Statistical Disclosure Limitation with Released Marginals and Conditionals for Contingency Tables

Aleksandra B. Slavkovic
Department of Statistics
The Pennsylvania State University
University Park, PA 16802
sesa@stat.psu.edu

Abstract

The goal of statistical disclosure limitation is to develop methods and tools that while preserving confidentiality can provide access to useful statistical data, not just a few numbers. In this paper we consider releases from contingency tables in the form of marginal counts and observed conditional frequencies. We link data utility to log-linear models, and evaluation of disclosure risk to bounds on cell entries in the table. We illustrate some of the selected ideas from [8, 18, 17, 12] with 2^4 *example.*

1 Introduction

Historically, government agencies and social science and public health researchers have collected observational and experimental information from respondents at the individual level (microdata). Yet, often these agencies and researchers have reported the results in the form of marginal cross-classification tables of counts by aggregating over different categories, or as tables with proportions or percentages adding to one for a key explanatory variable. In a table of counts, a cell entry is a non-negative integer value representing the number of individuals sharing the same attribute. In tables of rates, i.e., conditional observed frequencies, a cell entry is a non-negative rational number between zero and one representing a proportion of individuals who share an attribute with respect to the marginal count. The latter has been a standard form of reporting results of sample surveys, typically in the form of two-way and three-way tables.

The goal of statistical disclosure limitation is to develop methods and tools that while preserving confidentiality can provide access to useful statistical data, not just a few numbers. Sufficient information should be available such that multivariate inferences should be the same as if we had original complete data. Well-designed statistical disclosure limitation methods require the ability to reverse disclosure protection mechanism, not for individual identification, but for inferences about parameters in statistical models (e.g, likelihood function for disclosure procedure). Data utility in the tables of counts is typically tied to usefulness of marginal totals especially in connection with log-linear models.

The risk of identity disclosure in a table of counts is usually associated with small cell values. Small counts such as “1”, “2”, and potentially “3” allow an intruder to match characteristics in table with other databases and learn confidential information. But if we report selected margins and/or conditional probability values from a multi-way table with such small values, can that information be used to infer values in the cells of the full table?

In this paper, we explore the questions: (1) What data are releasable from a table with small counts that will not raise confidentiality concerns?, and (2) Will the released data be useful for statistical inference? Slavkovic [17] describe several complete and incomplete characterizations of probability distributions for two-way and k-way contingency tables using marginals, conditionals, and odds ratios, and technical tools from algebraic geometry, probability, directed acyclic graphs, and log-linear models. Complete specifications are associated with unique identification of the full joint distribution, i.e., full disclosure. The partial specifications involve dropping of components from complete specifications. Here there is more than one joint distribution or tables of counts. Tools from algebraic geometry help describe the space of solutions. Upper and lower bounds on cell entries are calculated via linear and integer programming. Statistical underpinnings for partial specifications offer insights for developing methodology for disclosure limitation.

The above mentioned ideas and concepts from statistical disclosure limitation are relevant for computer security as well. The usual notions of privacy preserving data-mining typically do not allow users to do proper statistical analyses on information stored in various databases. A part of the longer-term model of privacy preserving data-mining should be to have some parts of databases be more broadly accessible, and thus increase the data utility component in the computer security framework. Complete and incomplete specifications are relevant for distributed databases where each component is available on a different machine or system; this may include having different information on the same individual spread across different databases as well. The question then is whether the combination of pieces of information from different sites, machines or systems compromises the confidentiality of individual information.

In this paper we illustrate selected ideas from [8, 11, 18, 17, 12] with a special emphasis on using conditionals in place of selected marginals using a 2^4 example. In the next section, we introduce a notation and give a brief overview of log-linear models. In Section 3, we present data from [15] and describe potential confidentiality concerns. This draws on the discussion in [12] and extends it with extra technical details. We use this example to illustrate how our results relate to log-linear model analysis and discuss some implications for statistical disclosure limitation. We also briefly discuss the relevance of this work to data-mining association rules.

2 Contingency Tables and Log-Linear Models

Let $X = (X_1, X_2, ..., X_k)$ be a discrete random vector with probability function

$$p(x) = P(X = x) = P(X_1 = x_1, ..., X_k = x_k)$$

where $x = (x_1, ..., x_k)$. Each X_i is defined on a finite set of integers $[d_i] = \{1, 2, ..., d_i\}, d_i \geq 1, i = 1, ..., K$, with $\mathcal{D} = [d_1] \times ... \times [d_k]$. The data can be represented as counts in a k-way contingency table, $d_1 \times d_2 \times \cdots \times d_k$. Defined in this way, a table of counts is a point in a simplex of dimension equal to $\mathcal{D} - 1$, i.e., the number of cells-1. The values of X_i are lattice points in the convex polytope. Parameter sets lie in a related simplex. This sets up a link between contingency tables and algebraic geometry and allows us to use tools from algebraic geometry to describe the space of tables all satisfying some constraints or a model.

Consider disjoint subsets A and B of $K = \{1, ..., k\}$. The marginal table X_A with probabilities is defined as $p(x_A) = \sum_{K \setminus A} p(x_K)$, or equivalently $x_A = (x_j : j \in A)$. For example, if $A = \{1, 4\}$, then $x_A = (x_1, x_4)$. We define a conditional table $X_{A|B}$ with conditional probability values as a multi-conditional array $p(x_A|x_B) = \frac{p(x_{AB})}{p(x_B)}$ (e.g., Table 6).

Margins are a linear map of the original cell counts or cell probabilities. Conditional probability values are a linear-fractional map of the cell values. Any k-way table satisfying a set of compatible marginals and/or conditionals is a point in a convex polytope defined by a system of linear equations induced by released conditionals and marginals. Certain sets of conditionals and marginals are sufficient to uniquely determine the original table. Slavkovic and Fienberg [18, 17] gives technical details of possible complete specifications. This work builds on and extends the uniqueness theorem described by [14, 3].

When the table does not satisfy the uniqueness theorem, there is more than one possible realization of the joint distribution for X; i.e., there is more than one table that satisfies constraints imposed by conditionals and marginals. Given a set of margins only, [9, 8, 7] used undirected graphical representation of log-linear models as a framework to compute bounds on cell entries as input to assessing disclosure risk. Slavkovic and Fienberg [18, 17] extend this work by describing incomplete specification of the joint and calculation of bounds given an arbitrary collection of marginals and conditionals. They also discuss some inadequacies in treating conditional constraints via linear programming.

Log-linear models are classical statistical tools for modeling multivariate discrete data. Log-linear model theory explains how to do the estimation and how to assess the fit of the models to the data in a multi-way table [11]. Every major statistical package either includes specific programs for carrying out the calculations or has a generalized linear model program that can be used for this purpose.

We present here a symmetric parameterizations of log-linear expansion adopted by [4] and others by analogy with analysis of variance (ANOVA) models:

$$\log f(x) = \sum_{A \subset K} u_A(x)$$

where $A \subset K = \{1, ..., k\}$ and u's satisfy:

1. $u_0(x)$ is a constant;
2. For every $A \subset K$, $u_A(x)$ is only a function of x_A;
3. If $i \in A$ and $x_i = 0$, then $u_A(x) = 0$.

Suppose for a four-dimensional table $I \times J \times K \times L$ and $X = (X_1, X_2, X_3, X_4)$, that the total of the counts is N and that the count for the (i, j, k, l) cell is x_{ijkl}. When the cell frequencies are added over a particular variable, we replace the subscript for that variable by a "+". For example, x_{i++l} is a cell in a marginal table $x_A = (x_1, x_4)$. The simplest model for a four dimensional table corresponds to the complete independence of all four variables:

$$\log p_{ijkl} = u_0 + u_1(i) + u_2(j) + u_3(k) + u_4(l),$$

with the usual ANOVA-like constraints:

$$\sum_i u_1(i) = \sum_j u_2(j) = \sum_k u_3(k) = \sum_l u_4(l) = 0$$

The log-linear model can be described by the highest order u-terms (or model generators). In the abbreviated notation, we refer to the model of complete independence as [1][2][3][4].

More complex models are need to represent dependence between 4 variables, and they may include two-factor and higher-order interaction terms. For 4 variables there are $\binom{4}{2}$ = 6 possible sets of two-factor terms such as $\{u_{12}(i,j)\}$, $\binom{4}{3}$=4 possible sets of three-factor terms such as $\{u_{123}(i,j,k)\}$, and one four-factor term $\{u_{1234}(i,j,k,l)\}$. All models are are special cases of the the general (saturated) log-linear model [1234] which imposes no restrictions on cell values and includes all possible lower-order u-terms. If we include a term that corresponds to a margin in the model, then a log-linear model is effectively equal to a model holding that margin fixed.

To check the goodness-of-fit of a model we can use the likelihood ratio test based on the deviance statistic:

$$\Delta G^2 = 2(\hat{l}_{\text{sat}} - \hat{l}_M)$$

where $\hat{l}_{\text{sat}}$ is the log-likelihood of the saturated model evaluated at the maximum likelihood estimates (MLE), and $\hat{l}_M$ is the log-likelihood of the model M evaluated at its own MLE. The statistic ΔG^2 has an approximate χ^2 distribution with degrees of freedom equal to the difference in the degrees of freedom between the two models.

A key theoretical result is that the "minimal sufficient statistics" or "data summaries needed for efficient estimation" associated with a log-linear model corresponded to the highest order terms or interactions in the model, e.g., a two-may margin corresponds to a first-order interaction for the corresponding variables, and a three-way margin corresponds to a 2nd-order interaction. The ideas on log-linear models make it clear, however, that an analyst had to use the information in all of the minimal sufficient statistics simultaneously for estimation purposes and not simply proceed piecemeal by looking at the association margin by margin (e.g., see [4, 2]). Otherwise one might mistakenly infer dependencies among variables that in effect are explained by other dependencies, or even get a reversal of the "sign" associated with the association, as in the phenomenon known as Simpson's or Yule's paradox.

The confluence of log-linear model theory and the desire to report marginals means that a statistical agency or the researchers carrying out a clinical trial or epidemiological investigation could possibly share partial information in the form of marginals with users (researchers) and still protect the confidentiality of the data in a multi-way table.

		R	1	2	3
C	S	T			
1	1	1	3	20	5
1	1	2	11	14	8
1	2	1	3	14	12
1	2	2	6	13	5
2	1	1	12	12	0
2	1	2	11	10	0
2	2	1	3	9	4
2	2	2	6	9	3

Table 1. Results of clinical trial for the effectiveness of an analgesic drug. Source: Koch et al. (1983).

3 2^4 Example

In Table 1, we present data from Koch et al. (1983) on the results of a clinical trial on the effectiveness of an analgesic drug for patients of two different statuses and from two different centers. Here, the example draws on the discussion in [12] and extends it with extra technical details. For shorthand, denote Status as [S], Center as [C], Treatment as [T] with Active=1 and Placebo=2, and Response [R] with Poor=1, Moderate=2, Excellent=3. Given that individuals in the clinical trial form a "population", confidentiality questions will focus on the potential harm associated with the release of information on the four cells with counts of "3" in the table, corresponding to two sets of three individuals in 'Center 1', and two sets of three individuals in 'Center 2.' The following analytical question is of interest: What is the effect of the treatment on the response, controlling for the other two variables? More specifically, we are interested in answering: What data can we release from this table that would allow an analyst to make proper inferences about the substantive question of interest, without fully disclosing the four cells containing counts of 3.

In particular, the analyst needs the margins to go with a "good" log-linear model that fits the data well. To check on the model fit we need more data than the minimal sufficient statistics for the model itself, i.e., more margins or at least some of higher dimension. This more elaborate data release that corresponds to a more complex log-linear model and we can then compare the expected values under the simpler model with those under the more complex one.

We first illustrate analysis and evaluation of disclosure risk where "partial data releases" are marginal tables of counts. Then in Section 3.2 we consider extension of these partial data releases that include tables of conditional probability values corresponding to selected marginals.

3.1 Release of Margins

For the data in Table 1, because this is a randomized clinical trial we need to include the margin for the three explanatory variables, i.e., Center by Status by Treatment—we use the notation [CST] as a shorthand for this three-way margin. And virtually all model search procedures would narrow the focus to the two models:

1. [CST] [CSR],
2. [CST][CSR][TR].

both of which fit the data well. Model 1 is a special case of model 2 and the likelihood ratio test for the difference between them takes the value $\Delta G^2 = 5.4$ with 2 degrees of freedom, a value that is not significant at the 0.10 level when compared with a chi-squared distribution. Thus one might reasonably conclude that the effect of the treatment on the response is explained through the interactive effect of Center and Status. A key point for the present purposes is that we need three sets of marginal totals to make this inference: [CST], [CSR], and [TR].

For an $I \times J$ table with table entries x_{ij}, given row margins x_{i+} and column margins x_{+j}, the bounds have the following form:

$$\min\{x_{i+}, x_{+j}\} \geq x_{ij} \geq \max\{0, x_{i+} + x_{+j} - x_{++}\}. \quad (1)$$

First we treat the data in Table 1 as if they come from an 8×3 table and compute the Fréchet bounds, given in Table 2. There are 6,718,227,637,086,252 tables with the same sets of marginal totals and across all of them these are the maximum and minimum values for each of the cell counts. We note that all of the lower bounds in this example are 0 even though this need not be the case in general. Since the uppers bounds are far from the lower bounds and since these bounds correspond to an extremely large collection of tables, an intruder cannot use them to make strong inferences about potentially small cell entries.

The rows of Table 1 correspond to three variables and thus we have computed the bounds for a four-way table given the margins [CST] and [R]. The Fréchet bounds then for x_{ijkl} cells are bounded below by,

$$x_{ijkl} \geq \max\{0, x_{ijk+} + x_{+++l} - x_{++++}\}, \quad (2)$$

and above by,

$$\min\{x_{ijk+}, x_{+++l}\} \geq x_{ijkl}. \quad (3)$$

Over the past decade, these bounds and the ideas on bounds have been extended to multi-way tables given two or more, possibly overlapping margins, and not surprisingly these extensions are linked to the theory of log-linear models. Many special cases have explicit formulas like those in equations (2) and (3). For other cases various numerical procedures can produce bounds example by example. [9, 8, 7, 11] give many of the details.

		R	Poor	2	3
C	S	T			
1	1	1	[0,28]	[0,28]	[0,28]
1	1	2	[0,33]	[0,33]	[0,33]
1	2	1	[0,29]	[0,29]	[0,29]
1	2	2	[0,24]	[0,24]	[0,24]
2	1	1	[0,24]	[0,24]	[0,24]
2	1	2	[0,21]	[0,21]	[0,21]
2	2	1	[0,16]	[0,16]	[0,16]
2	2	2	[0,18]	[0,18]	[0,18]

Table 2. Upper and lower bounds for cell entries in Table 1 given the [CST] and [R] margins.

		R	1	2	3
C	S	T			
1	1	1	[0,14]	[1,28]	[0,13]
1	1	2	[0,14]	[6,33]	[0,13]
1	2	1	[0,9]	[3,27]	[1,17]
1	2	2	[0,9]	[0,24]	[0,16]
2	1	1	[2,21]	[3,22]	[0,0]
2	1	2	[2,21]	[0,19]	[0,0]
2	2	1	[0,9]	[0,16]	[0,7]
2	2	2	[0,9]	[2,18]	[0,7]

Table 3. Upper and lower bounds for entries in Table 1 given the [CST] , [CSR], and [TR] margins.

If a cell count is small and the upper bound is close to the lower bound, the intruder knows with a high degree of certainty that there is only a small number of individuals possessing the characteristics corresponding to the cell. This may pose a risk of disclosure of the identity of these individuals.

For the data in Table 1, we observed earlier that the four cell entries of “3” pose potential disclosure risk and we would like to protect them by releasing only subsets of the data in the form of marginal totals. We have explored the possible bounds associated with the release of the [CST] margin and all other possible sets of margins. Table 3 contains the bounds for the sets of margins needed to fit and compare the two log-linear models of analytical interest, [CST][CSR] and [CST][CSR][TR] and now we clearly see several cells with positive lower bounds. As before, all of the upper bounds are reasonably far from the

		R	1	2	3
C	S	T			
1	1	1	[0,13]	[10,23]	[5,5]
1	1	2	[1,14]	[11,24]	[8,8]
1	2	1	[0,6]	[7,20]	[9,16]
1	2	2	[3,9]	[7,20]	[1,8]
2	1	1	[2,15]	[9,22]	[0,0]
2	1	2	[8,21]	[0,13]	[0,0]
2	2	1	[0,6]	[3,16]	[0,7]
2	2	2	[3,9]	[2,15]	[0,7]

Table 4. Upper and lower bounds for entries in Table 1 given the [CST], [CSR], and [STR] margins.

		R	1	2	3
C	S	T			
1	1	1	[0,6]	[9,28]	[0,13]
1	1	2	[8,14]	[6,25]	[0,13]
1	2	1	[0,6]	[6,25]	[4,17]
1	2	2	[3,9]	[2,21]	[0,13]
2	1	1	[6,15]	[9,18]	[0,0]
2	1	2	[8,17]	[4,13]	[0,0]
2	2	1	[0,9]	[3,12]	[4,4]
2	2	2	[0,9]	[6,15]	[3,3]

Table 5. Upper and lower bounds for cell entries in Table 1 given the [CST], [CSR], and [CTR] margins.

		R	1	2	3
C	S	T			
1	1	1	0.107	0.714	0.179
1	1	2	0.333	0.424	0.242
1	2	1	0.103	0.483	0.414
1	2	2	0.250	0.542	0.208
2	1	1	0.500	0.500	0
2	1	2	0.524	0.476	0
2	2	1	0.188	0.563	0.250
2	2	2	0.333	0.500	0.167

Table 6. Observed conditional probability values for [R|CST] from data in Table 1.

lower bounds except for the (2,1,2,3) cell where the upper and lower bounds are now 0, and perhaps the (2,2,1,3) and (2,2,2,3) cells where the bounds are [0,7] in both tables. If we released the [CST], [CSR], and [TR] margins an intruder would be far from certain what entries belonged in the 4 cells that actually contain the value 3.

While it is true that releasing the [CST], [CSR], and [TR] margins allows others to carry out the likelihood ratio test to assess the effect of the treatment on the response (c.f. [11]), releasing even more information would be desirable. There are two additional three-way margins to consider: [CTR] and [STR]. If we also release [STR], then we get the bounds in Table 4, which show that the (1,1,1,3) cell which contains a count of 5 and the (1,1,2,3) cell which contains a count of 8 are identified with certainty. If instead we add the three-way margin [CTR], from Table 5 we see that the count of 4 in the (2,2,1,3) cell and the count of 3 in the (2,2,2,3) are revealed with certainty. So it may be possible to release a bit more information in the form of the the [STR] margin but the release of [CTR] is potentially problematic.

3.2 Release of Marginals and Conditionals

The idea of partial releases in the form of sets of margins can be extended to other types of data summaries such as marginal tables of rates, that is conditional or relative observed frequencies for a margin. Until recently nothing was known, however, about the effect of their release on confidentiality. Furthermore, releasing of conditional distributions for higher-dimensional contingency tables could be useful for researchers interested in assessing causal inference using directed acyclic graphs while still maintaining confidentiality. We are currently exploring the theory associated with such releases (see [18, 17]) and illustrate a few of the ideas here.

If we return to our example, it is clear that we can explore the question of treatment effect by using the empirical conditional probability values from a full conditional distribution of R given C, S, and T — we use the notation [R|CST] to represent this information (see Table 6). For example, the observed conditional probability value in the (1,1,1,1) cell is $\frac{3}{28} = 0.107$. If we also have the margin [CST], we can clearly reconstruct the full four-way table! Given [R|CST] there are 7,703,002 tables all having the same conditional probability values [R|CST]. The number of tables is calculated by using tools from algebraic geometry and LattE software [6]. Linear programming relaxation bounds are given in Table 7. The bounds seem wide enough for the small cell counts that the agency might be tempted to conclude that it is safe to release this information. The smallest bound of [1, 17.03] is for the cell count of “3” in the cell (1,1,1,1). It is important to notice that this single release reveals the zero counts in the table unlike the margins, where we need 3 three-way margins to learn the position of zeros. While the disclosure of zero, for this particular example, does not have much impact on an overall confiden-

		R	1	2	3
C	S	T			
1	1	1	[1, 17.03]	[6.67, 113.55]	[1.67, 28.39]
1	1	2	[1.38, 51.26]	[1.75, 65.23]	[1, 37.28]
1	2	1	[1, 16.48]	[4.67, 76.91]	[4, 65.92]
1	2	2	[1.2, 38.61]	[2.60, 83.66]	[1, 32.18]
2	1	1	[1.10, 79.44]	[1, 72.26]	0
2	1	2	[1.10, 79.48]	[1, 72.26]	0
2	2	1	[1, 29.06]	[3, 87.17]	[1, 38.74]
2	2	2	[2, 51.89]	[3, 77.83]	[1, 25.94]

Table 7. Linear programming relaxation bounds for cell entries in Table 1 given [R|CST] conditional probability values.

tiality risk, for larger and sparser k-way tables the presence of a large fraction of zero cells that are identified as such this may substantially increase the risk by constraining the nonzero cell values even more.

Suppose we try to relax bounds by releasing slightly more information via conditionals [RT] and [CRS]. Releasing either [R|T] and [CST] is equivalent to releasing [RT] and [CST]. Next suppose that [CSR] and [TR] are available and that the researchers also release [T|CS] believing that the relative frequencies offer more protection than the three-way marginal [CST]. It is easy to see that this is equivalent to publishing the [CST], [CSR] and [TR] margins; from [CSR] we can get the [CS] margin which together with [T|CS] gives the [CST] margin. We also get the same bounds by publishing [CS|T] along with [CSR] and [TR]! What is happening in this example is that the release of the margin [CSR] allows for the reconstruction of other margins from their corresponding conditionals.

If we release [R|T] and [R], we get [RT] because the number of levels in [R], three, is greater than in [T] which is two. This result comes from a theorem described in [17]. On the other hand, theoretically, [R|CS] and [R] will not uniquely identify [RCS] because the number of levels in [R] is not greater than in [CS] which is four. The number of tables for [RCS] is 31,081,397,760,000, and for [R|CS] is 31,081,579,235,840. As we see in Table 8 the linear programming relaxation bounds for releasing the conditional [R|CS] instead of the margin [RCS] are much wider. For example, the upper linear programming bound for (1,1,1,1) cell for [R|CS] is 37.42 while for [RCS] is 14. If we relied on these bounds, the agency would mistakenly conclude that it is safer to release the conditional. However, after computing the sharp bounds for [R|CS] we find the same bounds as for the margin! We believe that the bounds coincide because of the small sample size. The gap between the linear programming bound and the sharp bound, for example, for the cell (1,1,1,1) is 23.42. For some of the other cells the gap is much wider. This example demonstrates that it would be a mistake to use linear programming in our setting.

		R	1	2	3
C	S	T			
1	1	1	[0,14]	[0,34]	[0,13]
			[1, 37.42]	[1, 92.31]	[1, 34.68]
1	1	2	[0,14]	[0,34]	[0,13]
			[1,37.42]	[1,74.73]	[1,34.68]
1	2	1	[0,9]	[0,27]	[1,17]
			[1, 27.84]	[0, 57.10]	[0,53.47]
1	2	2	[0,9]	[0,27]	[0,17]
			[1, 27.84]	[1,85.51]	[1,53.48]
2	1	1	[0,23]	[0,22]	[0,0]
			[1, 32.22]	[1,78.36]	0
2	1	2	[0,23]	[0,22]	[0,0]
			[1,75.04]	[1,11.23]	0
2	2	1	[0,9]	[0,18]	[0,7]
			[1,43.40]	[1, 87.81]	[1, 33.54]
2	2	2	[0,9]	[2,18]	[0,7]
			[1,43.40]	[1, 87.81]	[1, 33.54]

Table 8. Sharp upper and lower bounds for cell entries in Table 1 given the [CSR] margin, and linear programming relaxation bounds given [R|CS] conditional probability values.

In this example, the larger space of tables for conditionals did not produce larger sharp bounds. However, the difference in the number of tables, can have potential implications for estimating distributions over the space of solutions. We are beginning to explore this question (e.g., see [17]).

3.3 Association Rules and Marginal and Conditional Releases

A goal of data mining is knowledge and information discovery in large, complex databases. Association rules are type of knowledge discovery method [1]. If A and B are two disjoint attributes from a database D, then an association rule is an implication of the form $A \longrightarrow B$ parametrized by two measures: *confidence* and *support*.

More recently in privacy preserving data mining literature, the problem of sharing the data while concealing some association rules, has been addressed ([10, 16, 5]). The question is what rules should we hide? Our results offer some insight as to what association rule(s) should not be released.These insights include an explicit link, which we often find lacking in data-mining literature, between association rules and some statistical concepts such as joint and conditional distributions. Here we only briefly describe a link between association rules and marginal and conditional tables, and refer reader to [11] for more details.

Consider the two subsets CS and R from the joint table [CRST]. Equivalently, we can think of CS and R as disjoint sets of attributes from a database. The marginal table with probability values [CSR] is an example of the *support* $s = \frac{\#(CS \cap R)}{\#(CSRT)} = P(CS, R)$, while the table with conditional probability values [R|CS] is the *confidence* $c = \frac{\#(CS \cap R)}{\#(CS)}$ for a rule $R \longrightarrow CS$. Thus the type of calculations and analyses described in the previous section can be applied to association rules as well. The width of the bounds on the sensitive cells (Table 8) or the knowledge about the space of tables satisfying this association rule, can be used to determine if it is safe to release the rule. The analyses can be extended to compatible rules across databases, and help address the question whether the combination of pieces of information from different systems compromises the confidentiality of individual information.

4 Conclusions

The moral of this example is that: when we are faced with a relatively sparse multi-way contingency table containing small counts that might disclose sensitive information about individuals with reasonably high probability, we still are able to release enough of the marginal totals from the table to allow a statistician to explore relevant questions of inference. The example also demonstrates that conditionals can become a useful tool for releasing more data than one might otherwise have considered based on marginals alone.

This example illustrate that full conditionals (corresponding to the full table) and some small conditionals (corresponding to a marginal table) reveal the zero counts. While the following is not the case in example presented here, [17] demonstrate that depending on the position of the zero counts in relation to other small counts, the sensitive cells could be easily identified; despite the large number of tables it could be potentially too risky to release certain conditionals.

The theoretical results of [17] imply that the space of tables for given small conditional is at least as large as the space for the corresponding margin. They also imply that the bounds for a given conditional are at least as large as for the corresponding margin. The example in this paper illustrate this point; however, it also illustrates that in most cases the sharp bounds are the same, and that linear programming grossly overestimates these bounds. This observation argues for caution in using linear programming bounds to estimate disclosure risk for these type of non-linear constraints.

The simple example in this paper offers some insights as to how to evaluate release of association rules. Our results are also relevant to assessing possible inferences an intruder can make about confidential categorical data following the release of information on one or more association rules. It is important to recognize here that decision not to release certain information is also informative; thus the assessment of disclosure risk and of possible inferences should be made conditional on released association rules and the discarded ones.

Acknowledgments

The author would like to thank Prof. Stephen E. Fienberg for his advice and support. The research reported here was supported in part by NSF grant IIS–0131884 to the National Institute of Statistical Sciences, while the author was in the Department of Statistics, Carnegie Mellon University.

References

[1] R. Aggrawal, T. Imielinski, and A. Swami. Mining association rules between sets in large databases. In *Proceedings of the ACM-SIGMOD International Conference on Management of Data*. Washington, DC., 1993.

[2] A. Agresti. *Categorical Data Analysis*. John Wiley & Sons, New York, 1990.

[3] B. Arnold, E. Castillo, and J. Sarabia. *Conditional Specification of Statistical Models*. Springer-Verlag, 1999.

[4] Y. Bishop, S. Fienberg, and P. Holland. *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Cambridge MA, 1975.

[5] C. Clifton, M. Kantarcioglu, X. Lin, J. Vaidya, and M. Zhu. Tools for privacy preserving distributed data mining, 2003.

[6] J. De Loera, D. Haws, R. Hemmecke, P. Huggins, J. Tauzer, and R. Yoshida. *A User's Guide for LattE v1.1*. University of California, Davis, 2003.

[7] A. Dobra. *Statistical Tools for Disclosure Limitation in Multi-Way Contingency Tables*. Ph.D. Thesis, Department of Statistics, Carnegie Mellon University, 2002.

[8] A. Dobra and S. Fienberg. Bounding entries in multi-way contingency tables given a set of marginal totals. In *Proceedings of Conference on Foundation of Statistical Inference and Its Applications, Jerusalem, Israel*. Springer-Verlag, to appear, 2002.

[9] A. Dobra, S. Fienberg, and M. Trottini. Assessing the risk of disclosure of confidential categorical data. In J. Bernardo, editor, *Bayesian Statistics 7, Proceedings of the Seventh Valencia International Meeting on Bayesian Statistics*. Oxford University Press., 2003.

[10] A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke. Privacy preserving mining of association rules. In *Proc. of* $8th$ *ACM IGKDD Intl. Conf. on Knowledge Discovery and Data Mining (KDD)*, July 2002.

[11] S. Fienberg. Data mining and disclosure limitation for categorical statistical databases. *Unpublished Manuscript.*

[12] S. Fienberg and A. Slavkovic. Making the release of confidential data from mutli-way tables count. *Chance*, 17(3):5–10, 2004.

[13] S. Fienberg and A. Slavkovic. Preserving the confidentiality of categorical statistical data bases when releasing association rules. *Data Mining and Knowledge Discovery*, 2004 to appear.

[14] A. Gelman and T. Speed. Characterizing a joint probability distribution by conditionals. *Journal of the Royal Statistical Society. Series B*, 55(1):185–188, 1993.

[15] G. Koch, J. Amara, S. Atkinson, and W. Stanish. Overview of categorical analysis methods. *SAS-SUGI*, 8:785–795, 1983.

[16] Y. Saygin, V. S. Verykios, and A. K. Elmagarmid. Privacy preserving association rule mining. In *RIDE*, 2002.

[17] A. Slavkovic. *Statistical Disclosure Limitation Beyond the Margins: Charcterization of Joint Discrete Distributions for Contingency Tables.* Ph.D. Thesis, Department of Statistics, Carnegie Mellon University, 2004.

[18] A. B. Slavkovic and S. E. Fienberg. Bounds for cell entries in two-way tables given conditional relative frequencies. In J. Domingo-Ferrer and V. Torra, editors, *Privacy in Statistical Databases– PSD'2004, Lecture Notes in Computer Science No. 3050*, pages 30–43. Springer-Verlag, 2004.