large number of linear combinations by providing a general upper bound for the \( R^2_{21} \) measure derived by Tendick (1991, Equation (2)). As shown earlier, \( R^2 \), which is used to measure the security for an individual confidential attribute (Palley and Simonoff 1987), is a special case of \( R^2_{21} \) and hence of \( \lambda \). Thus \( \lambda \) represents the most general measure of inferential value disclosure, and \((1 - \lambda)\) represents the worst-case security. The value of \( \lambda \) is usually computed as the primary eigenvalue resulting from the eigenvalue equation:

\[
\Sigma_{XX}^{-1}\Sigma_{XS}\Sigma_{SS}^{-1}\Sigma_{SX} - \lambda I = 0. \tag{11}
\]

Canonical correlation analysis also provides the DBA with the ability to estimate the level of accuracy with which a snooper can predict a confidential attribute. In general, given a variable \( Z \) whose variance is \( \sigma^2_Z \), a rough estimate of the smallest standard error can be determined as \( [(1 - \lambda) \sigma^2_Z]^{0.5} \). Based on this estimate of standard error, a rough 95% confidence interval for \( Z \) can be written as:

\[
\hat{Z} \pm 2 \sqrt{(1 - \lambda)\sigma^2_Z}^{0.5}. \tag{12}
\]

### 3.2. CCA and Categorical Attributes

Organizational databases typically consist of both numerical and categorical attributes. Categorical attributes present in the database will have two types of impacts on the structure of the database:

1. The mean value of one or more confidential numerical attributes is different for at least one of the subsets defined by one or more categorical attributes, as compared to other subsets, and

2. The correlation between numerical attributes is different for at least one of the subsets defined by one or more categorical attributes, as compared to other subsets.

When either or both of the above scenarios are present in the database, a snooper may be able to predict linear combinations of confidential attributes with greater accuracy (than when categorical attributes are not present or used in analysis) by using one of the following approaches:

1. The snooper can convert any meaningful categorical attribute (excluding simple identifiers such as social security numbers) with \( D \) possible values into \((D - 1)\) binary, numerical attributes. These binary variables can then be used, in addition to other numerical nonconfidential attributes, to predict the confidential attributes.

2. The snooper could divide the database into \( D \) subsets of data corresponding to a categorical attribute, and attempt to predict confidential attributes (using other numerical nonconfidential attributes) for each of the \( D \) subsets.

3. Where multiple categorical attributes are present, the snooper could combine both of the approaches above converting some to binary variables and dividing the database based on the values of one or more of the other categorical attributes.

Under these conditions CCA can be applied to every subset defined by the values of one or more categorical attributes. If the data has been divided into a subset based on a given categorical variable, then this categorical variable cannot be used for prediction because the value of this variable is the same for all observations in this subset. Nonconfidential numerical attributes, and those binary variables representing categorical attributes that are not used to divide the database into subsets, are used as predictor variables.

Note that dividing the numerical attributes into categories, ranges, or intervals will result in loss of information and will not improve the predictive ability of the snooper (Palley and Simonoff 1987, Neter et al. 1996) when the assumptions underlying linear models are satisfied. Linear models assume that, for the entire range of data values, the relationship between the numerical variables is linear and that the error terms have a constant variance (Neter et al. 1996). When these two assumptions are satisfied, then even if the regression model is estimated using only the data in an arbitrary subset defined by nonconfidential numerical data, the proportion of variability explained is approximately the same as that of the entire data set (subject to sampling error). However, if these assumptions are not satisfied it is possible that the proportion of variability explained may be higher. A further discussion of this issue can be found in §5.4.

Assume that in addition to the variables (\( X \) and \( S \))
described earlier, $D$ represents the $M$ nonconfidential categorical binary attributes. Let $R$ consist of $S$ and $D$. Let $\Sigma_{RR}$ represent the covariance matrix of $R$ and let $\Sigma_{XR}$ represent the covariance between $X$ and $R$. If all the categorical attributes are converted to binary predictive variables (i.e., the database is not divided into any subsets at all), then in order to evaluate the canonical correlation, $\Sigma_{XR}$ would be computed for the entire database, and will replace $\Sigma_{XS}$ in Equation (11). If one (or more) of the categorical attribute values are used to obtain subsets of the database, the corresponding binary variables are removed from the set $D$, and canonical correlation analysis is carried out for each of these subsets using the appropriate covariance matrices resulting from the data in the subsets. Then, the security provided can be determined as $(1 - \lambda^*)$, where $\lambda^*$ is the largest $\lambda$ obtained for any subset. The level of security $(1 - \lambda^*)$ represents the minimum proportion of variability in any linear combination of confidential attributes that cannot be explained by a snooper. The exact type, number, and/or sequence of queries cannot result in a lower level of security as long as the assumption of constant variance of error terms and linear relationship between variables hold.

Thus, the CCA approach provides the DBA with both a measure of the level of security as well as the accuracy with which a snooper can predict a confidential attribute, taking into account the presence of categorical attributes. An illustration of this approach using a simulated database is provided in the following section.

4. A Security Evaluation Example

Consider a payroll database with nine attributes. The numerical confidential attributes ($X = X_1, X_2, X_3$) are annual salary (SALARY), annual bonus (BONUS), and the value of stock options (STOCKS), while the numerical nonconfidential attributes ($S = S_1, S_2, S_3$) are age (AGE), number of years of formal education (EDU), and number of years of experience (EXP). There are also three nonconfidential categorical attributes that represent the location of the employee (LOC), department that they belong to (DEPT), and their gender (GEN). The LOC attribute has three possible values (LA, CHI, and NY); the DEPT attribute has four possible values (ACC, MKT, MIS, and PRO); and the GEN attribute has two possible values (M and F). If all categorical attributes are converted into binary variables, this results in a total of six dummy binary variables ($D = D_1, D_2, \ldots, D_6$). The following specifications were added to simulate the impact of the categorical attributes on the numerical attributes:

(1) The salary of GEN = M was specified to be (on average) $2,000 higher than GEN = F.

(2) The average bonus in DEPT = MKT, DEPT = PRO, and DEPT = MIS are, on average, $250, $500, and $750 more than the average bonus in DEPT = ACC.

(3) The correlation between the EXP and STOCK attributes was specified to be 0.60 in DEPT = MIS, LOC = NY (while the same correlation in the rest of the data set was 0.15).

(4) The relationships between STOCKS and AGE, EDU, EXP were specified to be higher in LOC = CHI, DEPT = PRO, GEN = M.

As discussed earlier, the above specifications represent all possible types of impacts that categorical attributes can have on the structure of the database. If the CCA method is able to capture the impact of the above specifications on security, then it is also capable of capturing multiple instances of such changes. A database consisting of 50,000 records was simulated using the above characteristics. The covariance matrix of the numerical attributes is shown in Table 1. Further, other than security mechanisms that prevent access to individual data points of confidential attributes and to prevent exact disclosure, no other query or access restrictions are assumed. The impact of these restrictions would be to allow analysis at the aggregate level without access to individual values.

CCA reveals that a snooper following Palley and

<table>
<thead>
<tr>
<th>Table 1 Covariance Matrix of Numerical Attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Salary</td>
</tr>
<tr>
<td>--------</td>
</tr>
<tr>
<td>Salary</td>
</tr>
<tr>
<td>Bonus</td>
</tr>
<tr>
<td>Stocks</td>
</tr>
<tr>
<td>Age</td>
</tr>
<tr>
<td>Education</td>
</tr>
<tr>
<td>Experience</td>
</tr>
</tbody>
</table>
Simonoff's (1987) approach could use the nonconfidential numerical attributes AGE, EDU, and EXP, to explain:

1. 17% of the variability in the confidential attribute SALARY alone,
2. 27% of the variability in the confidential attribute BONUS alone, and
3. 16% of the variability in the confidential attribute STOCKS alone.

The importance of considering linear combinations of attributes is evident in this data. Assume that in this organization the total incentives received by employees averaged 15% of salary, with higher performers receiving more and lower performers receiving less. Employees often receive such information either through official channels or through the organizational “grapevine.” Using this information, a snooper may attempt to gain information regarding the “above/below average incentives” by predicting the following linear combination: BONUS + STOCKS - (0.15 X SALARY), using AGE, EDU, and EXP. In this case, the proportion of variability explained is about 32%, which is greater than the maximum proportion of variability explained in an individual confidential attribute (27%). In practice, it would be impossible for the DBA to visualize every such combination of confidential attributes, let alone evaluate the level of security provided.

As discussed in the previous section, in order to evaluate security for this database it is necessary to consider the predictive ability of the snooper for each of the subsets that result when records are selected by categorical attributes. CCA analysis on the entire database reveals that the largest eigenvalue is 0.572 (the maximum proportion of variability explained in a linear combination of confidential attributes using a linear combination of both the numerical and categorical nonconfidential attributes), resulting in a security level of 42.8%. When subsets of records are selected by a single category, a total of nine subsets result (three locations + four departments + two genders). When subsets are defined by two categories, 26 subsets result (12 for LOC and DEPT combinations + 6 for LOC and GEN combinations + 8 for DEPT and GEN combinations). Finally, a total of 24 (3 x 4 x 2) subsets result for LOC, DEPT, and GEN combinations. This results in a total of 60 subsets (the whole database is also considered as a subset). For each of these subsets CCA was performed, taking into account the confidential attributes and the nonconfidential attributes (both numerical attributes and binary variables representing those categorical attributes not used to define the subset). The largest of the 60 eigenvalues (λ = 0.77) occurs in the subset (LOC = CHI, DEPT = PRO, GEN = M). This eigenvalue represents the most information that a snooper can gain regarding any linear combination of confidential attributes, in any subset. Hence, the worst-case security for the entire database is approximately 23% (1 - λ).

It is important to note that for this payroll database the security provided for an individual query based on categorical attributes cannot be lower than the security level that resulted from the CCA approach (namely 23%), no matter the nature and number of queries employed by a snooper. In a business environment it is not unusual, for example, to provide users with considerable flexibility in the queries that they can issue. Many of these queries could take advantage of the structure of the categorical attributes. For example, performance comparisons across divisions and departments would require queries relating to subsets of data based on categorical values. CCA provides the database administrator with the assurance that, even if it is not possible to exactly measure security for such queries, there is a certain minimum level of security that the CCA approach is able to quantify. As observed earlier, in cases where the assumptions of linear relationship and/or constant variance do not hold, creating arbitrary subsets using numerical nonconfidential attributes may result in lower security.

This security level can also be used by the DBA to estimate the accuracy with which a snooper can predict confidential attributes. Assume that a snooper is interested in predicting the confidential attribute SALARY (Z). A rough 95% confidence interval for the predicted value of SALARY is (from Table 1, the variance of SALARY = 87.78):

\[
\hat{Z} \pm 2([0.27 \times 87.78]^{0.5}) = \hat{Z} \pm 9.74. \quad (13)
\]

The DBA is guaranteed that, irrespective of the number and/or types of queries issued, a confidence interval for SALARY constructed by a snooper using linear
models will have an error margin of at least \( \pm 9.74K \). Since \((1 - \lambda^*)\) represents the worst-case security, it is more likely that the confidence interval constructed by a snooper is actually wider than the one above (resulting in error margins larger than \( \pm 9.74K \)).

The confidence interval above allows the DBA to assess the level of security afforded for individual values of confidential attributes. In some cases, the DBA may already have established disclosure intervals for individual values (Gopal et al. 1998). If the confidence intervals for individual values constructed using CCA are narrower than the disclosure intervals, then the DBA can conclude that the disclosure requirements are violated. The confidence interval above may also provide the DBA with some sense of whether the existing level of security is adequate or not. It can be seen from the above that, using linear models, the best predictor of the attribute SALARY will have an error margin of just under \( \pm 10,000 \). In most cases, such an interval would be considered to be an adequate protection of the SALARY attribute. Similar evaluations can be performed for the other confidential attributes to determine if the level of security is adequate or not. For a large database, constructing "disclosure intervals" around each individual value could be difficult and may be time consuming. Further, since there are a multitude of linear combinations of attributes, it would be difficult to establish disclosure intervals around individual values for each one of these. In these cases, the CCA approach could be used in place of specifying a disclosure interval for each individual value. Finally, if the number of records in a subset is less than or equal to the number of variables in the model, then the resulting eigenvalue will be 1.0 and security will be zero. This represents a situation in which, unless corrective action is taken, exact disclosure may result. This is one of the benefits of using the CCA approach; it also identifies subsets that could potentially lead to serious security breaches. In addition, security control mechanisms used to prevent exact disclosure would also restrict the snooper's prediction ability in such cases.

In summary, inferential security measured through CCA

- provides the DBA with the ability to quantify the basic, worst-case measure of security for any database on a standardized (zero to one) scale, where zero represents no security and one represents the highest possible level of security,
- is based on all the information available to the snooper and includes both numerical and meaningful categorical nonconfidential attributes,
- is independent of the queries issued by the snooper, and
- depends on the structure of the database.

In some situations, the DBA may wish to restrict or enhance the information available by considering additional types of inference control mechanisms, as discussed in the following section.

5. Evaluating Inference Control Mechanisms

One of the major reasons for evaluating the level of security is to maximize security (or minimize disclosure of confidential information to spooers) while providing maximum access to legitimate users. In this section, we provide the DBA with a methodology that can be used to evaluate the impact of different types of inference control mechanisms that are available for implementation. The reason for such an evaluation could be twofold. First, the DBA may consider the level of security provided as inadequate for a given case, and may wish to consider implementing inference control mechanisms that increase the level of security. Alternatively, a DBA may be satisfied with the level of security provided and may wish to consider increasing the level of access that legitimate users have to the database.

Palley and Simonoff (1987) and Adam and Wortmann (1989) identify several inference control mechanisms that can be used to control the disclosure of confidential numerical and categorical information in statistical databases. In most cases, these mechanisms are also applicable to controlling inferential security in general databases where access to individual values of confidential attributes is disallowed. These mechanisms can be broadly classified into two types, namely, Query Restriction Methods and Perturbation Methods. In this study, we evaluate those mechanisms that can be used in the context of a general database against disclosure to a snooper using linear models to estimate confidential numerical attributes using a
framework similar to that of Adam and Wortmann (1989).

5.1. Query-Restriction Methods
As the name implies, Query-Restriction (QR) methods attempt to increase security by restricting access to users. QR methods that can be used in the general database context to prevent disclosure of confidential information are discussed below. While there are other techniques such as Partitioning and Cell Suppression (Adam and Wortmann 1989), they cannot be directly applied to protect numerical attributes from a snooper using linear models.

5.1.1. Query Set-Size Control. One of the simplest methods that can be implemented to prevent disclosure of confidential information is to impose a restriction such that if the response to a query consists of less than $k$ records, then no response will be provided. An important benefit of CCA is the identification of that subset in the database, which results in the maximum information provided to sniffers. Therefore, any query whose record-selection criterion matches the values of the categorical attributes in this subset would result in the lowest security. If this subset contains $k$ records, then, to increase the level of security provided, it is necessary to have a query set size control of at least $k + 1$.

In the example database illustrated earlier, the minimum security occurred in the subset $(LOC = CHI, DEPT = PRO, GEN = M)$. The number of records in that subset is 441. If the value of $k$ was specified as a value larger than 441 (say 500), then this would ensure that no response is provided to queries smaller than 500. If $k$ is specified with any value less than 441, it would not prevent the snooper from gaining information on the subset with the highest canonical correlation, and there would be no improvement in security. When $k$ is specified as 500, the resulting level of security can be determined as $(1 - \lambda)$, where $\lambda$ is the largest eigenvalue for all subsets larger than 500. Using the prior computations, this value can be determined as 38% (subset defined by $LOC = NY$). Note that this information requires no additional computations.

It is also obvious that by specifying the minimum query set size as 500, the DBA restricts access to a large section of the database, and also limits flexibility in performing legitimate analysis. Hence, for the example database, this is a serious limitation of the query set-size approach. This approach could, however, be effective in situations where the low security occurs in a subset with a very small number of records. In such cases, query set-size control may provide an increase in security with only a limited restriction of access. Another problem with the query set-size restriction is that, if used as the only inference control mechanism, it may result in exact disclosure (Adam and Wortmann 1989).

5.1.2. Restricting the Number of Attributes Used in Queries. The ability of a snooper to use linear models arises from the fact that the snooper is able to reconstruct the structure of the database for subsets of data by using statistical queries. In the example database, there were three categorical attributes, and in order to reconstruct the structure of the database for every subset, the snooper must be able to use queries with five attributes such as:

\[
\text{CORRELATION BETWEEN (AGE AND BONUS)}
\text{WHERE (LOC = NY) AND (DEPT = MIS)}
\text{AND (GEN = M).}
\]

Suppose the DBA imposes the restriction that all queries involving confidential attributes will be limited to having four attributes in the query. In this case, the snooper can only reconstruct the structure of the database for subsets defined by two categorical attributes using queries such as:

\[
\text{CORRELATION BETWEEN (AGE AND BONUS)}
\text{WHERE (LOC = NY) AND (DEPT = MIS).}
\]

Such restrictions may increase the level of security provided. For the example database, imposing such a restriction increases security from 23 to 38% (for the subset $LOC = NY$). This method does restrict some access to data as well as reducing the flexibility that users have in performing data analysis. Overall, compared to other methods, this type of restriction may present an excellent compromise between access and security and can be easily implemented in practice.

5.1.3. Restricting Access to Attributes. The DBA may also attempt to increase security by limiting access
to one or more confidential attributes. The selection of specific attribute(s) to which access is denied can be based on CCA or the judgment of the DBA concerning the usefulness of the attribute(s). In the example database, the DBA could deny access to the STOCKS attribute (that represents the value of the stock options provided to an employee) on the basis that it is not "necessary" for analysis purposes. The snooper is now restricted to predicting SALARY or BONUS or linear combinations thereof. When CCA is performed under this condition, the resulting level of security is approximately 57%, a significant increase from the original level of 23%. The DBA could also evaluate the impact of removing access to other confidential attributes either in place of, or in addition to, the STOCKS attribute. The trade-off is that legitimate users have no access to such attributes, thereby eliminating their ability to perform analysis on this attribute. While this method is also easy to implement, denying access to attributes is very restrictive and should be considered only when security is deemed very low.

5.1.4. Restricting the Types of Queries. Another approach available to the DBA is to restrict the types of queries that involve confidential attributes to simple queries such as SUM, COUNT, and PERCENTILES. Responses to all other queries, specifically statistical queries such as VARIANCE and CORRELATION that enable the snooper to reconstruct the structure of the database, may be denied. This would prevent snoopers from using linear models (Palley and Simonoff 1987). Such restrictions can be easily implemented in practice. The best predictor of the individual confidential attributes under this condition would be the mean of individual attributes ($R^2 = 0$), resulting in the highest possible security of 100%. However, as long as any analysis on confidential attributes is allowed, this method is also susceptible to exact disclosure. While this method does not directly eliminate access to data, not allowing users to perform even simple statistical analyses is extremely restrictive, providing no flexibility in legitimate data analysis. Hence, this method should be considered only as a last resort.

5.1.5. Query Auditing. This approach involves keeping up-to-date logs of all queries made by each user and checking for possible compromise whenever a new query is issued (Chin and Ozsoyoglu 1982). Recently, Gopal et al. (1998) proposed a modification to the basic query-auditing procedure that allows the DBA to specify ranges for each value of the confidential attribute(s). As Palley and Simonoff (1987) have illustrated, query auditing has little or no impact on a snooper attempting to estimate confidential attributes by reconstructing the structure of a database using random queries. Hence, for the example database, query audit will not increase the security from the original level of 23%. Query audit may also potentially restrict access to data and may restrict the analysis flexibility by not responding to legitimate queries. In addition, query audit requires extensive computational effort and may not be practically feasible, except for very small databases (Adam and Wortmann 1989, Gopal et al. 1998). Collusion between different users increases the complexity of the problem.

5.2. Data Perturbation

An alternative approach that can be used in some databases is to use masking (see Adam and Wortmann 1989 for a comprehensive review of masking methods). Of the different methods of masking, (fixed-data) perturbation is best suited in the context of an organizational database (Muralidhar et al. 1995). When perturbation is applied to a database, the original values of the confidential attributes are replaced by a new set of "perturbed" values. Users are then provided with complete access to the perturbed values (and no access to the original values). The responses to all queries are based on the perturbed values as well. Since the individual values have been perturbed, there is no fear of exact disclosure. The advantage of the data-perturbation method is that users can be provided with unlimited access, even to individual (albeit perturbed) values of confidential attributes. This provides legitimate users with unlimited flexibility in performing analyses on the data. For example, consider the case where the user requests a "scatter plot" of two numerical attributes, at least one of which is a confidential attribute. If a query-restriction method were used, it would be necessary to deny this request because such a plot will result in exact disclosure. However, with perturbation, because perturbed values are used in place of the original values, no such restriction...
Table 2  Evaluation of Inference Control Mechanisms

<table>
<thead>
<tr>
<th>Method</th>
<th>Implementation Difficulty</th>
<th>Prevents Exact Disclosure?</th>
<th>Data Access Flexibility</th>
<th>Analysis Flexibility</th>
<th>Bias</th>
<th>Level of Security Provided for Example Database</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Perturbation</td>
<td>Moderate</td>
<td>Yes</td>
<td>Complete</td>
<td>Complete</td>
<td>Moderate to None</td>
<td>23%</td>
</tr>
<tr>
<td>Queries Must Consist of Four or Fewer Attributes</td>
<td>Easy</td>
<td>No</td>
<td>Moderate</td>
<td>Low</td>
<td>None</td>
<td>38%</td>
</tr>
<tr>
<td>Query Set Size = 500</td>
<td>Easy</td>
<td>No</td>
<td>Low</td>
<td>Low</td>
<td>None</td>
<td>38%</td>
</tr>
<tr>
<td>No Access to STOCKS</td>
<td>Easy</td>
<td>No</td>
<td>Low</td>
<td>Low</td>
<td>None</td>
<td>56%</td>
</tr>
<tr>
<td>No Statistical Queries Allowed on Confidential Attributes</td>
<td>Easy</td>
<td>Yes</td>
<td>Low</td>
<td>None</td>
<td>None</td>
<td>100%</td>
</tr>
</tbody>
</table>

is necessary. When perturbation methods are used, no other security mechanism is required to prevent exact disclosure (Duncan and Mukherjee 2000). In terms of implementation, perturbation requires more effort compared to other QR methods, except query audit. Compared to query audit, perturbation methods are much easier to implement.

Muralidhar et al. (1999) recently proposed a perturbation method that was shown to be more effective than other methods and will be considered in this study. In terms of security, when the perturbation method suggested by Muralidhar et al. (1999) is used; the proportion of variability explained remains $\lambda^*$. Hence, for the example database, the level of security remains 23%. It is important to note that perturbation cannot be used to increase the level of security provided.

A disadvantage of data perturbation is that it may modify the statistical characteristics of the database. If this occurs, then responses to certain queries using the perturbed values in the database may be different from responses to the same queries using the original database. This is often referred to as the “bias” resulting from perturbation. Such bias may affect even such simple statistical properties as PERCENTILLES, conditional SUMS, COUNTS, etc. In early perturbation methods, the “bias” problem was quite severe, rendering the database practically useless for statistical analyses. However, for databases described by a multivariate normal distribution, the method proposed by Muralidhar et al. (1999) completely eliminates all bias resulting from perturbation. Even when the database is not described by a multivariate normal distribution, the method proposed by Muralidhar et al. (1999) has the ability to maintain linear relationships between all attributes.

However, for such databases, responses to queries such as PERCENTILLES would result in bias.

5.3. Comparison of Inference Control Mechanisms

Table 2 provides a summary of the effectiveness of the different inference control mechanisms and is based on the general criteria used by Adam and Wortmann (1989). Note that since the structure of a given database will affect the level of security provided by an inference control mechanism, it cannot be determined for the general case and must be determined on an individual basis. Table 2 also helps to illustrate the importance of the CCA methodology. Without CCA, the DBA will be required to choose an inference control mechanism without knowing the impact of the chosen mechanism on the most important aspect of the evaluation, namely, the level of security provided. CCA fills this void and provides the DBA with useful information in making this important choice.

When the characteristics of a specific database are available (as in the case of the example database), it is possible to derive complete results for evaluating inference control mechanisms using the CCA approach, as summarized in Table 2. The results provided in Table 2 allow the DBA to make an informed choice regarding the most appropriate inference control mechanism to implement in this specific situation. For instance,

(1) If the DBA is satisfied with the level of security provided (23%) but wishes to provide greater access to users, then it may be appropriate to consider the perturbation method suggested by Muralidhar et al. (1999).

(2) If the DBA wishes to increase security to (say)
SARATHY AND MURALIDHAR

Security of Confidential Numerical Data

while providing the highest possible level of access, then it will be necessary to restrict queries to those having not more than four attributes.

(3) If the DBA wishes to increase security to (say) 50%, then it would be necessary to deny all access to the STOCKS attributes.

The above results clearly illustrate the trade-off between security, access, accuracy, and flexibility for analysis. The CCA approach allows the DBA to quantify this trade-off and evaluate alternatives.

5.4. Methodology for Evaluating Security and Choosing Inference Control Mechanisms

Combining the discussion in §3 along with the discussion presented in this section provides the DBA with a complete methodology not only for evaluating security, but also for selecting the appropriate inference control mechanism that is appropriate for a given situation. Figure 1 summarizes the methodology that was developed in this study for evaluating the basic level of security and for evaluating the security provided by inference control mechanisms. The only issue that the DBA will have to evaluate is whether the level of security provided is adequate. It is difficult to provide a general cut-off value for security. However, the DBA can be guided both by the minimum level of security provided and the width of the confidence interval presented in Equation (13). The sequence of steps presented in Figure 1 reflects the fact that inference control mechanisms that are least restrictive in terms of data access should be considered prior to more restrictive mechanisms and that denying access to the confidential attributes must be considered only as a last step. Both Figure 1 and Table 2 show that by providing access to all data, data perturbation is least restrictive, while denying access to any statistical queries represents the most restrictive data access. Figure 1 illustrates that the methodology developed in this study can be used directly by the DBA for implementing security procedures to prevent disclosure of confidential, numerical information to snoopers.

Figure 1 also reflects the fact that in practice, regardless of the level of security, it would be necessary to implement some basic inference control mechanism(s) to prevent disclosure (top box in the second column). This is necessary not only to prevent exact disclosure (as assumed in §2), but also to prevent disclosure if a snooper selects arbitrary subsets based on numerical attributes (rather than subsets based on categorical attributes), using a query such as:

\[
\text{MEAN OF SALARY WHERE (AGE > 68) AND (LOC = NY) AND (DEPT = MIS).}
\]

The proportion of variability explained in this subset may be higher than that resulting from CCA. First, the size of the subset may be so small that the number of records in the subset is less than the number of attributes. As observed earlier, the resulting level of security for this subset is 0.0. Second, it is possible that the variance of the error term is smaller than that of other subsets, resulting in a higher proportion of variability explained (and therefore lower security). Third, there may exist nonlinear relationships between the variables allowing a snooper to also estimate a nonlinear prediction equation using higher-order moments to predict the confidential attributes, again resulting in lower security. Because CCA is based on linear models, the security level of \((1 - \lambda^*)\) obtained through CCA is applicable under the assumptions of linear models. In practice, it is possible that these assumptions are not satisfied and the resulting level of security may be lower than \((1 - \lambda^*)\) in some subsets.

A simple security mechanism such as query set-size restriction can be used to eliminate the situation where the resulting query set size is smaller than the number of attributes. As shown in Figure 1, it is generally a good idea to institute basic security control mechanisms to prevent disclosure arising from small subsets. Selecting a minimum query set size judiciously would also alleviate the problem associated with a snooper selecting a very small arbitrary subset resulting in lower security.

A thorough preanalysis by the DBA is also an important step towards preserving security. The DBA may employ such tools as scatter plots, piecewise linear regression, and converting some or all numerical nonconfidential attributes to categorical attributes to identify those subsets that do not satisfy the assumptions of linear models (Neter et al. 1996), and to institute security mechanisms to prevent the snooper from gaining additional information. However, even with
Number of categorical attributes = \( L \).

For \( l = 0 \) to \( L \):

Identify all subsets defined by \( l \) categorical attributes

Let total number of subsets = \( K \)

For \( j = 1 \) to \( K \):

Convert \((L - l)\) categorical attributes not used to create subsets into binary variables (Sec 3.2)

Compute and store canonical correlation using all numerical and binary variables

Security Adequate?

No

Restrict query set size (Sec 5.1.1) or the number of attributes used in a query (Sec 5.1.2) and re-evaluate security using CCA

Security Adequate?

No

Restrict access to one or more confidential attributes (Sec 5.1.3) and re-evaluate security using CCA

Security Adequate?

No

Deny all statistical queries relating to confidential attributes (Sec 5.1.4)

Yes

Consider perturbation to allow access to individual (perturbed) values of confidential attributes by evaluating the trade-off between access and accuracy (Sec 5.2)

Yes

STOP
preanalysis, it may not be possible for the DBA to identify every arbitrary subset that results in lower security. In conclusion, the development of a general measure of security evaluation for both nonlinear and linear relationships and arbitrary subsets remains a difficult yet important goal. The CCA approach described in this study, by providing the DBA with a general measure to evaluate security provided for linear combinations, represents an important step towards the achievement of this goal.

6. Computational Issues Relating to the CCA Approach

In this section we discuss certain important issues relating to the successful use of the CCA methodology in practice. A key implementation issue is the level of computational complexity associated with the CCA approach. For the example database, the computation time required for the entire security analysis was just over 10 seconds using SAS® System for Windows (Version Eight). In the example database, it was necessary to perform computations for 60 subsets. Even for a very large number of subsets, the computation time is reasonable.

Extensive experiments were performed to estimate the computation time required for implementing the CCA approach using a database with 250,000 records. A select set of results is provided in Table 3, which shows the categorical attributes, the number of possible values for each categorical attribute, the number of possible subsets, and the computation time. In Experiment 1, CCA was used to evaluate the security provided with only categorical attribute A (with 15 possible values). The experiments were repeated by incrementally including the other categorical attributes. Even in the worst case, when the number of possible subsets is very large (close to 35,000), the computation time required is less than 800 seconds. Faster computers and/or specialized software will result in a significant reduction in the computation time. In summary, we believe that the benefit derived from the CCA approach (preventing disclosure of confidential information) far outweighs the required computational effort.

Finally, it is appropriate for the database administrator to perform CCA at reasonable intervals of time, to account for changes (if any) to the database structure, and to ensure that the security mechanisms that are in place are adequate. Note that it is not necessary to perform CCA every time an update takes place, since it is unlikely (though not impossible) for individual updates to change the inherent structure of the database, especially for large databases. It is necessary to perform CCA only when the changes to the database result in changing the structure of the database. Where such changes are frequent, the need to perform security evaluation often may present a problem.

7. Conclusions

In recent years there has been an increasing demand for protecting the privacy of confidential data residing in organizational databases. At the same time, the need for legitimate user access to data is also increasing. Duncan and Pearson (1991) term this as a "growing tension between confidentiality and data access." A key requirement for alleviating this tension is to determine the extent to which confidential attributes may be vulnerable to attacks from snoopers. Dutta
Chowdhury et al. (1999) recently provided a method for evaluating the potential for disclosure of categorical confidential attributes. An adequate and appropriate method for evaluating security of numerical confidential attributes in an organizational context is presently not available.

In this study, we have shown that canonical correlation analysis (CCA) provides a means to measure security against snoopers, using linear models that can be used in the context of an organizational database containing both numerical and categorical attributes, and is also a general form of prior security measures. The methodology described in this study can also be used to evaluate the impact of different inference control mechanisms on security, and to select the most appropriate mechanism. Our methodology for numerical confidential attributes, along with Chowdhury et al.'s (1999) approach for categorical confidential attributes, provides a practical framework for the DBA to evaluate the potential for disclosure of all confidential attributes.

References


