Joint UNECE/Eurostat work session on statistical data confidentiality
(Bilbao, Spain, 2-4 December 2009)

Topic (vii): Risk/benefit analysis and new directions for statistical disclosure limitation

AN ENHANCED FRAMEWORK AND DECISION SYSTEM FOR PROTECTING THE CONFIDENTIALITY OF TABULAR DATA

Invited Paper

Prepared by Marco Better and James P. Kelly, OptTek Systems
An enhanced framework and decision system for protecting the confidentiality of tabular data

Marco Better and James P. Kelly
OptTek Systems, Inc., 1919 Seventh Street, Boulder, CO, USA, better@opttek.com, kelly@opttek.com

Abstract: Statistical databases for public use pose a critical problem: how to make tabular data available for analysis without disclosing information that would infringe on privacy, violate confidentiality, or endanger national security. The challenge is to represent data in a form that permits accurate analysis for supporting research, decision-making, and policy initiatives, while preventing an unscrupulous or ill-intentioned party from exploiting the data for harmful consequences.

The objective of this research is to develop a practical framework for assessing, measuring, and mitigating disclosure risk in tabular public-use data. Our proposed framework, called OptShield, overcomes the disadvantages found in currently deployed disclosure limitation methods. We achieve this by combining perturbation and suppression methods with optimal switching of sensitive records at the micro-data level, to produce a method that protects confidentiality while preserving data integrity. Our system consists of a three-phased approach to the disclosure limitation problem: (1) Assess a user’s qualitative and quantitative disclosure risks inherent in the organization’s data publishing and sharing plans; (2) Measure disclosure risks in a user’s data products; and (3) Protect the user’s data by applying appropriate disclosure limitation techniques. This work is being supported by a grant from the U.S. National Institute of Mental Health for application within the area of HIV treatment and prevention.

1 Introduction

Preserving privacy in statistical databases has captured the interest of scientific and practitioner communities for many years. In an internal notice, the U.S. Census Bureau (2003) states that “criteria for avoiding disclosure in data … for public use have been the subject of internal research, consultations with stakeholders, and input from experts in disclosure limitation techniques over the past few decades.”

The tension between the desires of disclosing party and analyst is conspicuously growing. On one hand, the publishing party is bound by ethical and legal constraints to preserve the confidentiality of the data; on the other hand, increasing computational power motivates the analyst to obtain the data in its purest form. A common compromise is to publish data in tabular form where entries of a table represent sums or counts of individual records with specific characteristics (a database query). The popularity of this solution is twofold: first, it provides data in a compact, summarized format that lends itself to useful analysis; second, if done correctly, it protects the privacy of the individual records.

Over the years, many disclosure limitation techniques have been developed, each with inherent advantages and disadvantages. Some techniques, like data swapping (see Moore (1996) for a discussion), focus on protecting the micro-data directly, but can result in large distortions at both the micro- and aggregate-data levels. Other
techniques such as controlled rounding (Cox and Ernst, 1982), cell suppression (Cox, 1980), and controlled tabular adjustment (CTA) (Cox et al, 2004) focus on protecting data appearing in tabular form. These methods suffer the disadvantage that they can induce severe inconsistencies among various cuts of the data, reducing their analytical usefulness and, sometimes, making it possible to infer confidential information from these inconsistencies. A third type of methodology that is based on modifying the micro-data in order to generate data tables that minimize disclosure. Traditional instances employ data perturbation methods that add noise to the micro-data, or employ locally greedy methods for record switching. While these methods prevent inconsistencies among data tables, they have been known to allow partial disclosure of the data (Muralidhar and Sarathy, 1999).

Under the auspices of the National Institute of Mental Health (NIMH) SBIR program, OptTek Systems, Inc. (OptTek) developed an innovative methodology called Optimal Switching (OS) that overcomes the disadvantages detailed above. OS differs from existing weighting and imputation techniques by embodying optimization algorithms that facilitate the selection of globally optimal weights to protect data while minimizing data distortion. We have found that OS affords a significant improvement over other approaches. Nevertheless, special situations can arise where OS will not be able to find a solution that ensures 100% confidentiality protection. In these cases, we believe that OS would work well as part of a phased approach, where it is followed by other disclosure limitation methods.

Our prior experience in the area of confidentiality protection and disclosure limitation, in addition to insights gained from conversations with potential users and partners for the commercialization of this technology, lead us to believe that OS would be most effective as part of an overall framework for the assessment, measurement and mitigation of disclosure risk.

This paper discusses our proposed framework. Section 2 provides a high-level view of the framework, and a detailed description of each of the framework’s components. Section 3 provides results of a computational study performed with our proof-of-concept software developed in Phase I of the SBIR research program. Finally, Section 4 contains our conclusions and a discussion of the issues related to our future research and development phase of the project.

2 OptShield™: A Disclosure Risk Management Framework

Our proposed disclosure risk management framework is graphically depicted in Figure 1. The proposed framework consists of three components: Assess, Measure, and Protect, which form a disclosure risk management cycle. The framework functions as a combination of disclosure risk analysis software and services which, as a whole, we call OptShield.
The first component, **Assess**, involves the deployment of services in order to identify potential sources of disclosure risk in a client’s current data management practices.

The second component, **Measure**, involves the implementation of disclosure risk measures, based on existing sensitivity rules, and the development of new rules, imputation techniques, and bounding algorithms.

The third component, **Protect**, will provide users with a suite of disclosure limitation methods that will work in concert to help mitigate the disclosure risks identified during the assessment and measurement phases.

The successful development of such a framework has generated interest from many types of organizations. In particular, within the Colorado Department of Public Health and the Environment (CDPHE), the staff of its Disease Control and Environmental Epidemiology Division has agreed to collaborate with us in creating a version of the software and diagnostic service to meet the division’s requirements. Other potential users have expressed their interest in the framework, and have offered to continue to share their data with us for the purpose of testing our framework.

![Figure 1: Proposed framework](image)

### 2.1 Assessing Disclosure Risks

The task of assessing disclosure risks in sensitive data can be done both qualitatively and quantitatively. In our framework, we propose to engage experts in the field of statistical disclosure to assess a client organization’s risks from a qualitative perspective, by conducting an audit of their internal data collection, management, and data-sharing policies and practices. This will be offered as a service during an initial phase of the implementation of our framework. This activity will not only help the organization to update its data sharing policies and procedures, but it will aid
in the selection of the disclosure limitation methods within our framework that are best suited for the situation.

### 2.2 Measuring Disclosure Risks

Measuring disclosure risk in data consists of two steps: (1) Identifying sensitive records in data products so that they can be effectively protected; and (2) Measuring how accurately one can estimate the true value of a record in the data product after it has been altered by a disclosure limitation method. The first can be achieved by a set of “sensitivity rules” that define what a sensitive record is. The second can be achieved by imputation techniques, or solving maximization and minimization problems to obtain upper and lower bounds of the possible values. (Cox and Better, 2009)

#### 2.2.1 Identifying sensitive records

A considerable amount of research has been devoted to identifying sensitive cells in tabular data. Once a sensitive cell is identified, all records contributing to the sensitivity of that cell must be considered sensitive as well. As a result, a number of sensitivity rules have been developed to identify sensitive records in tables, for both count data and magnitude data. In general, we can define disclosure risk as the risk that a user can estimate the reported value of an individual respondent’s confidential information too precisely. Sensitivity rules are therefore designed to identify cells in which this risk is present. The specific rules we will use to identify sensitive cells include, but are not limited to, the following:

#### 2.2.1.1 Rules for magnitude data

- **The Number Rule:** If the number of respondents contributing to a cell or sum is less than or equal to some threshold $N$, then the cell is sensitive.
- **The Percent Rule:** Let $LC$ represent the largest contributor to a cell $C$. Let $P$ define a parameter that represents a percentage. $E = \{(P/100 \times LC) - \text{sum of all other cells}\}$. If $E > 0$, then the cell is sensitive.
- **The p-percent Rule:** If upper and lower estimates for a respondent’s value are closer to the reported value of the cell than a pre-specified percentage, $p$, then the cell is sensitive.
- **The pq Rule:** In the $p$-Percent rule, it is assumed that there is limited prior knowledge about respondents’ values. Agencies should not make this assumption. According to the $pq$ rule, a value $q$ represents how accurately respondents can estimate another respondent's value before any data are published. Let $LC1$ and $LC2$ represent the two highest contributors to a cell where $LC1$ is the largest contributor. Let $SUM = (\text{Sum of all contributors}) – LC1 – LC2$. Let $E = (p/100 \times LC1) – (q/100 \times SUM)$. If $E > 0$, then the cell is sensitive.
• **The (n,k) Rule:** If a small number (n or fewer) of respondents contribute a large percentage (k or more) of the total cell value, then the so-called n respondent, k percent rule of cell dominance defines this cell as sensitive.

### 2.2.1.2 Rules for frequency (count) data

• **The Number Rule:** If the number of respondents contributing to a cell is less than or equal to N, then the cell is sensitive.

Our framework includes experimental testing of different parameter values for p, q, n, and k as factors that can have a “high,” “medium” or “low” value, to assess the impact of the different methods on disclosure risk and data quality. These values are also influenced by an organization’s stance on the relative importance of data quality vis-à-vis data protection.

While all of the above sensitivity rules apply to cells in tables, very little is known about identifying potentially sensitive records directly in the microdata. First, it may be possible to generate all “meaningful” tables of up to \( n \) dimensions from a data set. This will permit us to apply the sensitivity rules described above to identify all sensitive records. However, if \( n \) is large, then this approach may be computationally too intensive to be practical. A second approach is to use data mining techniques to identify outliers in the microdata. An outlying record will be more likely to be isolated in a table cell, either by itself or with a very small complement in its row or column, making it likely to represent a disclosure risk.

### 2.2.2 Estimating the true value of a protected record

Our recent work with the National Center for Health Statistics (NCHS) has involved the study of efficient methods to estimate bounds on suppressed or perturbed entries in data tables and the development of a specialized sampling procedure in tables with missing data (Cox, 2007), in an effort to develop effective imputation methods.

### 2.3 Protecting Sensitive Data

Once a sensitive record has been identified, the question becomes how to best protect its confidential information without damaging the quality of the data. Most disclosure control and disclosure limitation methods in use today focus on modifying the values of cells in tables, while some focus on modifying the microdata so that the resulting tables are protected. Of the former kind, complementary cell suppression and controlled rounding are notably the most popular, and controlled tabular adjustment is considered a relatively new arrival in the field with excellent prospects. Of the latter kind, microdata perturbation methods and microdata switching are two front-running methods. Our Optimal Switching (OS) approach is the best-in-class of switching methods, as is shown in Section 3 of this paper.
2.3.1 Complementary Cell Suppression

Although widely popular, cell suppression methods do exhibit considerable drawbacks. First, finding the optimal solution to the complementary cell suppression problem requires solving a mixed-integer problem that is NP-hard. Thus, only the smallest instances (generally, small 2-dimensional tables) can be efficiently solved. Furthermore, the linear programming relaxation of the integer problem frequently produces fractional values for table entries. Second, cell suppressions have a “cascading” effect, resulting in rapid degradation of the quality of the data.

2.3.2 Controlled Rounding

Rounding in tabular data involves presenting data where the actual numerical values in the cells are rounded to the nearest integer value, or the nearest multiple of some pre-specified integer base. In controlled rounding, changes to the values of the sums associated with the columns and the rows of a table are minimized. Like the suppression problem, controlled rounding can also be stated as an optimization problem. A notable drawback of the method is that there is no guarantee that a feasible solution exists beyond tables of two dimensions. In addition, in the case of magnitude data, controlled rounding does not guarantee protection. Furthermore, the problem is also NP-complete, so that instances of moderate size beyond three dimensions cannot be efficiently solved.

2.3.3 Controlled Tabular Adjustment

The objective of controlled tabular adjustment (CTA) “is to closely mimic the original data, subject to obscuring sensitive cell values to a sufficient extent.” (Glover et al, 2009) The value of each sensitive cell is replaced by an adjusted value, either up or down. These adjusted values are selected to be at a safe distance from the original value and are known as a cell’s lower and upper protection limits, or minimum perturbations required to add uncertainty in the data. Some or all non-sensitive cells are then adjusted by small amounts to restore additivity to column and row sums. Consider the following example of a 2-way table as shown in Table 1. Cells (3,1), (1,2), and (3,2) shown in bold have been identified as sensitive cells, and the associated protection limits are shown in the brackets. The upper and lower bounds for the non-sensitive cells are set at 20% of the original cell value. Table 2 shows the tabular data after solving the mathematical program. Cells with * indicate that they have been adjusted.

<table>
<thead>
<tr>
<th>74</th>
<th>17[0,37]</th>
<th>85</th>
<th>176</th>
</tr>
</thead>
<tbody>
<tr>
<td>71</td>
<td>51</td>
<td>30</td>
<td>152</td>
</tr>
<tr>
<td>1[0,21]</td>
<td>9[0,29]</td>
<td>36</td>
<td>46</td>
</tr>
<tr>
<td>146</td>
<td>77</td>
<td>151</td>
<td>374</td>
</tr>
</tbody>
</table>

Table 1: data before CTA

<table>
<thead>
<tr>
<th>75*</th>
<th>0*</th>
<th>85</th>
<th>160*</th>
</tr>
</thead>
<tbody>
<tr>
<td>71</td>
<td>51</td>
<td>30</td>
<td>152</td>
</tr>
<tr>
<td>0*</td>
<td>29*</td>
<td>36</td>
<td>65*</td>
</tr>
<tr>
<td>146</td>
<td>80*</td>
<td>151</td>
<td>377*</td>
</tr>
</tbody>
</table>

Table 2: data after CTA
Finding an optimal solution for CTA requires solving a mixed-integer program with binary variables, as is the case with controlled rounding, although efficient heuristic methods have been developed. (Glover et al, 2009)

In general, tabular adjustment methods have various drawbacks in common. First, when requests involve more than one table (i.e., different “cuts” of the data), these methods can create inconsistencies from one table to another. For example, in controlled rounding, a certain cell in one table could be rounded up, while in another table the same cell is rounded down, not only degrading the quality of the data, but also providing a potential intruder with an indication that the cell in question is sensitive, and its value falls between two consecutive multiples of the base.

OptTek has implemented efficient algorithms for complementary cell suppression, controlled rounding, and controlled tabular adjustment. We propose to enhance these algorithms so they can be seamlessly integrated into the overall OptShield disclosure risk management framework. Regardless of their limitations, techniques like the ones described above are widely used in practice. We recognize that many of the potential users of our framework will want to have such techniques available, and we believe that many of their limitations can be avoided by working in combination with our optimal switching method, which we now describe in detail.

3 Optimal Switching

Optimal Switching (OS) is a method that provides an optimal mechanism for switching records at the microdata level in order to achieve disclosure protection while minimizing data distortion in tabular data. The OS procedure is as follows:

Step 1: Construct the desired data tables to be published or provided for a particular data request.
Step 2: Identify any sensitive cells in the tables and mark the records that contribute to those cells as sensitive.
Step 3: Perform OS to find a globally optimal solution that eliminates all disclosure risks and maximizes data quality, by iterating over the following substeps:
   3.1 Use a metaheuristic optimizer to select a set of weights, W, such that an affinity function value, $W \Delta$, can be computed between two records by multiplying the selected vector of weights by the vector of the differences in the values of the records’ attributes.
   3.2 Compute the affinity function value for each sensitive record and all candidate switching records.
   3.3 Solve an assignment problem that switches each sensitive record with its optimal candidate switching record.
Step 4: Reconstruct the desired data tables after switching the records.
Step 5: Compare the original tables with the switched tables in terms of data quality.
The critical step in the process involves the optimization problems solved in Steps 3.1 and 3.3. In Step 3.1 we make use of a global optimization engine designed to conduct an intelligent search for rapidly improving solutions. OS makes use of the advanced search algorithms to find an optimal weighting $W^*$ of the set of attributes $A$ of two records in the data set, in order to find a switching of microdata records that achieves protection $p$. At each iteration the optimizer suggests a new set of weights. This is used in Step 3.2 to compute the affinity function values, $W_{ij}$, between each sensitive record $i$ and all of its candidate switching records. Once these have been calculated for all possible switching pairs, in Step 3.3 we solve an assignment problem that optimally switches each sensitive record with its best switching candidate. The result of each solution is a data quality measure $Q$ that is maximized.

In order to develop and test our software, we generated various sets of simulated data. These data sets were designed to be diverse in terms of their size, data types and descriptive statistics, thus providing a basis to gauge the robustness of our method. The testing procedure for each data set is as follows: First, construct a set of data tables. Next, identify sensitive cells in the tables according to specific sensitivity rules. Once sensitive cells are identified, tag the corresponding microdata records as sensitive, and apply two data switching methods to the underlying microdata: (1) proximity-ranked data switching, currently in use by a number of Federal agencies; and (2) the optimal data switching method. Finally, reconstruct the data tables and compare the performance of the two methods according to disclosure risk and data quality criteria.

### 3.1 Testing on simulated data

For our testing of count data, we created three different data tables, corresponding to Test 1, Test 2, and Test 3 respectively. Tables were created from microdata generated on the basis of a HIV behavioural survey conducted in 2002 by the Centers for Disease Control and Prevention. The data consists of subjects in six ethnicity categories, three types of health insurance, two HIV status categories, and four education levels. Table 3 shows results for Tests 1, 2, and 3.

<table>
<thead>
<tr>
<th>Method</th>
<th>TEST 1 (HIV by Ethnicity and Insurance Type)</th>
<th>TEST 2 (HIV by Education and Ethnicity)</th>
<th>TEST 3 (Sex ID by Education and Ethnicity)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$S$</td>
<td>$R^2$</td>
<td>$\chi^2$</td>
</tr>
<tr>
<td>Original table</td>
<td>98.22</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Optimal Switching</td>
<td>100</td>
<td>0.999</td>
<td>0.987</td>
</tr>
<tr>
<td>Ranked Proximity (P = 5%)</td>
<td>98.22</td>
<td>0.998</td>
<td>0.931</td>
</tr>
<tr>
<td>Ranked Proximity (P = 10%)</td>
<td>96.44</td>
<td>0.990</td>
<td>0.091</td>
</tr>
</tbody>
</table>

Table 3: Results for Tests 1, 2 and 3 of Simulated HIV data set
The first column of each Test shows the protection score $S$, while the next two columns show measures of data quality, where $R^2$ is the correlation coefficient between the modified cells in the protected table and the corresponding cells in the original table, and $\chi^2$ $p$-value is the probability that the modified cells in the protected table belong to the same distribution as the corresponding cells in the original table.

For magnitude data, we used simulated data provided by the Bureau of Transportation Statistics. This data set consists of 667 records, corresponding to fictitious shipping companies. The data contains four identifying variables and one sensitive variable related to companies’ revenue. The sensitivity rule used for these tests was the $p$-percent rule with $P = 80\%$. Table 4 shows the results of this testing.

![Table 4: Tests 1 and 2 of Simulated BTS data set](image)

The results show that OS is able to find a solution with 100% protection in all cases, whereas the ranked-proximity method fails to do so in every instance. The results also show that as proximity parameter $P$ increases, the data quality measures in ranked-proximity degrade very quickly. OS is a more robust method, producing results with high quality of data across all instances.

Although OS is well-suited for protecting data regardless of type (count, magnitude, continuous, discrete, etc.), there are special cases where it does not find a solution that protects 100% of the data. In these cases, OS can be used as a “first pass” during protection, followed by a tabular method, such as CTA, cell suppression, or controlled rounding. Using OS in combination with a tabular method is preferable to applying the tabular method by itself. First, by initially applying OS, we would avoid, to a great extent, the inconsistencies that tabular methods can introduce to data products that involve more than one table. Second, applying OS in a first pass will guarantee that the amount of distortion caused by the tabular method is minimized. Third, because of the nature of optimal switching in cases involving count data, the magnitude of the modifications or suppressions will become small, thus minimizing the distortion to the table cells as well as the column and row sums.
4 Conclusions and Implementation Issues

Our potential customers, potential partners, and subject matter experts all concur that there is no comprehensive framework today that manages each aspect of the disclosure risk problem like our framework will. The strength of our approach lies in offering the data provider the flexibility to choose a method or combination of methods that best suits their data, their risk-acceptance profile, and their customers’ requirements. Another advantage of our approach is that we combine qualitative assessments by experts with quantitative techniques to measure overall disclosure risk and risk-aversion attitudes prior to selecting a disclosure limitation method that best addresses these measurements.

References

"Revised Confidentiality Criteria for Bureau of the Census Public Use Data Products" (2003), General Release, Bureau of the Census, Department of Commerce, Wash., DC.


