| From: | Ramesh Dandekar [ramesh.dandekar@gmail.com](mailto:ramesh.dandekar@gmail.com) |
| :--- | :--- |
| Sent: | Thursday, June 28, 20076:27 PM |
| To: | boaz@cs.princeton.edu; kamalika@cs.berkeley.edu; dwork@microsoft.com; |
|  | satyen@cs.princeton.edu; mcsherry@microsoft.com; kunal@microsoft.com |
| Cc: | Dandekar, Ramesh |
| Subject: | About Your Research Paper on Holistic Solution to Contingency Table Release .. |
| Attachments: | BarakEtPODS07Conting.pdf |

## Hi Friends

I am with U. S. Department of Energy and have been actively doing research on Synthetic Tabular Data. My Research papers are on

## http://mysite.verizon.net/vze7w8vk/

I do have a proprietary procedure to process a high dimensional tabular data (count and magnitude) using LP setup. I think, we need to join our forces to test your new procedure against what I have proposed using a real life CPS counts data
"Maximum Utility-Minimum Information Loss Table Server Design for Statistical Disclosure Control of Tabular Data", Ramesh A. Dandekar, June 9-11, 2004, Barcelona, Spain. Related LOG-LINEAR analysis and 3-D sectional data from Eight Dimensional CPS Counts Data.

Please let me know if there is any interest in collaborative work.

- sincerely

Ramesh A Dandekar

# Maximum Utility-Minimum Information Loss Table Server Design for Statistical Disclosure Control of Tabular Data 

Ramesh A. Dandekar<br>Energy Information Administration, U.S. Department of Energy<br>Washington DC 20585, USA<br>Ramesh.Dandekar@eia.doe.gov


#### Abstract

Statistical agencies typically serve a diverse group of end users with varying information needs. Accommodating the conflicting needs for information in combination with stringent rules for statistical disclosure limitation (SDL) of statistical information creates a special challenge. We provide a generic table server design for SDL of tabular data to meet this challenge. Our table server design works equally well with counts data and magnitude data, and is compatible with commonly used cell perturbation methods and cell suppression methods used for the statistical disclosure control of sensitive tabular data. We demonstrate the scope and the effectiveness of our table server design on counts and magnitude data by using a simplified controlled tabular adjustment procedure proposed by Dandekar (2003). In addition to ad hoc queries, the information compiled using our table server design could be used to capture multi-way interactions of counts data and magnitude data either in a static environment or in dynamic mode.


## 1 Introduction

Procedures to protect sensitive cells in tabular data have evolved over the last few decades. In recent years there has been an increased realization that "one table at a time" tabular data protection procedures, though easy to implement, could be error prone and pose serious disclosure risks. As a result, system designers are increasingly tasked to develop a unified strategy to simultaneously protect multiple different tables resulting from the same survey data collection form. The complexity of the data collection form and resulting tabular data structures in related publications varies from survey to survey. This makes it difficult to develop a unified design strategy that will accommodate multiple variations in table structures as well as inter-table relations.

In the first part of the paper, we propose a unified method to simultaneously capture and process multiple different tables containing complex hierarchies and linked structures resulting from the same data collection efforts. As an integral part of our proposed method, we introduce the concept of a "guidance matrix" to capture complex interaction effects present within and across table structures.

We first explain the role of the guidance matrix concept in capturing complexities of seven different test cases of magnitude data. These test cases are currently available in the public domain to all researchers of tabular data protection from the website: http://webpages.ull.es/users/casc/ficheros.csp.htm. We also demonstrate the effectiveness of the guidance matrix concept in capturing multi-way interaction terms by illustrating its use on real life eight-dimensional counts data.

In the second part of the paper, we demonstrate the unique characteristics associated with simplified Controlled Tabular Adjustment (CTA) procedure by Dandekar (2003). This technique when applied to any generic N dimensional counts data structure in its entirety, offers complete protection from statistical disclosure. We utilize the guidance matrix concept, developed in the first part of the paper, to demonstrate the statistical validity of the CTA-adjusted counts table structure by performing cross section by cross section analysis of the SDL protected eight-dimensional counts data. We conclude by providing basic details on how to design a generic web-based table server to function in a dynamic mode by using the simplified CTA procedure. Our proposed web-based table server design, after some fine tuning, is capable of serving diverse user groups with conflicting information needs.

## 2 Concept of Guidance Matrix

A typical survey data collection form consists of multiple fields containing various categorical variables and related quantitative information. Categorical variables are mainly used to define table structures for the publication of cumulative statistics related to the quantitative information. Often, various intervals of quantitative information fields are used to create additional categories in the table structure.

For our discussion, we assume that the N dimensional space defined by the N different categorical variables has already been identified a priori by the statistical office. In our proposed implementation method, we first construct an N columns by M rows boolean guidance matrix as follows:

- $\mathrm{J}=1$ to M rows of the guidance matrix define M different tables to be created by the publication system.
- Within the Jth row of the guidance matrix, the column position assumes the value of zero if the related categorical variable details are not included in the Jth table. The column value of one indicates that the related categorical variable details are available in the Jth table.

We use the contents of the guidance matrix as a tool to connect, on a consistent basis, all different cross sections the statistical agency intends to display to the general public either through a conventional table publication system or through web access. The contents of the guidance matrix serve as a filter to extract only the relevant aggregates for publication purposes from a generic N dimensional table structure. Tables generated by using the guidance matrix concept could be protected by either cell perturbation or cell suppression methods.

## 3 Guidance Matrices for Seven Test Cases

Currently, seven complex test cases of magnitude data are available to all SDL researchers of tabular data. Three of the seven test cases (HIER13, HIER16, and BTS4) consist of single tables with complex hierarchical structures. For these three test cases, the guidance matrix consists of only a single row. Due to the fact that all the column details are included in the table structure, all column values are assigned the value of one in the guidance matrix. The remaining four test cases each consist of multiple different multi-dimensional tables with complex linked structures. By following the guideline above, the guidance matrix for these test cases, namely NINENEW, NINE5D, TWO5IN6 and NINE12, can be displayed by using a consistent procedure. The guidance matrices for the seven test cases are in Table 1.

## Table 1. Guidance Matrices for Seven Test Cases

| HIER13 |  | NINENEW |  |  |  |  |  |  |  |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| 0 |  |  |  |  |  |  |  |  |  |

## BTS4

1111
NINE12

| 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 1 |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 |
| 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 1 |
| 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 1 |
| 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 1 |
| 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 1 |
| 0 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 0 |
| 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 |
| 0 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 0 |
| 1 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 |
| 1 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 |
| 1 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 |

Two of the three 9-D test cases, namely NINENEW and NINE12, are created from the same micro data file. In these two test cases, only a limited number of distinct 4-D linked tables are captured in 9-D variable space. In a real data situation, the statistical office would have decided on the substructure, including a dimensionality, for each table by taking into account user needs for published information.

In an extreme situation, the statistical office could have decided on capturing all the possible 4-D linked tables resulting from the 9 variables and making them available to the general public after statistical disclosure control of the tabular data. Under such a scenario, the guidance matrix will consist of 126 rows derived
by using a combination of nine items, taken 4 items at-a-time combinatorial logic. Similarly, in the event that there is a need to publish all possible 5-D or 6 -D tables resulting from the nine variables for public use, the guidance matrix consisting respectively of 126 and 84 rows could be constructed to satisfy user needs.

## 4 Complex Interactions in Hierarchical Structures

The guidance matrix concept offers considerable flexibility in capturing multiway complex interactions in table publication systems consisting of multiple hierarchical structures. This is easy to demonstrate by using a hypothetical example. We assume that the statistical agency collects manufacturing industry data at the state level and three different Standard Industry Classification (SIC) levels, namely 2 digits, 3 digits and 4 digits SIC codes. We further assume that the individual states are further grouped into various regions. Complex requirements, such as the statistical agency deciding to publish (a) 2 digits SIC level data at the state level (b) 2 digits and 3 digits SIC level data at the regional level and (c) 2 digits, 3 digits, and 4 digits SIC level data at the national level, are easy to implement by using the guidance matrix concept. In the following guidance matrix implementation we assume that the first five columns of the guidance matrix represent state code, region code, 2 digits SIC code, 3 digits SIC code and 4 digits SIC code respectively. The rest of the column positions in the guidance matrix are assumed to relate to other categorical variables. To capture the complex data publication requirement above, the guidance matrix is implemented as shown in Table 2:

Table 2. Complex Hierarchies Using Guidance Matrix

```
10100...... 2 digits SIC level state data
01110..... 2 and 3 digits SIC level regional data
00111..... 2, 3, and 4 digits SIC level national data
```


## 5 Eight Dimensional Counts Data Using CPS File from UC Irvine

Multi-way interaction terms are of importance in an analysis of counts and magnitude data. In this section, we demonstrate the versatility of the use of a guidance matrix to capture multi-way interactions on eight dimensional counts data. The file contains information on 48,842 individuals. The file was obtained from Dr. Alan Karr of the National Institute of Statistical Sciences (NISS) in August $2001{ }^{1}$. Dr. Karr believes this to be real data created by using a public use file with the original source of Current Population Survey (CPS) file from UC Irvine.

[^0]As stated by Dr. Karr, there were originally more variables and more categories for some of the variables and that NISS staff did some collapsing of the variables. After collapsing, the test data has the characteristics shown in Table 3.

Table 3. Characteristics of Eight Dimensional Counts Data

| Variable | \# of Categories | Category Label |
| :---: | :---: | :---: |
| Age | 3 | <25,25-55, >55 |
| Employer Type | 4 | Govt, Priv, SE, Other |
| Education Completed | 5 | <HS, HS, Bach, $\mathrm{Bach}^{+}$, Coll ${ }^{+}$ |
| Marital Status | 2 | Married, Unmarried |
| Race | 2 | White, Non White |
| Sex | 2 | Male, Female |
| Hours Worked | 3 | <40, 40, >40 |
| Annual Salary | 2 | <\$50K, >=\$50K |

There are multiple ways this data could be processed for statistical disclosure control and released for general public use. As an option, based on the data users' needs the statistical office could decide a priori on the total number of cross sections and dimensionalities of each cross section for public release by using the concept of a guidance matrix; and thereafter simultaneously apply an appropriate SDL method to those cross sections prior to the release of the information for general public use. Both cell suppression methods and cell perturbation methods could be used for statistical disclosure control of related tables.

Alternatively, the statistical office could decide on public release of table structures by using all the possible combinations of eight variables. Table 4 shows eight such options available to the statistical agency. In the first option, the statistical office could decide to release one cross section containing the entire 8 -D structure. Such a structure will contain 33,860 non-zero cells, of which 3,874 cells could be considered as sensitive cells because those cells contain 2 or fewer respondents. To protect such a table from statistical disclosure by using linear programming based techniques, a total of 113,538 equality constraints will need to be evaluated for a feasible solution. Similarly, the statistical agency could decide on publishing eight 7-D structures or twenty-eight 6 -D structures and so on. The total number of cross sections for release for each option is determined by using combinatorial logic.

From Table 4, it is obvious that in the option to release all possible 3-D sections, there are no sensitive cells involved. As a result, the statistical office could safely release those 56 cross sections to the general public without any concern for statistical disclosure. The last two options in the table, namely 28 2-D sections and eight 1-D sections, are subsets of fifty-six 3-D sections.

It is possible that some of the 704 -D sections might not contain any sensitive cells. The 4-D sections without any sensitive cells could also be released to the general public without any concern for statistical disclosure. The drawback of such a data release strategy is that only a small fraction of the total data collection effort is published. In addition, cross sections released using such a strategy might not be what the majority of data users need. Considering the

Table 4. Characteristics of Multidimensional Cross Sections of 8-D Counts Data

| Dimension <br> of the <br> Cross <br> Section | Total <br> Number of <br> Cross <br> Sections | Total <br> Number of <br> Non-Zero <br> Cells | Total <br> Number of <br> Sensitive <br> Cells | Total <br> Number of <br> Equality <br> Constraints |
| :---: | :---: | :---: | :---: | :---: |
| 8 | 1 | 33,860 | 3,874 | 113,538 |
| 7 | 8 | 32,165 | 3,327 | 105,045 |
| 6 | 28 | 25,367 | 1,832 | 76,731 |
| 5 | 56 | 14,609 | 498 | 39,411 |
| 4 | 70 | 5,755 | 56 | 13,561 |
| 3 | 56 | 1,506 | 0 | N/A |
| 2 | 28 | 251 | 0 | N/A |
| 1 | 8 | 24 | 0 | N/A |

Table 5. Characteristics of Eight 7-D Cross Sections from 8-D Counts Data

| Section \# | \# of Non-Zero <br> Cells | \# of <br> Sensitive <br> Cells | \# of Equality <br> Constraints |
| :---: | :---: | :---: | :---: |
| 1 | $\mathbf{1 2 , 2 0 0}$ | 830 | 35,777 |
| 2 | 8,961 | 685 | 27,229 |
| 3 | $\mathbf{1 1 , 7 4 8}$ | 1,063 | 34,911 |
| 4 | $\mathbf{1 1 , 9 2 1}$ | 915 | 35,232 |
| 5 | 11,829 | 970 | 35,076 |
| 6 | 6,179 | 305 | 19,119 |
| 7 | 7,446 | 437 | 22,755 |
| 8 | 9,394 | 508 | 28,139 |

cost involved in conducting surveys, the statistical office needs to concentrate on developing a strategy that will publish as much of the information as possible after adequately protecting data from statistical disclosure. We tackle that issue in later sections.

The classical literature related to SDL of tabular data often relies on a sequential processing of related cross sections followed by a backtracking procedure to ensure consistent protection across all cross sections. Apart from the ease of software implementation, such a strategy reduces the technical problem to a manageable size. Such a practice was desired and was considered acceptable in the past due to the practical resource constraints dictated by limited computational power. We believe that such an approach, in some instances, could be error prone and is not justified in an era where computational resources are of minor concern. However, as a matter of general research interest, in Table 5 we summarize the characteristics of one such option namely, generating and processing eight 7-D structures sequentially - option 2 from Table 4. The overall summary statistics of these eight sections are shown using a format similar to the one used in Table 4.

The contents of Table 5 are based on the eight 7-D cross sections derived by excluding one variable at a time in 7-D cross sections, starting with section
number 1 which excludes the 8 th variable, followed by section number 2 which excludes the 7 th variable and so on. The guidance matrix associated with these eight sections is in Table 6.

Table 6. Guidance Matrix for Eight 7-D Cross Sections

| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 |
| 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 |
| 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 |
| 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 |
| 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 |
| 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 |
| 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |

As can be seen collectively from Tables 4,5 , and 6 , when processed as a single table consisting of eight 7-D sections, there are 32,165 non-zero cells in a table and over 105 K equality constraints to work with. However, when these eight sections are identified and processed separately, the non-zero cell count and the sensitive cell count varies widely from section to section. The total number of equality constraints in each section reduces significantly, reducing the computational complexity and computation time by an order of magnitude for LP-based data protection methods. The sum of the cell count over all the sections does not add up to 32,165 when each section is processed separately. This is due to complex 7-way interaction effects captured through the guidance matrix.

## 6 Statistical Disclosure Control by Controlled Tabular Adjustments

To minimize the information loss, as an option, the statistical office could use controlled tabular adjustments (CTA) on counts data and magnitude data ${ }^{2}$. Dandekar/Cox (2002) have proposed using a LP- based CTA implementation procedure. Dandekar (2003) has proposed using a simplified CTA procedure, which attempts to minimize percentage deviation from the true cell value for non-sensitive cells without using any special purpose software. In Table 7, we summarize the outcome from both the CTA procedures when used on NINE12 magnitude test data. The LP-based procedure uses the reciprocal of cell values as a cost function to minimize the overall deviation of non-sensitive cells from the true cell value. The reciprocal of the cell value-based cost function closely captures the minimum percent deviation criteria used in the simplified CTA procedure.

[^1]Table 7. Comparison of LP-Based CTA and Simplified CTA Method
LP-Based CTA $12 \times 4-\mathrm{D}$ Cross sections

| Cell Size | 0 | 1 | 11 | 26 | 51 | 101 | 151 | 201 | 301 |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| = | To | To | To | то | To | To | To | то | To |
| From To | 0 | 10 | 25 | 50 | 100 | 150 | 200 | 300 | 600 |
| 1-32 | 34 | 1 | 2 | 0 | 1 | 0 | 0 | 0 | 0 |
| $33-64$ | 61 | 6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| $65-128$ | 36 | 63 | 18 | 0 | 2 | 0 | 0 | 0 | 0 |
| 129-250 | 99 | 31 | 51 | 0 | 0 | 4 | 0 | 0 | 0 |
| 251-500 | 124 | 1 | 86 | 95 | 1 | 0 | 0 | 0 | 0 |
| 501-1000 | 366 | 43 | 29 | 250 | 230 | 7 | 0 | 0 | 0 |
| 1001 - 2000 | 514 | 277 | 99 | 101 | 447 | 25 | 3 | 1 | 0 |
| 2001 - 4000 | 583 | 880 | 326 | 299 | 258 | 56 | 19 | 0 | 0 |
| 4001 - 8000 | 200 | 881 | 382 | 455 | 513 | 98 | 30 | 4 | 0 |
| 8001 - 16000 | 30 | 391 | 262 | 312 | 331 | 97 | 17 | 7 | 0 |
| 16001-32000 | 2 | 55 | 67 | 92 | 138 | 83 | 40 | 13 | 3 |
| 32001-64000 | 0 | 39 | 38 | 44 | 84 | 71 | 25 | 13 | 2 |
| 64001-820000 | 1 | 1 | 4 | 3 | 6 | 7 | 10 | 14 | 5 |
| T O T A L | 2051 | 2669 | 1364 | 1651 | 2011 | 448 | 144 | 52 | 10 |

## Simplified CTA $12 \times 4-D$ Cross sections

| Cell size | 0 | 1 | 11 | 26 | 51 | 101 | 151 | 201 | 301 |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| ============= | To | To | To | To | To | To | To | To | To |
| From To | 0 | 10 | 25 | 50 | 100 | 150 | 200 | 300 | 600 |
| $1-32$ | 38 | 0 | 0 | $\begin{array}{r} === \\ 0 \end{array}$ | $0$ | 0 | 0 | 0 | 0 |
| $33-64$ | 61 | 6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 65-128 | 36 | 73 | 10 | 0 | 0 | 0 | 0 | 0 | 0 |
| 129-250 | 101 | 30 | 54 | 0 | 0 | 0 | 0 | 0 | 0 |
| 251-500 | 125 | 4 | 86 | 92 | 0 | 0 | 0 | 0 | 0 |
| 501 - 1000 | 429 | 9 | 37 | 227 | 223 | 0 | 0 | 0 | 0 |
| 1001 - 2000 | 936 | 25 | 38 | 45 | 413 | 10 | 0 | 0 | 0 |
| 2001 - 4000 | 1936 | 49 | 94 | 128 | 169 | 28 | 16 | 1 | 0 |
| 4001 - 8000 | 1695 | 76 | 109 | 220 | 385 | 61 | 15 | 2 | 0 |
| $8001-16000$ | 821 | 34 | 89 | 166 | 284 | 35 | 15 | 3 | 0 |
| 16001-32000 | 77 | 35 | 48 | 77 | 135 | 64 | 31 | 23 | 3 |
| 32001-64000 | 20 | 13 | 36 | 59 | 85 | 56 | 22 | 24 | 1 |
| 64001-820000 | 0 | 1 | 2 | 7 | 8 | 8 | 1 | 9 | 15 |
| T ○ T A L | 6275 | 355 | 603 | 1021 | 1702 | 262 | 100 | 62 | 19 |

In Table 7, we use a two-dimensional histogram of a non-zero cell count to show how the cell value changes are distributed by the cell size. The rows in the table are classified into various cell size categories. The table columns are used to classify the changes in the cell value in various categories after the CTA procedure is used on the NINE12 test case. The comparative analysis of the outcome from the two CTA methods reveals that the simplified CTA procedure makes overall small changes to the magnitude data relative to the LP-based procedure proposed by Dandekar/Cox (2002). Approximately $62 \%$ of cells (6,275
out of 10,399 cells) in the simplified CTA method are unchanged, but only $20 \%$ of cells ( 2,051 out of 10,399 cells) in the LP-based CTA procedure are unchanged. Based on the percentage change in cell value criteria, $73 \%$ of cells in the LP-based CTA procedure and $79 \%$ of cells in the simplified CTA procedure are within $1 \%$ of the true cell value. Based on these statistics, it is clear that the simplified CTA procedure offers a low cost, computationally efficient, practical alternative to the LP-based CTA procedure.

As mentioned earlier, in the NINE12 test case only 12 out of the possible 126 4-D cross sections were included to generate 4-D linked tables. In Appendix A, we provide similar two-dimensional histograms based summary statistics for all possible 4-D, 5-D and 6-D sections protected by CTA procedure. Databases created using these three test cases result in (a) 1264 -D sections containing 60,588 non-zero cells (b) 1265 -D sections containing 120,161 non-zero cells and (c) 846 -D sections containing 161,783 non-zero cells. In all three test runs, $68 \%$ of total cell values are consistently within $1 \%$ of the true cell value.

## 7 Simplified CTA Procedure for Counts Data

We have used the simplified CTA procedure on the entire 8-D cross section of the counts data ${ }^{3}$. In our CTA implementation, the cell values of less than three are selected for an adjustment at random. The values for these cells are changed either to zero or to three in such a way that the net adjustment to the table total always stays close to zero after each adjustment. The CTA procedure changes the table total from 48,842 to 48,844 . Table 8 summarizes the outcome from the simplified CTA procedure by using the two-dimensional histogram constructed in a format similar to Table 7. Approximately $10 \%$ of 8 -D cross section data $(3,874$ out of 33,860$)$ contain sensitive cells. Yet, approximately one third of the total cells $(10,986$ out of 33,860$)$ are unchanged. Another third of the cells $(10,266$ out of 33,860$)$ are changed by only a single unit. The remaining one third of the cells undergo varying amounts of change ${ }^{4}$. Overall, only 18 percent of the cells are altered by more than 2 count.

By using the strategy proposed by Dandekar/Cox (2002) to convey the overall cell value accuracy to the general public, the CTA protected 8-D section of the counts data could be released for public use in its entirety. As a part of the recommended strategy, the statistical agency could make the Table 8 format summary statistics publicly available to all data users. In Table 8, some collapsing of small cell categories might be required to avoid an indirect disclosure of the data protection strategy.

[^2]Table 8. Histogram By Cell Size and Change in Cell Value

| 8-D CROSS SECTIONS - CELL COUNT: 33860 |  |  |  |  |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  |  | Abso | ute | Change | in Cell Value |  |  |  |  |
| Cell Size |  | 0 | 1 | 2 | 3 | 5 | 7 | 11 | 21 | 31 |
|  | $======$ | To | то | то | To | To | то | то | то | то |
| From | To | 0 | 1 | 2 | 4 | 6 | 10 | 20 | 30 | 40 |
| 1 | 1 | 0 | 1589 | 798 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 2 | 0 | 835 | 563 | 89 | 0 |  | 0 | 0 | 0 |
| 3 | 5 | 1351 | 752 | 422 | 362 | 65 | 3 | 0 | 0 | 0 |
| 6 | 10 | 1147 | 884 | 530 | 320 | 50 | 16 | 0 | 0 | 0 |
| 11 | 25 | 1752 | 1339 | 896 | 509 | 125 | 32 | 2 | 0 | 0 |
| 26 | 50 | 1456 | 1032 | 629 | 468 | 139 | 39 | 3 | 0 | 0 |
| 51 | 100 | 1426 | 976 | 651 | 518 | 167 | 78 | 9 | 0 | 0 |
| 101 | 500 | 2480 | 1744 | 1214 | 1028 | 391 | 189 | 34 | 2 | 0 |
| 501 | - 1000 | 630 | 491 | 317 | 301 | 144 | 75 | 25 | 3 | 0 |
| 1001 | - 5000 | 670 | 548 | 359 | 382 | 201 | 130 | 48 | 10 | 3 |
| 5001 | - 10000 | 58 | 60 | 51 | 55 | 35 | 26 | 8 | 4 | 1 |
| 10001 | - 25000 | 16 | 15 | 13 | 27 | 13 | 14 | 9 | 2 | 0 |
| 25001 | - 50000 | 0 | 1 | 3 | 4 | 0 | 2 | 2 | 0 | 0 |
| T 0 T | A L | 10986 | 10266 | 6446 | 4063 | 1330 | 604 | 140 | 21 | 4 |

To convey cell level data quality, the statistical agency could use the quality indicator field in connection with cell values to reflect "poor" or "good" data quality. The other option could be to simply not publish the cell values which are beyond previously accepted threshold accuracy criteria. In the later option, if a user decides to estimate the missing cell value by making use of related equality constraints, that will be the choice for those who need the "ball park" estimate.

Statistical agencies typically serve a diverse group of end users with varying information needs. Once CTA-adjusted data is in the public domain, different users are interested in different cross sections of these data. For example, for individuals interested in developing log-linear models on these data, the first few orders of interaction terms captured by all combinations of 3-D cross sections is all that might be required. As mentioned earlier, there are no sensitive cells in any of the 3-D sections of the database. However, the challenge for the statistical agency is to balance the need for the data for log-linear model developers with potential other users with different information needs.

In Table 9 we explain how CTA attempts to balance data utility and information loss. Table 9 is in 3 parts. Part A of the table shows cell size distribution for different cross section dimensions ${ }^{5}$. As illustrated in the table, the lower dimensional cross sections are typically larger in cell size. In our example, the upper limit for the cell size is 48,842 , which is the total number of individuals in the database. At the bottom of the first part of the table, we provide a cumulative cell count for various cross sections. If log-linear modelers were the only users

[^3]Table 9. Cross Sectional Details of Simplified CTA Protected Counts Data
Part (A) CELL COUNT BY CELL SIZE

Part(B) AVERAGE CELL VALUE CHANGE BY CELL SIZE

| $\begin{aligned} & \text { CELL } \\ & \text { From } \end{aligned}$ | $\begin{gathered} \text { SIZE }< \\ \text { To } \end{gathered}$ | 1-D | 2-D | $\begin{aligned} & \text { SECTION } \\ & 4-D \end{aligned}$ |  | 5-D | 6-D | 7-D | 8-D |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 1. | 1. | . 00 | . 00 | . 00 | 1.30 | 1.30 | 1.32 | 1.33 | 1.33 |
| 2. | 2. | . 00 | . 00 | . 00 | 1.48 | 1.65 | 1.64 | 1.59 | 1.56 |
| 3. | 5. | . 00 | . 00 | . 67 | 1.68 | 1.64 | 1.43 | 1.18 | 1.06 |
| 6. | 10. | . 00 | . 00 | . 50 | 2.19 | 1.85 | 1.50 | 1.25 | 1.15 |
| 11. | 25. | . 00 | . 00 | 2.28 | 2.20 | 1.93 | 1.57 | 1.31 | 1.24 |
| 26. | 50. | . 00 | . 00 | 2.68 | 2.57 | 2.07 | 1.63 | 1.36 | 1.31 |
| 51. | 100. | . 00 | 4.00 | 3.68 | 2.90 | 2.26 | 1.76 | 1.51 | 1.47 |
| 101. | 500. | . 00 | 4.80 | 3.62 | 2.99 | 2.30 | 1.84 | 1.68 | 1.66 |
| 501. | 1000. | . 00 | 2.56 | 3.91 | 3.17 | 2.49 | 2.09 | 1.97 | 1.95 |
| 1001. | 5000. | 2.50 | 5.19 | 4.41 | 3.42 | 2.71 | 2.42 | 2.37 | 2.37 |
| 5001 | 10000. | 1.33 | 4.42 | 4.03 | 3.52 | 3.19 | 3.17 | 3.17 | 3.17 |
| 10001. | 25000. | 5.44 | 4.75 | 4.78 | 4.26 | 4.26 | 4.26 | 4.26 | 4.26 |
| 25001. - | 50000. | 3.14 | 5.08 | 5.08 | 5.08 | 5.08 | 5.08 | 5.08 | 5.08 |

Part (C) AVERAGE PERCENT CELL VALUE CHANGE BY CELL SIZE

| $\begin{aligned} & \text { CELI } \\ & \text { From } \end{aligned}$ | $\begin{gathered} \text { SIZE }< \\ \text { To } \end{gathered}$ | 1-D | 2-D | $\begin{array}{r} \text { ROSS } \\ 3-D \end{array}$ | $\begin{gathered} \text { SECTION } \\ 4-\mathrm{D} \end{gathered}$ | DIMENS | 6-D | 7-D | 8-D |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 1 | 1. | . 00 | . 00 | . 00 | 129.63 | 130.50 | 132.06 | 133.27 | 133.43 |
| 2. | 2. | . 00 | . 00 | . 00 | 74.14 | 82.41 | 82.21 | 79.40 | 77.91 |
| 3. | 5. | . 00 | . 00 | 16.67 | 42.08 | 41.10 | 35.66 | 29.46 | 26.39 |
| 6. | 10. | . 00 | . 00 | 6.25 | 27.32 | 23.07 | 18.70 | 15.62 | 14.32 |
| 11. | 25. | . 00 | . 00 | 12.65 | 12.24 | 10.73 | 8.71 | 7.28 | 6.88 |
| 26. | 50. | . 00 | . 00 | 7.05 | 6.76 | 5.46 | 4.29 | 3.59 | 3.46 |
| 51. | 100. | . 00 | 5.30 | 4.87 | 3.85 | 3.00 | 2.33 | 2.01 | 1.95 |
| 101. | 500. | . 00 | 1.60 | 1.20 | 1.00 | . 76 | . 61 | . 56 | . 55 |
| 501. | 1000. | . 00 | . 34 | . 52 | . 42 | . 33 | . 28 | . 26 | . 26 |
| 1001. | - 5000. | . 08 | . 17 | . 15 | . 11 | . 09 | . 08 | . 08 | . 08 |
| 5001. | - 10000. | . 02 | . 06 | . 05 | . 05 | . 04 | . 04 | . 04 | . 04 |
| 10001. | - 25000. | . 03 | . 03 | . 03 | . 02 | . 02 | . 02 | . 02 | . 02 |
| 25001. | - 50000. | . 01 | . 01 | . 01 | . 01 | . 01 | . 01 | . 01 | . 01 |

of this database, the statistical agency could release the information for a total of 1,506 cells associated with all the 3-D cross sections in the database (without confidentiality concern) and choose not to publish the rest of the cells in the database. By adopting such a policy, however, the statistical agency would have published only $4 \%$ ( 1,506 out of 33,860 cells) of data collected by the survey operation. The challenge facing the statistical agency is how to balance the data needs for diverse user groups.

Part B of Table 9 shows the average change in the cell value within various cross sections of CTA adjusted 8-D data by using the same size categories as in the first part of the table. Part C of the Table 9 shows the corresponding percentage changes in cell values in different cross sections by related size categories. Due to the fact that the majority of the lower dimensional cross sections are relatively large in cell size, the resulting percentage adjustments from the CTA procedure are generally smaller than one percentage point. Such a small percentage change in cell values is therefore likely to be statistically insignificant for hypothesis testing using log-linear model analysis. At the same time, a diverse group of other data users have access to all other relevant information from various cells of different dimensionalities resulting from the 8-D database. To demonstrate the suitability of the CTA protected data for log-linear analysis, in Appendix B, we provide a comparative evaluation of the actual and CTA protected data on the 3-D section of age, education and annual salary from the 8-D data.

## 8 Static vs. Dynamic Table Server Design

In recent years, there has been some discussion on operating SDL protected table servers in dynamic mode through interactive web access. The guidance matrix concept, in combination with commonly used cell suppression and/or cell perturbation methods, offers wide flexibility to statistical agencies in generating a wide variety of table structures for public use. Except for one, all tabular data protection methods require (a) a number of categorical variables for publication, (b) a number of table structures, and (c) dimensionality of the table structure for public release to be defined a priori to ensure adequate statistical disclosure control.

For the simplified CTA procedure, however, there is no need to know the total number of table structures and table dimensionalities for public release a priori in the system design. This allows the table server to function more or less in a dynamic mode without concern for statistical disclosure. Whenever dynamic table server capability is required through interactive web access, the statistical agency has two separate options. In the first option, the statistical agency could store in a database only the internal tabular cell values after adjusting internal sensitive table cells by plus or minus adjustment factors. The statistical agency could then provide user specific table generation capability in real time. In the second option, the statistical agency could store in the database management system all SDL protected internal cell values along with all possible marginal cell values. Users will be allowed to retrieve necessary information by accessing the database through a query system.

## 9 Conclusion

The concept of a guidance matrix allows statistical agencies a flexible option to simultaneously capture and process multiple multi-dimensional cross sections
containing complex hierarchies in tabular data release. The guidance matrix tool is equally effective on counts data and on magnitude data and is compatible with all cell suppression and cell perturbation methods used to protect sensitive tabular data. Table servers, based on the CTA procedure for SDL control, in combination with flexibility of data release offered by the guidance matrix concept, further enhances table server capabilities.

Log-linear models are association models. Inferences based on any model are subject to some uncertainty. The uncertainty bounds for model-based inferences are easy to estimate and quantify by performing sensitivity/uncertainty analysis. Latin hypercube sampling based experimental design studies are routinely conducted to simultaneously capture the effects of uncertainties in multiple model input parameters on the model inference(s). We look forward to such evaluations of log-linear and other relevant models by using CTA protected tabular data.

## Appendix A

## Simplified CTA on 9-D Magnitude Data

| $126 \times 4-D$ | Cross | sections |  |  | CEL工 | COUNT : |  | 60,588 |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| $-\gg$ | $<----$ | Absol | ute | Change | in Cel | 1 Va | ue | - - |  |
| Cell Size | 0 | 1 | 11 | 26 | 51 | 101 | 151 | 201 | 301 |
| = = = = = = = = = = = = = = | To | To | To | To | TO | To | To | To | To |
| From To | 0 | 10 | 25 | 50 | 100 | 150 | 200 | 300 | 600 |
| 1 - 32 | 502 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| $33-64$ | 606 | 97 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 65-128 | 363 | 713 | 86 | 0 | 0 | 0 | 0 | 0 | 0 |
| 129-250 | 941 | 331 | 575 | 0 | 0 | 0 | 0 | 0 | 0 |
| 251 - 500 | 1074 | 49 | 809 | 906 | 0 | 0 | 0 | 0 | 0 |
| 501 - 1000 | 3689 | 79 | 239 | 2267 | 2242 | 0 | 0 | 0 | 0 |
| 1001-2000 | 7045 | 239 | 326 | 629 | 4143 | 110 | 11 | 0 | 0 |
| 2001 - 4000 | 12819 | 405 | 765 | 1096 | 1849 | 278 | 108 | 6 | 0 |
| $4001-8000$ | 5962 | 358 | 634 | 1155 | 2302 | 345 | 99 | 23 | 1 |
| $8001-16000$ | 959 | 202 | 307 | 488 | 905 | 241 | 79 | 32 | 2 |
| 16001-32000 | 34 | 41 | 73 | 95 | 193 | 98 | 66 | 54 | 11 |
| $32001-64000$ | 0 | 20 | 32 | 47 | 120 | 66 | 43 | 39 | 14 |
| 64001-820000 | 0 | 3 | 2 | 3 | 8 | 5 | 6 | 11 | 12 |
| T O T A L | 33994 | 2537 | $===$ 3848 | 6686 | 11762 | 1143 | 412 | 165 | 40 |

$126 \times 5-D$ Cross sections - CELL COUNT: 120,161

| Cell Size | 0 | 1 | 11 | 26 | 51 | 101 | 151 | 201 | 301 |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| = = = = = = = = = = = = | T0 | To | T0 | T0 | TO | TO | T0 | T0 | T0 |
| From To | 0 | 10 | 25 | 50 | 100 | 150 | 200 | 300 | 600 |
|  |  |  |  |  |  |  |  |  | = |
| $1-32$ | 1289 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| $33-64$ | 1527 | 278 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| $65-128$ | 936 | 1822 | 202 | 0 | 0 | 0 | 0 | 0 | 0 |
| 129 - 250 | 2402 | 802 | 1555 | 0 | 0 | 0 | 0 | 0 | 0 |
| 251 - 500 | 2666 | 61 | 2036 | 2339 | 0 | 0 | 0 | 0 | 0 |
| 501 - 1000 | 9094 | 98 | 496 | 5338 | 5523 | 0 | 0 | 0 | 0 |
| 1001 - 2000 | 17062 | 283 | 388 | 777 | 9451 | 131 | 12 | 0 | 0 |
| 2001 - 4000 | 29464 | 489 | 933 | 1251 | 2149 | 311 | 116 | 6 | 0 |
| 4001 - 8000 | 9248 | 379 | 676 | 1226 | 2521 | 350 | 99 | 23 | 1 |
| 8001-16000 | 995 | 202 | 307 | 489 | 907 | 241 | 79 | 32 | 2 |
| 16001-32000 | 34 | 41 | 73 | 95 | 193 | 98 | 66 | 54 | 11 |
| $32001-64000$ | 0 | 20 | 32 | 47 | 120 | 66 | 43 | 39 | 14 |
| 64001-820000 | 0 | 3 | 2 | 3 | 8 | 5 | 6 | 11 | 12 |
| T O T A L | 74717 | 4478 | 6700 | 11565 | 20872 | 1202 | 421 | 165 | 40 |

84 x 6-D Cross sections - CELL COUNT: 161,783

| Cell size | 0 | 1 | 11 | 26 | 51 | 101 | 151 | 201 | 301 |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| $======$ = | To | то | To | To | To | To | To | To | To |
| From To | 0 | 10 | 25 | 50 | 100 | 150 | 200 | 300 | 600 |
| 1 - 32 | 1868 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 33-64 | 2194 | 428 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 65-128 | 1353 | 2643 | 285 | 0 | 0 | 0 | 0 | 0 | 0 |
| 129-250 | 3474 | 1135 | 2292 | 0 | 0 | 0 | 0 | 0 | 0 |
| 251 - 500 | 3827 | 62 | 2940 | 3403 | 0 | 0 | 0 | 0 | 0 |
| 501 - 1000 | 13038 | 100 | 666 | 7507 | 7913 | 0 | 0 | 0 | 0 |
| 1001 - 2000 | 24276 | 287 | 394 | 802 | 13196 | 133 | 12 | 0 | 0 |
| 2001 - 4000 | 41315 | 508 | 964 | 1268 | 2173 | 314 | 116 | 6 | 0 |
| 4001 - 8000 | 11224 | 381 | 679 | 1233 | 2546 | 350 | 99 | 23 | 1 |
| 8001 - 16000 | 999 | 202 | 307 | 489 | 907 | 241 | 79 | 32 | 2 |
| 16001-32000 | 34 | 41 | 73 | 95 | 193 | 98 | 66 | 54 | 11 |
| $32001-64000$ | 0 | 20 | 32 | 47 | 120 | 66 | 43 | 39 | 14 |
| 64001-820000 | 0 | 3 | 2 | 3 | 8 | 5 | 6 | 11 | 12 |
| T O T A L | $\begin{array}{r} ===== \\ 103602 \end{array}$ | 5810 | 8634 | $\begin{array}{r} ==== \\ 14847 \end{array}$ | = 27056 | $\begin{aligned} & ==== \\ & 1207 \end{aligned}$ | = = = | = $==$ 165 | = = |

## Appendix B1

## ACTUAL DATA - LOG LINEAR ANALYSIS

PARTIAL ASSOCIATION STATISTICS USING IMSL ROUTINE "CTASC"
3-D SECTION X ( $A=A G E, ~ B=E D U C A T I O N, ~ C=A N N U A L ~ S A L A R Y)$
Table 1: C = 1
A (row) by B (column)

|  | 1 | 2 | 3 | 4 | 5 |
| :---: | :---: | :---: | :---: | :---: | :---: |
| 1 | 1738.0 | 2444.0 | 1088.0 | 33.0 | 3036.0 |
| 2 | 3051.0 | 8930.0 | 5676.0 | 1245.0 | 5010.0 |
| 3 | 1253.0 | 1907.0 | 675.0 | 300.0 | 769.0 |

A (row) by B (column)

|  |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | 1 | 2 | 3 | 4 | 5 |
|  | 1 | 11.0 | 20.0 | 28.0 | 4.0 | 0.0 |
|  | 2 | 221.0 | 2020.0 | 3657.0 | 2037.0169 | 4.0 |
|  | 3 | 134.0 | 463.0 | 563.0 | 466.0 | 9.0 |
|  |  | Partial Association Statistics |  |  |  |  |
| Omitted |  | Degrees of |  |  |  | Marginal |
| Effect |  | Chi-S | are | Freedom | P-value | Zeros |
| A |  | 255 | 5.30 | 2.0 | . 0000 | . 0 |
| B |  |  | 3.81 | 4.0 | . 0000 | . 0 |
| C |  | 139 | 8.71 | 1.0 | . 0000 | . 0 |
| A*B |  |  | 4.77 | 8.0 | . 0000 | . 0 |
| A*C |  |  | 9.65 | 2.0 | . 0000 | . 0 |
| B*C |  |  | 6.03 | 4.0 | . 0000 | . 0 |
| A*B*C |  |  | 6.51 | 8.0 | . 0009 | . 0 |

Chi-square statistics for testing that all $k$ and higher interactions are zero.


## Appendix B2

CTA PROTECTED DATA - LOG LINEAR ANALYSIS
PARTIAL ASSOCIATION STATISTICS USING IMSL ROUTINE "CTASC" 3-D SECTION X (A=AGE, B=EDUCATION, C=ANNUAL SALARY )

Table 1: $\mathrm{C}=1$
A (row) by B (column)


A (row) by B (column)

|  | 1 | 2 | 3 | 4 | 5 |
| ---: | ---: | ---: | ---: | ---: | ---: |
| 1 | 6.0 | 17.0 | 31.0 | 6.0 | 29.0 |
| 2 | 226.0 | 2016.0 | 3653.0 | 2043.0 | 1691.0 |
| 3 | 127.0 | 467.0 | 562.0 | 471.0 | 343.0 |


| Omitted | Partial Association Statistics |  |  |  |
| :--- | ---: | ---: | ---: | ---: |
| Degrees of |  |  |  |  |
| Effect | Chi-Square | Freedom | P-value | Marginal <br> Zeros |
| A | 25539.01 | 2.0 | .0000 | .0 |
| B | 9150.00 | 4.0 | .0000 | .0 |
| C | 13958.07 | 1.0 | .0000 | .0 |
| A*B | 2572.01 | 8.0 | .0000 | .0 |
| A*C | 3110.17 | 2.0 | .0000 | .0 |
| B*C | 4655.57 | 4.0 | .0000 | .0 |
| A*B*C | 24.14 | 8.0 | .0022 | .0 |



## References

1. Dandekar R. A. and Cox L. H. (2002), "Synthetic Tabular Data: An Alternative to Complementary Cell Suppression", manuscript available from ramesh.dandekar@eia.doe.gov or ramesh.dandekar@verizon.net
2. Dandekar, Ramesh A. (2003), "Cost Effective Implementation of Synthetic Tabulation (a.k.a. Controlled Tabular Adjustments) in Legacy and New Statistical Data Publication Systems", working paper 40, UNECE Work session on statistical data confidentiality (Luxembourg, 7-9 April 2003)
3. Saltelli A., Chan K., Scott E. M. (2001), "Sensitivity Analysis", Wiley Series in Probability and Statistics

# Privacy, Accuracy, and Consistency Too: A Holistic Solution to Contingency Table Release 

Boaz Barak ${ }^{*}$<br>Princeton University<br>boaz@cs.princeton.edu<br>Satyen Kale ${ }^{*}$<br>Princeton University<br>satyen@cs.princeton.edu

Kamalika Chaudhuri* UC Berkeley kamalika@cs.berkeley.edu

Frank McSherry Microsoft Research, SVC mcsherry@microsoft.com

Cynthia Dwork
Microsoft Research, SVC
dwork@microsoft.com

Kunal Talwar Microsoft Research, SVC kunal@microsoft.com


#### Abstract

The contingency table is a work horse of official statistics, the format of reported data for the US Census, Bureau of Labor Statistics, and the Internal Revenue Service. In many settings such as these privacy is not only ethically mandated, but frequently legally as well. Consequently there is an extensive and diverse literature dedicated to the problems of statistical disclosure control in contingency table release. However, all current techniques for reporting contingency tables fall short on at least one of privacy, accuracy, and consistency (among multiple released tables). We propose a solution that provides strong guarantees for all three desiderata simultaneously.

Our approach can be viewed as a special case of a more general approach for producing synthetic data: Any privacypreserving mechanism for contingency table release begins with raw data and produces a (possibly inconsistent) privacypreserving set of marginals. From these tables alone - and hence without weakening privacy - we will find and output the "nearest" consistent set of marginals. Interestingly, this set is no farther than the tables of the raw data, and consequently the additional error introduced by the imposition of consistency is no more than the error introduced by the privacy mechanism itself.

The privacy mechanism of [20] gives the strongest known privacy guarantees, with very little error. Combined with the techniques of the current paper, we therefore obtain excellent privacy, accuracy, and consistency among the tables. Moreover, our techniques are surprisingly efficient.

Our techniques apply equally well to the logical cousin of the contingency table, the OLAP cube.


[^4]Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
PODS'07, June 11-14, 2007, Beijing, China.
Copyright 2007 ACM 978-1-59593-685-1/07/0006 ...\$5.00.

## Categories and Subject Descriptors

G. 3 [PROBABILITY AND STATISTICS]: [Contingency table analysis]

## General Terms

Algorithms, Theory

## Keywords

Privacy, OLAP, Contingency table release

## 1. INTRODUCTION

Privacy-preserving data-mining, also known as statistical disclosure control, has historically been the purview of statisticians, both in practice, for example, at the Census Bureau, the Internal Revenue Service, and the Bureau of Labor Statistics, and in the research community. (see, for example, $[12,24,17,34,1]$ and the references in Section 1.1 below). In recent years the topic has experienced a resurgence in the computer science community (see, for example, $[2,3,4,23,14,9,10,6,20,18,37,35])$. In this work we focus on contingency tables, also known as frequency tables, and their logical cousins, On-Line Analytical Processing (OLAP) cubes.

### 1.1 Contingency Table Release

Informally, a contingency table is a table of counts. From a database consisting of $n$ rows, each comprising values for a fixed set of, say, binary attributes $a_{1}, \ldots, a_{k}$, the contingency table is the histogram of counts for each of the $2^{k}$ possible settings of these attributes. Contingency tables are essentially equivalent to OLAP cubes, which cast traditional relational databases as a high-dimensional cube with dimensions corresponding to the attributes. While we stay with the notation of statisticians, we stress that this is simply notational, and the results can be directly mapped to privacypreserving OLAP. ${ }^{1}$

What is commonly released is not a contingency table itself, but the projection of the cube onto a subset of the at-

[^5]tributes: the counts for each of the possible settings of the restricted set of attributes. These counts are called marginals, each marginal associated with a subset of the attributes, and called and $j$-way marginals when at most $j \leq k$ attributes are used. The data curator will typically release many sets of low-order marginals for a single contingency table, with the goal of revealing correlations between many different, and possibly overlapping sets of attributes.

At first glance, it might seem that low-order marginals are "naturally" privacy-preserving: after all, they are aggregations over many of database rows. This, however, is not the case. For example, small counts are considered disclosive: if a given pair of attribute values corresponds to a unique individual, then these fields can be used as a key in other databases to reveal further information about an individual. Empty cells are also potentially disclosive: while they do not point to a specific individual, they can permit the rejection of claims, eg: that student X received all 'A's by virtue of the fact that no student received such marks. Access to large counts over time can permit "differencing attacks", where the changes to the data set can serve as the basis for inference and privacy violation. Several papers examine the degree to which individual cell entries are revealed by marginals; see, eg, [15]. Finally, recent results of de Loera and Onn [30,29, 31] are particularly discouraging.

The disclosure risks inherent in contingency tables have given rise to an extensive and diverse literature on techniques for altering the true tables. There are two broad classes of techniques: non-perturbative (specifically, cell suppression) and perturbative (eg, controlled rounding and controlled tabular adjustment). In cell suppression, so-called "sensitive" cells are identified (the primary cells; see [16, 33] for a discussion of sensitivity rules). These are suppressed, together with a set of complementary secondary cells (to avoid the disclosure of the primary cells). The typical goal is to suppress as few secondary cells as possible, leading to difficult combinatorial problems with impractical execution times on large instances. Controlled rounding, initially introduced in [5], also suffers from combinatorial explosion, and is NP-hard even for the case of three-dimensional tables [27] Controlled Tabular Adjustment, due to Dandekar and Cox [13], and the use of quadratic interior-point methods, due to Castro [7], were introduced to address these difficulties. This is an active area of research; see for example, the discussion in [8],

Over the last 5 years or so, the database and cryptography communities have provided rigorous definitions of privacy and introduced techniques that provably satisfy the given definitions [23, 9, 4, 14, 22, 20].

The most general and robust of these, and the notion used in this work, captures the following intuition: the adversary learns nothing more about an individual when her data are included in the database than the adversary can learn about the individual when her data are not included in the database (see Section 1.2 for the formal definition and its motivation) [20, 18]. Combined with the algorithmic techniques developed in a series of papers ([14, 22, 6] and particularly $[20]$ ), these yield a simple approach to contingency table release, with excellent accuracy and strong privacy guarantees, independent of any auxiliary information available to the adversary and regardless of the adversary's computational power. At a high level, this approach
involves adding a small amount of independently and identically distributed noise to each cell in the released marginals. However, the small errors introduced to ensure privacy will cause distinct breakdowns of the data to yield slightly different counts (not to mention possibly negative and non-integer cell counts). The current work addresses this, adding consistency (and positivity and integrality) to privacy and accuracy.

### 1.2 A Formal Statement of Our Contribution

Our contribution, as suggested by the paper's title, comes in the parts of privacy, accuracy, and consistency, each of which are critical components of any data analysis system. At an intuitive level, which we soon formalize, we are concerned with

- Privacy: The presence or absence of any one data element should not substantially influence the distribution over outcomes of the computation.
- Accuracy: The difference between the reported marginals and true marginals should be bounded, preferably independent of the size of the data set.
- Consistency: There should exist a contingency table whose marginals equal the reported marginals.

We now formally discuss each, in the context of prior work.

### 1.2.1 Privacy

Since a rigorous claim about privacy is integral to our result we begin by recalling the definition of differential privacy $[18,20]$, which our algorithms will ensure.

Definition 1. [18, 20]. A randomized function $\mathcal{K}$ gives $\epsilon$-differential privacy if for all databases $D_{1}$ and $D_{2}$ differing on at most one element, and all measurable $S \subseteq \operatorname{Range}(\mathcal{K})$,

$$
\begin{equation*}
\operatorname{Pr}\left[\mathcal{K}\left(D_{1}\right) \in S\right] \leq \exp (\epsilon) \times \operatorname{Pr}\left[\mathcal{K}\left(D_{2}\right) \in S\right] \tag{1}
\end{equation*}
$$

A randomized function satisfying this definition addresses any concern that a participant might have about the use of his or her data. In a formal sense, the distribution over outcomes is almost as if the participant had opted out of the data set; no event is made substantially more or less likely by the use of her data. These "events" can be viewed mathematically, perhaps as outputs leading to a substantial shift between prior and posterior probabilities, or pragmatically, as actual objectionable events, eg: outputs leading to telemarketing calls or denial of credit.

Remark: Differential privacy has several consequences that follow from the definition but may not be immediately apparent. Notably, the definition is agnostic to auxiliary information an adversary may possess, and provides guarantees against arbitrary attacks. Moreover, any function with $\epsilon$ differential privacy also ensures $(\epsilon t)$-differential privacy for groups of size up to $t$, and the composition of $s$ functions with $\epsilon$-differential privacy ensures ( $\epsilon s$ )-differential privacy. See [20] for further discussion.

## Comparison with Other Definitions.

Differential Privacy provides much stronger guarantees than other privacy definitions of which we are aware. For example, $k$-anonymity [35, 37, 36] and its extension $l$-diversity[32] impose syntactic constraints on the outputs, requiring that
many groups of tuples appear indistinguishable, or uninformative about specific values. Nonetheless, neither of these definitions protects against even simple background knowledge of the form "My colleague Mr. R., who works in zip code $2770^{*}$, is in the database". For example, if, cumulatively, the people in the database suffer from a small set of ailments, then the adversary learns that Mr. R suffers from one of these ailments. This may be worse than embarassing; it may result in Mr. R. being dropped from consideration for promotion, say. In addition, even two $k$-anonymous or even $\ell$-diverse tables taken together may be completely disclosive.

Similarly, [23] promotes the concept of the ( $\rho_{1}, \rho_{2}$ )-privacy breach. Very roughly, such a breach represents a substantial change in the adversary's belief that an individual data item satisfies some particular property $P$. By definition, this notion is entangled with the adversary's prior knowledge about the data and seems to have forced some awkward assumptions (eg, independence among data items, adversary's knowledge of true prior). Comparable definitional awkwardness appears in $[22,9,10,6]$ (eg, the "informed adversary"). Interestingly, in hindsight we find that the algorithmic techniques in [23, 22, 6] yield stronger privacy than is proved in the papers themselves. This occurs because all three papers provide statistical guarantees regarding the outputs of the privacy mechanism, yielding, to differing extents, approximations to differential privacy.

Protection against ( $\rho_{1}, \rho_{2}$ )-privacy breaches in [23] comes as a consequence of the $\gamma$-amplification statistical guarantees. Other algorithmic approaches, such as in [4], prevent different sets of ( $\rho_{1}, \rho_{2}$ ) privacy breaches than does ensuring at most $\gamma$-amplification, and are less satisfactory. For example, the guarantees in [4] may fail to protect Mr. R.'s recent purchase of "herbal supplements" - a not uncommon event for the population in general but again, embarassing to Mr. R.

One might - erroneously - conclude that no technique can protect individuals against adversaries with arbitrary background knowledge. After all, it is formally proven in [18] that, for essentially any non-trivial mechanism and definition of privacy compromise, there exists auxiliary information for which the output of the mechanism enables a privacy compromise that would not otherwise be possible ${ }^{2}$.

This result, and its underlying intuition, led to Differential Privacy, which does provide guarantees against arbitrary auxiliary information. It succeeds because it makes the only fair comparison: the probabilities of disclosure (or any event at all, for that matter) with, versus without, the sensitive data. If a disclosure will happen even without a participant's data, perhaps because it is known beforehand that the participant is in the majority, say, then it is unfair to cast blame on the privacy mechanism: any mechanism that reports the majority would lead to the breach. This is a key distinction: comparing with and without a participant's data, rather than with and without the output of the mechanism, and it is what allows Differential Privacy to give such strong bounds. Since the definition talks about the statis-

[^6]tical distribution of the outcome, it obviates any discussion of the adversary's auxiliary information.

### 1.2.2 Accuracy

Privacy guarantees are of course meaningless without accompanying accuracy guarantees. We could easily erase the data if the former were all we cared about. We now detail guarantees that our algorithm makes about the accuracy of the counts in the released marginals, while ensuring $\epsilon$ differential privacy.

Our theorem statement is necessarily loose at the moment, for notational reasons. The full version appears as Theorem 7, and is tighter than what is presented now:

Theorem 1. (Rough Version): Let $C$ be a set of marginals of the contingency table, each on at most $j$ attributes. We compute marginals $C^{\prime}$ of a positive, integral contingency table, preserving $\epsilon$-differential privacy, such that with probability $1-\delta$ for any marginal $c \in C$,

$$
\begin{equation*}
\left\|c-c^{\prime}\right\|_{1} \leq 2^{j+3}|C| \log (|C| / \delta) / \epsilon+|C| \tag{2}
\end{equation*}
$$

This result does not depend on the total number of attributes in the data set, nor on the total number of elements in the data set, but rather only on the "complexity" of the query, in terms of the number and order of the marginals. Our result is the first we are aware of where the error in the marginals falls below statistical error due to sampling. Note also that while one might be concerned that $2^{j}$ is a large number, it is the number of elements that are reported by each marginal, and a natural scale for the L1 norm.

The most natural comparison to make is with the recent work of [4], on privacy preserving OLAP. In this work, which provides a limited form of ( $\rho_{1}, \rho_{2}$ )-privacy, the data are randomized with a constant probability, resulting in each count being reconstructible to within roughly $\sqrt{\mid \text { dataset } \mid}$. Our approach improves the error by exploiting the property that it is the number of marginals requested, $|C|$, that determines a sufficient amount of noise.

Remark: Randomized response, or any other mechanism that allows the user to learn answers to too many counting queries of the form "How many of the subset S of tuples satisfy property P?" must necessarily introduce large amounts of noise. This follows from results of [14], originally obtained for the interactive case, but applying here as well. For example, given any mechanism for which the magnitude of the error on all $2^{n}$ counting queries is bounded by $E$, the adversary can produce a candidate vector that agrees with $\left(P\left(\right.\right.$ tuple $\left._{1}\right), \ldots, P\left(\right.$ tuple $\left.\left._{n}\right)\right)$ on all but $4 E$ entries. So for $E \in o(n)$, say, $E=n^{1-\epsilon}$, the adversary learns more than $99.99 \%$ of the $P$ values. An efficient version of this requires only that the adversary obtain responses to $n \log ^{2} n$ randomly chosen subset counting queries with $o(\sqrt{n})$ error (in fact, an efficient attack may be carried out even if more than $20 \%$ of $O(n)$ queries have wild error, while the remaining suffer from at most $o(\sqrt{n})$ error [21]). In some sense the problem is that the randomized response mechanism reveals the (very roughly approximate) answers to many more queries than the user may actually want to pose. By focusing on interactive mechanisms we can add just enough noise to ensure privacy for a given number of queries. Whenever the curator knows the questions in advance, a "transcript" can be prepared with the desired queries and responses, so for the case of contingency tables, or OLAP cubes, we are not restricting ourselves by focusing on the interactive model.

### 1.2.3 Consistency

The matter of consistency among the released marginals might appear trivial; indeed most previous approaches, which produced actual randomized data sets, it is a non-issue, as their tables are produced from these specific data sets. However, there is previous work, namely [20], that assures differential privacy and strong accuracy simply by adding noise to relased cell values. It is unlikely that there exists a single data set that yields all of the released marginals, and this potential inconsistency in the released data can be the source of many technical frustrations.

As we will base our privacy and accuracy around the techniques in [20], we take this section to introduce their results and approaches, while also distinguishing our current work from theirs.

Definition 2. [20]. For $f: \mathcal{D} \rightarrow R^{d}$, the L1-sensitivity of $f$ is

$$
\begin{equation*}
\Delta f=\max _{D_{1}, D_{2}}\left\|f\left(D_{1}\right)-f\left(D_{2}\right)\right\|_{1} \tag{3}
\end{equation*}
$$

for all $D_{1}, D_{2}$ differing in at most one element.
Note that sensitivity is a property of the function alone, and is independent of the database. In the particular case of marginals of contingency tables, which integrate counts over disjoint regions of attribute space, the $L 1$-sensitivity is always two: changing a single participant's data can alter at most two counts, one old and one new.

Our interest in sensitivity is summarized by Theorem 2 below, connecting sensitivity to the amount of noise that suffices to ensure $\epsilon$-differential privacy.

Theorem 2. [20]. For any $f: \mathcal{D} \rightarrow R^{d}$, the addition of Laplace noise ${ }^{3}$ with variance $2 \sigma^{2}$ preserves $(\Delta f / \sigma)$-differential privacy.

Proof. Using the definition of the Laplace density, the density at any $a$ is

$$
\begin{equation*}
\mu[a \mid D] \propto \exp \left(\|f(D)-a\|_{1} / \sigma\right) \tag{4}
\end{equation*}
$$

Applying the triangle inequality, we bound the ratio

$$
\begin{align*}
\frac{\mu\left[a \mid D_{1}\right]}{\mu\left[a \mid D_{2}\right]} & =\frac{\exp \left(\left\|f\left(D_{1}\right)-a\right\|_{1} / \sigma\right)}{\exp \left(\left\|f\left(D_{2}\right)-a\right\|_{1} / \sigma\right)}  \tag{5}\\
& \leq \exp \left(\left\|f\left(D_{1}\right)-f\left(D_{2}\right)\right\|_{1} / \sigma\right) \tag{6}
\end{align*}
$$

The last term is bounded by $\exp (\Delta f / \sigma)$, by the definition of $\Delta f$. Thus (1) holds for singleton sets $S=\{a\}$, and the theorem follows by integrating over $S$.

Remark: To ensure $\epsilon$-differential privacy for a query of sensitivity $\Delta$ we take $\sigma=\Delta / \epsilon$.

This perturbation approach directly leads to a mechanism for releasing approximations to the marginals of the contingency table: Assume the curator wishes to release the set of marginals $C$. One privacy-preserving approach applies Theorem 2 to the $|C|$ marginals (adding noise to each cell in the collection of tables independently), with sensitivity $\Delta f=|C|$. This yields $\epsilon$-differential privacy, which is a very strong guarantee. When $n$ (the number of rows in the database) is large compared to $|C|$ this also yields excellent

[^7]accuracy. Thus we would be done, and there would be no need for the current paper, if the small table-to-table inconsistencies caused by independent randomization of each (cell in each) table are not of concern, and if the user is comfortable with occasionally negative and typically non-integer cell counts.

We have no philosophical or mathematical objection to these artifacts of the privacy-enhancing technology, but in practice they can be problematic. For example, the cell counts may be used as input to other, possibly off-the-shelf, programs that anticipate positive integers, giving rise to type mismatch. We know of one real-life test case in which poor communication with the system prototypers caused users who experienced inconsistencies among OLAP cubes to question the validity of the data.[19] And while such a problem can be solved with better communication and education, it may be difficult to arrange, say, when users are ordinary citizens accessing the United States Census public interface.

### 1.3 Key Steps in Our Solution

## Apply Theorem 2 and Never Look Back.

In this paper we always obtain privacy by applying Theorem 2 to the raw data or a possibly reversible transformation of the raw data. This gives us an intermediate object, on which we operate further, but we never again access the raw data. Since anything obtained via Theorem 2 is privacypreserving, any quantity computed from the intermediate object is still safe: the curator could equally well release the privacy-protective intermediate object and the user can carry out the rest of the compuatations. The results would be the same.

## Move to the Fourier Domain.

When adding noise, two natural approaches present themselves: add noise to entries of the source table and compromise on accuracy, or add noise to the reported marginals and violate consistency. A third approach transforms the data into the Fourier domain, which serves as a non-redundant encoding of the information in the marginals. Adding noise in this domain will not violate consistency, because any set of Fourier coefficients corresponds to a (fractional and possibly negative) contingency table. Morever, as we will show, very few Fourier coefficients are required to compute low-order marginals, and consequently the magnitude of the noise we must add to them is small.

## Use Linear Programming.

We employ linear programming to obtain a non-negative, but likely non-integer, contingency table with (almost) the given Fourier coefficients, and then round the results to obtain integrality. Interestingly, the marginals obtained from the linear program are no "farther" (made precise below) from those of the noisy measurements than are the marginals of the raw data. Consequently, the additional error introduced to impose consistency is no more than the error introduced by the privacy mechanism itself.

Strictly speaking, we don't really need to move to the Fourier domain: we can perturb the marginals directly and then use linear programming to find a positive fractional data set, which can then be rounded as above. The accuracy in this case suffers slightly.

## When $k$ is Large.

The linear program requires time polynomial in $2^{k}$, which is the size of the contingency table (because that is what the linear program is solving for). When $k$ is large this is not satisfactory. However we show, somewhat surprisingly, that non-negativity (but not integrality) can be achieved by adding a relatively small amount to the first Fourier coefficient before moving back to the data domain. No linear program is required, and the error introduced is pleasantly small. Thus if $2^{k}$ is an unbearable cost and one can live with non-integrality then this approach serves well. We note that non-integrality was a non-issue in the prototyped system mentioned above, since answers were anyway converted to percentages.

## 2. NOTATION AND PRELIMINARIES

Our formal treatment of contingency table release begins by casting our data set as a vector $x$ in a high-dimensional space, indexed by attribute tuples. Formally, imagine $k$ binary attributes, and for each $\alpha \in\{0,1\}^{k}$ there is a count $x_{\alpha}$ of the number of data elements with this setting of attributes. We let $n=\|x\|_{1}$ be the total number of tuples in our data set. As it is likely that $x$ will be sparse, with $n \ll 2^{k}$, we will be mindful of $n$ and $2^{k}$ independently.

For any $\alpha \in\{0,1\}^{k}$, we use $\|\alpha\|_{1}$ for the weight of $\alpha$, the number of non-zero locations. We write $\alpha \preceq \beta$ for $\alpha, \beta \in$ $\{0,1\}^{k}$ if every non-zero location in $\alpha$ is also non-zero in $\beta$.

### 2.1 The Marginal Operator

Central to our discussion are the operators $C^{\alpha}: \mathbb{R}^{2^{k}} \rightarrow$ $\mathbb{R}^{2\|\alpha\|_{1}}$ for $\alpha \in\{0,1\}^{k}$ mapping contingency tables to the marginals of the attributes that are positively set in $\alpha$. For any $\beta \preceq \alpha$, the outcome of $C^{\alpha} x$ at position $\beta$ is the sum over those coordinates of $x$ that agree with $\beta$ on the coordinates described by $\alpha$ :

$$
\begin{equation*}
\left(C^{\alpha}(x)\right)_{\beta}=\sum_{\gamma: \gamma \wedge \alpha=\beta} x_{\gamma} \tag{7}
\end{equation*}
$$

Notice that we are abusing notation, and only defining $C^{\alpha} x$ at those locations $\beta$ for which $\beta \preceq \alpha$.

Theorem 3. The operator $C^{\alpha}$ is linear for all $\alpha$.
Proof. As each output coordinate of $C_{i}$ is a sum over predetermined input coordinates, scaling and addition of its inputs translate to equivalent scaling and addition of outputs.

It is common to consider the ensemble of marginals $C^{\alpha}$ for all $\alpha$ with a fixed value of $\|\alpha\|_{1}=i$, referred to as the $i$-way marginals.

### 2.2 The Fourier Basis

We will find it helpful to view our contingency table $x$ in an alternate basis; rather than a value for each position $\alpha$, we will project onto a set of $2^{k}$ so-called Fourier basis vectors that each aggregate across the table in various ways. Our motivation lies in the observation, made formally soon, that while a low-order marginal needs access to all coordinates of the contingency table, it will need only a few of the new coordinates in the Fourier basis.

The Fourier basis for real vectors defined over the Boolean hypercube is the set of vectors $f^{\alpha}$ for each $\alpha \in\{0,1\}^{k}$,
defined coordinate-wise as

$$
\begin{equation*}
f_{\beta}^{\alpha}=(-1)^{\langle\alpha, \beta\rangle} / 2^{k / 2} \tag{8}
\end{equation*}
$$

That is, each Fourier basis vector is comprised of coordinates of the form $\pm 1 / 2^{k / 2}$, with the sign alternating based on the parity of the intersection between $\alpha$ and $\beta$. In fact, it will occosionally be helpful to view the vectors $f^{\alpha}$ as contingency tables themselves, as we will want to apply the marginal operators $C^{\beta}$ to them.

## ThEOREM 4. The $f^{\alpha}$ form an orthonormal basis for $\mathbb{R}^{2^{k}}$.

Proof. This is a standard result from the theory of Fourier analysis. See for example, [28]

A change of basis allows us to rewrite a vector $x$ as a sum of its projections onto the basis vectors, each of which is referred to as a Fourier coefficient. For our purposes, we will want to rewrite $x$ in this basis just before it is supplied as input to a marginal computation $C^{\beta}$, which by linearity is

$$
\begin{equation*}
C^{\beta} x=C^{\beta} \sum_{\alpha}\left\langle f^{\alpha}, x\right\rangle f^{\alpha}=\sum_{\alpha}\left\langle f^{\alpha}, x\right\rangle C^{\beta} f^{\alpha} \tag{9}
\end{equation*}
$$

As promised, the motivation for this transformation comes from the following theorem, that any marginal over few attributes requires only a few Fourier coefficients.

Theorem 5. $C^{\beta} f^{\alpha} \neq 0$ if and only if $\alpha \preceq \beta$.
Proof. For any coordinate $\gamma \preceq \beta$ of the output

$$
\begin{equation*}
\left(C^{\beta}\left(f^{\alpha}\right)\right)_{\gamma}=\sum_{\eta: \eta \wedge \beta=\gamma} f_{\eta}^{\alpha}=\sum_{\eta: \eta \wedge \beta=\gamma}(-1)^{\langle\alpha, \eta\rangle} / 2^{k / 2} \tag{10}
\end{equation*}
$$

If $\alpha \npreceq \beta$, then there is a coordinate for which $\alpha$ is one and $\beta$ zero. For every $\eta$ in the sum above, the same string with this bit flipped is also in the sum, as $\eta \wedge \beta$ is ignorant of this bit. However, their coordinates in $f^{\alpha}$ have opposite sign, and their contributions to the sum cancel exactly. This holds for all $\eta$, making the total sum zero.

If $\alpha \preceq \beta$, then $\left(C^{\beta}\left(f^{\alpha}\right)\right)_{\alpha}$ is non-zero, as the sum is taken over $\eta$ with $\eta \wedge \beta=\alpha$, causing $\langle\eta, \alpha\rangle=\langle\alpha, \alpha\rangle$. Thus all terms contributing to the summation are positive.

Consequently, we are able to write any marginal as the small summation over relevant Fourier coefficients:

$$
\begin{equation*}
C^{\beta} x=\sum_{\alpha \preceq \beta}\left\langle f^{\alpha}, x\right\rangle C^{\beta} f^{\alpha} \tag{11}
\end{equation*}
$$

The coefficients $\left\langle f^{\alpha}, x\right\rangle$ are necessary and sufficient data from $x$ for the computation of $C^{\beta} x$.

## 3. ALGORITHMS AND THEOREMS

We now delve into the details of our algorithm, which comes in two parts. We first show how to create consistent marginals by applying a privacy-preserving mechanism to the Fourier coefficients rather than directly to the marginals. The resulting Fourier coefficients may correspond to a contingency table whose entries are negative and fractional, and we then give a linear program which, after rounding, returns a positive integral contingency table, from which we compute marginals.

### 3.1 Consistency

Rather than perturb the marginals, a naive, but effective, manner of ensuring privacy and consistency is to simply perturb and release each coordinate of the contingency table. As low-order marginals are sums over many entries in the contingency table, their entries will have noise that is Binomially distributed with variance $\Theta\left(2^{k}\right)$.

Instead, we will isolate and perturb those features of the data set relevant to the marginal computation, the Fourier coefficients. Because we are taking substantially fewer measurements, as compared with $2^{k}$ above, we can add substantially less noise to each measurement. For example, we need only $2^{i}$ coefficients for a $i$-way marginal, and only $\sum_{j \leq i}\binom{k}{j}$ coefficients for the full set of $i$-way marginals. While these numbers may seem large, recall that a $i$-way marginal releases $2^{i}$ counts, making this the natural scale.

We use the privacy mechanism of [20], based on the addition of additive noise, to ensure $\epsilon$-differential privacy. Formally, we let $\operatorname{Lap}(\sigma)$ be a random variable with density at $y$ proportional to $\exp (-|y| / \sigma)$. The following theorem describes the amount of noise we must add to each Fourier coefficient, as a function of the number of coefficients we require.

Theorem 6. Let $A \subseteq\{0,1\}^{k}$ describe a set of Fourier basis vectors, and let $x$ be the contingency table that results from a data set $D$. Releasing the set $\phi_{\alpha}=\left\langle f^{\alpha}, x\right\rangle+$ $\operatorname{Lap}\left(2|A| / \epsilon 2^{k / 2}\right)$ for $\alpha \in A$ preserves $\epsilon$-differential privacy of $D$.

Proof. Each tuple of the data set $D$ contributes exactly $\pm 1 / 2^{k / 2}$ to each output coordinate, and consequently the $L 1$ sensitivity of the set of $|A|$ outputs is at most $2|A| / 2^{k / 2}$. By Theorem 2, the addition of Laplace noise with parameter $2|A| / \epsilon 2^{k / 2}$ gives $\epsilon$-differential privacy.

Remark: Note that $n=|D|$ does not appear in Theorem 6. To get a sense of scale for the error, we could achieve a similar perturbation to each coordinate by randomly relocating $4|A|^{2} / \epsilon$ individuals in the data set, which can be much smaller than $n$.

### 3.2 Non-Negative Integrality

While there is certainly a real valued contingency table whose Fourier coefficients equal the perturbed values, e.g.: by returning the perturbed values to the original space, it is unlikely that there is a non-negative, integral contingency table with these coefficients. We now use linear programming to find a non-negative, but likely fractional, contingency table with nearly the correct Fourier coefficients, which we round to an integral contingency table with little additional error.

Letting $B \subset\{0,1\}^{k}$, suppose that we observed Fourier coefficients $\phi_{\beta}$ for $\beta \in B$. The following linear program minimizes, over all contingency tables $w$, the largest error $b$ error between its Fourier coefficients $\left\langle f^{\beta}, w\right\rangle$ and the ob-
served $\phi_{\beta}$ :

$$
\begin{aligned}
\text { minimize } & b \\
\text { subject to: } & \\
w_{\alpha} & \geq 0 \quad \forall \alpha \\
\phi_{\beta}-\sum_{\alpha} w_{\alpha} f_{\alpha}^{\beta} & \leq b \quad \forall \beta \in B \\
\phi_{\beta}-\sum_{\alpha} w_{\alpha} f_{\alpha}^{\beta} & \geq-b \quad \forall \beta \in B
\end{aligned}
$$

This optimization occurs in a $2^{k}+1$ dimensional space, and any vertex of the feasible polytope must intersect $2^{k}+1$ constraints. At most $|B|$ of these can relate to Fourier coefficients (since for each $\beta$, only one of the two constraints corresponding to $\beta$ can be satisfied by any point). Thus at least $2^{k}-|B|+1$ must be non-negativity constraints. This means that at any vertex of the polytope, all but at most $|B|$ weights are zero. Without loss of generality, the linear program will return a vertex solution[25], and rounding to the nearest integral point will result in at most an $L 1$ change of $|B|$.

### 3.3 Algorithmic Recap

To bring things together, we now collect the various steps we have taken into a single algorithm.

## $\operatorname{Marginals}\left(A \subseteq\{0,1\}^{k}, x\right)$ :

1. Let $B$ be the downward closure of $A$ under $\preceq$.
2. For $\beta \in B$, compute $\phi_{\beta}=\left\langle f^{\beta}, x\right\rangle+\operatorname{Lap}\left(2|B| / \epsilon 2^{k / 2}\right)$.
3. Solve for $w_{\alpha}$ in the following linear program, and round to the nearest integral weights, $w_{\alpha}^{\prime}$.

$$
\begin{aligned}
\text { minimize } & b \\
\text { subject to: } & \\
w_{\alpha} & \geq 0 \quad \forall \alpha \\
\phi_{\beta}-\sum_{\alpha} w_{\alpha} f_{\alpha}^{\beta} & \leq b \quad \forall \beta \in B \\
\phi_{\beta}-\sum_{\alpha} w_{\alpha} f_{\alpha}^{\beta} & \geq-b \quad \forall \beta \in B
\end{aligned}
$$

4. Using the contingency table $w_{\alpha}^{\prime}$, compute and return the marginals for $A$.
Theorem 7. Using the notation of Marginals(A), for all $\delta \in[0,1]$ with probability $1-\delta$, for all $\alpha \in A$,

$$
\begin{equation*}
\left\|C^{\alpha} x-C^{\alpha} w^{\prime}\right\|_{1} \leq 2^{\|\alpha\|_{1}} 8|B| \log (|B| / \delta) / \epsilon+|B| \tag{12}
\end{equation*}
$$

Proof. Each Fourier coefficient has Laplace noise with parameter $2|B| / \epsilon 2^{k / 2}$ added to it, and with probability $1-\delta$ none of these exceeds $4|B| \log (|B| / \delta) / \epsilon 2^{k / 2}$. In solving the linear program, the error associated with each Fourier coefficient is at most this bound as well, as the original contingency table $x$ is at least as close. Mapping the perturbation of a single Fourier coefficient back to the contingency table domain increases the L 1 norm of the perturbation by at most $2^{k / 2}$, up to at most $8|B| \log (|B| / \delta) / \epsilon$.

Consequently, for any marginal $C^{\alpha}$, the error $C^{\alpha} x-C^{\alpha} w^{\prime}$ is a result of noise in the $2^{\|\alpha\|_{1}}$ Fourier coefficients that contribute to the table, as well as the rounding that occurs. Multiplying the number of coefficients, $2^{\|\alpha\|_{1}}$ by the bound above, and adding the $|B|$ error due to rounding, gives the stated bound.

Even tighter bounds can be placed on sub-marginals of a marginal $C^{\alpha}$, by noting that the bounds hold for the marginals $C^{\beta}$ for $\beta \preceq \alpha$ at no additional cost. No more Fourier coefficients are used, so $|B|$ is not increased, but $\|\beta\|_{1} \leq\|\alpha\|_{1}$.

## 4. ALTERNATE APPROACHES

We now describe a few variants on the previous approaches that trade some of the accuracy of the previous approach for some conceptual or computational simplicity.

### 4.1 Alternate Linear Programs

The linear program we chose to use minimizes the largest error in any Fourier coefficient. There are other linear programs that one could write, for example minimizing the total error in Fourier coefficients, the largest error in reported marginals, the total error in the reported marginals, or several hybrids thereof.

This flexibility allows the data analyst with more specific accuracy concerns (eg: per cell accuracy) to address them. The perturbed Fourier coefficients can be released, and the specific linear program can be run to arrive at an integral, non-negative solution. Bounds similar to Theorem 7 can be proven, using the same methodology: the noise added perturbs the measurements by some distance in the norm of choice, and the linear program finds a non-negative solution at no greater distance from the perturbed measurements.

### 4.2 Non-Fourier Linear Programming

Our conversion to the Fourier domain is done because the Fourier coefficients exactly describe the information required by the marinals. By measuring exactly what we need, we add the least amount of noise possible using the techniques of [20].

Instead, we could apply the techniques of [20] directly to the true marginals, producing a set of noisy marginals that preserve privacy but not consistency. To these noisy marginals we apply a linear program to find a non-negative contingency table with nearest marginals. Imagining we have observed the noisy marginals $c^{\beta}$, the linear program is

$$
\text { minimize } \quad b
$$

subject to:

$$
\begin{aligned}
w_{\alpha} & \geq 0 \quad \forall \alpha \in\{0,1\}^{k} \\
\left(c^{\beta}-C^{\beta} w\right)_{\gamma} & \leq b \quad \forall \beta \in A, \gamma \preceq \beta \\
\left(c^{\beta}-C^{\beta} w\right)_{\gamma} & \geq-b \quad \forall \beta \in A, \gamma \preceq \beta
\end{aligned}
$$

As before, we are likely to discover a fractional contingency table $w$. However, the number of cell constraints is at most $2 \sum_{\beta \in A} 2^{\|\beta\|_{1}}$, and at most $\sum_{\beta \in A} 2^{\|\beta\|_{1}}$ of the $w_{\alpha}$ variables are non-zero. By the reasoning above, any rounding to integers introduces error at most this much in the contingency table.

Theorem 8. Using the above approach, with probability at least $1-\delta$, for all $\beta \in A$, then $\left\|C^{\beta} x-C^{\beta} w^{\prime}\right\|_{1}$ is at most

$$
\begin{equation*}
2^{\|\beta\|_{1}} 8|A| \log \left(\sum_{\beta \in A} 2^{\|\beta\|_{1}} / \delta\right) / \epsilon+\sum_{\beta \in A} 2^{\|\beta\|_{1}} . \tag{13}
\end{equation*}
$$

Proof. The reasoning is the same as before: the difference in the marginals is no more than twice the difference caused by the additive noise, which is a Laplacian with parameter $2|A| / \epsilon$. We introduce the $\log \left(\sum_{\beta \in A} 2^{\|\beta\|_{1}} / \delta\right)$ term
to give the high probability guarantee, and the additive term to account for rounding error.
Remark: This theorem mirrors Theorem 7, using $|A|$ and $\sum_{\beta \in A} 2^{\|\beta\|_{1}}$ in place of $|B|$. Depending on the situation, these bounds can actually be tighter than in Theorem 7, though only when a single multi-attribute marginal is desired. The tighter bounds given by Theorem 7 through subtables also would not apply here.

### 4.3 Simple Non-Negativity

The solution of the linear programs we have described is an expensive process, taking time polynomial in $2^{k}$. In many settings, but not all, this is an excessive amount that must be avoided. We now describe a very simple technique for arriving at Fourier coefficients corresponding to a non-negative, but fractional, contingency table with high probability, without the solution of a linear program. We construct the output marginals directly from the Fourier coefficients, rather than reconstructing the contingency table, which could take time $2^{k}$.

To ensure the existence of a non-negative contingency table with the observed Fourier coefficients turns out to be a simple task, we simply add a small amount to the first Fourier coefficient. Intuitively, any negativity due to the small perturbation we have made to the Fourier coefficients is spread uniformly across all elements of the contingency table. Consequently, very little needs to be added to make the elements non-negative.

Theorem 9. Let $B \subset\{0,1\}^{k}$, and let $x$ be a non-negative contingency table with Fourier coefficients $\phi_{\beta}$ for $\beta \in B$. If the Fourier coefficients are perturbed to $\phi_{\beta}^{\prime}$, then the contingency table

$$
\begin{equation*}
x^{\prime}=x+\sum_{\beta}\left(\phi_{\beta}^{\prime}-\phi_{\beta}\right) f^{\beta}+\left\|\phi^{\prime}-\phi\right\|_{1} f^{\overline{0}} \tag{14}
\end{equation*}
$$

is non-negative, and has $\left\langle f^{\beta}, x^{\prime}\right\rangle=\phi_{\beta}^{\prime}$ for $\beta \neq \overline{0}$.
Proof. Each of the coordinates of $f^{\beta}$ are $\pm 1 / 2^{k / 2}$, and the most negative an entry could become due to the perturbation is $-\left\|\phi^{\prime}-\phi\right\|_{1} / 2^{k / 2}$. By increasing the Fourier coefficient of the zero vector by $\left\|\phi^{\prime}-\phi\right\|_{1}$, we increase every entry in the contingency table by this much, making them all non-negative.

Our perturbation to the Fourier coefficients has $L 1$ norm distributed exponentially with standard deviation $2^{3 / 2}|B|^{2} / \epsilon$. It is critical that we not disclose the actual $L 1$ norm of the perturbation, but we can add a value for which the negativity probability is arbitrarily low:

Corollary 1. By adding $t \times 4|B|^{2} / \epsilon 2^{k / 2}$ to the first Fourier coefficient, the resulting contingency table is non-negative with probability at least $1-\exp (-t)$.
The addition of $4 t|B|^{2} / \epsilon 2^{k / 2}$ to the first Fourier coefficient corresponds to the introduction of $4 t|B|^{2} / \epsilon$ individuals at random locations in the table; a relatively minor accuracy compromise.

## 5. CONCLUSIONS

We have shown a holistic solution to the problem of contingency table release, that outputs an accurate and consistent set of tables while guaranteeing in a very strong sense
that privacy of individuals is preserved. We also show how to construct a positive and integral synthetic database that corresponds to these tables-thus, e.g., one can output a synthetic database that preserves all low-order marginals up to a small error. Moreover, we can get a gracefully degrading version of the results: we can compute a synthetic database such that the error in the low-order marginals is small, and increases smoothly with the order of the marginal.

One of the main algorithmic questions left open from this work is that of efficiency. In particular, solving the linear program could be a bottleneck when the number of attributes is large, and it seems possible that one could devise more efficient algorithms for this step. We remark that the simplex algorithm is already space efficient in this setting, since each vertex of the polytope that simplex traverses has a sparse description. We leave open the question of devising faster combinatorial algorithms for this problem.

We have optimized for a specific measure of data quality, the distance between the reported and true marginals. It would also be useful to analyze the effect of our techniques or variants thereof on statistical properties of the marginals, such as means, variances, covariances, regressions. See the related work on controlled tabular adjustment [11].

## 6. ACKNOWLEDGEMENTS

We would like to thank Larry Cox and Steve Fienberg for comments on an earlier version of this paper and general guidance though the statistics literature.

## 7. REFERENCES

[1] Special Issue on Statistical Disclosure Control, volume 14(4) of Journal of Official Statistics. 1998.
[2] D. Agrawal and C. C. Aggarwal. On the design and quantification of privacy preserving data mining algorithms. In PODS. ACM, 2001.
[3] R. Agrawal and R. Srikant. Privacy-preserving data mining. In W. Chen, J. F. Naughton, and P. A. Bernstein, editors, SIGMOD Conference, pages 439-450. ACM, 2000.
[4] R. Agrawal, R. Srikant, and D. Thomas. Privacy preserving OLAP. In SIGMOD '05: Proceedings of the 2005 ACM SIGMOD international conference on Management of data, pages 251-262, New York, NY, USA, 2005. ACM Press.
[5] M. Bacharach. Matrix rounding problems. Management Science, 9:732-742, 1966.
[6] A. Blum, C. Dwork, F. McSherry, and K. Nissim. Practical privacy: the SuLQ framework. In C. Li, editor, PODS, pages 128-138. ACM, 2005.
[7] J. Castro. Quadratic interior-point methods in statistical disclosure control. Computational Management Science, 2:pages 107-121, 2005.
[8] J. Castro. Minimum-distance controlled perturbation methods for large-scale tabular data protection. Euorpean Journal of Operantional Research, 171:39-52, 2006.
[9] S. Chawla, C. Dwork, F. McSherry, A. Smith, and H. Wee. Toward privacy in public databases. In J. Kilian, editor, TCC, volume 3378 of Lecture Notes in Computer Science, pages 363-385. Springer, 2005.
[10] S. Chawla, C. Dwork, F. McSherry, and K. Talwar. On privacy-preserving histograms. In Proceedings of
the 21th Annual Conference on Uncertainty in Artificial Intelligence (UAI-05), Arlington, Virginia, 2005. AUAI Press.
[11] L. Cox, J. Kelly, and R. Patil. Balancing quality and confidentiality in multivariate tabular data. Privacy in Statistical Databases, 3080:87-98, 2004.
[12] T. Dalenius. Towards a methodology for statistical disclosure control. Statistisk. tidskrift, 3:213-225, 1977.
[13] R. A. Dandekar and L. Cox. Synthetic tabular data: An alternative to complementary cell suppression, 2002. manuscript, energy Information Administration,
[14] I. Dinur and K. Nissim. Revealing information while preserving privacy. In Milo [26], pages 202-210.
[15] A. Dobra and S. Fienberg. Bounding entries in multi-way contingency tables given a set of marginal totals, 2002. Proceedings of Conference on Foundation of Statistical Inference and its Applicaitons.
[16] J. Domingo-Ferrer and V. Torra. A critique of the sensitivity rules usually employed for statistical table protection. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(5):545-556, 2002.
[17] G. Duncan. Confidentiality and statistical disclosure limitation. In N. Smelser and P. Baltes, editors, International Encyclopedia of the Social and Behavioral Sciences. Elsevier, 2001.
[18] C. Dwork. Differential privacy. In M. Bugliesi, B. Preneel, V. Sassone, and I. Wegener, editors, ICALP (2), volume 4052 of Lecture Notes in Computer Science, pages 1-12. Springer, 2006.
[19] C. Dwork, D. Lee, and F. McSherry. Privacy preserving histogram case study, 2007. Manuscript.
[20] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In S. Halevi and T. Rabin, editors, TCC, volume 3876 of Lecture Notes in Computer Science, pages 265-284. Springer, 2006.
[21] C. Dwork, F. McSherry, and K. Talwar. The price of privacy and the limits of LP decoding. In Proceedings of the 39th annual Symposium on the Theory of Computation., 2007.
[22] C. Dwork and K. Nissim. Privacy-preserving datamining on vertically partitioned databases. In M. K. Franklin, editor, CRYPTO, volume 3152 of Lecture Notes in Computer Science, pages 528-544. Springer, 2004.
[23] A. V. Evfimievski, J. Gehrke, and R. Srikant. Limiting privacy breaches in privacy preserving data mining. In Milo [26], pages 211-222.
[24] I. Fellegi. On the question of statistical confidentiality. Journal of the American Statistical Association, pages 7-18, 1972.
[25] M. Grötschel, L. Lovász, and A. Schrijver. Geometric Algorithms and Combinatorial Optimization (Algorithms and Combinatorics). Springer, December 1994.
[26] T. Milo, editor. Proceedings of the Twenty-Second ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, June 9-12, 2003, San Diego, CA, USA. ACM, 2003.
[27] J. Kelly, A. Assad, and B. Golden. The controlled rounding problem: Relaxations and complexity issues. OR Spektrum, 12:129-138, 1990.
[28] T. W. Körner. Fourier Analysis. Cambridge University Press, Cambridge, UK, 1988.
[29] J. A. D. Loera and S. Onn. All rational polytopes are transportation polytopes and all polytopal integer sets are contingency tables. Proceedings of the 10th Ann. Math. Prog. Soc. Symp. Integ. Prog. Combin. Optim., LNCS, 3064:338-351, 2004.
[30] J. A. D. Loera and S. Onn. The complexity of three-way statistical tables. SIAM J. Comput., 33:819-836, 2004.
[31] J. A. D. Loera and S. Onn. Markov bases of three-way tables are arbitrarily complicated. J. Symb. Comput., 41(2):173-181, 2006.
[32] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. In L. Liu, A. Reuter, K.-Y. Whang, and J. Zhang, editors, $I C D E$, page 24. IEEE Computer Society, 2006.
[33] D. A. Robertson and R. Ethier. Cell suppression: Experience and theory. In J. Domingo-Ferrer, editor, Inference Control in Statistical Databases, volume 2316 of Lecture Notes in Computer Science, pages 8-20. Springer, 2002.
[34] D. Rubin. Discussion: Statistical disclosure limitation. Journal of Official Statistics, 9:461-469, 1993.
[35] P. Samarati and L. Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression.
[36] L. Sweeney. Achieving k-anonymity privacy protection using generalization and suppression. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(5):571-588, 2002.
[37] L. Sweeney. k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(5):557-570, 2002.

## COMPARATIVE

 STATISTICAL EVALUATION OF

## FROM 8-D COUNTS

DATA
(AGE, EDUCATION, ANNUAL SALARY)

## ACTUAL DATA VS <br> CTA PROTECTED DATA <br> USING IMSL ROUTINES

REAL LIFE 8-D DATA USED

```
3-D SECTION X ( AGE, EDUCATION, ANNUAL SALARY )
section # 46
```

A: Age - three categories
B: Education - five categories
C: Annual Salary - two categories

## ACTUAL DATA

Table 1: $C=1$
A (row) by B (column)

|  | 1 | 2 | 3 | 4 | 5 |
| ---: | ---: | ---: | ---: | ---: | ---: |
| 1 | 1738.0 | 2444.0 | 1088.0 | 33.0 | 3036.0 |
| 2 | 3051.0 | 8930.0 | 5676.0 | 1245.0 | 5010.0 |
| 3 | 1253.0 | 1907.0 | 675.0 | 300.0 | 769.0 |

## CTA PROTECTED DATA

Table 1: C = 1
A (row) by B (column)

|  | 1 | 2 | 3 | 4 | 5 |
| ---: | ---: | ---: | ---: | ---: | ---: |
| 1 | 1745.0 | 2453.0 | 1082.0 | 31.0 | 3034.0 |
| 2 | 3048.0 | 8933.0 | 5681.0 | 1240.0 | 5012.0 |
| 3 | 1255.0 | 1901.0 | 680.0 | 298.0 | 763.0 |

Table 2: C = 2
A (row) by B (column)

| 2 | 3 | 4 | 5 |
| ---: | ---: | ---: | ---: |
| 17.0 | 31.0 | 6.0 | 29.0 |
| 2016.0 | 3653.0 | 2043.0 | 1691.0 |
| 467.0 | 562.0 | 471.0 | 343.0 |

REAL LIFE NISS 8-D DATA USED

3-D SECTION X ( AGE, EDUCATION, ANNUAL SALARY )

## ACTUAL DATA

| Variable | Number of Levels |
| :--- | :---: |
| 1 A | 3 |
| 2 B | 5 |
| 3 C | 2 |


|  | A (row) by B (column) |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: |
|  | 1 | 2 | 3 | 4 | 5 |
| 1 | 1738.0 | 2444.0 | 1088.0 | 33.0 | 3036.0 |
| 2 | 3051.0 | 8930.0 | 5676.0 | 1245.0 | 5010.0 |
| 3 | 1253.0 | 1907.0 | 675.0 | 300.0 | 769.0 |

Table 2: C = 2
A (row) by B (column)

|  | 1 | 2 | 3 | 4 | 5 |
| ---: | ---: | ---: | ---: | ---: | ---: |
| 1 | 11.0 | 20.0 | 28.0 | 4.0 | 30.0 |
| 2 | 221.0 | 2020.0 | 3657.0 | 2037.0 | 1694.0 |
| 3 | 134.0 | 463.0 | 563.0 | 466.0 | 339.0 |

Partial Association Statistics

| Omitted | Chi-Square | Degrees of <br> Freedom | P-value | Marginal |
| :--- | ---: | ---: | ---: | ---: |
| Effect | 25535.30 | 2.0 | .0000 | Zeros |
| A | 9153.81 | 4.0 | .0000 | .0 |
| B | 13958.71 | 1.0 | .0000 | .0 |
| C | 2584.77 | 8.0 | .0000 | .0 |
| A*B | 3089.65 | 2.0 | .0000 | .0 |
| A*C | 4606.03 | 4.0 | .0000 | .0 |
| B*C | 26.51 | 8.0 | .0009 | .0 |

Chi-square statistics for testing that all $\mathbf{k}$ and higher interactions are zero.
Likelihood Degrees of

| k | Ratio | P-Value | Freedom | Pearson | P-Value |
| :--- | ---: | ---: | ---: | ---: | ---: |
| 1 | 61349.58 | .0000 | 29.0 | 73429.19 | .0000 |
| 2 | 12701.76 | .0000 | 22.0 | 12810.05 | .0000 |
| 3 | 26.51 | .0009 | 8.0 | 28.43 | .0004 |


| hi-square statistics for testing that all k-factor interactions are simultaneously zero. |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: |
|  | Likelihood |  | ees of |  |  |
| k | Ratio | P-Value | Freedom | Pearson | P-Value |
| 1 | 48647.82 | . 0000 | 7.0 | 60619.14 | . 0000 |
| 2 | 12675.24 | . 0000 | 14.0 | 12781.62 | . 0000 |
| 3 | 26.51 | . 0009 | 8.0 | 28.43 | 0004 |

REAL LIFE NISS 8-D DATA USED

3-D SECTION X ( AGE, EDUCATION, ANNUAL SALARY )

## CTA PROTECTED DATA

| Variable | Number of Levels |
| :--- | :---: |
| 1 A | 3 |
| 2 B | 5 |
| 3 C | 2 |

Table 1: $\mathrm{C}=1$
A (row) by B (column)

|  | 1 | 2 | 3 | 4 | 5 |
| ---: | ---: | ---: | ---: | ---: | ---: |
| 1 | 1745.0 | 2453.0 | 1082.0 | 31.0 | 3034.0 |
| 2 | 3048.0 | 8933.0 | 5681.0 | 1240.0 | 5012.0 |
| 3 | 1255.0 | 1901.0 | 680.0 | 298.0 | 763.0 |


|  | Table 2: C = 2 (row) by B (column) |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: |
|  | 1 | 2 | 3 | 4 | 5 |
| 1 | 6.0 | 17.0 | 31.0 | 6.0 | 29.0 |
| 2 | 226.0 | 2016.0 | 3653.0 | 2043.0 | 1691.0 |
| 3 | 127.0 | 467.0 | 562.0 | 471.0 | 343.0 |

Partial Association Statistics

| Omitted | Chi-Square | Degrees of <br> Freedom | P-value | Marginal <br> Zeros |
| :--- | ---: | ---: | ---: | ---: |
| Effect | 25539.01 | 2.0 | .0000 | .0 |
| A | 9150.00 | 4.0 | .0000 | .0 |
| C | 13958.07 | 1.0 | .0000 | .0 |
| A*B | 2572.01 | 8.0 | .0000 | .0 |
| A*C | 3110.17 | 2.0 | .0000 | .0 |
| B*C | 4655.57 | 4.0 | .0000 | .0 |
| A*B*C | 24.14 | 8.0 | .0022 | .0 |



## ITERATIVE PROPORTIONAL FIT USING IMSL ROUTINE "PRPFT"

## REAL LIFE NISS 8-D DATA USED

3-D SECTION X ( AGE, EDUCATION, ANNUAL SALARY )

| i | j | k | CTA value | IPF fit | Actual Value | IPF fit | CTA\% Diff | Fitted \% Diff |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 1 | 1 | 1 | 1745.0 | 1746.1 | 1738.0 | 1743.8 | . 40 | . 13 |
| 2 | 1 | 1 | 3048.0 | 3036.8 | 3051.0 | 3030.4 | -. 10 | . 21 |
| 3 | 1 | 1 | 1255.0 | 1265.1 | 1253.0 | 1267.8 | . 16 | -. 21 |
| 1 | 2 | 1 | 2453.0 | 2450.5 | 2444.0 | 2443.6 | . 37 | . 28 |
| 2 | 2 | 1 | 8933.0 | 8960.8 | 8930.0 | 8959.3 | . 03 | . 02 |
| 3 | 2 | 1 | 1901.0 | 1875.7 | 1907.0 | 1878.0 | -. 31 | -. 12 |
| 1 | 3 | 1 | 1082.0 | 1087.6 | 1088.0 | 1089.4 | -. 55 | -. 17 |
| 2 | 3 | 1 | 5681.0 | 5653.7 | 5676.0 | 5649.7 | . 09 | . 07 |
| 3 | 3 | 1 | 680.0 | 701.7 | 675.0 | 700.0 | . 74 | . 25 |
| 1 | 4 | 1 | 31.0 | 35.0 | 33.0 | 34.9 | -6.06 | . 18 |
| 2 | 4 | 1 | 1240.0 | 1267.1 | 1245.0 | 1274.9 | -. 40 | -. 61 |
| 3 | 4 | 1 | 298.0 | 266.9 | 300.0 | 268.2 | -. 67 | -. 50 |
| 1 | 5 | 1 | 3034.0 | 3025.9 | 3036.0 | 3027.3 | -. 07 | -. 05 |
| 2 | 5 | 1 | 5012.0 | 4995.6 | 5010.0 | 4997.7 | . 04 | -. 04 |
| 3 | 5 | 1 | 763.0 | 787.6 | 769.0 | 790.0 | -. 78 | -. 31 |
| 1 | 1 | 2 | 6.0 | 4.9 | 11.0 | 5.2 | -45.45 | -6.01 |
| 2 | 1 | 2 | 226.0 | 237.2 | 221.0 | 241.6 | 2.26 | -1.82 |
| 3 | 1 | 2 | 127.0 | 116.9 | 134.0 | 119.2 | -5.22 | -1.92 |
| 1 | 2 | 2 | 17.0 | 19.5 | 20.0 | 20.4 | -15.00 | -4.06 |
| 2 | 2 | 2 | 2016.0 | 1988.2 | 2020.0 | 1990.7 | -. 20 | -. 12 |
| 3 | 2 | 2 | 467.0 | 492.3 | 463.0 | 492.0 | . 86 | . 06 |
| 1 | 3 | 2 | 31.0 | 25.4 | 28.0 | 26.6 | 10.71 | -4.50 |
| 2 | 3 | 2 | 3653.0 | 3680.3 | 3657.0 | 3683.3 | -. 11 | -. 08 |
| 3 | 3 | 2 | 562.0 | 540.3 | 563.0 | 538.0 | -. 18 | . 42 |
| 1 | 4 | 2 | 6.0 | 2.0 | 4.0 | 2.1 | 50.00 | -3.02 |
| 2 | 4 | 2 | 2043.0 | 2015.9 | 2037.0 | 2007.1 | . 29 | . 43 |
| 3 | 4 | 2 | 471.0 | 502.1 | 466.0 | 497.8 | 1.07 | . 87 |
| 1 | 5 | 2 | 29.0 | 37.1 | 30.0 | 38.7 | -3.33 | -4.13 |
| 2 | 5 | 2 | 1691.0 | 1707.4 | 1694.0 | 1706.3 | -. 18 | . 07 |
| 3 | 5 | 2 | 343.0 | 318.4 | 339.0 | 318.0 | 1.18 | . 13 |

FITTED MODEL: (A*B, A*C, B*C)

REAL LIFE NISS 8-D DATA USED

3-D SECTION X ( AGE, EDUCATION, ANNUAL SALARY )

## ACTUAL DATA

Fitted Model: (A*B, A*C, B*C)

| Variable | Number of Levels |
| :--- | :---: |
| 1 A | 3 |
| 2 B | 5 |
| 3 C | 2 |


| Model Statistics |  |
| :--- | ---: |
| Log-likelihood | 8.0229 |
| Likelihood ratio | 26.51 |
| Degrees of freedom | 8.0 |
| P-value | .0009 |


|  |  | Coeff | Statistic Standard Error | Asymptotic Z-statistic | P-value |
| :---: | :---: | :---: | :---: | :---: | :---: |
| 1 | intercept | 6.2219 | . 0213 | 292.76 | . 0000 |
| 2 | A (1) | -1.6065 | . 0409 | -39.29 | . 0000 |
| 3 | A (2) | 1.5546 | . 0213 | 73.11 | . 0000 |
| 4 | B (1) | -. 4645 | . 0253 | -18.34 | . 0000 |
| 5 | B (2) | . 6527 | . 0169 | 38.70 | . 0000 |
| 6 | B (3) | . 4390 | . 0177 | 24.78 | . 0000 |
| 7 | B (4) | -1.0828 | . 0458 | -23.64 | . 0000 |
| 8 | C | 1.0280 | . 0187 | 54.91 | . 0000 |
| 9 | A*B (1) | . 4065 | . 0280 | 14.53 | . 0000 |
| 10 | A*B (2) | . 1391 | . 0265 | 5.26 | . 0008 |
| 11 | A*B (3) | . 0832 | . 0295 | 2.82 | . 0223 |
| 12 | A*B (4) | -1.3939 | . 0893 | -15.62 | . 0000 |
| 13 | A*B (5) | -. 5601 | . 0186 | -30.19 | . 0000 |
| 14 | A*B (6) | -. 0809 | . 0158 | -5.12 | . 0009 |
| 15 | A*B(7) | . 2100 | . 0177 | 11.86 | . 0000 |
| 16 | A*B(8) | . 6838 | . 0466 | 14.67 | . 0000 |
| 17 | A* C (1) | 1.1220 | . 0353 | 31.78 | . 0000 |
| 18 | A*C (2) | -. 5198 | . 0186 | -27.94 | . 0000 |
| 19 | B*C (1) | . 7564 | . 0224 | 33.70 | . 0000 |
| 20 | B*C (2) | . 2439 | . 0114 | 21.48 | . 0000 |
| 21 | B*C (3) | -. 2943 | . 0107 | -27.40 | . 0000 |
| 22 | B*C(4) | -. 7351 | . 0146 | -50.41 | 0000 |

Table 1: C = 1
$A=1$ by $B$ (column)

|  | 1 | 2 | 3 | 4 | 5 |
| :--- | ---: | ---: | ---: | ---: | ---: |
| Observed | 1738.00 | 2444.00 | 1088.00 | 33.00 | 3036.00 |
| Fit | 1743.79 | 2443.65 | 1089.37 | 34.94 | 3027.25 |
| Root chi-square | -.14 | .01 | -.04 | -.33 | .16 |
| Likelihood | -11.56 | .71 | -2.75 | -3.77 | 17.52 |
| Freeman-Tukey | -.13 | .01 | -.03 | -.29 | .16 |

Observed
Fit
Root chi-square
Likelihood
Freeman-Tukey
Residual
Observed
Fit
Root chi-square
Likelihood
Freeman-Tukey
Residual

| A $=2$ by B (column) |  |  |  |  |
| ---: | ---: | ---: | ---: | ---: |
| 1 | 2 | 3 | 4 | 5 |
| 3051.00 | 8930.00 | 5676.00 | 1245.00 | 5010.00 |
| 3030.39 | 8959.33 | 5649.66 | 1274.87 | 4997.74 |
| .37 | -.31 | .35 | -.84 | .17 |
| 41.36 | -58.57 | 52.80 | -59.04 | 24.55 |
| .38 | -.31 | .35 | -.83 | .18 |
| 20.61 | -29.33 | 26.34 | -29.87 | 12.26 |

$A=3$ by B (column)

| 1 | 2 | 3 | 4 | 5 |
| ---: | ---: | ---: | ---: | ---: |
| 1253.00 | 1907.00 | 675.00 | 300.00 | 769.00 |
| 1267.82 | 1878.02 | 699.96 | 268.19 | 790.00 |
| -.42 | .67 | -.94 | 1.94 | -.75 |
| -29.47 | 58.40 | -49.02 | 67.26 | -41.45 |
| -.41 | .67 | -.94 | 1.90 | -.74 |
| -14.82 | 28.98 | -24.96 | 31.81 | -21.00 |

```
Table 2: C = 2
\(A=1\) by \(B\) (column)
```

| 1 | 2 | 3 |
| ---: | ---: | ---: |
| 11.00 | 20.00 | 28.00 |
| 5.21 | 20.35 | 26.63 |
| 2.54 | -.08 | .27 |
| 16.43 | -.70 | 2.82 |
| 2.11 | -.02 | .31 |
| 5.79 | -.35 | 1.37 |

$A=2$ by $B$ (column)
1
221.00
241.61
-1.33
-39.41
-1.34
-20.61

| 2 | 3 |
| ---: | ---: |
| 2020.00 | 3657.00 |
| 1990.67 | 3683.34 |
| .66 | -.43 |
| 59.09 | -52.48 |
| .66 | -.43 |
| 29.33 | -26.34 |

$A=3$ by $B$ (column)

| 1 | 2 | 3 | 4 | 5 |
| ---: | ---: | ---: | ---: | ---: |
| 134.00 | 463.00 | 563.00 | 466.00 | 339.00 |
| 119.18 | 491.98 | 538.04 | 497.81 | 318.00 |
| 1.36 | -1.31 | 1.08 | -1.43 | 1.18 |
| 31.42 | -56.21 | 51.06 | -61.55 | 43.37 |
| 1.34 | -1.31 | 1.07 | -1.44 | 1.17 |
| 14.82 | -28.98 | 24.96 | -31.81 | 21.00 |

Asymptotic Coefficient Covariance

|  | 1 |
| :--- | ---: |
| 1 | $4.5166 \mathrm{E}-04$ |
| 2 |  |
| 3 |  |
| 4 |  |
| 5 |  |


| 2 | 3 | 4 | 5 |
| ---: | ---: | ---: | ---: |
| $7.9186 \mathrm{E}-04$ | $-4.1877 \mathrm{E}-04$ | $-2.3688 \mathrm{E}-05$ | $-1.3991 \mathrm{E}-04$ |
| $1.6721 \mathrm{E}-03$ | $-8.1154 \mathrm{E}-04$ | $-2.1653 \mathrm{E}-04$ | $-2.1766 \mathrm{E}-04$ |
|  | $4.5220 \mathrm{E}-04$ | $1.1681 \mathrm{E}-04$ | $1.1619 \mathrm{E}-04$ |
|  |  | $6.4176 \mathrm{E}-04$ | $-4.3974 \mathrm{E}-06$ |
|  |  |  | $2.8454 \mathrm{E}-04$ |


|  | 6 |
| ---: | ---: |
| 1 | $-1.3687 \mathrm{E}-04$ |
| 2 | $-2.1692 \mathrm{E}-04$ |
| 3 | $1.0611 \mathrm{E}-04$ |
| 4 | $-1.4692 \mathrm{E}-05$ |
| 5 | $1.0478 \mathrm{E}-04$ |

$-4.9884 \mathrm{E}-04$
$2.0979 \mathrm{E}-03$

| $2.7038 \mathrm{E}-05$ | $2.0404 \mathrm{E}-04$ | $2.0901 \mathrm{E}-04$ |
| ---: | ---: | ---: |
| $5.0042 \mathrm{E}-05$ | $-9.3986 \mathrm{E}-04$ | $-9.3644 \mathrm{E}-04$ |
| $3.5050 \mathrm{E}-04$ | $-1.5535 \mathrm{E}-05$ | $-2.5832 \mathrm{E}-06$ |
|  | $7.8243 \mathrm{E}-04$ | $4.3000 \mathrm{E}-04$ |
|  |  | $6.9977 \mathrm{E}-04$ |


|  | 11 | 12 | 13 | 14 | 15 |
| :---: | :---: | :---: | :---: | :---: | :---: |
| 1 | -2.2412E-04 | 8.8946E-04 | 1.0928E-04 | 1.1780E-04 | 1.1077E-04 |
| 2 | -4.5689E-04 | $1.8373 \mathrm{E}-03$ | $2.1090 \mathrm{E}-04$ | 2.2635E-04 | $2.2831 \mathrm{E}-04$ |
| 3 | 2.2724E-04 | -9.1125E-04 | -9.3444E-05 | -1.2924E-04 | -1.2340E-04 |
| 4 | 2.0377E-04 | -9.4338E-04 | -2.0368E-04 | -1.0630E-04 | -8.7364E-05 |
| 5 | 2.0993E-04 | -9.3654E-04 | -1.0946E-04 | -1.9272E-04 | -9.0437E-05 |
| 6 | 3.1555E-04 | -9.5048E-04 | -8.7438E-05 | -9.1617E-05 | -2.4750E-04 |
| 7 | -9.5214E-04 | $3.7529 \mathrm{E}-03$ | 5.0188E-04 | $4.9491 \mathrm{E}-04$ | $5.0852 \mathrm{E}-04$ |
| 8 | $7.8917 \mathrm{E}-06$ | 9.6029E-06 | 1.9062E-05 | $7.9302 \mathrm{E}-06$ | -8.3844E-06 |
| 9 | $3.6775 \mathrm{E}-04$ | -2.0006E-03 | -3.1937E-04 | -2.1432E-04 | -2.0307E-04 |
| 10 | 3.9525E-04 | -1.9713E-03 | -2.1451E-04 | -2.9253E-04 | -2.1228E-04 |
| 11 | 8.6815E-04 | -2.0202E-03 | -2.0424E-04 | -2.1296E-04 | -3.3909E-04 |
| 12 |  | $7.9675 \mathrm{E}-03$ | 9.6509E-04 | 9.5727E-04 | 9.7670E-04 |
| 13 |  |  | 3.4428E-04 | 8.8858E-05 | $5.9464 \mathrm{E}-05$ |
| 14 |  |  |  | 2.4979E-04 | 9.0332E-05 |
| 15 |  |  |  |  | $3.1320 \mathrm{E}-04$ |
|  | 16 | 17 | 18 | 19 | 20 |
| 1 | -4.5352E-04 | -5.7350E-04 | 2.9218E-04 | -9.0941E-05 | $1.6788 \mathrm{E}-05$ |
| 2 | -9.1109E-04 | -1.1795E-03 | 5.8032E-04 | 4.0375E-06 | 5.9516E-06 |
| 3 | $4.6755 \mathrm{E}-04$ | 5.8062E-04 | -3.0731E-04 | -8.2219E-06 | -2.2811E-06 |
| 4 | $4.9617 \mathrm{E}-04$ | -1.1927E-05 | $5.6127 \mathrm{E}-06$ | -4.3284E-04 | $1.1285 \mathrm{E}-04$ |
| 5 | $4.9632 \mathrm{E}-04$ | -1.3553E-05 | $1.4786 \mathrm{E}-05$ | $1.1505 \mathrm{E}-04$ | -9.1212E-05 |
| 6 | 5.1071E-04 | -3.2709E-06 | 1.6594E-06 | $1.0331 \mathrm{E}-04$ | -1.0617E-05 |
| 7 | -2.0058E-03 | 2.1742E-05 | -2.5645E-05 | $9.7265 \mathrm{E}-05$ | -1.2493E-05 |
| 8 | -2.1134E-05 | 5.8272E-04 | -3.1571E-04 | $8.3659 \mathrm{E}-05$ | -2.8479E-05 |
| 9 | $9.6316 \mathrm{E}-04$ | -5.5195E-05 | 1.1578E-05 | -6.6035E-05 | $4.5883 \mathrm{E}-06$ |
| 10 | 9.5654E-04 | -3.6818E-05 | 1.0635E-05 | $2.8380 \mathrm{E}-06$ | -3.4939E-05 |
| 11 | 9.7901E-04 | $1.4458 \mathrm{E}-05$ | 1.7958E-06 | 1.1109E-05 | -2.6554E-07 |
| 12 | -3.8468E-03 | 8.9567E-05 | -2.8268E-05 | 4.2674E-05 | 2.9911E-05 |
| 13 | -5.6426E-04 | $1.3331 \mathrm{E}-05$ | -2.9682E-05 | $1.8311 \mathrm{E}-05$ | -1.0243E-06 |
| 14 | -5.2911E-04 | $1.1124 \mathrm{E}-05$ | -1.7481E-05 | -4.3727E-06 | 1.4372E-05 |
| 15 | -5.3985E-04 | -1.3298E-06 | $9.0132 \mathrm{E}-06$ | -8.8527E-07 | 2.5063E-07 |
| 16 | 2.1729E-03 | -2.7384E-05 | 4.5185E-05 | -8.4273E-06 | -1.3704E-05 |
| 17 |  | $1.2463 \mathrm{E}-03$ | -5.9475E-04 | 2.0302E-06 | 4.6625E-06 |
| 18 |  |  | $3.4626 \mathrm{E}-04$ | $1.5440 \mathrm{E}-05$ | -2.5409E-06 |
| 19 |  |  |  | $5.0360 \mathrm{E}-04$ | -1.1720E-04 |
| 20 |  |  |  |  | $1.2896 \mathrm{E}-04$ |

18
19
20
21
22

$$
\begin{array}{rr}
-9.3811 \mathrm{E}-06 & -1.0810 \mathrm{E}-05 \\
-1.1458 \mathrm{E}-04 & -1.4538 \mathrm{E}-04 \\
1.1364 \mathrm{E}-05 & -2.1378 \mathrm{E}-05 \\
1.1537 \mathrm{E}-04 & -1.6178 \mathrm{E}-05 \\
& 2.1263 \mathrm{E}-04
\end{array}
$$

## 'CTLLN'

REAL LIFE NISS 8-D DATA USED

3-D SECTION X ( AGE, EDUCATION, ANNUAL SALARY )

## CTA PROTECTED DATA

Fitted Model: (A*B, A*C, B*C)

| Variable | Number of Levels |
| :--- | :---: |
| 1 A | 3 |
| 2 B | 5 |
| 3 C | 2 |


| Model Statistics |  |
| :--- | ---: |
| Log-likelihood | 8.0237 |
| Likelihood ratio | 24.14 |
| Degrees of freedom | 8.0 |
| P-value | .0022 |


|  |  | Coeff | Statistic Standard Error | Asymptotic Z-statistic | P-value |
| :---: | :---: | :---: | :---: | :---: | :---: |
| 1 | intercept | 6.2135 | . 0216 | 288.02 | . 0000 |
| 2 | A (1) | -1.6199 | . 0415 | -39.00 | . 0000 |
| 3 | A (2) | 1.5611 | . 0216 | 72.36 | . 0000 |
| 4 | B (1) | -. 4726 | . 0255 | -18.53 | . 0000 |
| 5 | B (2) | . 6544 | . 0169 | 38.76 | . 0000 |
| 6 | B (3) | . 4405 | . 0177 | 24.84 | . 0000 |
| 7 | B (4) | -1.0790 | . 0458 | -23.56 | . 0000 |
| 8 | C | 1.0358 | . 0191 | 54.26 | . 0000 |
| 9 | A*B (1) | . 4060 | . 0280 | 14.52 | . 0000 |
| 10 | A*B(2) | . 1400 | . 0265 | 5.29 | . 0007 |
| 11 | A*B (3) | . 0797 | . 0295 | 2.70 | . 0269 |
| 12 | A*B(4) | -1.3904 | . 0893 | -15.58 | . 0000 |
| 13 | A*B (5) | -. 5583 | . 0186 | -30.08 | . 0000 |
| 14 | A*B (6) | -. 0812 | . 0158 | -5.14 | . 0009 |
| 15 | A*B (7) | . 2103 | . 0177 | 11.88 | . 0000 |
| 16 | A*B (8) | . 6810 | . 0466 | 14.61 | . 0000 |
| 17 | A* C (1) | 1.1368 | . 0361 | 31.53 | . 0000 |
| 18 | A*C (2) | -. 5264 | . 0190 | -27.76 | . 0000 |
| 19 | B*C (1) | . 7654 | . 0226 | 33.82 | . 0000 |
| 20 | B*C (2) | . 2434 | . 0114 | 21.38 | . 0000 |
| 21 | B*C (3) | -. 2947 | . 0108 | -27.37 | . 0000 |
| 22 | B*C(4) | -. 7415 | . 0146 | -50.75 | . 0000 |


2.53
Observed
Fit
Root chi-square
Likelihood
Freeman-Tukey
Residual
Observed
Fit
Root chi-square
Likelihood
Freeman-Tukey
Residual

| A $=2$ by B (column) |  |  |  |  |
| ---: | ---: | ---: | ---: | ---: |
| 1 | 2 | 3 | 4 | 5 |
| 3048.00 | 8933.00 | 5681.00 | 1240.00 | 5012.00 |
| 3036.79 | 8960.80 | 5653.72 | 1267.14 | 4995.55 |
| .20 | -.29 | .36 | -.76 | .23 |
| 22.47 | -55.51 | 54.70 | -53.70 | 32.95 |
| .21 | -.29 | .37 | -.76 | .24 |
| 11.21 | -27.80 | 27.28 | -27.14 | 16.45 |

$A=3$ by $B$ (column)
122

| 1 | 2 | 3 | 4 | 5 |
| ---: | ---: | ---: | ---: | ---: |
| 1255.00 | 1901.00 | 680.00 | 298.00 | 763.00 |
| 1265.11 | 1875.73 | 701.71 | 266.86 | 787.59 |
| -.28 | .58 | -.82 | 1.91 | -.88 |
| -20.14 | 50.88 | -42.74 | 65.79 | -48.41 |
| -.28 | .59 | -.82 | 1.87 | -.87 |
| -10.11 | 25.27 | -21.71 | 31.14 | -24.59 |


| $A=1 \text { by } B \text { (column) }$ |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: |
|  |  |  |  |  |
| 1 | 2 | 3 | 4 | 5 |
| 6.00 | 17.00 | 31.00 | 6.00 | 29.00 |
| 4.90 | 19.53 | 25.43 | 2.00 | 37.15 |
| . 50 | -. 57 | 1.11 | 2.83 | -1.34 |
| 2.43 | -4.71 | 12.29 | 13.18 | -14.36 |
| . 56 | -. 53 | 1.09 | 2.10 | -1.37 |
| 1.10 | -2.53 | 5.57 | 4.00 | -8.15 |

$A=2$ by $B$ (column)
Observed
Fit
Root chi-square
Likelihood
Freeman-Tukey
Residual
Observed
Fit
Root chi-square
Likelihood
Freeman-Tukey
Residual
$-4.9819 \mathrm{E}-04$
$2.0980 \mathrm{E}-03$

| $2.7444 \mathrm{E}-05$ | $2.0391 \mathrm{E}-04$ | $2.0890 \mathrm{E}-04$ |
| ---: | ---: | ---: |
| $5.1073 \mathrm{E}-05$ | $-9.3980 \mathrm{E}-04$ | $-9.3640 \mathrm{E}-04$ |
| $3.6446 \mathrm{E}-04$ | $-1.5726 \mathrm{E}-05$ | $-2.6745 \mathrm{E}-06$ |
|  | $7.8249 \mathrm{E}-04$ | $4.3021 \mathrm{E}-04$ |
|  |  | $6.9966 \mathrm{E}-04$ |


|  | 11 | 12 | 13 | 14 | 15 |
| ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| 1 | $-2.2382 \mathrm{E}-04$ | $8.8871 \mathrm{E}-04$ | $1.0908 \mathrm{E}-04$ | $1.1772 \mathrm{E}-04$ | $1.1067 \mathrm{E}-04$ |
| 2 | $-4.5649 \mathrm{E}-04$ | $1.8361 \mathrm{E}-03$ | $2.1081 \mathrm{E}-04$ | $2.2628 \mathrm{E}-04$ | $2.2805 \mathrm{E}-04$ |
| 3 | $2.2698 \mathrm{E}-04$ | $-9.1080 \mathrm{E}-04$ | $-9.3249 \mathrm{E}-05$ | $-1.2916 \mathrm{E}-04$ | $-1.2332 \mathrm{E}-04$ |
| 4 | $2.0366 \mathrm{E}-04$ | $-9.4335 \mathrm{E}-04$ | $-2.0383 \mathrm{E}-04$ | $-1.0628 \mathrm{E}-04$ | $-8.7296 \mathrm{E}-05$ |
| 5 | $2.0983 \mathrm{E}-04$ | $-9.3649 \mathrm{E}-04$ | $-1.0945 \mathrm{E}-04$ | $-1.9266 \mathrm{E}-04$ | $-9.0447 \mathrm{E}-05$ |
| 6 | $3.1614 \mathrm{E}-04$ | $-9.5069 \mathrm{E}-04$ | $-8.7385 \mathrm{E}-05$ | $-9.1623 \mathrm{E}-05$ | $-2.4748 \mathrm{E}-04$ |
| 7 | $-9.5237 \mathrm{E}-04$ | $3.7530 \mathrm{E}-03$ | $5.0189 \mathrm{E}-04$ | $4.9480 \mathrm{E}-04$ | $5.0839 \mathrm{E}-04$ |
| 8 | $7.8094 \mathrm{E}-06$ | $1.0025 \mathrm{E}-05$ | $1.9231 \mathrm{E}-05$ | $8.0047 \mathrm{E}-06$ | $-8.3188 \mathrm{E}-06$ |
| 9 | $3.6773 \mathrm{E}-04$ | $-2.0009 \mathrm{E}-03$ | $-3.1902 \mathrm{E}-04$ | $-2.1450 \mathrm{E}-04$ | $-2.0309 \mathrm{E}-04$ |
| 10 | $3.9527 \mathrm{E}-04$ | $-1.9715 \mathrm{E}-03$ | $-2.1468 \mathrm{E}-04$ | $-2.9243 \mathrm{E}-04$ | $-2.1224 \mathrm{E}-04$ |
| 11 | $8.6882 \mathrm{E}-04$ | $-2.0207 \mathrm{E}-03$ | $-2.0428 \mathrm{E}-04$ | $-2.1292 \mathrm{E}-04$ | $-3.3967 \mathrm{E}-04$ |
| 12 |  | $7.9689 \mathrm{E}-03$ | $9.6514 \mathrm{E}-04$ | $9.5744 \mathrm{E}-04$ | $9.7717 \mathrm{E}-04$ |
| 13 |  |  | $3.4449 \mathrm{E}-04$ | $8.8903 \mathrm{E}-05$ | $5.9440 \mathrm{E}-05$ |
| 14 |  |  |  | $2.4976 \mathrm{E}-04$ | $9.0360 \mathrm{E}-05$ |
| 15 |  |  |  | $3.1318 \mathrm{E}-04$ |  |


|  | 16 |
| ---: | ---: |
| 1 | $-4.5301 \mathrm{E}-04$ |
| 2 | $-9.1063 \mathrm{E}-04$ |
| 3 | $4.6710 \mathrm{E}-04$ |
| 4 | $4.9614 \mathrm{E}-04$ |
| 5 | $4.9621 \mathrm{E}-04$ |
| 6 | $5.1059 \mathrm{E}-04$ |
| 7 | $-2.0055 \mathrm{E}-03$ |
| 8 | $-2.1474 \mathrm{E}-05$ |
| 9 | $9.6319 \mathrm{E}-04$ |
| 10 | $9.5669 \mathrm{E}-04$ |
| 11 | $9.7950 \mathrm{E}-04$ |
| 12 | $-3.8477 \mathrm{E}-03$ |
| 13 | $-5.6444 \mathrm{E}-04$ |
| 14 | $-5.2913 \mathrm{E}-04$ |
| 15 | $-5.3981 \mathrm{E}-04$ |
| 16 | $2.1731 \mathrm{E}-03$ |

17
$-6.0024 \mathrm{E}-04$
$-1.2329 \mathrm{E}-03$
$6.0737 \mathrm{E}-04$
$-1.2097 \mathrm{E}-05$
$-1.3684 \mathrm{E}-05$
$-3.4600 \mathrm{E}-06$
$2.2343 \mathrm{E}-05$
$6.0958 \mathrm{E}-04$
$-5.5760 \mathrm{E}-05$
$-3.7161 \mathrm{E}-05$
$1.4088 \mathrm{E}-05$
$9.1022 \mathrm{E}-05$
$1.3483 \mathrm{E}-05$
$1.1267 \mathrm{E}-05$
$-1.1569 \mathrm{E}-06$
$-2.7973 \mathrm{E}-05$
$1.3001 \mathrm{E}-03$

> 18 $3.0549 \mathrm{E}-04$ $6.0706 \mathrm{E}-04$ $-3.2064 \mathrm{E}-04$ $5.8035 \mathrm{E}-06$

19
$-9.3134 \mathrm{E}-05$

- 20
1.7289E-05
$6.0527 \mathrm{E}-06$
-2.3227E-06
1.1503E-04
-9.1801E-05
-1.1095E-05
-1.3115E-05
-2.9027E-05
$4.5449 \mathrm{E}-06$
-3.5100E-05
-3.2256E-07
$3.0201 \mathrm{E}-05$
-9.7921E-07
1.4385E-05
2.8429E-07
-1.3821E-05
4.6386E-06
-2.5208E-06
$-1.1933 \mathrm{E}-04$
$1.2961 \mathrm{E}-04$

| 21 | 22 |
| ---: | ---: |
| $2.6300 \mathrm{E}-05$ | $2.2170 \mathrm{E}-05$ |
| $3.6804 \mathrm{E}-06$ | $-4.2005 \mathrm{E}-05$ |
| $1.8395 \mathrm{E}-06$ | $2.0755 \mathrm{E}-05$ |
| $1.0546 \mathrm{E}-04$ | $1.0310 \mathrm{E}-04$ |
| $-1.2325 \mathrm{E}-05$ | $-1.3716 \mathrm{E}-05$ |
| $-6.2898 \mathrm{E}-05$ | $-2.2932 \mathrm{E}-05$ |
| $-2.0582 \mathrm{E}-05$ | $-5.5537 \mathrm{E}-05$ |
| $-3.0710 \mathrm{E}-05$ | $6.3211 \mathrm{E}-06$ |
| $1.0974 \mathrm{E}-05$ | $3.8690 \mathrm{E}-05$ |
| $7.0653 \mathrm{E}-07$ | $2.9626 \mathrm{E}-05$ |
| $-4.8566 \mathrm{E}-05$ | $3.3977 \mathrm{E}-05$ |
| $3.1867 \mathrm{E}-05$ | $-1.3724 \mathrm{E}-04$ |
| $4.7227 \mathrm{E}-07$ | $-1.6029 \mathrm{E}-05$ |
| $2.1501 \mathrm{E}-06$ | $-1.3036 \mathrm{E}-05$ |
| $1.8725 \mathrm{E}-05$ | $-1.5728 \mathrm{E}-05$ |
| $-2.0288 \mathrm{E}-05$ | $6.0277 \mathrm{E}-05$ |
| $1.5954 \mathrm{E}-07$ | $1.8607 \mathrm{E}-05$ |

22
2.2170E-05
-4.2005E-05
2.0755E-05

310E-04
-2.2932E-05
-5.5537E-05
$6.3211 \mathrm{E}-06$
3.8690E-05
3.3977E-05
$-1.3724 \mathrm{E}-04$
-1.6029E-05
-1.3036E-05
-1.5728E-05
1.8607E-05

18
19
20
21
22
-9.3932E-06 -1.0778E-05
-1.1668E-04 -1.4757E-04
$1.1894 \mathrm{E}-05$-2.0940E-05
$1.1594 \mathrm{E}-04 \quad-1.5717 \mathrm{E}-05$
2.1346E-04

## CTA ADJUSTED

## TWENTY-EIGHT 2-D

## AND

## FIFTY-SIX 3-D

## SECTIONS OF

 8-D COUNTS DATA
## File Format

Field 1: CTA adjusted Cell Count
Field 2: Actual Cell Count
Field 3: deviation from true cell value
Field 4: Percent Deviation from true cell value
Fields 5 through 12: Numeric Subscripts



| 3359 | 3368 | -9 | -. 26 * * * 1 * * 1 |
| :---: | :---: | :---: | :---: |
| 8323 | 8319 | 4 | . 04 * * * 2 * * 1 |
| 10831 | 10840 | -9 | -.08 * * * 1 * * 2 |
| 11963 | 11963 |  | . 00 * * * 2 |
| 8854 | 8836 | 18 | 20 * * * 1 * * 3 |
| 5514 | 5516 | -2 | -.03 * * * 2 * |
| 9850 | 9859 | -9 | . 09 * * * * 1 |
| 1832 | 1828 | 4 | . 21 * * * * 2 * 1 |
| 18807 | 18794 | 13 | . 06 * * * * 1 * 2 |
| 3987 | 4009 | -22 | -. 54 * * * * 2 * 2 |
| 13108 | 13109 | -1 | -.00 * * * * 1 * 3 |
| 1260 | 1243 | 17 | 1.36 * * * * 2 * 3 |
| 5493 | 5498 | -5 | -. 09 |
| 6189 | 6189 |  | . 00 * * * * * 21 |
| 15437 | 15428 | 9 | . 05 * * * * * 12 |
| 7357 | 7375 | -18 | -.24 * * * * * 22 |
| 11729 | 11724 | 5 | . 04 * * * * * 13 |
| 2639 | 2628 | 11 | . 41 * * * * 23 * |
| 8345 | 8339 | 6 | 071 * * |
| 23914 | 23912 | 2 | . 002 * * * * * * 1 |
| 4897 | 4904 | -7 | -. 143 * |
| 89 | 93 | -4 | -4.30 1 * * * * * * 2 |
| 9629 | 9629 |  | 002 * * * * * * 2 |
| 1970 | 1965 | 5 | .25 3 * * * * * * 2 |
| 4551 | 4531 | 20 | 44 * 1 |
| 26507 | 26519 | -12 | -. 04 * 2 * * * |
| 3550 | 3542 | 8 | . 22 * 3 * * * * * 1 |
| 2548 | 2563 | -15 | -.58 * 4 * * * * * |
| 1999 | 2018 | -19 | -.94 * 1 * * * * * 2 |
| 7402 | 7387 | 15 | . 20 * 2 * * * * * 2 |
| 2006 | 2015 | -9 | -.44 * 3 * * * * * 2 |
| 281 | 267 | 14 | 5.24 * 4 * * * * * 2 |
| 6048 | 6042 | 6 | . 09 * * 1 |
| 13287 | 13281 | 6 | . 04 * * 2 * * * * 1 |
| 7443 | 7439 | 4 | . 05 * * 3 * * * * 1 |
| 1569 | 1578 | -9 | -.57 * * 4 * * * * |
| 8809 | 8815 | -6 | -. 06 * * 5 * * * * |
| 359 | 366 | -7 | -1.91 * * 1 * * * * 2 |
| 2500 | 2503 | -3 | -. 12 * * 2 * * * * 2 |
| 4246 | 4248 | -2 | -. 04 * * 3 * * * * 2 |
| 2520 | 2507 | 13 | . 51 * * 4 * * * * 2 |
| 2063 | 2063 |  | . 00 * * 5 * * * * 2 |
| 12988 | 12988 |  | . 00 * * * 1 * |
| 24168 | 24167 | 1 | . 00 * * * 2 * * * 1 |
| 10056 | 10056 |  | . 00 * * * 1 * * * 2 |
| 1632 | 1631 | 1 | . 06 * * * 2 * * * 2 |
| 31169 | 31155 | 14 | . 04 * * * * 1 * * 1 |
| 5987 | 6000 | -13 | -21 * * * * 2 * * 1 |
| 10596 | 10607 | -11 | -. 10 * * * * 1 * * 2 |
| 1092 | 1080 | 12 | 1.11 * * * * 2 * * 2 |



## FIFTY-SIX 3-D SECTIONS FROM 8-D COUNTS DATA



| Section: 2 ---- $x(5,7,8)$ |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: |
| 8865 | 8870 -5 | -. 05 | * * * * 1 | * 11 |
| 1726 | 17224 | . 23 | * * * * 2 | * 11 |
| 14527 | 14523 4 | . 02 | * * * 1 | * 21 |
| 3410 | 3430-20 | -. 58 | * * * * 2 | * 21 |
| 7777 | 776215 | . 19 | * * * 1 | * 31 |
| 851 | 8483 | . 35 | * * * 2 | * 3 |
| 985 | 989 -4 | -. 40 | * * * 1 | * 12 |
| 106 | 106 | . 00 | * * * 2 | * 12 |
| 4280 | 42719 | . 21 | * * * 1 | * 22 |
| 577 | 579 -2 | -. 34 | * * * 2 | * 22 |
| 5331 | 5347-16 | -. 29 | * * * 1 | * 32 |
| 409 | 39514 | 3.54 | * * * 2 | * 32 |



| Section: 4 ---- $x(5,6,7)$ |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 4750 | 4748 | 2 | . 04 | * * * * 1 | 11 | * |
| 743 | 750 | -7 | -. 93 | * * * 2 | 11 | * |
| 5100 | 5111 | -11 | -. 21 | * * * 1 | 21 | * |
| 1089 | 1078 | 11 | 1.02 | * * * 2 | 21 | * |
| 13162 | 13148 | 14 | . 10 | * * * 1 | 12 | * |
| 2275 | 2280 | -5 | -. 21 | * * * 2 | 12 | * |
| 5645 | 5646 | -1 | -. 01 | * * * 1 | 22 | * |
| 1712 | 1729 | -17 | -. 98 | * * 2 | 22 | * |
| 10830 | 10839 | -9 | -. 08 | 1 | 13 | * |
| 899 | 885 | 14 | 1.58 | * * * 2 | 13 |  |
| 2278 | 2270 | 8 | . 35 | * * * 1 | 23 |  |
| 361 | 358 | 3 | . 83 | * * * * 2 | 23 |  |



|  | Section: 6 ---- $x(4,6,8)$ |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 11322 | 11318 | 4 | 03 | * * 1 | * 1 | * |
| 11418 | 11414 | 4 | . 03 | * * 2 | 1 | * 1 |
| 1666 | 1670 | -4 | -. 24 | * * 1 | 2 | * 1 |
| 12750 | 12753 | -3 | -. 02 | * * 2 | * 2 | * 1 |
| 8910 | 8917 | -7 | -. 07 | * * 1 | 1 | * 2 |
| 1009 | 1001 | 8 | . 79 | * * 2 | 1 | * |
| 1146 | 1139 | 7 | . 61 | * * 1 | * 2 | * |
| 623 | 630 | -7 | -1.11 | * * 2 | * 2 | * |


| Section: 7 ---- $x(4,6,7)$ |  |  |
| :---: | :---: | :---: |
| 2299 | 2310-11 | -. 47 * * * 1 * 11 |
| 3194 | 31886 | 18 * * * 2 * 1 |
| 1060 | 10582 | 18 |
| 5129 | 5131 -2 | . 03 * * * 2 * 2 |
| 9580 | 95791 | . 01 * * * 1 * 1 |
| 5857 | 58498 | 13 * * * 2 * 1 |
| 1251 | 1261-10 | . 79 * * * 1 * 2 |
| 6106 | 6114 -8 | -. 13 * * * 2 * 22 |
| 8353 | 83467 | . 08 * * * 1 * 13 |
| 3376 | 3378 -2 | $-.05 * * * 2 * 13$ |
| 501 | 49011 | 2.24 * * * 1 * 23 |
| 2138 | 2138 | 00 * * * 2 * 23 |



Section: 9 ---- $x(4,5,7)$

| 2938 | 2945 | -7 | -.23 | $*$ | $*$ | 1 | 1 | $*$ | 1 | $*$ |
| ---: | ---: | ---: | ---: | :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| 6912 | 6914 | -2 | -.02 | $*$ | $*$ | $*$ | 2 | 1 | $*$ | 1 |$*$







Section:13 ---- $x(3,6,7)$


|  | 996 | 991 | 5 | $.50 * * 3 * * 1$ | 1 |
| ---: | ---: | ---: | ---: | ---: | ---: |
| 398 | 400 | -2 | $-50 *$ |  |  |

$13841387-3-21$ * * 5 * * 11 1
$10481052-4-38$ * * 1 * * 21 *
189518905 . 26 * * 2 * * 2 1 *
$11031114-11 \quad-.98 * * 3 * * 21$ *
4.25 * * 4 * * 21 *
-10 * * 5 * * 21 * .04 * * 1 * * 12 * -.03 * * 2 * * 12 * .29 * * 3 * * 12 *
$-.21 * * 4 * * 12 *$ .06 * * 5 * * 12 * . 26 * * 1 * * 2 2 -27 * * 2 * * 22 * $.10 * * 3 * * 22 *$
$-2.79 * * 2 * *$ -16 * * 5 * * 22 * $-.31 * * 1 * * 13 *$ . 20 * * 2 * * 13 * $-.08 * * 3 * * 13 *$ $.18 * * 4 * * 13 *$ $.04 * * 5 * * 13 *$
3.46 * * 1 * * 23 * $.31 * * 2 * * 23 *$ -11 * * 3 * * 23 *
1.32 * * 4 * * 23 *
-18 * 5 * * 23 *

Section:14 ---- $x(3,5,8)$


| 1372 | 1382 | -10 | -.72 | $*$ | 5 | $*$ | 2 | $*$ | 1 |  |
| ---: | ---: | ---: | ---: | ---: | :--- | :--- | :--- | :--- | :--- | :--- |
| 309 | 318 | -9 | -2.83 | $*$ | 1 | $*$ | 1 | $*$ | $*$ | 2 |
| 2299 | 2301 | -2 | -.08 | $*$ | 2 | $*$ | 1 | $*$ | $*$ | 2 |
| 3860 | 3862 | -2 | -.05 | $*$ | 3 | $*$ | 1 | $*$ | $*$ | 2 |
| 2283 | 2270 | 13 | .57 | $*$ | 4 | $*$ | 1 | $*$ | $*$ | 2 |
| 1845 | 1856 | -11 | -.59 | $*$ | 5 | $*$ | 1 | $*$ | $*$ | 2 |
| 50 | 48 | 2 | 4.16 | $*$ | 1 | $*$ | 2 | $*$ | $*$ | 2 |
| 201 | 202 | -1 | -.49 | $*$ | 2 | $*$ | 2 | $*$ | $*$ | 2 |
| 386 | 386 |  | .00 | $*$ | $*$ | $*$ | 2 | $*$ | $*$ | 2 |
| 237 | 237 |  | .00 | $*$ | 4 | $*$ | 2 | $*$ | $*$ | 2 |
| 218 | 207 | 11 | 5.31 | $*$ | $*$ | 5 | 2 | $*$ | $*$ | 2 |




| 3144 | 3153 | -9 | -.28 | $*$ | $*$ | $*$ | 1 | 2 | $*$ | $*$ |  |
| ---: | ---: | ---: | ---: | ---: | :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| 951 | 950 | 1 | .10 | $*$ | $*$ | 4 | $*$ | 1 | 2 | $*$ | $*$ |
| 3398 | 3400 | -2 | -.05 | $*$ | $*$ | 5 | $*$ | 1 | 2 | $*$ | $*$ |
| 493 | 491 | 2 | .40 | $*$ | $*$ | 1 | $*$ | 2 | 2 | $*$ | $*$ |
| 1067 | 1071 | -4 | -.37 | $*$ | $*$ | 2 | $*$ | 2 | 2 | $*$ | $*$ |
| 684 | 685 | -1 | -.14 | $*$ | $*$ | 3 | $*$ | 2 | 2 | $*$ | $*$ |
| 144 | 140 | 4 | 2.85 | $*$ | 4 | 4 | 2 | 2 | $*$ | $*$ |  |
| 774 | 778 | -4 | -.51 | $*$ | $*$ | 5 | $*$ | 2 | 2 | $*$ | $*$ |

Section:17 ---- $\times(3,4,8)$


Section:18 ---- $\times(3,4,7)$
$\left.\begin{array}{rrrrrlllll}517 & 520 & -3 & -.57 & * & 1 & 1 & * & * & 1\end{array}\right]$

| 452 | 452 |  | .00 | $*$ | $*$ | 1 | 2 | $*$ | $*$ | 3 |
| ---: | ---: | ---: | ---: | :--- | :--- | :--- | :--- | :--- | :--- | :--- |



| Section: 20 ---- $\times(3,4,5)$ |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: |
| 2349 | 2348 | 1 | . 04 | * * 111 | * * * |
| 6657 | 6653 | 4 | . 06 | * * 211 | * * * |
| 5409 | 5406 | 3 | . 05 | * 311 | * * * |
| 2294 | 2294 |  | . 00 | * 411 | * * * |
| 3827 | 3832 | -5 | -. 13 | * 511 | * * * |
| 2908 | 2911 | -3 | -. 10 | * 121 | * * * |
| 6732 | 6734 | -2 | -. 03 | * * 221 | * * * |
| 4796 | 4799 | -3 | -. 06 | * * 321 | * * * |
| 1338 | 1328 | 10 | . 75 | * 421 | * |
| 5455 | 5457 | -2 | -. 03 | * 521 | * * * |
| 378 | 378 |  | . 00 | * 112 | * |
| 803 | 804 | -1 | -. 12 | * 212 | * * |
| 589 | 588 | 1 | . 17 | * 312 | * * |
| 277 | 276 | 1 | . 36 | * 412 | * * |
| 461 | 465 | -4 | $-.86$ | * 512 | * * * |
| 772 | 771 | 1 | . 13 | * 122 | * * |
| 1595 | 1593 | 2 | . 12 | * 222 | * |
| 895 | 894 | 1 | . 11 | * 322 | * * |
| 180 | 187 | -7 | -3.74 | * 422 | * * * |
| 1129 | 1124 | 5 | . 44 | * 522 | * * * |



| 997 | 997 |  | .00 | $*$ | 3 | $*$ | $*$ | $*$ | 2 | 1 |  |
| ---: | ---: | ---: | ---: | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| 938 | 949 | -11 | -1.15 | $*$ | 4 | $*$ | $*$ | $*$ | 2 | 1 |  |
| 852 | 835 | 17 | 2.03 | $*$ | 1 | $*$ | $*$ | $*$ | 3 | 1 |  |
| 5894 | 5892 | 2 | .03 | $*$ | 2 | $*$ | $*$ | $*$ | 3 | 1 |  |
| 1616 | 1621 | -5 | -.30 | $*$ | 3 | $*$ | $*$ | $*$ | 3 | 1 |  |
| 266 | 262 | 4 | 1.52 | $*$ | 4 | $*$ | $*$ | $*$ | 3 | 1 |  |
| 192 | 197 | -5 | -2.53 | $*$ | 1 | $*$ | $*$ | $*$ | 1 | 1 | 2 |
| 540 | 538 | 2 | .37 | $*$ | 2 | $*$ | $*$ | $*$ | 1 | 2 |  |
| 247 | 256 | -9 | -3.51 | $*$ | 3 | $*$ | $*$ | $*$ | 1 | 2 |  |
| 112 | 104 | 8 | 7.69 | $*$ | 4 | $*$ | $*$ | $*$ | 1 | 2 |  |
| 1076 | 1081 | -5 | -.46 | $*$ | 1 | $*$ | $*$ | $*$ | 2 | 2 |  |
| 3204 | 3194 | 10 | .31 | $*$ | 2 | $*$ | $*$ | $*$ | 2 | 2 |  |
| 481 | 483 | -2 | -.41 | $*$ | 3 | $*$ | $*$ | $*$ | 2 | 2 |  |
| 96 | 92 | 4 | 4.34 | $*$ | 4 | $*$ | $*$ | $*$ | 2 | 2 |  |
| 731 | 740 | -9 | -1.21 | $*$ | 1 | $*$ | $*$ | $*$ | 3 | 2 |  |
| 3658 | 3655 | 3 | .08 | $*$ | 2 | $*$ | $*$ | $*$ | 3 | 2 |  |
| 1278 | 1276 | 2 | .15 | $*$ | 3 | $*$ | $*$ | $*$ | $*$ | 3 | 2 |
| 73 | 71 | 2 | 2.81 | $*$ | 4 | $*$ | $*$ | $*$ | $*$ | 3 | 2 |


| Section:22 ---- $\times(2,6,8)$ |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 2490 | 2468 | 22 | . 89 | * 1 | * * 1 | * 1 |
| 16011 | 16015 | -4 | -. 02 | * 2 * |  |  |
| 2894 | 2890 | 4 | 13 | 3 |  |  |
| 1345 | 1359 | -14 | -1.03 | 4 |  |  |
| 2061 | 2063 | -2 | -. 09 | 1 | * 2 |  |
| 10496 | 10504 | -8 | -. 07 | 2 | * 2 | * 1 |
| 656 | 652 | 4 | . 61 | * 3 * | * 2 |  |
| 1203 | 1204 | -1 | -. 08 | 4 | * 2 |  |
| 1605 | 1608 | -3 | -. 18 | 1 | * 1 | * 2 |
| 6299 | 6292 | 7 | . 11 | * 2 |  |  |
| 1824 | 1827 | -3 | -. 16 | 3 | * 1 |  |
| 191 | 191 |  | . 00 | 4 | * 1 | * 2 |
| 394 | 410 | -16 | -3.90 | 1 | * 2 |  |
| 1103 | 1095 | 8 | 73 | 2 | 2 |  |
| 182 | 188 | -6 | -3.19 | * 3 | * 2 | * 2 |
| 90 | 76 | 14 | 18 |  | * * 2 | * |



| 236 | 240 | -4 | -1.66 | $*$ | 4 | $*$ | $*$ | 3 | $*$ |  |
| ---: | ---: | ---: | ---: | :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| 464 | 463 | 1 | .21 | $*$ | 1 | $*$ | $*$ | $*$ | 3 | $*$ |
| 1805 | 1803 | 2 | .11 | $*$ | 2 | $*$ | $*$ | $*$ | 3 | $*$ |
| 267 | 269 | -2 | -.74 | $*$ | 3 | $*$ | $*$ | 2 | 3 | $*$ |
| 103 | 93 | 10 | 10.75 | $*$ | 4 | $*$ | $*$ | $*$ | 3 | $*$ |






Section:27 ---- $x(2,4,8)$

| 1519 | 1505 | 14 | .93 | $*$ | 1 | $*$ | 1 | $*$ | $*$ | $*$ |
| ---: | ---: | ---: | ---: | :--- | :--- | :--- | :--- | :--- | :--- | :--- |





| Section:33 ---- $\times(2,3,6)$ |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 240 | 234 | 6 | 2.56 | 1 | * 1 | * * |
| 3296 | 3302 | -6 | -. 18 | * 21 * | * 1 | * * |
| 508 | 508 |  | . 00 | * 31 * | * 1 | * * |
| 370 | 375 | -5 | -1.33 | * 41 | * 1 |  |
| 1001 | 1000 | 1 | . 10 | 12 | * 1 | * * |
| 7761 | 7759 | 2 | . 02 | 22 | * 1 | * * |
| 1484 | 1482 | 2 | . 13 | 32 | * 1 | * * |
| 444 | 446 | -2 | -. 44 | 2 | * 1 | * * |
| 1227 | 1221 | 6 | . 49 | 3 | * 1 | * * |
| 5216 | 5210 | 6 | . 11 | 3 | * 1 | * * |
| 1197 | 1193 | 4 | . 33 | * 33 | * 1 | * * |
| 221 | 225 | -4 | -1.77 | 43 | * 1 | * * |
| 753 | 752 | 1 | . 13 | 4 | * 1 | * * |
| 1491 | 1491 |  | . 00 | * 24 | * 1 | * * |
| 670 | 671 | -1 | -. 14 | 34 | * 1 | * * |
| 80 | 81 | -1 | -1.23 | 4 | * 1 |  |
| 874 | 869 | 5 | . 57 | 15 | * 1 | * * |
| 4546 | 4545 | 1 | . 02 | 25 | * 1 |  |
| 859 | 863 | -4 | -. 46 | 35 | * 1 | * * |
| 421 | 423 | -2 | -. 47 | * 45 | * 1 | * * |
| 120 | 124 | -4 | -3.22 | * 11 * | * 2 | * * |
| 1516 | 1511 | 5 | . 33 | * 21 * | * 2 | * * |
| 87 | 89 | -2 | -2.24 | * 3 | * 2 | * * |
| 270 | 265 | 5 | 1.88 | 41 | * 2 |  |
| 572 | 571 | 1 | . 17 | 2 | * 2 | * * |
| 3923 | 3923 |  | . 00 | * 22 | * 2 | * |
| 222 | 223 | -1 | -. 44 | 32 | * 2 | * * |
| 380 | 380 |  | . 00 | * 42 | * 2 | * * |
| 749 | 756 | -7 | -. 92 | * 13 | * 2 | * * |
| 2663 | 2666 | -3 | 11 | 23 | 2 | * * |


| 228 | 231 | -3 | -1. 29 | 3 | 3 |  | * 2 |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 188 | 185 | 3 | 1.62 | 4 | 3 |  | * 2 |  |  |
| 448 | 449 | -1 | -. 22 | * 1 | 4 |  |  |  |  |
| 509 | 509 |  | . 00 | 2 | 4 |  | * 2 |  |  |
| 107 | 104 | 3 | 2.88 | 3 | 4 |  | + |  |  |
| 31 | 28 | 3 | 10.71 | 4 | 4 |  | + 2 |  |  |
| 566 | 573 | -7 | -1.22 | 1 | 5 |  | * 2 |  |  |
| 2988 | 2990 | -2 | -. 06 | 2 | 5 |  | + 2 |  |  |
| 194 | 193 | 1 | . 51 | 3 | 5 |  | * 2 |  |  |
| 424 | 422 | 2 | 47 | 4 | 5 |  | * 2 |  |  |

Section:34 ---- $\times(2,3,5)$

| 271 | 267 | 4 | 1.49 |  |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 3932 | 3935 | -3 | -. 07 |  |  |  |  |  |  |  |
| 552 | 550 | 2 | . 36 | * |  |  |  |  |  |  |
| 502 | 507 | -5 | -. 98 | * |  |  |  |  |  |  |
| 1234 | 1233 | 1 | . 08 |  |  |  |  |  |  |  |

15851585 . 00 * 32 * 1 * * *

$1320-1321-1 \quad-.07$ * 33 * 1 * * *

| 348 | 348 |  | .00 | $*$ | 4 | 3 | $*$ | 1 |
| ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | * $*$ *

$7257205 \quad .69$ * 34 * 1 * * *
3.12 * 4 * 1 * * *
$11051110-5-.45$ * 15 * 1 * * *
647164701 . 01 * 25 * 1 * * *


| 339 | 338 | 1 | $.29 * 122 * 2 * *$ |
| ---: | ---: | ---: | ---: |

1211201.83 * 32 * 2 * * *
$163164-1 \quad-.61$ * 42 * 2 * * *
378 379-1 -.26 * 13 * 2 * * *
9409382 . 21 * 2 3 * 2 * * *
10510321.94 * 3 3 * 2 * * *
$6162-1-1.61$ * 43 * 2 * * *
169169.00 * 14 * 2 * * *
$224-226-2-.88$ * 24 * 2 * * *
$5255-3-5.45$ * 34 * 2 * * *
12 13-1 -7.69 * 44 * 2 * * *

3353323 . 90 * 15 * 2 * * *
$10631065-2-.18$ * 25 * 2 * * *
$6971-2-2.81$ * 35 * 2 * * *

| Section:35 ---- $\times(2,3,4)$ |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: |
| 171 | 168 | 3 | 1.78 | 111 | * * * * |
| 1955 | 1955 |  | . 00 | 2 | * * * * |
| 406 | 406 |  | . 00 | * 3 | * * * * |
| 195 | 197 | -2 | -1.01 | 4 | * * * |
| 781 | 776 | 5 | . 64 | 12 | * * * * |
| 5135 | 5137 | -2 | -. 03 | 22 |  |
| 1200 | 1199 | 1 | . 08 | 3 |  |
| 344 | 345 | -1 | -. 29 | 42 | * * * * |
| 972 | 971 | 1 | . 10 | 13 |  |
| 3828 | 3826 | 2 | . 05 | * 23 |  |
| 995 | 993 | 2 | . 20 | * 33 |  |
| 203 | 204 | -1 | -. 49 | 43 |  |
| 689 | 689 |  | . 00 | 1 | * * * * |
| 1230 | 1231 | -1 | -. 08 | * 24 | * * * |
| 569 | 568 | 1 | . 17 | * 3 |  |
| 83 | 82 | 1 | 1.22 | * 4 | * * * * |
| 621 | 621 |  | . 00 | * 151 | * * * * |
| 2799 | 2801 | -2 | -. 07 | 2 |  |
| 707 | 711 | -4 | -. 56 | * 35 | * * * * |
| 161 | 164 | -3 | -1.82 | * 451 | * * * * |
| 189 | 190 | -1 | -. 52 | 2 |  |
| 2857 | 2858 | -1 | -. 03 | * 212 | * * * |
| 189 | 191 | -2 | -1.04 | * 312 | * * * * |
| 445 | 443 | 2 | . 45 | 412 |  |
| 792 | 795 | -3 | -. 37 | 122 | * * * |
| 6549 | 6545 | 4 | . 06 | * 222 | * * * * |
| 506 | 506 |  | . 00 | * 322 |  |
| 480 | 481 | -1 | -. 20 | * 422 | * * * * |
| 1004 | 1006 | -2 | -. 19 | * 132 | * * * * |
| 4051 | 4050 | 1 | . 02 | * 232 |  |
| 430 | 431 | -1 | -. 23 | * 332 | * * * * |
| 206 | 206 |  | . 00 | * 432 | * * * * |
| 512 | 512 |  | . 00 | 142 |  |
| 770 | 769 | 1 | . 13 | * 242 | * * * |
| 208 | 207 | 1 | . 48 | * 342 |  |
| 28 | 27 | 1 | 3.70 | * 442 |  |
| 819 | 821 | -2 | -. 24 | * 152 | * * * |
| 4735 | 4734 | 1 | . 02 | * 25 |  |
| 346 | 345 | 1 | . 29 | * 352 | * * * * |
| 684 | 681 | 3 | . 44 | * 452 | * * * * |
| Section: 36 ---- $\times(1,7,8)$ |  |  |  |  |  |
| 4280 | 4281 | -1 | -. 02 | 1 * * | * 11 |
| 4321 | 4320 | 1 | . 02 | 2 | 1 |
| 1990 | 1991 | -1 | -. 05 | 3 * * * | * 11 |
| 3114 | 3117 | -3 | -. 09 | 1 * * * | * 21 |
| 12802 | 12801 | 1 | . 00 | 2 | 21 |
| 2021 | 2035 | -14 | -. 68 | 3 | * 21 |
| 951 | 941 | 10 | 1.06 | 1 * * * | * 31 |
| 6791 | 6791 |  | . 00 | 2 | * 31 |
| 886 | 878 | 8 | . 91 | 3 * * * | * 31 |
| 9 | 11 | -2 | -18.18 | 1 | * * 12 |
| 734 | 734 |  | . 00 | 2 | * * 12 |
| 348 | 350 | -2 | -. 57 | * | * 12 |
| 43 | 41 | 2 | 4.87 | 1 * * * | * * 22 |
| 3957 | 3955 | 2 | . 05 | 2 * * * | * 22 |


| 857 | 854 | 3 | .35 | 3 | $*$ | $*$ | $*$ | 2 | 2 |  |
| ---: | ---: | ---: | ---: | ---: | :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| 37 | 41 | -4 | -9.75 | 1 | $*$ | $*$ | $*$ | $*$ | 3 | 2 |
| 4938 | 4940 | -2 | -.04 | 2 | $*$ | $*$ | $*$ | $*$ | 3 | 2 |
| 765 | 761 | 4 | .52 | 3 | $*$ | $*$ | $*$ | $*$ | 3 | 2 |


| Section:37 ---- $\times(1,6,8)$ |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 4558 | 4546 | 12 | . 26 | 1 | * 1 | * 1 |
| 15149 | 15158 | -9 | -. 05 | 2 | + 1 |  |
| 3033 | 3028 | 5 | . 16 | 3 | * 1 | * 1 |
| 3787 | 3793 | -6 | -. 15 | 1 | * 2 | * 1 |
| 8765 | 8754 | 11 | . 12 | 2 | * 2 | * 1 |
| 1864 | 1876 | -12 | -. 64 | 3 | * 2 |  |
| 68 | 67 | 1 | 1.49 | 1 * | 1 |  |
| 8109 | 8111 | -2 | -. 02 | 2 | * 1 | * 2 |
| 1742 | 1740 | 2 | . 11 | 3 | * 1 | * 2 |
| 21 | 26 | -5 | -19.23 | 1 * * * | 2 | * 2 |
| 1520 | 1518 | 2 | . 13 | 2 | 2 | * 2 |
| 228 | 225 | 3 | 1.33 | 3 * * * | 2 | * |


| Section:38 --- $x(1,6,7)$ |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 1954 | 1952 | 2 | 10 | 1 * | * * 1 |  |
| 2189 | 2198 | -9 | -. 40 | 2 * | * * 1 |  |
| 1350 | 1348 | 2 | . 14 | 3 * | * 1 |  |
| 2335 | 2340 | -5 | -. 21 | 1 * * | * * 2 |  |
| 2866 | 2856 | 10 | 35 | 2 * | * 2 |  |
| 988 | 993 | -5 | -. 50 | 3 | * 2 |  |
| 1942 | 1929 | 13 | . 67 | 1 * * | * * 1 |  |
| 11441 | 11446 | -5 | -. 04 | 2 * * | * 1 |  |
| 2054 | 2053 | 1 | 04 | 3 | * * 1 |  |
| 1215 | 1229 | -14 | -1.13 | 1 * * | * * 2 |  |
| 5318 | 5310 | 8 | 15 | 2 * * | * 2 |  |
| 824 | 836 | -12 | -1.43 | 3 * * | * * 2 |  |
| 730 | 732 | -2 | -. 27 | 1 * * | * * 1 |  |
| 9628 | 9625 | 3 | 03 | 2 * * | * * 1 |  |
| 1371 | 1367 | 4 | . 29 | 3 | * * 1 |  |
| 258 | 250 |  | 3.20 | 1 * * | * * 2 |  |
| 2101 | 2106 | -5 | -. 23 | 2 | * * 2 |  |
| 280 | 272 | 8 | 2.94 | 3 * * | * * 2 | 3 |







| 1792 | 1813 | -21 | -1.15 | 3 | $*$ | $*$ | 1 | $*$ | $*$ | 2 |
| ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: |


| Section:44 |  |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 562 | 558 | 4 | 71 | 1 | * * 1 | * 1 |  |
| 15725 | 15736 | -11 | -. 07 | 2 | * * 1 | * 1 |  |
| 3945 | 3941 | 4 | . 10 | 3 | * * 1 | * 1 |  |
| 4064 | 4055 | 9 | . 22 | 1 | 2 | * 1 |  |
| 7533 | 7533 |  | . 00 | 2 | * * 2 | * 1 |  |
| 830 | 827 | 3 | . 36 | 3 | * * 2 | * 1 |  |
| 216 | 221 | -5 | -2.26 | 1 | * * 1 | * 2 |  |
| 2265 | 2251 | 14 | . 62 | 2 | * * 1 | * 2 |  |
| 331 | 337 | -6 | -1.78 | 3 | * * 1 | * 2 |  |
| 3592 | 3598 | -6 | -. 16 | 1 | * 2 | * 2 |  |
| 8020 | 8021 | -1 | -. 01 | 2 | * | * 2 |  |
| 761 | 1764 | -3 | -. 17 | 3 | * * 2 | * 2 |  |


| Section:45 ---- $x(1,4,5)$ |  |  |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 686 | 687 | -1 | -. 14 | 1 | * * 1 | 1 |  |  |
| 15950 | 15944 | 6 | . 03 | 2 | * * 1 | 1 |  |  |
| 3900 | 3902 | -2 | -. 05 | 3 | * * 1 | 1 |  |  |
| 6565 | 6566 | -1 | -. 01 | 1 | 2 | 1 |  |  |
| 12485 | 12488 | -3 | -. 02 | 2 | * * 2 | 1 |  |  |
| 2179 | 2175 | 4 | . 18 | 3 | * 2 | 1 |  |  |
| 92 | 92 |  | . 00 | 1 | * * 1 |  |  |  |
| 2040 | 2043 | -3 | -. 14 | 2 | * * 1 | 2 |  |  |
| 376 | 376 |  | . 00 | 3 | * * 1 | 2 |  |  |
| 1091 | 1087 | 4 | . 36 | 1 | 2 | 2 |  |  |
| 3068 | 3066 | 2 |  | 2 | 2 |  |  |  |
| 412 | 416 | -4 | -. 96 | 3 | 2 | 2 |  |  |






----------------------------

| 320 | 316 | 4 | 1.26 | 1 | 1 | $*$ | $*$ | 1 | $*$ |  |  |
| :--- | ---: | ---: | ---: | ---: | ---: | ---: | ---: | :--- | :--- | :--- | :--- |
| 727 | 732 | -5 | -.68 | 2 | 1 | $*$ | $*$ | $*$ | 1 | $*$ |  |
| 269 | 271 | -2 | -.73 | 3 | 1 | $*$ | $*$ | $*$ | 1 |  |  |

$32923297-5 \quad-.1512$ * * * * 1 *
$33763377-1 \quad-.0322$ * * * * 1 *
10581058.0032 * * * * 1 *
100955.2613 * * * * 1 *

|  | 591 | 16 |
| :---: | :---: | :---: |
| 49 | 494 | 4033 * * * * 1 |

$577584-7-1.1914$ * * * * 1 *

204 208-4 -1.92 1 1 * * * * 2 *
$293529341 \quad .0321$ * * * * 2 *
$512513-1 \quad-.1931$ * * * * 2 *
249324903 . 1212 * * * * 2 *
12405124041 . 0022 * * * * 2 *
17331733.0032 * * * * 2 *

| 81 | 84 | -3 | -3.57 | 1 | 3 | $*$ | $*$ |
| ---: | ---: | ---: | ---: | ---: | ---: | ---: | :--- |
| 1035 | 1034 | 1 | .09 | 2 | 3 | $*$ | $*$ |
|  | * | 2 | * |  |  |  |  |

362362.0033 * * * * 2 *
3793763.7914 * * * * 2 *
384384 . 0024 * * * * 2 * 271 281-10 -3.5534 * * * * 2 *
636123.2711 * * * * 3 *
135813544.2921 * * * * 3 *
16216021.2531 * * * * 3 *
7937903.3812 * * * * 3 *
791779161.0122 * * * * 3 *
$842841 \quad 1 \quad .1132$ * * * * 3 *

| 50 | 52 | -2 | -3.84 | 1 | 3 | $*$ | $*$ | $*$ | 3 | $*$ |
| ---: | ---: | ---: | ---: | ---: | ---: | ---: | :--- | :--- | :--- | :--- | :--- |
| 2304 | 2305 | -1 | -.04 | 2 | 3 | $*$ | $*$ | $*$ | $*$ | $*$ |
| 540 | 540 |  | .00 | 3 | 3 | $*$ | $*$ | $*$ | 3 | $*$ |
| 82 | 79 | 3 | 3.79 | 1 | 4 | $*$ | $*$ | $*$ | 3 | $*$ |
| 150 | 156 | -6 | -3.84 | 2 | 4 | $*$ | $*$ | $*$ | 3 | $*$ |
| 107 | 98 | 9 | 9.18 | 3 | 4 | $*$ | $*$ | $*$ | 3 | $*$ |





| Section:55 ---- $x(1,2,4)$ |  |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 60 | 53 | 7 | 13.20 | 11 | * 1 | * * | * * |
| 2651 | 2648 | 3 | 11 | 21 | * 1 | * * | * |
| 523 | 524 | -1 | -. 19 | 31 | * 1 | * * | * |
| 643 | 647 | -4 | -. 61 | 12 | * 1 | * * | * |
| 12213 | 12210 | 3 | . 02 | 22 | * 1 | * | * * |
| 2091 | 2093 | -2 | -. 09 | 32 | * 1 | * * | * * |
| 25 | 26 | -1 | -3.84 | 13 | * 1 | * * | * * |
| 2756 | 2756 |  | . 00 | 23 | * 1 | * * | * * |
| 1096 | 1095 | 1 | . 09 | 33 | * 1 | * * | * * |
| 50 | 53 | -3 | -5.66 | 14 | * 1 | * * | * |
| 370 | 373 | -3 | -. 80 | 24 | * 1 | * * | * * |






[^0]:    ${ }^{1}$ In the SDL literature, the 8-D Counts data has been used by NISS to justify partial release of a select few summary tables as a statistical disclosure control strategy.

[^1]:    ${ }^{2}$ Some iterative refinement of the CTA protected counts data might be required to eliminate fractional adjustments.

[^2]:    ${ }^{3}$ The CTA protected data and related log-linear model analysis is available by contacting the author. We look forward to comparative evaluation of CTA protected data with other sensitive tabular data protection methods.
    ${ }^{4}$ The preliminary analysis reveals that for a large population, relatively small changes in the actual cell count does not hinder the statistical inference based on log-linear models. Uncertainty bounds of log-linear model based statistical inferences resulting from the small changes in actual cell counts are easy to establish by using Latin Hypercube Sampling technique.

[^3]:    ${ }^{5}$ In multi-variate data analysis, the cell values from a 1-D cross section are used to develop a main effects model. The cell values from a 2-D cross section are used to capture two-way interaction terms, and so on for the higher dimensional cross sections.

[^4]:    *Research conducted while visiting Microsoft Research.
    ${ }^{\dagger}$ Supported by NSF grants CNS-0627526 and CCF-0426582 and US-Israel BSF grant 2004288.

[^5]:    ${ }^{1}$ Typically, attributes are non-binary. While our exposition uses binary attributes, any attribute with $m$ possible values can be decomposed into $\log (m)$ binary attributes. This is even natural in many OLAP settings, where the attributes are hierarchically organized.

[^6]:    ${ }^{2}$ Intuitively, the utility of the database provides a cryptographic one-time pad which can be combined with auxiliary information to yield a devastating privacy compromise. The user of the system learns the utility and can therefore subtract out the one-time pad, revealing the privacy compromise. Anyone not having access to the system's utility cannot "decode" the auxiliary information.

[^7]:    ${ }^{3}$ The Laplace distribution is centered at zero, with exponential tails in each direction.

