Chapter 12

A Survey of Statistical Approaches to Preserving Confidentiality of Contingency Table Entries

Stephen E. Fienberg

Department of Statistics, Machine Learning Department, and Cylab,
Carnegie Mellon University
Pittsburgh PA 15213-3890, U.S.A.

fienberg@stat.cmu.edu

Aleksandra B. Slavkovic

Department of Statistics
Pennsylvania State University
University Park PA 16802, U.S.A.

sesa@stat.psu.edu

Abstract   In the statistical literature, there has been considerable development of methods of data releases for multivariate categorical data sets, where the releases come in the form of marginal and conditional tables corresponding to subsets of the categorical variables. In this chapter we provide an overview of this methodology and we relate it to the literature on the release of association rules which can be viewed as conditional tables. We illustrate this with two examples. A related problem, "association rule hiding" is often independently studied in the database community.

Keywords: Algebraic geometry, association rules, conditional tables, contingency tables, disclosure limitation, marginal tables, privacy preservation.

12.1   Introduction

The cross-classification of individuals or other units according to multiple categorical variables produces multi-way tables of counts, better known as contingency tables. There is an extensive statistical literature on the analysis of such tables, e.g., see [1], [4], [15], and [25]. When the number of variables is large, the cells of the resulting contingency tables often contain a substantial

number of small counts. These pose potential problems of disclosure risk. One strategy for protecting the confidentiality of the entries in such circumstances has been the release of subsets of the data in the form of marginal and conditional tables. In this chapter we provide a survey of the literature that explains the effectiveness of this strategy both for the protection of confidentiality and utility in connection with log-linear and logit model methods.

The search for association rules in datamining focuses on the detection of relationships or "associations" between specific values of categorical variables in large data sets, i.e., multi-way contingency table. This search requires working with observed conditional distributions for an outcome variable or feature given one or more explanatory variables. Thus the search for association rules requires the construction of marginal and then conditional tables from the full contingency table, i.e., datamining for association rules in effect involve the efficient construction and storage of marginal and conditional tables, e.g., see [2] and [27]. Different datamining methods use these marginal and conditional tables in different ways. Some approach the problem by focusing solely on low-dimensional marginal tables while others utilize the full power of log-linear and logit models and use higher-dimensional marginal tables. The methods we describe here are relevant to both approaches.

Our methods described here relate to "association rule hiding" problem studied by the privacy-preserving data mining and database community. In this volume, Verykios et al. [39] give a survey of association rule hiding methods. They do not describe any related statistical disclosure limitation methods. What they refer to as "data hiding" in SDL literature is labeled usually as data masking. They point out that in general the sensitivity of the rules is determined by security administrator, while the focus is on efficiency and algorithmic approaches for hiding of the rules rather than the usability. Our methodology offers a way for detecting a sensitivity of a rule based on the data utility relevant for valid statistical analysis.

12.2   The Statistical Approach Privacy Protection

Statisticians have approached this search problem in the following fashion. Suppose we have k-way cross-classification of counts arising from a sample of size n from a large population of size N, e.g., the size of the US adult population, or that from California. We want to report as much information from this table as possible without releasing data that would allow an intruder to identify one or more individuals with substantial probability. For the release to be useful, an analyst needs to be able to use what is released to reach some statistical conclusions that she would have tried to reach with the full k-way array.

Statisticians often define usefulness in this case in terms of fitting and interpreting the parameters in a log-linear model. The relevant quantities for doing this are marginal totals that correspond to the highest order interaction terms—these are the "best" data summaries, or *minimal sufficient statistics* for the mdel. The difficulty is: which log-linear model? To understand this we must do some form of model search, e.g., based on a search of model space and using some criterion like the Bayesian Information Criterion (BIC), e.g., see the paper by [26]. Releasing just those minimal sufficient margins that correspond to the model which minimizes BIC does not let the analyst check the fit of the model relative to others so we may wish to release even more i.e., higher-order margins that include these. When we fit the model we begin with the presence of certain interaction terms and we estimate their value along with asymptotic standard errors. The latter typical involve functions that are sums of inverses of the values in the minimal sufficient margins. This is extremely important since BIC and other criteria pick models where the asymptotic variance of the discarded terms are the same order of magnitude as the estimates. The implication is that for "good" log-linear models the minimal sufficient margins tend to have substantial sized counts typically on the order of 10 or more, and sometimes 100 or more! They will almost never have zeros in them, because that yields special estimability and fit problems and they will rarely include very small counts.

To check on privacy protection, we ask whether the information in the marginal and conditional tables used in the construction of association rules discloses confidential data about individuals or units represented in the full multi-way contingency table. Much of the statistical focus has tended to be on identification of small cell counts, e.g., "1" and "2." The first order of business is to assess the contribution from sampling. Roughly speaking, the probability that an individual record that is unique in the sample is also unique in the population from which the sample was drawn equals the sampling fraction, n/N, e.g., see [18]. Thus for a sample of size 2,000 drawn from a population of 200,000,000 adults the sampling fraction is 2,000/200,000,000 or 0.00001. The bottom line therefore is that sampling protects, just not absolutely or even in the formal sense that computer scientists have suggested, e.g., see [13]. Thus we go further and look directly at the table and compute several quantities, such as upper and lower bounds for the cell counts in the k-way table, or the number of possible tables satisfying the marginal or possibly marginal and conditional constraints, or we might look at the distribution over these possible tables to assure themselves that the probabilities don't lump up on just a few of the values between the bounds, e.g., see [9]. We provide some details in the remainder of the chapter.

12.3   Datamining Algorithms, Association Rules, and Disclosure Limitation

Association rules are often described using a market-basket metaphor that assumes that there are a large number of products that can be purchased by the customer, either in a single transaction, or over time in a sequence of transactions. Customers fill their basket with only a fraction of what is on display—i.e., with a sample. Association rules can be extracted from a database of transactions, to determine which products are frequently purchased together. For example, one might find that A ="purchases of diapers" typically coincide with B ="purchases of dog food" in the same basket. We then evaluate the usefulness of the rule using some form of statistical summary such as "support" and "confidence". For example,

Rule form: A $\Rightarrow$ B [support, confidence]

Example: buys(x, "diapers") $\Rightarrow$ buys(x, "dog food") [0.55%, 68%]

More generally, we have k-tuples based on k possible product types and the transactions or market baskets produce counts for a k-way contingency table with attributes corresponding to the presence or absence of the product types. Our new goal is to discover association rules involving the variables that make up this contingency table. For an association rule of the form: $\{A, B, C, \dots\} \Rightarrow \{E, F, G, \dots\}$, we define:

Confidence (accuracy) of $A \Rightarrow B$**:** $P(B|A)$ = (# of transactions containing both A and B) / (# of transactions containing A).

Support (coverage) of $A \Rightarrow B$**:** $P(A, B)$ = (# of transactions containing both A and B) / (total # of transactions)

There are many other possible criteria for assessing the usefulness of rules, e.g., [38] uses a variation on support and confidence while [29] and [30] use chi-square statistics for independence and conditional independence computed on the marginal tables.

Machine learning approaches often attempt to treat every possible combination of attribute values as a separate class, learn rules using the rest of attributes as input and then evaluate them for "support" and "confidence". This essentially involves examining all possible marginal tables corresponding to the attributes. The problem is that this approach tends to be computationally intractable, i.e., there are too many classes and consequently, too many rules. Alternatively criteria involve looking for rules that exceed pre-defined support (minimum support) and have high confidence. If we include among the objects of interest the negations of the items, or in statistical terms all of the categories of the variables, then in fact we are simply relying on full marginal and conditional tables for empirical evaluation and rule search. We reiterate this key

point: *Support is a marginal table, and confidence is a conditional table, both corresponding to a subset of variables making up the full table.*

There is a major issue about what we mean by "the release of association rules." Many of the authors in the datamining literature have taken this notion to simply mean announcing or releasing the form of the rule, i.e., the variables involved. We believe that this is essentially a vacuous approach, since using the association rule requires the data that allow one to make predictions. To us, releasing a rule means releasing the data on which it is based, i.e., the corresponding conditional and/or marginal table. The more complex the rules and the more rules the greater the risk of disclosure of individual information and thus the violation of confidentiality promised to and the privacy of those whose data are represented in the table. The real differences between between the machine learning literature on association rules and the statistical literature on contingency tables is how they deal with the marginal and conditional tables, and what is reported or shared with others. We address the latter point in the next section.

Fienberg and Slavkovic [20] describe results based on release of exact marginals and conditionals that can help us determine which rules to hide in order to preserve privacy but to allow sufficient information for statistical inference; in this paper we highlight some of those results. In the computer science literature there are a number of alternative approaches, e.g., perturbing the full data array as proposed by [14], [28], and [23].

12.4 Estimation and Disclosure Limitation for Multi-way Contingency Tables

There is a separate literature on privacy and confidentiality in categorical statistical data bases that approaches a number of the issues raised directly or indirectly in the datamining literature but with a different and heavier emphasis on the tradeoff between preserving confidentiality and assuring utility of the released data in the sense of allowing for proper statistical inferences.

For the present purposes we can group the approaches in the statistical literature into perturbational and aggregation or collapsing. For continuous data, aggregation methods go under names such as micro-aggregation and k-anonymity. For categorical data, aggregation typically involve combining categories of variables with more than two values, but a special example of collapsing involves summing over variables to produce marginal tables. Thus instead of reporting the full muti-way contingency table we might report multiple collapsed versions of it. The release of multiple sets of marginal totals has the virtue of allowing statistical inferences about the relationships among the variables in the original table using log-linear model methods. Barak et al. [3]

present a novel approach to contingency tables using perturbation and aggregation ideas.

Notation and Definitions.   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 $\mathcal{D} = [d_1] \times ... \times [d_k]$. A k-way contingency table of counts, $\mathbf{n} = \mathbf{n}(i), i \in \mathcal{D}$, is a k-way dimensional array of non-negative integers such that each cell entry $\mathbf{n}(i) = \#\{X = i\}$ represents the number of times the configuration i is observed in a series of independent realizations of $X_1, ..., X_k$. The data of interest are 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 a 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 $p(x_A) = \sum_{K \backslash A} p(x_K)$, 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 12.1).

Suppose that that we observe an arbitrary set of conditional and marginal tables, $\mathcal{T}$. We define the *fiber* $\mathcal{F}_\mathbf{t}$ as a set of all k-way non-negative integer tables that satisfy the constraints $\mathcal{T} = \mathbf{t}$. Consider a sublattice $\mathcal{L}_\mathbf{t}$ of $\mathbb{Z}^\mathcal{D}$ that depends on a collection $\mathcal{T}$ and a finite subset $\mathcal{B}_\mathbf{t}$ (e.g., a Markov basis is the smallest such subset) of $\mathcal{L}_\mathbf{t}$.

Each element of $\mathcal{B}_\mathbf{t}$, $\mathbf{z}$, can be thought of as a contingency table with values in $\mathbb{Z}^\mathcal{D}$, and each is called a *move* that satisfies $A_\mathbf{t}(\mathbf{n} + \mathbf{z}) = A_\mathbf{t}\mathbf{n}$, where $A_\mathbf{t}$ is a matrix that defines the constraints $\mathcal{T} = \mathbf{t}$ imposed on table $\mathbf{n}$. The most important property of Markov bases, for our purposes, is that they *connect* all tables satisfying the same set of constraints; thus they can be used for data swaps and for building a connected Markov chain. Helpful references for tools on algebraic statistics, including the calculation and use of Markov and Gröbner bases, are [6], [34], and [31].

Log-linear Models.   Consider an $I \times J \times K$ table of observed counts $\{n_{ijk}\}$, with corresponding estimated expected values, $\{m_{ijk}\}$ under a multinomial

sampling model. The saturated log-linear model for $\{m_{ijk}\}$ takes the form

$$
\begin{aligned}
\log m_{ijk} \; =\; & u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} \\
& + u_{13(ik)} + u_{23(jk)} + u_{123(ijk)},
\end{aligned}
\tag{12.1}
$$

where each subscripted u-term sums to zero over any subscript, e.g.,

$$
\sum_i u_{123(ijk)} = \sum_j u_{123(ijk)} = \sum_k u_{123(ijk)} = 0.
$$

We get *unsaturated* models from (12.1) by setting sets of u-terms equal to zero, e.g., if we set

$$
u_{123} = 0 \text{ for all } i, j, k,
\tag{12.2}
$$

we have the model of no second-order interaction. A logit model involves conditioning on a marginal total and for all practical purposes can be thought of as equivalent for the present purposes to the corresponding log-linear model which includes the u-terms that correspond to the marginal conditioned upon. These ideas and the definition of log-linear models generalize naturally from 3 to k dimensions.

Estimation and Assessing Goodness-of-Fit.   We have the following key features associated with inference for log-linear models:

- The relevant statistical models focus on simultaneous interactions among sets of variables that define the contingency table.

- Special subsets of these models include the family of conditional independence models and the family of graphical models, which involve simultaneous occurrence of conditional independencies. For more details on graphical models in statistics see [25], and in machine learning see [22].

- The minimal sufficient statistics (i.e., sufficient data summaries) for a log-linear model are the marginal totals corresponding to the highest-order interaction terms in the model. For example, for the no second-order interaction model for three-way tables in equation (12.2) above, the minimal sufficient statistics are the three sets of two-way marginal totals, $\{n_{ij+}\}$, $\{n_{i+k}\}$ and $\{n_{+jk}\}$ corresponding to $\{u_{12(ij)}\}$, $\{u_{13(ik)}\}$, and $\{u_{23(jk)}\}$, respectively.

- The maximum likelihood estimates for the expected cell values are found by setting the minimal sufficient statistics equal to their expectations. For example, for the no-second-order interaction model for

three-way tables in equation (12.2) above:

$$\begin{aligned} \hat{m}_{ij+} &= n_{ij+} \text{ for all } i, j, \\ \hat{m}_{i+k} &= n_{i+k} \text{ for all } i, k, \\ \hat{m}_{+jk} &= n_{+jk} \text{ for all } j, k. \end{aligned}$$

- Maximum likelihood estimates for expected cell values under logit models are the same as corresponding log-linear models which include terms associated with the fixed margins that the logit model conditions upon, e.g., see the discussion in [4] and [15].

- Decomposable log-linear models are graphical models for which the maximum likelihood estimates have an explicit closed-form expression. They correspond to triangulated graphs. See [25].

- Standard methods of goodness-of-fit allow the user to assess how well the model and its minimal sufficient statistical margins can explain or reconstruct the original cell counts. These include goodness-of-fit criteria such as likelihood ratio statistics for separate models or for comparing nested models, and penalized criteria such as the BIC, e.g., see Madigan and Raftery [26]. In particular, the likelihood ratio test for comparing a pair of nested log-linear models is expressible in terms of the minimal sufficient marginals of the more complex model, a result implicit in formulae in [4] and [25], and made explicit in [16].

Disclosure Limitation and Bounds on Cell Counts.   To check on the disclosure limitation provided by releasing only a subset of marginal totals one can consider the information in the margins for the construction of bounds for the individual cell entries. Consider an $I \times J$ table with entries $\{n_{ij}\}$ and row margins $\{n_{i+}\}$ and column margins $\{n_{+j}\}$. Then it is well-known that

$$\min\{n_{i+}, n_{+j}\} \geq n_{ij} \geq \max\{0, n_{i+} + n_{+j} - n_{++}\}, \tag{12.3}$$

and that these bounds, also known as Fréchet bounds, are sharp. Now consider the situation where instead of releasing a full k-way contingency table, we release a set of lower-dimensional marginal totals from it. Any contingency table with non-negative integer entries and fixed marginal totals is a lattice point in the convex polytope defined by the linear system of equations induced by the released marginals. The constraints given by the values in the released marginals induce upper and lower bounds on the interior cells of the initial table. In principle, we can obtain these bounds by solving the corresponding linear programming (LP) problem, but in general this is an NP-hard problem. Dobra and Fienberg [7, 8] have derived explicit formulas for several interesting sets of margins corresponding to special subsets of graphical log-linear models

and they have proposed strategies for using these methods to find sets of margins that would not allow an intruder to make sharp inferences about the entries in the original table. In particular, [7] provide simple and explicit bounds formulas that are generalizations of equation (12.3) when the margins correspond to the minimal sufficient statistics of decomposable log-linear models.

It is important to recognize that as the number and size of the released margins grow, we tighten the bounds on the cells in the table (based on in increasing amount of information available) and the tightening takes on subtly complex forms because of the interlocking structure of the margins. Slavkovic [31] explored the form of linear and integer programming (IP) bounds for given conditionals. We illustrate the bounds approach in the present paper and describe some extensions to it involving combinations of margins and conditionals.

A major theme in the literature on disclosure limitation deals with the trade off between disclosure risk and data utility. See especially [36], and selected papers in [10]. Duncan with a variety of coauthors has stressed a graphical representation for this trade-off which they call the R-U map, e.g., see [12] for a discussion in the context of categorical data. Trottini and Fienberg [35] take the trade-off formalism several steps further and embeds it in a fully Bayesian decision-theoretic framework. Following [16] we adopt a somewhat more informal assessment process by considering maximal releases of marginal and conditional tables subject to limited disclosure risk in terms of bounds on cell entries in the table.

Releasing Marginal and Conditional Tables. Because data from both marginal and conditional tables are potentially of interest in assessing and reporting association rules, we need to understand how they differ in terms of the information they convey about the entries in multi-way contingency tables. For example, we want to do is check to see whether or not sets of marginal and conditional distributions for a contingency table are sufficient to uniquely identify the existing joint distribution. If so, we might as well release the full table!

The joint distribution for any two-way table is uniquely identified by any of the following sets of distributions: (1) $P(X_1|X_2)$ and $P(X_2|X_1)$, (2) $P(X_1|X_2)$ and $P(X_2)$, or (3) $P(X_2|X_1)$ and $P(X_1)$. Cell entries are allowed to be zero as long as we do not condition on an event of zero probability. Sometimes the sets $P(X_1|X_2), P(X_1)$ and $P(X_2|X_1), P(X_2)$ uniquely identify the joint distribution. The following result, due to [33] and [31], describes this situation and a generalization for a k-way table.

THEOREM 12.1 *(Slavkovic(2004)) Consider a k-way table and a collection* $\mathcal{T} = \{p_{A|B}, p_A\}$, *where $A, B \subset K$. If given matrices with conditional*

probability values have a full rank, and $d_A \geq d_B$, then $\mathcal{T}$ uniquely identifies marginal table p_{AB}.

Trivially, for bivariate tables, the joint probability distribution is the *support*, and thus along with the knowledge of sample size n, an association rule will reveal all cell counts. The above results also imply that releasing the *confidence* of a rule along with some marginal information, again will identify all entries in a table, although we are concerned primarily with the identification of cells with small counts.

Often, there are multiple realizations of the joint distribution for X, i.e., there is more than one table that satisfies the constraints imposed by them. Slavkovic [33], and [31] describe the calculation of bounds given an arbitrary collection of marginals and conditionals. They use LP and IP and discuss potential inadequacies in treating conditional constraints via LP. These results rely on the fact that 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.

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 example, equation (12.3) gives the bounds when all that is released are the two one-way marginals in a two-way table. When a single marginal or a single conditional is given, the cell's probability is bounded below by zero and above by a corresponding marginal or a conditional value. This translates into bounds for cell counts as long as we have the knowledge of sample size n which is implicitly given by releasing the observed margins, while it must be provided as an additional piece of information for the released conditional probabilites.

When the conditions of Theorem 12.1 are not satisfied, we can obtain bounds for cell entries, and in some two-way cases there are closed form solutions. These bounds are sharp for a set of low dimensional tables with nicely rounded conditional probability values. For higher dimensions linear approximations of the bounds could be very far off from the true solution for the table of counts, and thus these bounds may mask the true disclosure risk. To calculate sharp IP bounds, we need either nicely rounded conditional probability values, which rarely occur in practice, or we need the observed cell counts. The latter implies that in practice the database owner is the only one which can produce the "true" bounds in the case of the conditionals; see [32].

Using the tools of computational commutative algebra such as Gröbner and Markov bases in statistics, we can find feasible solutions to the constrained maximization/minimization problem. Some advantages of this approach are that (1) we obtain sharp bounds when the linear or integer program approach fails, and (2) we can use it to describe all possible tables satisfying given

constraints. In particular, a set of minimal Markov bases (moves) allows us to build a connected Markov chain and perform a random walk over the space of tables of counts that have the same fixed marginals and/or conditionals. This will allow us to either enumerate or sample from the space of tables via Sequential Importance Sampling (SIS) or Markov Chain Monte Carlo sampling. Some disadvantages of algebraic approach are that (1) calculation of Markov bases can be computationally infeasible for k-way tables, and (2) for conditionals, Markov bases are extremely sensitive to rounding of cell probabilities. A technical description of calculation and structure of Markov bases given fixed conditionals for two-way tables can be found in [31]. The reported results in the examples below rely on use of this methodology.

In a two-way case, we only deal with so called *full* conditionals because they involve all variables in the data base. Theorem 1.1 also describes the relationship between a conditional and a marginal table that involves a subset of variables from the data base. In other words, it describes a relationship between confidence and support for a rule that involves a subset of characteristics from a data base. Related theorems, their heuristics and constructions are illustrated in [31], and [20] who also further elaborate on relationships between a Markov basis set and the confidence and support, and implications for privacy. Here we focus on some of the consequences of these theorems relevant to establishing bounds on cells for evaluating potential disclosure.

One result implies that given the full conditional and the sample size n, the value of the moves can be used to determine if we have a unique solution. Other results imply that, for the same sample size n, the number of solutions for a fixed small conditional, $p_{A|B}$, is greater than or equal to the number of solutions we obtain by fixing the margin X_{AB}. This in turn should lead to wider bounds on some of the cell entries. We can study a specific subsets of Markov basis and determine if we are in the situation where the bounds given the small conditionals are the same as given its corresponding marginal. In a number of examples that we have examined to date, however, we have obtained the exact same bounds. This observation has led us to consider a set of conditions and heuristics that we can use in practice to determine when the bounds on cells given these two sets of released information are the same.

To evaluate the effect of releasing an association rule has on disclosure, we want to evaluate both confidence and support of the rule. The results of this section imply that it is sufficient to evaluate the support.

12.5 Two Illustrative Examples

12.5.1 Example 1: Data from a Randomized Clinical Trial

Koch et al. [24] report the data in Table 12.1 on the results of a randomized clinical trial on the effectiveness of an analgesic drug for patients of two

different statuses and from two different centers. We use a shorthand notation to describe variables and marginals from the full tables, denoting Status as [S], Center as [C], Treatment as [T] with levels Active $= 1$ and Placebo $= 2$, and Response as [R] with levels Poor $= 1$, $|hboxModerate = 2$, Excellent $= 3$. Given that individuals in the clinical trial form a "population," confidentiality questions focus on the potential harm associated with the release of information on the four cells with counts of "3" in this table, corresponding to two sets of three individuals in 'Center 1,' and two sets of three individuals in 'Center 2.' In [19, 20] we analyzed these data with a focus on the links between the uniqueness and bounds results to association rules. Here we add to these earlier analyses and findings.

We are interested in the effect of the treatment on the response, controlling for the other two variables. More specifically, we are interested in answering: Which association rules are safe to release and provide enough information for an analyst to make proper inferences about the question of interest. We could be interested in evaluating the following association rules: $T \Rightarrow R$, $CS \Rightarrow R$, $CST \Rightarrow R$, and $CS \Rightarrow T$. In particular, the analyst needs the margins, or support, to go with a "good" log-linear model that fits the data well.

First, consider an association rule, $CST \Rightarrow R$. Support is the joint marginal distribution of $[CRST]$ and confidence $[R|CST]$ is a table with conditional probability values (see Table 12.1). It is trivial to see that release of the support of this rule results in full disclosure since it is the full four-way table. These probabilities along with the sample size n uniquely identify all cell counts.

If we just release the confidence associated with this rule we can explore an important inferential question of treatment effect by using the empirical conditional probability values from a full conditional distribution of $[R|CST]$. If we also have the 3-way margin $[CST]$, we can clearly reconstruct the full 4-way table! Given $[R|CST]$ with sample size n, there are 7,703,002 tables all having

Table 12.1. Results of clinical trial for the effectiveness of an analgesic drug. Source: Koch et al. [24]. The second panel contains observed counts, and the third panel has corresponding observed conditional probability values for $[R|CST]$.

C	S	T	R 1	2	3	1	2	3
1	1	1	3	20	5	0.107	0.714	0.179
1	1	2	11	14	8	0.333	0.424	0.242
1	2	1	3	14	12	0.103	0.483	0.414
1	2	2	6	13	5	0.250	0.542	0.208
2	1	1	12	12	0	0.500	0.500	0
2	1	2	11	10	0	0.524	0.476	0
2	2	1	3	9	4	0.188	0.563	0.250
2	2	2	6	9	3	0.333	0.500	0.167

the same conditional probability values. We give LP relaxation bounds in Table 12.2. The tightest bound for the count of "3" is [1, 16.48] in cell (1,2,1,1). We supplement these bounds by sharp integer bounds which in this case can be calculated only by using observed counts (see [32]). These bounds are much sharper than the LP bounds, with some cell counts being uniquely identified such as the above mentioned cell (1,2,1,1). Thus both the LP bounds and the number of possible tables can be misleading in evaluating the disclosure risk. More generally, [31] shows that with knowledge of the sample size n full conditionals are too risky to be released, and clearly in this example the release of confidence $[R|CST]$ is not safe! Fienberg and Slavkovic [20] demonstrate that we could potentially approximate "safely" the knowledge of the release of this association rule by treating the data in Table 12.1 as if they come from a two-way 8×3 table and compute the Fréchet bounds for margins $[CST]$ and $[R]$ (c.f., Table 1.5 in [20]).

We note that this single conditional release reveals the zero counts in the table unlike the release of margins, where we needed 3 3-way margins to learn the position of zeros. While the disclosure of zero in this example does not have much impact on an overall confidentiality risk, for larger and sparser k-way tables the presence of a large fraction of 0 cells that are identified as such may substantially increase the risk of disclosure of sensitive non-zero cells by constraining them even more than the constraints that come directly from the marginals.

Because this is a randomized clinical trial, in order to perform meaningful statistical analysis, we need to include the three-way margin for the three explanatory variables, i.e., $[CST]$. Most model search procedures would narrow the focus to two models, Model 1: $[CST] [CSR]$, or Model 2: $[CST][CSR][RT]$, 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

Table 12.2. Second panel has LP relaxation bounds, and third panel has sharp IP bounds for cell entries in Table 1.1 given $[R|CST]$ conditional probability values

		R	1	2	3	1	2	3
C	S	T						
1	1	1	[1,17.03]	[6.67,113.55]	[1.7,28.4]	[3,6]	[20,40]	[5,10]
1	1	2	[1.4,51.26]	[1.75,65.23]	[1,37.28]	[11,11]	[14,14]	[8,8]
1	2	1	[1,16.48]	[4.67,76.91]	[4,65.92]	[3, 3]	[14,14]	[12,12]
1	2	2	[1.2, 38.61]	[2.60,83.66]	[1,32.18]	[6,12]	[13,26]	[5,10]
2	1	1	[1.10,79.44]	[1,72.26]	0	[1,18]	[1,18]	[0]
2	1	2	[1.10,79.48]	[1,72.26]	0	[11,11]	[10,10]	[0]
2	2	1	[1,29.06]	[3,87.17]	[1,38.74]	[3,9]	[9,27]	[4,12]
2	2	2	[2,51.89]	[3,77.83]	[1,25.94]	[2,12]	[3,18]	[1,6]

at the 0.10 level when compared with a chi-squared distribution with 2 degrees of freedom. Thus one might reasonably conclude that the effect of the treatment on the response is explained through the interactive effect of Center and Status.

Note that we need three sets of marginal totals to make this inference: $[CST]$, $[CSR]$, and $[RT]$. We can think of these marginal tables as supports of the following association rules: $CS \Rightarrow T$, $CS \Rightarrow R$, and $T \Rightarrow R$. Thus we want to evaluate the release of these marginals in combination with appropriate confidences, that is conditional tables such as $[T|CS]$, $[R|CS]$ and $[R|T]$. By applying theorems mentioned in Section 3, we can draw a number of interesting conclusions. For example, bounds on cells given only the confidence $[R|T]$ will be as wide or wider than given only the rule's support $[RT]$. The same observation holds for the other association rules we are considering in this example. This result implies that for each rule it should be sufficient to evaluate only its support to determine if the release is safe.

Sometimes, however, we only have partial information on a rule, such as its confidence, and want to evaluate those along with other data summaries. For example, if we release $[R|T]$ and $[R]$, Theorem 1.1, tells us that we have $[RT]$. On the other hand, theoretically, $[R|CS]$ and $[R]$ will not uniquely identify $[CRS]$ because the number of levels in $[R]$ is not greater than in $[CS]$ which is four. The number of tables for $[CRS]$ is 31,081,397,760,000, and for $[R|CS]$ is 31,081,579,235,840. The LP relaxation bounds for releasing the conditional $[R|CS]$ instead of the margin $[CRS]$ are much wider, see Table 12.3. For example, the upper LP bound for (1,1,1,1) cell for $[R|CS]$ is 37.42 while for $[CRS]$ is 14. Based on these bounds, we could mistakenly conclude that it is safer to release the conditional, i.e., the confidence of the rule. The sharp bounds for $[R|CS]$ in place of $[CRS]$ are the same even though they produce a larger space of possible tables; however, the latter can have potential implications for estimating distributions over the space of solutions.

Table 12.3. Sharp upper and lower bounds for cell entries in Table 12.1 given the $[CSR]$ margin, and LP relaxation bounds given $[R|CS]$ conditional probability values

C	S	R T	1	2	3	1	2	3
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]

In our example, releasing the three association rules turns out to be safe based on an examination of the bounds given the rule's supports (c.f., [20], Table 1.7). As before, all of the upper bounds are reasonably far from the 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]. If we released the $[CST]$, $[CSR]$, and $[RT]$ margins an intruder would be far from certain what entries belonged in the four cells that actually contain the count of "3."

12.5.2 Example 2: Data from the 1993 U.S. Current Population Survey

Table 12.4 describes data extracted from the 1993 Current Population Survey. Versions of these data have been used previously to illustrate several other approaches to confidentiality protection. The resulting 8-way table contains 2880 cells and is based on 48,842 cases; 1185 cells approximately 41%, contain 0 count cells. This is an example of a sparse table, too often present in practice, which poses significant problems in the model fitting and estimation. Almost all lower level margins (e.g., 2-way margins) contain 0 counts. Thus the existence of maximum likelihood estimates is an issue. These zeros propagate into the corresponding conditional tables.

Table 12.4. Description of variables in CPS data extract

Variable	Label	Categories
Age (in years)	A	$< 25, 25 - 55, > 55$
Employer Type (*Empolment*)	B	Gov, Pvt, SE, Other
Education	C	<HS, HS, Bach, Bach+, Coll
Marital status (*Marital*)	D	Married, Other
Race	E	White, Non-White
Sex	F	Male, Female
Hours Worked (*HrsWorked*)	G	$< 40, 40, > 40$
Annual Salary (*Salary*)	H	$< \$50K, \$50K+$

From disclosure risk perspective we are interested in protecting cells with small counts such as "1" and "2". There are 361 cells with count of 1 and 186 with count of 2. Our task is to reduce a potential disclosure risk for at least 19% of our sample, while still providing sufficient information for a "valid" statistical analysis.

To alleviate estimation problems, we recoded variables B and G from 5 and 2 categories respectively to 2 categories each yielding a reduced 8-way table with 768 cells. This table is still sparse. There are 193 zero count cells, or about 25% of the cells. About 16% of cells have high potential disclosure risk; there are 73 cells with counts of 1 and 53 with counts of 2. For this table we find two reasonable log-liner models

Model 1: $[ABCFG][ACDFG][ACDGH][ADEFG]$,

Model 2: $[ACDGH][ABFG][ABCG][ADFG][BEFG][DEFG]$,

with goodness-of-fit statistics $G^2 = 1870.64$ with 600 degrees of freedom and $G^2 = 2058.91$ with 634 degrees of freedom, respectively.

Model 1 is a decomposable graphical log-linear model whose minimal sufficient statistics are the released margins. We first evaluate if these five-way marginal tables are safe to release by analyzing number of cells with small counts. Most of the cell counts are large and do not seem to present an immediate disclosure risk. Two of the margins are potentially problematic. Marginal table $[ABCFG]$ has 1 cell with count of "5" in (1,4,2,1,2) cell, while the margin $[ACDGH]$ has a low count of "4" and two cells with count of "8"; e.g., see Table 12.5. Even without out any further analysis, most agencies would not release such margins. Because we are fitting a decomposable models this initial exploratory analysis reveals that there will be at least one cell with a tight sharp upper bound of size "4". Bellow we investigate if these margins are indeed safe to release accounting for the log-linear model we can fit and the estimates they provide for the reduced and full eight-way tables.

Table 12.5. Marginal table $[ACDGH]$ from 8-way CPS table

		A	1		2		3	
		C	1	2	1	2	1	2
D	G	H						
1	1	1	198	139	943	567	2357	2225
		2	11	19	240	715	1009	3781
	2	1	246	144	765	294	3092	2018
		2	8	14	274	480	1040	2465
2	1	1	2327	2558	835	524	2794	3735
		2	8	14	51	105	114	770
	2	1	1411	1316	617	359	3738	3953
		2	4	15	32	68	78	372

Model 1 is easy to fit and evaluate: it is decomposable and there are closed-form solutions for bounds given the margins. Almost all lower bounds are 0. As expected from the analysis above, the smallest upper bound is 4 counts. There are 16 such cells, of which 4 contain counts of "1" and rest contain "0". The next smallest upper bound is 5, for 7 "0" cell counts and for 1 cell with a count of "5". The 5 cells with counts of "1" have the highest risk of disclosure. The next set of cells with a considerably high disclosure risk are cells with an upper bound of size 8. There are 32 such cells (23 contain counts of "0", 4 contain counts of "1", 3 contain counts of "2", and 2 contain counts of "3"). If we focus on count cells of "1" and "2", with the release of this model we directly identified 12 out of 126 sensitive cells.

Table 12.6. Summary of difference between upper and lower bounds for small cell counts in the full 8-way CPS table under Model 1 and under Model 2

Bound diff. Cell count	Model 1						Model 2					
	0	1	2	3	4	5	0	1	2	3	4	5
0	226	112	66	52	69	62	192	94	58	40	36	26
1	-	12	15	14	13	20	-	10	8	6	2	10
2	-	-	1	3	8	4	-	-	2	2	4	4
3	-	-	-	1	4	2	-	-	-	0	0	0

If we fit the same model to the full 8-way table with 2,880 cells, there are 660 cells with difference in bounds less than equal to 5, with all lower bounds being 0. Most of these are "0" cell counts; however, a high disclosure risk exists for 74 cells with count of "1", 16 cells with cell count equal "2", and 7 cells with counts of "3"; see the summary in Table 12.6. Thus releasing the margins corresponding to Model 1 poses a substantial risk of disclosure.

Model 2 is non-decomposable log-linear model and it requires an iterative algorithm for parameter estimation and extensive calculation for bounds. This model has 5 marginals as sufficient statistics. The 5-way margin $[ACDGH]$ is still problematic; however, the 4 4-way margins all appear to be safe to release with the smallest count of size "46" appearing in cell (1,4,1,1) of the margin $[ABFG]$.

We focus our discussion only on cells with small counts, as we did for the Model 1. Since Model 2 is non-decomposable, no closed-form solutions exist for cell bounds, and we must rely on LP and IP which sometimes may not produce sharp bounds. In this case this was not an issue. For the reduced 8-way table, all lower bounds are 0 and the minimum upper bound again is 4. There are 16 cells with upper bound of 4, of which four cells have count "1", and the rest are "0". The next smallest upper bound is 8, and there are 5 such cells with counts of "1", 4 cells with counts of "2", and 3 cells with counts of "3". With these margins, in comparison to the released margins under Model 1, we have eliminated the effect of the margin $[ABCFG]$, and reduced a disclosure risk for a subset of small cell counts; however, we did not reduced the disclosure risk for the small cell counts with the highest disclosure risk. For the full 8-way table, we compare the distribution of small cell bounds for the small cell counts under the two models; see Table 12.6. There are no cells with counts of "3" that have very tight bounds. For the cells with counts of "2", the number of tight bounds have not substantially decreased (e.g., 16 under Model 1 vs. 12 under Model 2), but there has been a significant decrease in the number of tight bounds for the cells with count of "1" (e.g., from 74 under Model 1 to 36 under Model 2).

In theory we could enumerate the number of possible tables utilizing algebraic techniques and software such as LattE [5], MCMC, or SIS. Due to large dimension of the solution polytope for this example, however, LattE is currently unable the execute the computation because the space of possible tables is extremely large. We have also been unable to fine-tune the SIS procedure to obtain a reasonable estimate except "infinity". While it is possible to find a Markov basis corresponding to the second log-linear model, utilizing those for calculating bounds and or sampling from the space of tables is also currently computationally infeasible. But the practicality of such calculations is likely to change with increased computer power and memory.

Based on Model 1, variables B and H are conditionally independent given the remaining 6 variables. Thus we can collapse the 8-way table to a 6-way table and carry out a disclosure risk analysis on it. The collapsed table has only 96 cells, and there is only one small cell count of size "2" that would raise an immediate privacy concern. Furthermore, we have collapsed over the two "most" sensitive and most interesting variables for statistical analysis: Type of Employer and Income. We do not pursue this analysis here but, if other variables are of interest, we could again focus on search for the best decomposable model. With various search algorithms and criteria, out of 32,768 possible decomposable models all searches converge to $[ACFG][ADEFG]$, a model with a likelihood ratio chi-square of $G^2 = 144.036$ and 36 degrees of freedom.

In this case, we could simply provide the margins of the above model to the user to construct association rules provided that they do not provide precise information on three sensitive cells. Numerous association rules can be derived from the given margins. Some interesting rules, for example could be $AFG \Rightarrow C$, and $AFG \Rightarrow DE$. As we did in in the clinical trial example, we can evaluate how safe the release of these rules are by determining the bounds on the cells given the marginal and conditional constraints, that is the rules' support and confidence.

12.6   Conclusions

The literature on datamining for association rules has focused on extracting rules with high predictive utility, measured by criteria such as support and confidence. For categorical data bases, coming in the form of multi-way contingency tables, these rules and criteria essentially are extracting marginal tables and linked conditionals. Some authors have recognized the relevance of log-linear and related models for this type of datamining activity, e.g., see [11], and [37], but few have addressed the issue of preserving the privacy of individuals represented in the data base being mined, with no links to date to ideas from log-linear and related models. In this chapter we have provided an overview of the totally separate statistical literature focused on protecting against disclosure limitation in contingency tables, while providing marginal and conditional tables for analysis and reporting.

From the perspective of privacy preservation the methods described in this chapter for bounds on cell counts provide an alternative approach to that found in most of the machine learning literature. These methods stress the link between the ensemble of data to be released, i.e., margins and conditionals, and their ability to characterize the data base through the use of log-linear and related statistical models and assessments of goodness-of-fit. Measures of privacy preservation based on bounds and other statistically related quantities may suggest that "the best association rules" may not be releasable without possibly compromising confidentiality.

New to this enterprise, and especially new to datamining are the tools from computational algebraic geometry. We have attempted to illustrate their applicability here largely through the examples. For more details we refer the interested reader to [6], [17], [31], and papers in a special 2006 issue of the *Journal of Symbolic Computation* devoted to problems at the interface of statistics and algebraic geometry.

Machine learning has made major progress in the efficient extraction of association rules from large data bases. The statistical literature has focused more heavily on understanding the utility of the the extracted information and on related methodologies for assessing disclosure limitation or privacy preservation. Our goal in reviewing the points of convergence in these two literatures has been to stimulate a fusion of the different methodologies and computational tools. Barak et al. [3] adds the element of perturbation to our toolkit and we hope to compare their methods with those described in this paper in the near future.

Acknowledgements

References

[1] Agresti, A. (2002). *Categorical Data Analysis.* 2nd Edition. New York: Wiley.

[2] Anderson, B. and Moore, A. (1998). AD-trees for Fast Counting and for Fast Learning of Association Rules, *Knowledge Discovery from Databases Conference*.

[3] Barak, B., Chaudhuri, K., Dwork, C., Kale, S., McSherry, M., and Talwar, K. (2007). Privacy, accuracy, and consistency too: a holistic solution to contingency table release, *PODS '07: Proceedings of 26th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, New York: ACM Press, 273–282.

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

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

[6] Diaconis, P. and Sturmfels, B. (1998). Algebraic Algorithms for Sampling From Conditional Distributions, *Annals of Statistics*, 26, 363–397.

[7] Dobra, A. and Fienberg, S. E. (2000). Bounds for Cell Entries in Contingency Tables Given Marginal Totals and Decomposable Graphs, *Proceedings of the National Academy of Sciences*, 97, 11885–11892.

[8] Dobra, A. and Fienberg, S. E. (2001). Bounds for Cell Entries in Contingency Tables Induced by Fixed Marginal Totals, *Statistical Journal of the United Nations ECE,* 18, 363–371.

[9] Dobra, A., Fienberg, S. E., and Trottini, M. (2003). Assessing the Risk of Disclosure of Confidential Categorical Data (with discussion), In J. Bernardo et al. eds., *Bayesian Statistics 7,* Clarendon: Oxford University Press, 125–144.

[10] Domingo-Ferrer, J. and Torra, V. (eds.) (2004). *Privacy in Statistical Databases, Lecture Notes in Computer Science No. 3050,* New York: Springer-Verlag.

[11] DuMouchel, W. and Pregibon, D. (2001). Empirical Bayes Screening for Multi-Item Associations, *Proceedings of the ACM SIGKDD Intentional Conference on Knowledge Discovery in Databases & Data Mining (KDD01)*, ACM Press, 67–76.

[12] Duncan, G. T., Fienberg, S. E., Krishnan, R., Padman, R., and Roehrig, S. F. (2001). Disclosure Limitation Methods and Information Loss for Tabular Data, In P. Doyle, J. Lane, J. Theeuwes, and L. Zayatz (eds.) *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies,* Amsterdam: Elsevier, 135–166.

[13] Dwork, C., McSherry, F., Nissim, K. and Smith, A. (2006). Calibrating Noise to Sensitivity of Functions in Private Data Analysis, *3rd Theory of Cryptography Conference (TCC) 2006*, 265–284.

[14] Evfimievski, A., Srikant, R., Agrawal, R., and Gehrke, J. (2002). Privacy Preserving Mining of Association Rules, *Proceedings of the 8th ACM*

SIGKDD International Conference on Knowledge Discovery in Databases and Data Mining, Edmonton, Canada, July 2002.

[15] Fienberg, S. E. (1980). *The Analysis of Cross-Classified Categorical Data.* 2nd edition. Cambridge, MA: MIT Press.

[16] Fienberg, S. E. (2004). Datamining and Disclosure Limitation for Categorical Statistical Databases, *Proceedings of Workshop on Privacy and Security Aspects of Data Mining, Fourth IEEE International Conference on Data Mining (ICDM 2004),* Brighton, UK, November 2004.

[17] Fienberg, S. E., Makov, U. E., Meyer, M. M., and Steele, R. J. (2001). Computing the Exact Distribution for a Multi-way Contingency Table Conditional on its Marginals Totals, In A. K. M. E. Saleh, ed. *Data Analysis from Statistical Foundations: Papers in Honor of D. A. S. Fraser*, Huntington, NY: Nova Science Publishing, 145–165.

[18] Fienberg, S. E. and Makov, U. E. (1998). Confidentiality, Uniqueness, and Disclosure Limitation for Categorical Data, *Journal of Official Statistics,* 14, 385–397.

[19] Fienberg, S. E. and Slavkovic, A. B. (2004). Making the Release of Confidential Data from Multi-Way Tables Count, *Chance*, 17(3), 5–10.

[20] Fienberg, S. E. and Slavkovic, A. B. (2005). Preserving the Confidentiality of Categorical Statistical Data Bases When Releasing Information for Association Rules, *Data Mining and Knowledge Discovery.* 11, 155–180.

[21] Hemmecke, R. and Hemmecke, R. (2003). 4ti2 Version 1.1— Computation of Hilbert bases, Graver bases, toric Gröbner bases, and more.
`http://www.4ti2.de`.

[22] Jordan, M. I. (ed.) (1998). *Learning in Graphical Models.* Cambridge MA: MIT Press.

[23] Kargupta, H., Datta, S., Wang, Q., and Sivakumar, K. (2003). Random Data Perturbation Techniques and Privacy Preserving Data Mining, *Proceedings of the 3rd IEEE International Conference on Data Mining (ICDM 2003),* Melbourn, Florida, USA, December 2003.

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

[25] Lauritzen, S. L. (1996). *Graphical Models.* Oxford: Oxford University Press.

[26] Madigan, D. and Raftery, A. E. (1994). Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occams Window, *Journal of the American Statistical Association*, 89: 1535–1546.

[27] Moore, A. and Schneider, J. (2002). Real-valued All-Dimensions Search: Low-overhead Rapid Searching Over Subsets of Attributes, *Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence, July, 2002*, San Francisco: Morgan Kaufmann Publishers, 360–369.

[28] Rizvi, S. and Haritsa, J. (2002). Maintaining Data Privacy in Association Rule Mining, *Proceedings of the 28th Conference on Very Large Data Base (VLDB'02)*.

[29] Silverstein, C., Brin, S., and Motwani, R. (1998). Beyond Market Baskets: Generalizing Association Rules to Dependence Rules, *Data Mining and Knowledge Discovery,* 2,39–68.

[30] Silverstein, C., Brin, S., Motwani, R. and Ullman, J. (2000). Scalable Techniques for Mining Causal Structures, *Data Mining and Knowledge Discovery*, 4, 163–192.

[31] Slavkovic, A. B. (2004). *Statistical Disclosure Limitation Beyond the Margins*. Ph.D. Thesis, Department of Statistics, Carnegie Mellon University.

[32] Slavkovic, A. B. and Smucker, B. (2007). *Calculating Cell Bounds in Contingency Tables Based on Conditional Frequencies*. Technical Report, Department of Statistics, Penn State University.

[33] Slavkovic, A. B. and Fienberg, S. E. (2004). Bounds for Cell Entries in Two-way Tables Given Conditional Relative Frequencies, In Domingo-Ferrer, J. and Torra, V. (eds.), *Privacy in Statistical Databases, Lecture Notes in Computer Science No. 3050*, 30–43. New York: Springer-Verlag.

[34] Sturmfels, B. (2003). *Algebra and Geometery of Statistical Models*. John von Neumann Lectures at Munich University.

[35] Trottini, M. and Fienberg, S. E. (2002). Modelling User Uncertainty for Disclosure Risk and Data Utility, *International Journal of Uncertainty, Fuzziness and Knowledge Based Systems,* 10, 511–528.

[36] Willenborg, L. C. R. J. and de Waal, T. (2000). *Elements of Statistical Disclosure Control.* Lecture Notes in Statistics, Volume 155, New York: Springer-Verlag.

[37] Wu, X., Barbará, D. and Ye, Y. (2003). Screening and Interpreting Multi-item Associations Based on Log-linear modeling, *Proceedings of the ACM SIGKDD Intentional Conference on Knowledge Discovery in Databases & Data Mining (KDD03)*, ACM Press, 276–285.

[38] Zaki M. J. (2004). *Mining Non-Redundant Association Rules, Data Mining and Knowledge Discovery*, 9, 223–248.

[39] Verykios S. Vassilios and Gkoulalas-Divani A.(2007) *A Survey of Association Rule Hiding Methods for Privacy,* in this volume .