Association for Information Systems
AIS Electronic Library (AISeL)

AMCIS 1996 Proceedings

Americas Conference on Information Systems (AMCIS)

8-16-1996

Bivariate Perturbations of Fixed Data Sets

Peeter J. Kirs
Information & Decision Sciences, The University of Texas at El Paso

Krishnamurthy Muralidhar
Flordia International University

Kurt Pflughoeft
Information & Decision Sciences, The University of Texas at El Paso

Bivariate Perturbations of Fixed Data Sets

Peeter J. Kirs, Information & Decision Sciences, The University of Texas at El Paso, El Paso, TX
Krishnamurthy Muralidhar, Florida International University, Miami, FL
Kurt Pflughoeft, Information & Decision Sciences, The University of Texas at El Paso, El Paso, TX

Introduction

In recent years, there has been an increase in the number of laws requiring organizations to make information available *and* simultaneously maintain confidentiality if the release of such information might violate individual or group rights to privacy. This is compounded by the issue of ethics; while the release of some potentially sensitive data may not be strictly illegal, it may not be behaviorally justifiable. For example, while it is not illegal to view employee electronic messages produced and stored on company-owned systems, invasion of privacy questions arise (Schultheis and Sumner, 1995).

Recently, in an effort to help clarify guidelines for the use of private information and to develop a National Information Infrastructure (NII), President Clinton established The Privacy Work Group of the Information Infrastructure (IITF) Task Force (Computer Privacy Digest, 1995). The IITF has focused its attention on three general sets of principles: General principles for all NII participants, Principles for users of Personal Information, and Principles of individuals who provide personal information. The focus of this paper is to further examine techniques which can be applied by *providers* of personal information. The IITF guidelines for this group of NII participants is directed toward (1) why the information is being collected, (2) what the information is expected to be used for, (3) what steps will be taken to protect its confidentiality, integrity, and quality, (4) the consequences of providing or withholding information, and (5) the rights of redress. (Computer Privacy Digest, 1995).

Our concern is with the third issue, *what steps will be taken to protect its confidentiality, integrity, and quality*. We focus our attention on a specific statistical database system approach, Fixed Data Perturbation. While this approach has been examined previously (Beck, 1980; Denning, 1980; Liew et al., 1985; Adam and Wortman, 1989; Adams and Jones, 1989; Muralidhar and Batra, 1995; Muralidhar et al., 1995), the present study provides an overview and extends previous research by including a bivariate analysis of perturbed data sets.

Background

In 1986, Mason noted that there are a number of precautions which must be taken in the building of intellectual capital. He specifically considered four ethical issues which producers of information must concern themselves with: (1) *Privacy*, or what information about one's self or one's associations must a person reveal to others, under what conditions and with what safeguards?, (2) *Accuracy*, or who is responsible for the authenticity, fidelity, and accuracy of information?, (3) *Property*, or who owns the information?, and (4) *Accessibility*, or what information does a person or organization have a right or a privilege to obtain, under what conditions, and with what safeguards? These issues, and the problems associated with them, remain of paramount importance to the Information Systems (IS) Manager, and their resolve has taken on an increased sense of urgency given the increased expansion, usage, and availability of corporate databases and the corresponding increases in database abuse (Wong, 1985; Straub, 1990; Muralidhar et al., 1995). The basic consensus is that an organization is charged with providing ***accurate*** data and information which is ***securely*** maintained but ***accessible*** to authorized users (Wysocki and Young, 1990; Laudon and Laudon, 1991; Schultheis and Sumner, 1995).

Attainment of this objective has proved difficult since it involves opposing goals. Information which is accessible in its original form is not secure; implementing security measures requires either limiting access or restricting data sets, thereby reducing the accuracy of the information generated from it, or both. As

Denning et al. (1979) have previously pointed out, "the requirement of complete secrecy of confidential information is not consistent with the requirement of producing exact statistical measures for arbitrary subsets of the population. At least one of these requirements must be relaxed" (pg. 92).

Fortunately, most queries made by users for decision making purposes generally do not involve specific observations (e.g., an individual's salary), but rather are directed toward the attainment of summary information (e.g., the average salary for a particular subset). In response a number of Statistical Database Systems (SDBS) have been suggested. These include *The Conceptual Approach* (Ozsoyoglu and Chin, 1982), in which the user is allowed access to only a subset of the database based upon predefined attributes, *Query Restriction* (Fellegi, 1972; Dobkin et al., 1979; Denning et al., 1979), whereby the user is allowed unrestricted access by programmatic controls are placed on data restriction which hinder compromise, *Output Perturbation* (Achugbue and Chin, 1979; Denning, 1980; Beck, 1980) which allows queries to made directly to the actual database, but programmatically perturbs all outcomes prior to disclosure, and *Data Perturbation* (Traub et al., 1984; Reiss, 1984) which presents the user with a dataset which has previously been perturbed in a manner so as to retain the same basic characteristics but make individual observations meaningless.

In a series of studies (Muralidhar et al. 1995; Muralidhar and Batra, 1995) contrasted the major SDBS approaches and concluded that the *Data Perturbation Approach*, particularly the *Fixed Data Perturbation* (FDP) approach, appears to offer some distinct advantages over the other methods. While it does have some disadvantages, most notably that it requires additional disk storage and that concurrency with the original dataset must be monitored and changed accordingly, it nonetheless (1) provides maximum disclosure against exact disclosure since individual observations have been altered, (2) allows maximum accessibility since users are given complete access to the full database, (3) does not require complex algorithms or programmatically elaborate instructions which must remain online to perturb the results, and (4) the actual data instance values cannot be derived through multiple queries.

Fixed Data Perturbation

By definition, perturbing a data observation means altering its value. Ideally, subjecting an existing data series to a perturbation approach should result in a disguised series which maintains the same characteristics of the original series, but which does not allow partial or exact disclosure of the original data occurrences. Statistical or summary queries made of the perturbed dataset should yield outcomes which, if not identical, should not be significantly different from those which would be obtained from the original dataset. Conversely, queries intended to reveal or implicate unique entity occurrences should prove fruitless. However, there is a distinct tradeoff between 'correctness' or accuracy of results obtained to statistical queries and the 'incompromisability' or security of the perturbed data series in safeguarding against queries which could divulge individual values.

There are a number of factors which affect the tradeoff between accuracy and security, and a number of measures which have been suggested for each. The accuracy of results obtained from queries on a perturbed data series are a function of how closely the perturbed dataseries maintains the characteristics of the original dataseries. However, since each value within a perturbed dataset has been altered, it is unreasonable to assume that there will not be some differences between the two series. At issue, then, is what attributes should be considered as representative, how perturbation affects them, and how they can be measured.

When provided perturbed data for statistical analysis, the primary concern is that the perturbed data series follow the same population distribution as the original series. However, the resultant form of the perturbed distribution, is a function of the original distribution form and the perturbation method applied. In some cases, the resultant distribution is known, can be derived, or can be shown empirically (see Muralidhar and Batra, 1995).

In the additive approach, $y_i = x_i + e_i$ where y_i is the ith observation of the perturbed series, x_i is the ith observation of the original series, and e_i is a random variable with =0 and prespecified variance (^2_e). Since E(e)=0, the expected mean of the original and the perturbed series are identical. For the multiplicative case, $y_i = x_i * e_i$ where e_i is a random variable with mean of one (1) and a prespecified variance. Again, since E(e)=1, the mean of the original series and the perturbed series are identical. In both approaches, since $^2_e >$ 0, the variance of the perturbed dataseries will be greater than that of the original dataseries.

If the original distribution can be maintained, regardless of method applied, the primary consideration remains how accurate the outcomes of the queries made upon the database are, from a reporting perspective, and how difficult it is to uncover the 'true' value of individual data observations, from a security perspective.

In practice, the proximity of outcomes from a perturbed dataset to those from the original series is dependent upon the analysis procedure applied. A number of measures have typically been used to this end, including the differences between the original series and the perturbed series in terms of the mean, variance, percentile values, and product moment and rank order correlations. Muralidhar and Batra (1995) also suggest using the Mean Absolute Perturbation (MAP), noting that it provides a direct measure of difference between the perturbed and original series whereas the standard deviation does not allow comparison across distributions.

Methodology

Previous research focused on univariate perturbation of a dataseries, and subsequently examined the security and accuracy provided by the perturbed series. While univariate statistical analysis does take place, it is more likely that such analyses are bivariate or multivariate in nature. It is our assumption that most queries made of databases consider the relationship between two or more variables than merely obtaining descriptive statistics about single series. Take for example an analysis of salaries. While the researcher would undoubtedly be concerned with overall mean salaries and variance, it seems likely that most of queries made would be about the relationship between salary, years experience, age, level of education, and other causal or mediating factors.

For the sake of illustration, assume that there are two dataseries, x and y, which are normally distributed. The Covariance matrix and correlation for the original series would be:

x^2 Cov(x,y)

Cov(x,y) y^2

Cov(x,y)

and r =

x^2 y^2

If the two series are additively, and independently, perturbed by random variables $e_x \sim N(0, _{ex})$ and $e_y \sim N(0, _{ey})$, the correlation matrix and correlation between the resultant series, x_1 and y_1, would be:

$x^2 + _ex^2$ Cov(x,y)

Cov(x,y) $y^2 + _ey^2$

Cov(x,y)

and r =

$(x^2 + {}_e x^2)(y^2 + {}_e y^2)$

Since $E(e_x)=E(e_y)=0$, and x and y are independent of e_x and e_y, the covariance will remain the same (see Neter, Wasserman and Kutner, 1983). However, since the variances of the perturbed series are increased by ${}_{ex}{}^2$ and ${}_{ey}{}^2$, respectively, the correlation will be decreased accordingly. For example, assume that two dataseries are both normally distributed with =0 and ^{2}=1, and the covariance and correlation between the original series is 1.0 and 0.8, respectively. If each series is perturbed by e˜ N(0,1), the covariance between the perturbed series would remain at 1.0, but the correlation would be 1/2 that for the original series (r=0.4), since the variance is twice that of the original series.

It is possible to generate random variables from a bivariate normal distribution (see Scheuer and Stoller, 1962; Lehman, 1977), maintaining the correlation (r) between the two series. In the simplest case, if ${}_x$=${}_y$=0.0 and ${}^2{}_x$=${}^2{}_y$=1, then the y value corresponding to x is obtained as a function of x, a second standard normal value (e˜N(0, ${}_e{}^2$)), and the desired r between them $y = rx + e(1-r^2)$. In the more general case, if the means and standard deviations are ${}_x$,${}_y$, ${}_x$, ${}_y$ and the correlation is r, then the relationships are $x = x_x + {}_x$ and $y = {}_y(rx + e(1 - r2)) + {}_y$

Some of the impacts of this technique are known, or can be derived. Tendick and Matloff (1994) have recently shown how a random number perturbation method which partially solves the bias problem for the multivariate normal case. They have further investigated nonparametric approaches to the problem (Tendick, 1988). However, as they note, given the number of distributions available (e.g., normal, log-normal, uniform, gamma) and the number of measures of accuracy and bias (e.g., , ¸ MAP, percentile estimates), many of the outcomes must be examined empirically.

We have conducted a number of Monte-Carlo simulations to analyze and illustrate the use of bivariate perturbation for statistical database explication. Given the page limitation for this proceedings, we cannot fully present our initial findings at this time

References available upon request