Preserving the Confidentiality of Categorical Statistical Data Bases When Releasing Information for Association Rules *

Stephen E. Fienberg (fienberg@stat.cmu.edu)
Department of Statistics, Cylab, and
Center for Automated Learning and Discovery
Carnegie Mellon University
Pittsburgh PA 15213-3890, U.S.A.

Aleksandra B. Slavkovic (sesa@stat.psu.edu)
Department of Statistics
Pennsylvania State University
University Park PA 16802, U.S.A.

February 20, 2005

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 tables corresponding to subsets of the categorical variables. Very recently some of the ideas have been extended to allow for the release of combinations of mixtures of marginal tables and conditional tables for subsets of variables. Association rules can be viewed as conditional tables. In this paper we consider possible inferences an intruder can make about confidential categorical data following the release of information on one or more association rules. We illustrate this with several examples.

Keywords: Algebraic geometry, Association rules, Conditional tables, Contingency tables, Disclosure limitation, Marginal tables, Privacy preservation.

1. Introduction

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., in what is commonly referred to in the statistical literature as a multi-way contingency table, see Agresti (2002), Bishop et al. (1975), Fienberg (1980). This 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

* The research reported here was supported in part by NSF grants EIA–9876619 and IIS–0131884 to the National Institute of Statistical Sciences, as well as by Grant R01-AG023141 from the NIH to the Department of Statistics and by Army contract DAAD19-02-1-3-0389 to CyLab, both at Carnegie Mellon University.

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 table, e.g., Anderson and Moore (1998), Komarek and Moore (2000), see Moore and Schneider (2002), **?** (**?**), Goldenberg and Moore (2004), Agrawal et al. (2000), Agrawal and Srikant (1994), Srikant and Agrawal (1995), and Agrawal and Srikant (2000). 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 modern statistical methods for categorical data such as loglinear and logit models and use higher-dimensional marginal tables. Here we do not focus on the specifics of the competing association search methods *per se* but rather we consider what they have in common in the sense of data needs to represent the rules and assess their "fit" to the data.

For datamining enterprises, a typical categorical database is large and sparse. When the data come from sources (i.e., individuals or enterprises) to which there have been promises of confidentiality, we need to understand whether approaches to mining for association rules possibly violate these promises. That is, we need to know the extent to which 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 multiway contingency table. This paper focuses on recent methods for detecting such potential disclosures and thus limiting the forms of marginals and conditionals to be made generally available. If such an approach to protection of data were implemented in practice, it would change some of the methods used to search for association rules.

In section 2, we explain why the association rule problem is actually a problem involving marginals and conditionals of a contingency table and then we provide an overview of some relevant literature on (i) association rules and privacy preservation and (ii) on statistical methods for disclosure limitation, especially methods linked to the selective release of margins and conditionals. In section 3, we discuss the differences between marginals and conditionals and the implications for their release in a simple two-way table setting. Then, in section 4, we describe the implications of the release of marginals and/or conditionals in multi-way tables. We illustrate the methods with a four-way and six-way examples, and, in the final section, we return to the implications for the tradeoff between confidentiality protection and data utility for mining of association rules.

2. Datamining Algorithms: Association Rules

2.1. Motivation and Terminology.

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: $\{A, B, C, \ldots\} \Rightarrow \{E, F, G, \ldots\}$, 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., Zaki (2004) uses a variation on support and confidence while Silverstein et al. (1998) and Silverstein et al. (1998) use chi-square statistics for independence and conditional independence computed on the marginal tables. We continue to use "support" and "confidence" here for illustrative purposes only.

A typical machine learning approach may set out 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, we can look for rules that exceed pre-defined support (minimum support) and have high confidence. But these criteria simply involve looking at observed marginal tables or observed conditional probabilities. 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 condition table, both corresponding to a subset of variable making up the full table.*

The release of marginal tables from a full table increases 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. It has been suggested by a reviewer that there is a semantic gap between the machine learning literature on association rules and the statistical literature on contingency tables. The real differences between the two literatures is how one deals with the marginal and conditional tables, and what is reported or shared with others. We address the latter point in the next subsection.

2.2. Privacy Preservation and Association Rules

The datamining literature now contains several suggestions for dealing with this potential disclosure problem associated with association rules. For example, some authors propose perturbing the full data array, e.g., see Evfimievski et al. (2002), Rizvi and Haritsa (2002), and Kargupta et al (2003). Others suggest distortion algorithms that involve data swapping (c.f. the statistical literature on swapping) or that involve deleting and/or hiding certain rules based on changing the level of support and confidence, e.g., Atallah et al. (1999), Oliveira and Zaïane (2003), and Pontikakis et al. (2004b). For deleting/hiding, the appropriate level is typically chosen a priori by a user who decides that certain items are sensitive to public without much consideration of the effect of these decisions on statistical inference and, because of the difficulty associated with the search through all possible association rules, most authors adopt heuristics of various sorts. In this paper, we in fact describe results that can help us determine which rules to hide in order to preserve privacy but to allow sufficient information for statistical inference.

Finally we note that Kantarcioglu et al. (2004) explore some broader issues of the privacy impact of datamining methods and their work is related to the literature on secure multi-party computation, e.g. see Kantarcioglu and Clifton (2004), and Vaidya and Clifton (2002).

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 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. So that the release of the form of the rule is not useful. To us, releasing a rule means releasing the data on which it is based, i.e., the corresponding conditional and/or marginal table. To understand the implications of such a release on confidentiality and privacy preservation we thus turn to the statistical literature that addresses this issue.

2.3. Disclosure Limitation for 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 extended and complementary reviews, see Duncan et al. (2001) and Fienberg (2004).

For the present purposes we can group the approaches in the statistical literature into perturbational and aggregation or collapsing.

Aggregation methods for categorical data 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 (e.g., see Agresti (2002), Bishop et al. (1975), and Fienberg (1980)).

Consider an $I \times J \times K$ table of observer counts $\{n_{ijk}\}$, with corresponding estimated expected values, $\{m_{ijk}\}$ under some sampling model. The saturated log-linear model for $\{m_{ijk}\}$ takes the form

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

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

$$\sum_i u_{123(ijk)} = \sum_j u_{123(ijk)} = \sum_k u_{123(ijk)} = 0. \tag{2}$$

We get *unsaturated* models from (1) by setting sets of u-terms equal to zero, e.g., if we set

$$u_{123} = 0 \text{ for all } i, j, k, \tag{3}$$

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.

For log-linear and logit models, the key features are the following:

- The relevant statistical models focus on simultaneous interactions among sets of variables that define the contingency table.
- Special subsets of 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 Lauritzen (1996), and in machine learning see Jordan (1998).)
- The minimal sufficient statistics (i.e., sufficient data summaries) for a log-linear model are the marginal totals corresponding to the highest interaction terms in the model. For example, for the no second-order interaction model for three-way tables in equation (3) 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.
- 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.

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

$$\min\{n_{i+}, n_{+j}\} \geq n_{ij} \geq \max\{0, n_{i+} + n_{+j} - n_{++}\}, \tag{4}$$

and that these bounds, also known as Fréchet bounds, are sharp. Now consider the situation where instead of releasing a full k-dimensional 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 problem, but in general this is an NP-hard problem. Dobra and Fienberg (2000; 2001; 2003) 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. 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. We illustrate this approach using bounds in the present paper and describe some extensions to it involving combinations of margins and conditionals.

The principal pertubational methods are data swapping as proposed by Dalenius and Reiss (1982) and set in an updated context by Fienberg and McIntyre (2004) and the post-randomization method (PRAM) of Gouweleeuw et al. (1998). The data swapping methods typically consider altering the counts in a k-way contingency table subject to the preservation of a set of lower-order marginal totals. Fienberg et al. (1998) generalize this notion to that of replacing the original data by a randomly generated one drawn from the "exact" distribution of the contingency table under a log-linear model given its minimal sufficient statistics, i.e., a set of marginal totals. The difficulty in this approach involves the generation of Markov bases, i.e., moves that allow us to go from one table to another "nearby" one, which can be used to construct the data swaps or moves involved in calculating the probabilities associated with the exact distribution. Methods from the algebraic geometry of polynomial rings can be used to generate Markov bases, e.g., see Diaconis and Sturmfels (1998), and they take relatively simple forms for some of the special subfamilies of log-linear models that are linked to closed form bounds of Dobra and Fienberg(2000; 2001; 2003). For additional discussion, see Fienberg (2004).

Finally we note that the PRAM methodology looks to carrying out independent perturbations of individual variables, by analogy with the statistical method of randomized response. If the parameters of the transformations are released along with the perturbed data, then an external user can make appropriate inferences from the altered data about association-like quantities, although with less accuracy than would be

possible using the original table. This approach seems very close to that advocated in Rizvi and Haritsa (2002) for association rules.

A major theme in the literature on disclosure limitation deals with the trade off between disclosure risk and data utility. See especially Willenborg and de Waal (2000), and selected papers in Domingo-Ferrer and Torra (2004). 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 Duncan et al. (2001) for a discussion in the context of categorical data. Trottini and Fienberg (2002) and Trottini (2003) take the trade-off formalism several steps further and embeds it in a fully Bayesian decision theoretic framework. Following Fienberg (2004) 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.

3. Marginal and Conditional Tables for Two-Way Tables

Because data from both marginal and conditional tables may be potentially of interest in assessing and reporting association rules, we need to understand the differences between them in terms of the information they convey regarding information about the entries in a multi-way contingency tables. In this section we discuss the difference between marginals and conditionals, and the implications for their release. After providing notation we focus initially on the two-way setting. Then we explain the multi-way generalizations in section 4.

3.1. 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 contigency 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 $p(x_A) = \sum_{K \setminus A} p(x_K)$, 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 III).

Given that we observe an arbitrary set of conditional and marginal tables, $\mathcal{T}$, define the *fiber* $\mathcal{F}_{\mathbf{t}}$ as a set of all k-way non-negative integer tables that satisfy these constraints. Consider a sublattice $\mathcal{L}_{\mathbf{t}}$ of $\mathbb{Z}^{\mathcal{D}}$ that depends on a collection $\mathcal{T}$ and a finite subset $\mathcal{B}_{\mathbf{t}}$ of $\mathcal{L}_{\mathbf{t}}$.

DEFINITION 3.1. *A Markov basis for $\mathcal{T}$ is the smallest subset $\mathcal{B}_{\mathbf{t}}$ of $\mathcal{L}_{\mathbf{t}}$ such that for any $\mathbf{x}, \mathbf{x}' \in \mathcal{F}_{\mathbf{t}}$, the difference $\mathbf{x} - \mathbf{x}'$ is in the linear hull of $\mathcal{B}_{\mathbf{t}}$.*

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 purporses, 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. Good references for tools on algebraic statistics including calculation and use of Markov and Gröbner bases are Diaconis and Sturmfels (1998), Pistone et al. (2001), Sturmfels (2003), and Slavkovic (2004).

3.2. Uniqueness and Bounds for Two-way Tables

There are numerous ways we can characterize bivariate distributions e.g., see Arnold et al. (1999). In the disclosure context, under the assumption that released sets of marginal and conditional distributions are compatible, we want to check whether or not they are sufficient to uniquely identify the existing joint distribution. This is because unique identification of the joint distribution may lead to unique identification of sensitive cells whose values we wish to protect.

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 $P(X_2)$,
3. $P(X_2|X_1)$ and $P(X_1)$.

(Gelman and Speed, 1993; Gelman and Speed, 1999; Arnold et al., 1999). 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 $P(X_2|X_1), P(X_1)$ uniquely identify the joint distribution (Arnold et al., 1999). The following theorem due to Slavkovic and Fienberg (2004), and Slavkovic (2004) describes this situation.

THEOREM 3.1. *For two discrete dependent random variables,* X_1 *and* X_2*, either collection* $\mathcal{T}_{x_1} = \{P(X_1|X_2), P(X_1)\}$ *or* $\mathcal{T}_{x_2} = \{P(X_2|X_1), P(X_2)\}$ *uniquely identifies the joint distribution if matrices* $(p(x_1|x_2))$ *and* $(p(x_2|x_1))$ *have full rank and* $d_{x_1} \geq d_{x_2}$ *for* $\mathcal{T}_{x_1}$ *and* $d_{x_2} \geq d_{x_1}$ *for* $\mathcal{T}_{x_2}$*.*

Proof. The proof relies on a fact that the cell probabilities are linear and linear-fractional functions of marginal and conditional probabilities. Given the full rank matrices $(p(x_1|x_2))$ and $(p(x_2|x_1),)$ the number of linearly independent constraints is greater than equal to the number of cells we are trying to estimate; thus each cell entry can be uniquely identified. More details can be found in Slavkovic (2004).

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.

When the released marginals and conditionals do not satisfy the uniqueness theorem, 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 and Fienberg (2004), and Slavkovic (2004) extend this work by describing incomplete specification of the joint and calculation of bounds given an arbitrary collection of marginals and conditionals. They use linear and integer programming and discuss some inadequacies in treating conditional constraints via linear programming. These results rely on the fact that any two-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 (4) gives the bounds when all

that is released are the two one-way marginals. When a single marginal or a single conditional is given, the cells are bounded below by zero and above by a corresponding marginal or a conditional value. When a marginal and a conditional are released but there is no uniquness, that is when conditions of Theorem 3.1 are not satisfied, we can obtain bounds, and in some two-way cases there are nice closed form solutions.

THEOREM 3.2. *For an $I \times 2$ table given $\mathcal{T} = \{P(X_1|X_2), P(X_1)\}$, let $UB_1 = p_{i|j}\frac{p_{i+}-\max\limits_{k\neq j}\{p_{i|k}\}}{p_{i|j}-\max\limits_{k\neq j}\{p_{i|k}\}}$, and $UB_2 = p_{i|j}\frac{p_{i+}-\min\limits_{k\neq j}\{p_{i|k}\}}{p_{i|j}-\min\limits_{k\neq j}\{p_{i|k}\}}$. Then there are sharp upper bounds (UB) and lower bounds (LB) on the cell probabilities, p_{ij} given by*

$$UB = \begin{cases} UB_1 & \text{if} \quad p_{i+} \geq p_{i|j} \\ \\ UB_2 & \text{if} \quad p_{i+} < p_{i|j} \end{cases} \tag{5}$$

and

$$LB = \begin{cases} \max\{0, UB_2\} \text{ s.t. } UB_2 \leq UB & \text{if} \quad p_{i+} \geq p_{i|j} \\ \\ \max\{0, UB_1\} \text{ s.t. } UB_1 \leq UB & \text{if} \quad p_{i+} < p_{i|j} \end{cases} \tag{6}$$

where $p_{ij} = P(X_1 = x_i, X_2 = x_j), i = 1, ..., I, j = 1, ..., J$, $p_{i+} = \sum_j p_{ij}$, $p_{+j} = \sum_i p_{ij}$, and $p_{i|j} = \frac{p_{ij}}{p_{+j}}$.

Proof. The following optimization problem describes relationships between the cell probabilities and the marginal and conditional probabilities given by $\mathcal{T}$. Let $P(X_1|X_2) = p_{i|j}$ and $P(X_1) = p_{i+}$.

$$\begin{aligned} \text{Max} \quad & p_{ij}, && (7) \\ \text{subject to} \quad & \sum_i \sum_j p_{ij} = 1 && (8) \\ & \sum_j p_{ij} = p_{i+}, && (9) \\ & p_{i|j} = \frac{p_{ij}}{p_{ij} + \sum_{k\neq i} p_{kj}}, \forall i = \{1, 2, ..., I\}, j = \{1, 2\}, && (10) \\ \text{and} \quad & p_{ij} \geq 0, \forall i = \{1, 2, ..., I\}, j = \{1, 2\}. && (11) \end{aligned}$$

The bounds are derived by solving the above linear programming problem via simplex method. The proof is an extension of the proof of Theorem 3.1 and uniqueness of the solutions in 2×2 tables from Slavkovic (2004).

These bounds are sharp for a set of low dimensional tables with nicely rounded conditional probability values. For higher dimensions

Table I. Delinquent children data by county and education level. The Fréchet bounds for released margins are given in square brackets.

Education / County	Low	Medium	High	Very High
Alpha	15[0,20]	1[0,20]	3[0,20]	1[0,20]
Beta	20[0,50]	10[0,35]	10[0,30]	15[0,20]
Gamma	3[0,25]	10[0,25]	10[0,25]	2[0,20]
Delta	12[0,35]	14[0,35]	7[0,30]	2[0,20]

the linear approximation 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.

Using the tools of computational commutative algebra such as Gröbner and Markov bases in statistics (Diaconis and Sturmfels, 1998), 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 costraints. 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. A technical description of calculation and structure of Markov bases given fixed conditional distributions for two-way tables can be found in Slavkovic (2004). In the following subsection we demonstrate the calculation and use of bases for a 4×4 table.

3.3. Example 1: A 4×4 Table on Delinquent Children Data

In this section we consider a 4×4 table of counts reported in Disclosure Limitation Methodology Federal Committee on Statistical Methodology (1994). Titles and row and column headings are fictitious. Table I shows the number of delinquent children by county and education level of household head. The joint distribution in this case is completely specified by following collections: $\{P(County|Education), P(Education)\}$, $\{P(Education|County), P(County)\}$, and $\{P(County|Education), P(Education|County)\}$. Based on Theorem 3.1, $\{P(County|Education), P(County)\}$ and $\{P(Education|County), P(Education)\}$ also uniquely identify the table of counts.

For the case of incomplete specification, we compare the outcome of releasing margins versus conditionals on disclosure. First consider

releasing both margins (row sums and column sums). Fréchet bounds are given in the Table I. Observe that the cells with small counts (i.e., sensitive cells) are well protected. For example, cell [Alpha, Medium] with count "1" is bounded below by 0 and above by 20. By using the tools from computational algebra we obtain the same sharp bounds, but we also learn that there are 18,272,363,056 tables having the fixed margins.

The 8×16 matrix $A_{\mathbf{t}}$ represents a linear relationship between cells of the 4×4 table and its row and column sums. A Markov basis is in the $kernel(A_{\mathbf{t}})$ of

$$A_{\mathbf{t}} = \begin{pmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

There are 36 primitive moves, that is, 36 vectors of $\{-1/0/+1\}$ that span the kernel of matrix $A_{\mathbf{t}}$. These are calcuated using algebraic software 4ti2 ((Hemmecke and Hemmecke, 2003)). They are of the following form:

$$\mathcal{B}_{\mathbf{t}} = \begin{pmatrix} 0 & 0 & 0 & 0 & -1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 \\ -1 & 0 & 1 & 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ \vdots & & & & & & & & & & & & & & & \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 \end{pmatrix}.$$

The first move presented as a binomial is $n_{24}n_{41} - n_{21}n_{44}$. These moves connect all the elements of a fiber:

$$\mathcal{F}_{\mathbf{t}} = \{\mathbf{n} \in I\!N^{4\times 4} : \mathbf{t} = (\text{column sums}(\mathbf{n}),\ \text{row sums}(\mathbf{n})\}.$$

One problem concerning this fiber is the *enumeration* problem. Given a table of counts, how many other tables of counts have the same marginals? This problem deals with counting the number of lattice points in a polytope. We used LattE software (De Loera et al., 2003). to determine that there are 18,272,363,056 4×4 possible tables in our example.

Next, suppose that instead of the pair of margins we release conditional frequencies which correspond to a *confidence* of an association rule $Education \Rightarrow County$:

$$P(Education|County) = \begin{pmatrix} 0.750 & 0.050 & 0.150 & 0.050 \\ 0.364 & 0.182 & 0.182 & 0.272 \\ 0.120 & 0.400 & 0.400 & 0.080 \\ 0.343 & 0.400 & 0.200 & 0.057 \end{pmatrix}$$

A surprising result comes from calculating Markov bases and using tools from algebra to determine the space of tables. We demonstrate the calculation of the Markov basis for this example.

Let $\mathbf{t} = \{P(Education|County)\}$. Then matrix $A_{\mathbf{t}}$ is:

$$\begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 5 & -15 & -15 & -15 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 19 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -3 & -3 & 17 & -3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 35 & -20 & -20 & -20 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -10 & 45 & -10 & -10 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -10 & -10 & 45 & -10 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 22 & -3- & 3 & -3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -10 & 15 & -10 & -10 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -10 & -10 & 15 & -10 & 0 & 00 & 0 & \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 23 & -12 & -12 & -12 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -14 & 21 & -14 & -14 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -7 & -7 & 28 & -7 \end{pmatrix}.$$

There are four generators or only eight possible moves compared to 36 when we fixed the margins:

$$\begin{pmatrix} -45 & -3 & -9 & -3 & 0 & 0 & 0 & 0 & 3 & 10 & 10 & 2 & 12 & 14 & 7 & 2 \\ -15 & -1 & -3 & -1 & 20 & 10 & 10 & 15 & 0 & 0 & 0 & 0 & -12 & -14 & -7 & -2 \\ -15 & -1 & -3 & -1 & 0 & 0 & 0 & 0 & -6 & -20 & -20 & -4 & 24 & 28 & 14 & 4 \\ -30 & -2 & -6 & -2 & 0 & 0 & 0 & 0 & 9 & 30 & 30 & 6 & -12 & -14 & -7 & -2 \end{pmatrix}.$$

The following binomial represents one of the moves:

$$n_{31}^3 n_{32}^{10} n_{33}^{10} n_{34}^2 n_{41}^{12} n_{42}^{14} n_{43}^7 n_{44}^2 - n_{11}^{45} n_{12}^3 n_{13}^9 n_{14}^3.$$

In this case, any of the eight moves applied to the original table of counts give tables with negative counts. Hence in the non-negative simplex there are no other solutions but the original table that satisfies this observed conditional distribution and the sample size $N = 135$. For the given conditional if the observed sample size is larger, there would be more than one observable joint distribution. In this case it is not safe to release the conditionals because they can be used to reconstruct

the full two-way array which itself is unsafe to release. More generally, while the release of a conditional table might appear to provide less information than the corresponding full table, in fact it either does not or the amount of information is so similar that we might as well release the full table. This notion generalizes to higher-dimensional tables.

4. Release of Marginal and Conditional Tables From a k-way Table

The following result generalizes the uniqueness Theorem 3.1 to k-way tables.

THEOREM 4.1. *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}.*

Proof. Follows the results for two-dimensional case.

In the two-way case, we dealt only with so called *full* conditionals because they involve all variables in the data base (e.g., table). The next two theorems describe the relationship between a conditional and a marginal table that involves a subset of variables from the data base. In other words, they describe a relationship between confidence and support for a rule that involves a subset of characteristics from a data base. The proofs of these theorems require tools and arguments from algebraic geometry and statistics that go beyond the scope of the purpose of this paper. Some heuristics and the constructions of these theorems are illustrated in Slavkovic (2004). Here we focus on their consequences relevant to establishing bounds on cells for evaluating potential disclosure.

THEOREM 4.2. *(Slavkovic (2004)) Let the released set be $\mathcal{T} = \{p_{A|B}\}$ such that $A, B \in K$ and $A \cup B \in K$, but $A \cup B \neq K$. Let $\mathcal{B}_{\mathbf{t}_{AB}}$ denote a Markov basis for a fixed margin p_{AB}. Let $\mathcal{B}_{\mathbf{t}_{\widehat{A|B}}}$ be a set of moves with fixed $p_{A|B}$ but for the marginal table AB. Markov basis set $\mathcal{B}_{\mathbf{t}}$ for a fixed $p_{A|B}$ describing all k-way tables that satisfy observed $\mathbf{t}$ contains two subsets of moves: $\mathcal{B}_{\mathbf{t}} = \mathcal{B}_{\mathbf{t}_{AB}} \cup \mathcal{B}_{\mathbf{t}_{\widehat{A|B}}}$.*

This result implies that, for the same sample size N, the number of solutions for a fixed $p_{A|B}$ 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. 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.

In the previous section we saw that, given a matrix of conditional values representing the full conditional, the value of the moves can be used to determine if we have a unique solution given the sample size N. A consequence of the above theorem is that for the small conditionals we can study the subset $\mathcal{B}_{\mathbf{t}_{\widehat{A|B}}}$ and determine if we are in the situation where the bounds given the small conditionals are the same as given its corresponding marginal.

THEOREM 4.3. *(Slavkovic (2004)) Fix $p_{A|B}$ such that $A, B \in K$, but $A \cup B \neq K$. Consider the generators in the set of generators in $\mathcal{B}_{\mathbf{t}_{\widehat{A|B}}} = \mathcal{B}_{p_{A|B} \backslash p_{AB}}$.*

- *If for every binomial representing the Markov basis, the sum of exponents of a monomial in the binomial representing is greater than or equal to sample size N,*
- *If for every binomial representing the Markov basis, any exponent in the binomial is greater than or equal to the sample size N, or*

then the space of solutions for fixed $p_{A|B}$ is the same as the space of solutions for fixed p_{AB}, and the bounds for cell entries 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. Technical results of this section for conditional and marginal distributions imply, that it is sufficient to evaluate the support. We demonstrate the application of these results on a simple example.

4.1. Example 2: Data from a Randomized Clinical Trial

The following example is adapted from material in Fienberg and Slavkovic (2004) and Slavkovic and Fienberg (2004) but here we link the uniqueness and bounds results explicitly to association rules.

Koch et al. (1983) report the data in Table II 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 will use a shorthand notation to decribe variables and marginals from the full tables. In particular, we denote Status as [S], Center as [C], Treatment as [T] with levels Active=1 and Placebo=2, and Response as [R]

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

C	S	R T	1	2	3
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

with levels 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: 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: $R \Rightarrow T, R \Rightarrow CS, R \Rightarrow CST$, and $T \Rightarrow CS$. 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, $R \Rightarrow CST$. Support is the joint marginal distribution of [CRST] and confidence [R|CST] is a table with conditional probability values (see Table III). For example, the observed conditional probability value in the (1,1,1,1) cell is $\frac{3}{28} = 0.107$. It is trivial to see that, in this example, release of this rule and more specifically its support 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

Table III. Observed conditional probability values for [R|CST] from data in Table II.

C	S	R T	1	2	3
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

the empirical conditional probability values from a full conditional distribution of [R|CST] to represent this information. If we also have the three-way margin [CST], we can clearly reconstruct the full four-way table! Given [R|CST] alone, there are 7,703,002 tables all having the same conditional probability values [R|CST]. We give linear programming relaxation bounds in Table IV. The tightest bound for the count of "3" is [1, 16.48] in cell (1,3,1,1).

We note that this single conditional release reveals the zero counts in the table unlike the release of margins, where we needed 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 confidentiality risk, for larger and sparser k-way tables the presence of a large fraction of zero 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.

We could potentially approximate the knowledge of the release of this association rule by treating the data in Table II as they come from a two-way 8×3 table and compute the Fréchet bounds for margins [CST] and [R], given in Table V. 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 upper bounds are far from the lower bounds and since these bounds correspond to an extremely

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

C	S	R T	1	2	3
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 V. Upper and lower bounds for cell entries in Table II given the [CST] and [R] margins.

C	S	R T	Poor	2	3
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]

large collection of tables, an intruder cannot use them to make strong inferences about potentially small cell entries.

Because this is a randomized clinical trial, in order to perform meaningful statistical analysis, we need to include the margin for the three explanatory variables, i.e., Center by Status by Treatment [CST]. Most model search procedures would narrow the focus to the two models:

1. [CST] [CSR],

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 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.

A key point here is 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: $T \Rightarrow CS, R \Rightarrow CS$, and $R \Rightarrow T$. 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 from this section, we can draw a number of interesting conclusions. For example, by Theorem 4.2, 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], based on Theorem 4.1, we get [RT] because the number of levels in [R], three, is greater than in [T] which is two. 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. As we see in Table VI the linear programming relaxation bounds for releasing the conditional [R|CS] instead of the margin [CRS] are much wider. For example, the upper linear programming bound for (1,1,1,1) cell for [R|CS] is 37.42 while for [CRS] is 14. If we relied on these bounds, we could mistakenly conclude that it is safer to release the conditional, and decide to release only the confidence of the rule. After computing the sharp bounds for [R|CS], however, we find the same bounds as for the corresponding support! Although in this example, the larger space of tables for conditionals did not produce larger sharp bounds, the difference in the number of tables, can have potential implications for estimating distributions over the space of solutions. We have begun to explore this question (e.g., see Slavkovic (2004)).

Next suppose that [CSR] and [RT] 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

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

C	S	RT	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]

that this is equivalent to publishing the [CST], [CSR] and [RT] 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 [RT]! 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.

In our example, release of the three association rules would be safe and can be evaluated by estimating bounds given the rule's supports. Table VII 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][RT]. 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".

Table VII. Upper and lower bounds for entries in Table II given the [CST], [CSR], and [RT] margins.

C	S	R T	1	2	3
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]

4.2. Example 3: Prognostic Risk Factors for Czech Auto Workers

The data in Table VIII come from a prospective epidemiological study of 1841 workers in a Czechoslovakian car factory, as part of an investigation of potential risk factors for coronary thrombosis (see Edwards and Havranek (1985)). Here we build on the analyses reported in Dobra and Fienberg (2003) and Fienberg (2004), interpreting them in the context of searching for association rules.

In Table VIII, A indicates whether or not the worker "smokes," B corresponds to "strenuous mental work," C corresponds to "strenuous physical work," D corresponds to "systolic blood pressure," E corresponds to "ratio of β and α lipoproteins," and F represents "family anamnesis of coronary heart disease." Our focus for disclosure limitation is on the three cells in the table with counts of "1" and "2". Because the data include all of the workers in the factory we in essence have a population and we *act* here as if the variables describing the dimensions of the table are key variables that could be used by intruders for record linkage, giving them possible access to other information on the workers not included in this particular table.

The log-linear model with minimal sufficient marginals

[ABCE] [ADE] [BF]

fits the data extremely well. In fact an even simpler model, corresponding to

[ABC][ABE] [ADE][BF]

Table VIII. Prognostic factors for coronary heart disease as measured on Czech autoworkers. Source: (Edwards and Havranek, 1985).

				B	no		yes	
F	E	D	C	A	no	yes	no	yes
neg	< 3	< 140	no		44	40	112	67
			yes		129	145	12	23
		≥ 140	no		35	12	80	33
			yes		109	67	7	9
	≥ 3	< 140	no		23	32	70	66
			yes		50	80	7	13
		≥ 140	no		24	25	73	57
			yes		51	63	7	16
pos	< 3	< 140	no		5	7	21	9
			yes		9	17	1	4
		≥ 140	no		4	3	11	8
			yes		14	17	5	2
	≥ 3	< 140	no		7	3	14	14
			yes		9	16	2	3
		≥ 140	no		4	0	13	11
			yes		5	14	4	4

Table IX. Marginal [ABCE] from Table VIII

		B	no		yes	
E	C	A	no	yes	no	yes
< 3	no		88	62	224	117
	yes		261	246	25	38
≥ 3	no		58	60	170	148
	yes		115	173	20	36

Table X. Marginal [BF] from Table VIII

F	B	no	yes
neg		929	134
pos		652	126

Table XI. Marginal [ADE] from Table VIII

E	D	A	no	yes
< 3	< 140		333	312
	≥ 140		265	151
≥ 3	< 140		182	227
	≥ 140		181	190

provides an excelent fit to the data, with a likelihood ratio chi-square value of $G^2 = 52.1$ with 46 degrees of freedom, as does the model which decomposes [ABCE] into [ACE] and [BCE]. Thus we could simply provide these to the user to construct association rules provided that they do not provide precise information through bounds on the three sensitive cells.

Numerous association rules can be derived from the given margins. We concentrate on the first model with the full four-way marginal [ABCE] because it includes most of the models a user would identify using the standard statistical model search criteria. Some interesting rules, for example, are $B \Rightarrow F$, $D \Rightarrow AE$, and $AE \Rightarrow BC$. As we did in Example 2, 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.

Based on the Theorem 4.2, as in Example 2, if we just release the confidence of each aforementioned rule, e.g., $[B|F]$, $[D|AE]$, and $[AE|BC]$, the bounds on each cell are going to be at least as wide as if the released information are corresponding supports. We give the three released marginals in Tables X, XI, and IX. We thus focus on these marginal tables in order to evaluate how safe are these rules. Each marginal table can be treated as a two-way table and thus the bounds are well known bounds given by equation (4). Given [BF], the bounds corresponding to the three cells with small counts of "1" and "2", are all [0,126]. For [ADE], the bound for "1" is [0,333], and for the two "2"'s the bounds are [0,151] and [0,182]. [ABCE] seems to release more information as the bounds are tighter, i.e., [0,25], [0,38], and [0,20], but not tight enough to compromise confidentiality.

We can evaluate separate pieces further. For example, for a fixed $[D|AE]$ we get 35 generators, or elements of a Markov basis. We are unable to enumerate the number of possible solutions with LattE software, however. There are 24 moves that come from the fixed margin $[ADE]$. The remaining ten moves all have either more than one exponent that

Table XII. The bounds for the marginals [ABCE],[ADE] and [BF] for the data in Table VIII

				B	no		yes	
F	E	D	C	A	no	yes	no	yes
neg	< 3	< 140	no		[0,88]	[0,62]	[0,224]	[0,117]
			yes		[0,261]	[0,246]	[0,25]	[0,38]
		≥ 140	no		[0,88]	[0,62]	[0,224]	[0,117]
			yes		[0,261]	[0,151]	[0,25]	[0,38]
	≥ 3	< 140	no		[0,58]	[0,60]	[0,170]	[0,148]
			yes		[0,115]	[0,173]	[0,20]	[0,36]
		≥ 140	no		[0,58]	[0,60]	[0,170]	[0,148]
			yes		[0,115]	[0,173]	[0,20]	[0,36]
pos	< 3	< 140	no		[0,88]	[0,62]	[0,126]	[0,117]
			yes		[0,134]	[0,134]	[0,25]	[0,38]
		≥ 140	no		[0,88]	[0,62]	[0,126]	[0,117]
			yes		[0,134]	[0,134]	[0,25]	[0,38]
	≥ 3	< 140	no		[0,58]	[0,60]	[0,126]	[0,126]
			yes		[0,115]	[0,134]	[0,20]	[0,36]
		≥ 140	no		[0,58]	[0,60]	[0,126]	[0,126]
			yes		[0,115]	[0,134]	[0,20]	[0,36]

is greater than the sample size $N = 1841$, or are such that the sum of exponents in a monomial is greater than N. Therefore given the sample size and based on Theorem 4.3, the space of tables defined by this conditional is the same as the space of tables defined by fixing the margin $[ADE]$ and the bounds are the same.

In Table XII we give the bounds for the cell counts when all these marginal tables are released. The bounds corresponding to the three small counts of "1" and "2" are [0,25], [0,38] and [0,20], and they correspond to the bounds given by [ABCE] only. All three of these pairs of bounds differ quite substantially and thus we might conclude that there is little chance of identifying the individuals in the small cells.

Fienberg (2004) notes that we could in fact release the marginals

[ACDE][ABCDF][ABCEF][BCDEF][ABDEF]

without compromising confidentiality. The likelihood ratio chi-square value for this model is $G^2 = 3.9$ with 3 degrees of freedom. The interested reader could attempt to search for more elaborate association rules than those base on [ABCE], [ADE], and [BF] using this larger

collection of marginal releases, but the interesting statistical question is whether any such rules are "more accurate" than those based on the simpler set of marginals we have focused on here.

5. 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 DuMouchel and Pregibon (2001), Pavlov et al. (2003), Goldenberg and Moore (2004), and Wu et al. (2003). The issue of preserving the privacy of individuals represented in the data base being mined is has only recently received focus attention, with no links to date to ideas from log-linear and related models. There has been a totally separate statistical literature focused on protecting against disclosure limitation in contingency tables, while providing marginal and conditional tables for analysis and reporting. In this paper, we have demonstrated that these literatures are addressing similar issues and using similar approaches to disclosure limitation, e.g., perturbation, data swapping, and restricted reporting.

From the perspective of privacy reservation the methods described in this paper for bounds on cell counts provide an alternative approach for dataminers. But these methods also 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. 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 Diaconis and Sturmfels (1998), Fienberg et al. (2001), Slavkovic (2004), and the forthcoming special issue of the *Journal of Symbolic Computation* devoted to problems at the interface of statistics and algebraic geometry.

One of the main issues in privacy preserving data mining for association rules is the limited scope in evaluation of the ensemble of released rules with respect to how it characterizes the data base. The typical datamining approach attempts to look at all possible association rules, but only only report a subset of them, sometimes only one or two. Even then, 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.

The datamining literature 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 demonstrating the points of convergence in these two literatures has been to stimulate a fusion of the different methodologies and computational tools.

Acknowledgements

We owe special thanks to Alan Karr for drawing our attention to the close correspondence between the confidentiality problems we have been working on and those associated with association rule mining. We are indebted to the comments of the referees for some references and suggestions that helped to emphasize the complementary nature of the statistical and datamining literatures. This research is part of several larger efforts focused on privacy and confidentiality, including a project coordinated by the National Institute of Statistical Sciences involving several U.S. federal statistical agencies.

References

Agrawal, R. and Imielinski, T. and Swami, A. Mining Association Rules Between Sets of Items in Large Databases. *Proceedings of the 1993 ACM SIGMOD Conference,* Washington, DC, 1993.

Agrawal, R. and Srikant, R. Fast Algorithms for Mining Association Rules. *Proceedings of the 20th VLDB Conference,* Santiago, Chile, 1994.

Agrawal, R. and Srikant, R. Privacy-Preserving Data Mining. *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data,* Dallas, May 2000.

Agresti, A. *Categorical Data Analysis.* 2nd Edition. New York: Wiley, 2002.

Anderson, B. and Moore, A. AD-trees for Fast Counting and for Fast Learning of Association Rules. Knowledge Discovery from Databases Conference, 1998.

Arnold, B. C. and Castillo, E. and Sarabia, J. M. *Conditional Specification of Statistical Models.* Springer–Verlag, 1999.

Arnold, B. C. and Press, J. S. Compatible Conditional Distributions. *Journal of the American Statistical Association*, 84, 405:152–156, 1998.

Atallah, M., Bertino, E., Elmagarmid, A., Ibrahim, M., and Verykios, V. Disclosure Limitation of Sensitive Rules. *Proceedings of the IEEE Knowledge and Data Engineering Exchange Workshop (KDEX'99),* November 1999, Chicago, IL, 45–52.

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

Chang, L. and Moskowitz, I. S. An Integrated Framework for Database Privacy Protection. *Proceedings of the IFIP TC11/ WG11.3 Fourteenth Annual Working Conference on Database Security*, Kluwer, B.V., 161–172, 2001.

Dalenius, T. and Reiss, S. P. Data-swapping: A Technique for Disclosure Control. *Journal of Statistical Planning and Inference,* 6, 73–85, 2004.

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

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

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

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

Dobra, A. and Fienberg, S. E. Bounding Entries in Multi-way Contingency Tables Given a Set of Marginal Totals", In Y. Haitovsky and H.R. Lerche and Y. Ritov, eds. *Foundations of Statistical Inference: Proceedings of the Shoresh Conference 2000*, Berlin: Springer-Verlag, 3–16, 2003.

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

DuMouchel, W. and Pregibon, D. 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, 2001.

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

Duncan, G. T. and Stokes, S. L. Disclosure Risk vs. Data Utility: The R-U Confidentiality Map as Applied to Topcoding. *Chance,* 17(3), 16–20, 2004.

Edwards, D. E. and Havranek, T. A Fast Procedure for Model Search in Multidimensional Contingency Tables. *Biometrika,* 72, 339–351, 1985.

Estivill-Castro, V., and Brankovic, Lj. Data Warehousing and Knowledge Discovery. In Mohania, M, K. and Min Tjoa, A. (eds.), *First International Conference, DaWaK '99, Lecture Notes in Computer Science No. 1676*, New York: Springer-Verlag, 389-398, 1999.

Evfimievski, A. and Srikant, R. and Agrawal, R. and Gehrke, J. 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.

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

Fienberg, S. E. 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, 2004.

Fienberg, S. E. and Makov, U. E. and Meyer, M. M. and Steele, R. J. 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, 2001.

Fienberg, S. E. and Makov, U. E. and Steele, R. J. Disclosure Limitation Using Perturbation and Related Methods for Categorical Data (with discussion). *Journal of Official Statistics,* 14, 485–502, 1998.

Fienberg, S. E. and McIntyre, J. Data Swapping: Variations on a Theme by Dalenius and Reiss. In Domingo-Ferrer, J. and Torra, V. (eds.), *Privacy in Statistical Databases– PSD'2004, Lecture Notes in Computer Science No. 3050*, New York: Springer-Verlag, 14–29, 2004.

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

Gelman, A. and Speed,T. S. Characterizing a Joint Probability Distribution by Conditionals. *Journal of the Royal Statistical Society. Series B*, 55 (1): 185–188, 1993.

Gelman, A. and Speed,T. S. Corrigendum: Characterizing a joint probability distribution by conditionals. *Journal of the Royal Statistical Society. Series B*, 61(2): 483, 1999.

Goldenberg, A. and Moore, A. Tractable Learning of Large Bayes Net Structures from Sparse Data. *Proceedings of the 21st International Conference on Machine Learning*, 2004.

Gouweleeuw, J. M. and Kooiman, P. and Willenborg, L. C. R. J. and Wolf, P. P. de., Post Randomization for Statistical Disclosure Control: Theory and Implementation. *Journal of Official Statistics,* 14 463–478, 1998.

Hemmecke, R. and Hemmecke, R. 4ti2 Version 1.1—Computation of Hilbert bases, Graver bases, toric Gröbner bases, and more. Available at www.4ti2.de, 2003.

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

Kantarcioglu, M. and Clifton, C. Privacy Preserving Data Mining of Association Rules on Horizontally Partitioned Data. *Transactions on Knowledge and Data Engineering*, to appear, 2004.

Kantarcioglu, M. and Jin, J. and Clifton, C. When Do Data Mining Results Violate Privacy? *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, Washington, USA, August 22-25, 2004* ACM Press, 599–604, 2004.

Kargupta, H., Datta, S., Wang, Q., and Sivakumar, K. 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.

Komarek, P. and Moore, A. A Dynamic Adaptation of AD-trees for Efficient Machine Learning on Large Data Sets. *Proceedings of the 17th International Conference on Machine Learning*, 495–502, 2000.

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

Lauritzen, S. L. *Graphical Models.* Oxford: Oxford University Press.

Moore, A. and Schneider, J.. 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.

Oliveira, S. R. M., and Zaïane, O. R. Algorithms for Balancing Privacy and Knowledge Discovery in Association Rule Mining. *In Proceedings of the 7th International Database Engineering and Applications Symposium (IDEAS 2003),* Hong Kong, China, July 2003, 54–63.

Pavlov, D. and Mannila, H. and Smyth, P. Beyond Independence: Probabilistic Models for Query Approximation on Binary Transaction Data. *IEEE Transactions on Knowledge and Data Engineering,* 15: 1409-1421, 2003.

Pelleg, D. and Moore, A. Using Tarjan's Red Rule for Fast Dependency Tree Construction. *Advances in Neural Information Processing Systems 15 (NIPS 2002),* Cambridge, MA: MIT Press, 2003, 801–808.

Pistone, J., Riccomagno, E., and Wynn, H. *Algebraic Statistics - Computational Commutative Algebra in Statistics*. Chapman and Hall/CRC, Boca Raton, FL. 2001.

Pontikakis, E. D. and Verykios, V. S. and Theodoridis, Y. On The Comparison of Association Rule Hiding Techniques *Hellenic Database Management Symposium,* Athens, Greece, June 2004.

Pontikakis, E. D. and Tsitsonis, A. A. and Verykios, V. S. A Quantitative Experimental Study of Distortion-based Techniques in Association Rule Hiding. *Conference in Database Security*, Sitges, Spain, July 2004.

Pontikakis, E. D. and Tsitsonis, A. A. and Verykios, V. S. and Theodoridis, Y. and Chang, L. A Quantitative and Qualitative Analysis of Blocking in Association Rules Hiding. *ACM Workshop on Privacy in Electronic Society*, Washington DC, USA, October 2004.

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

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

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

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

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

Srikant, R. and Agrawal, R. Mining Generalized Association Rules. *Proceedings of the 21st International Conference on Very Large Databases,* Zurich, Switzerland, September 1995.

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

Trottini, M. Decision Models for Data Disclosure Limitation. Ph.D. Thesis, Department of Statistics, Carnegie Mellon University, 2003.

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

Vaidya, J. and Clifton, C. Privacy Preserving Association Rule Mining in Vertically Partitioned Data. *The Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* Edmonton, Alberta, Canada, July 23 - 26, 2002.

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

Witten, I. H., and Frank, E. *Data Mining: Practical Machine Learning Tools and Techniques.* New York: Morgan Kaufmann, 2000.

Wu, X. and Barbar, D. and Ye, Y. 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, 2003.

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