ALGEBRAIC STATISTICS AND CONTINGENCY TABLE PROBLEMS: LOG-LINEAR MODELS, LIKELIHOOD ESTIMATION, AND DISCLOSURE LIMITATION

ADRIAN DOBRA*, STEPHEN E. FIENBERG†, ALESSANDRO RINALDO‡, ALEKSANDRA SLAVKOVIC§, AND YI ZHOU¶

Abstract. Contingency tables have provided a fertile ground for the growth of algebraic statistics. In this paper we briefly outline some features of this work and point to open research problems. We focus on the problem of maximum likelihood estimation for log-linear models and a related problem of disclosure limitation to protect the confidentiality of individual responses. Risk of disclosure has often been measured either formally or informally in terms of information contained in marginal tables linked to a log-linear model analysis and has focused on disclosure potential of small cell counts, especially those equal to 1 or 2. One way to assess risk is to compute bounds for cell entries given a set of released marginals. Both of these methodologies become complicated for large sparse tables. This paper revisits the problem of computing bounds for cell entries and picks up on a theme first suggested in Fienberg [21] that there is an intimate link between the ideas on bounds and the existence of maximum likelihood estimates, and shows how these ideas can be made rigorous through the underlying mathematics of the same geometric/algebraic framework. We illustrate the linkages through a series of examples. We also discuss the more complex problem of releasing marginal and conditional information. We illustrate the statistical features of the methodology on two examples and then conclude with a series of open problems.

Key words. Conditional tables, Marginal tables, Markov bases, Maximum likelihood estimate, Sharp bounds for cell entries, Toric ideals.

AMS(MOS) subject classifications. 13P10, 62B05, 62H17, 62P25.

1. Introduction.

Polynomials abound in the specification of statistical models and inferential methods. In particular, many common statistical procedures involve finding the solution to polynomial equations. Thus, in

*Department of Statistics, University of Washington, Box 354322, Seattle, WA 98195-4322 (adobra@u.washington.edu).

†Department of Statistics, Machine Learning Department and Cylab, Carnegie Mellon University, Pittsburgh, PA 15213-3890 (fienberg@stat.cmu.edu). Supported in part by NSF grants EIA9876619 and IIS0131884 to the National Institute of Statistical Sciences, and NSF Grant DMS-0631589 and Army contract DAAD19-02-1-3-0389 to Carnegie Mellon University.

‡Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213-3890 (arinaldo@stat.cmu.edu). Supported in part by NSF Grant DMS-0631589 and a grant from the Pennsylvania Department of Health through the Commonwealth Universal Research Enhancement Program to Carnegie Mellon University.

§Department of Statistics, Pennsylvania state University, State College, PA (sesa@stat.psu.edu). Supported in part by NSF grants EIA9876619 and IIS0131884 to the National Institute of Statistical Sciences and SES-0532407 to Pennsylvania State University.

¶Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA 15213-3890 (yizhou@cs.cmu.edu). Supported by Army contract DAAD19-02-1-3-0389 to Carnegie Mellon University.

retrospect, we should not be surprised at the sudden emergence of a wide array of papers linking statistical methodology to modern approaches to computational algebraic geometry. But the fact is that these connections are a relatively recent development in the statistical literature and they have led to the use of the terminology "algebraic statistics" to describe this linkage.

Contingency tables are arrays of non-negative integers arising from cross-classifying n objects based on a set of k criteria or categorical variables (see [1, 32]). Each entry of a contingency table is a non-negative integer indicating the number of times a given configuration of the classifying criteria, or *cell*, has been observed in the sample. Log-linear models form a class of statistical models for the joint probability of cell entries. Our work has focused on three interrelated classes of problems: (1) geometric characterization of log-linear models for cell probabilities for contingency tables, (2) estimation of cell probabilities under log-linear models, and (3) disclosure limitation strategies associated with contingency tables which protect against the identification of individuals associated with counts in the tables.

The disclosure limitation literature for contingency table data is highly varied, e.g., see [18], but over the past decade a substantial amount of it has focused on the risk-utility tradeoff where risk has been measured either formally or informally in terms of information contained in marginal tables and risk has focused on disclosure potential of small cell counts, especially those equal to 1 or 2 (for details, see [25, 26]). Among the ways considered for assessing risk have been the computation of bounds for cell entries, e.g., see [11, 12, 13, 14, 15], and counting of possible table realizations, e.g., see [26].

Recent advances in the field of algebraic statistics have provided novel and broader mathematical tools for log-linear models and, more generally, the analysis of categorical data. We outline below the most relevant aspects of the algebraic statistics formalism, which essentially involves a representation, through polynomials and polyhedral objects, of the interaction between the set of all possible configurations of cell probabilities, known as the *parameter space*, and the set of all observable arrays of non-negative entries summing to n and satisfying certain linear relationships to be described below, known as the *sample space*.

2. Some Technical Details for Bounds and MLEs. We can describe both the determination of cell bounds associated to the release of marginal tables and the problem of nonexistence of the MLE within the same geometric/algebraic framework.

2.1. Technical Specifications and Geometrical Objects. Consider k categorical random variables, $X_1, \ldots, X_k$, where each X_i takes value on the finite set of categories $[d_i] \equiv \{1, \ldots, d_i\}$. Letting $\mathcal{D} = \bigotimes_{i=1}^{k} [d_i]$, $\mathbb{R}^{\mathcal{D}}$ is the vector space of k-dimensional arrays of the format $d_1 \times \ldots \times d_k$,

with a total of $d = \prod_i d_i$ entries. The cross-classification of n independent and identically distributed realizations of $(X_1, \ldots, X_k)$ produces a random integer-valued array $\mathbf{n} \in \mathbb{R}^{\mathcal{D}}$, called a k-way *contingency table*, whose coordinate entry $n_{i_i,\ldots,i_k}$ is the number of times the label combination, or *cell*, $(i_1, \ldots, i_k)$ is observed in the sample (see [1, 32] for details). The probability that a given cell appears in the sample is

$$p_{i_1,\ldots,i_k} = Pr\{(X_1, \ldots, X_k) = (i_1, \ldots, i_k)\}, \qquad (i_1, \ldots, i_k) \in \mathcal{D},$$

and we denote the corresponding array in $\mathbb{R}^{\mathcal{D}}$ with $\mathbf{p}$. It will often be convenient to order the cells in some prespecified way (e.g., lexicographically) and to treat $\mathbf{n}$ and $\mathbf{p}$ as vectors in $\mathbb{R}^d$ rather than arrays. For example, for a 3-way contingency table $\mathbf{n}$ with $d_1 = d_2 = d_3 = 2$, or a $2 \times 2 \times 2$ table, we will use interchangeably the array notation $\mathbf{n} = (n_{111}, n_{112}, \ldots, n_{222})$ and the vector notation $\mathbf{n} = (n_1, n_2, \ldots, n_8)$. A hierarchical log-linear model is a probabilistic model specifying the set of dependencies, or maximal *interactions*, among the k variables of interest. One can think of a log-linear model as a simplicial complex Δ on $[k] = \{1, \ldots, k\}$, whose facets indicate the groups of interacting variables. For example, for a $2 \times 2 \times 2$ table, the model $\Delta = \{\{1, 2\}, \{3\}\}$ specifies an interaction between the first and second variable, while the third is independent of the other two. Similarly, $\Delta = \{\{1, 2\}, \{1, 3\}, \{2, 3\}\}$, the model of *no-2nd-order interaction*, postulates an interaction between all pairs of variables. In accordance to the notation adopted in the statistical literature, we will also write these models as [12][3] and [12][13][23], respectively.

A sub-class of log-linear models that enjoys remarkable properties and deserves special attention is the class of decomposable models: Δ is said to be decomposable if there exists a decomposition $\Delta = (\Delta_1, S, \Delta_2)$ with $\Delta_1 \cup \Delta_2 = \Delta$ and $\Delta_1 \cap \Delta_2 = 2^S$, and Δ_i is either a simplex or decomposable, for each $i = 1, 2$. Decomposable models are the simplest log-linear models for which the statistical tasks described in this article become straightforward. The smallest decomposable model with a non-trivial separation is $\Delta = \{\{1, 2\}, \{2, 3\}\}$, where $S = \{2\}$.

For any given log-linear model Δ, the vector of cell probabilities $\mathbf{p}$ is a point in the interior of the standard $(d-1)$-simplex such that $\log \mathbf{p}$ belongs to the row span of some $m \times d$ matrix A, called the *design matrix*, which depends only on Δ (and not on the random realization $\mathbf{n}$). Clearly, for every Δ, there are many choices for A, but we may always assume A to be 0-1. For an example, see Table 1.

Once we specify a model Δ through its design matrix A, we consider the vector $\mathbf{t} = \mathrm{A}\mathbf{n}$ of *margins* or *marginal tables*. From an inferential standpoint, the vector $\mathbf{t}$ is all that a statistician needs to know in order to study Δ: in statistics, $\mathbf{t}$ is called a *minimal sufficient statistics*. In fact, although many different tables may give rise to the same margins $\mathbf{t}$, they are indistinguishable in the sense that they all provide the same information on Δ. See, e.g. [1, 30, 32] for details. In general, we can choose the design

TABLE 1

A design matrix for the model of no-2nd-order interaction for a $2 \times 2 \times 2$ table. The first line displays the label combinations ordered lexicographically.

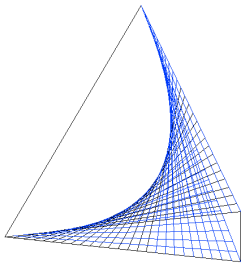
111	112	121	122	211	212	221	222
1	1	0	0	0	0	0	0
0	0	1	1	0	0	0	0
0	0	0	0	1	1	0	0
0	0	0	0	0	0	1	1
1	0	0	0	1	0	0	0
0	1	0	0	0	1	0	0
0	0	1	0	0	0	1	0
0	0	0	1	0	0	0	1
1	0	1	0	0	0	0	0
0	1	0	1	0	0	0	0
0	0	0	0	1	0	1	0
0	0	0	0	0	1	0	1

matrices in such a way that the coordinates of the vector $\mathbf{t}$ are the marginal sums of the array $\mathbf{n}$ with respect to the coordinates specified by the cofacets of Δ. For the example in Table 1, it is easy to see that the first coordinate of $\mathbf{t}$ is $n_{111} + n_{112}$, which we will write in marginal notation as n_{11+}, the "+" symbol referring to the variable over which the summation is taken.

Parameter Space. The parameter space refers to the set of all probability points $\mathbf{p}$ in the standard $(d-1)$-simplex such that $\log \mathbf{p}$ belongs to the row span of A. In algebraic statistics the parameter space is defined by the solution set of certain polynomial maps. Specifically, the parameter space is a smooth hyper-surface of points satisfying binomial equations, in fact a *toric variety* [41]. For a given design matrix A, the toric variety describing the associated log-linear model is the set of all probability vectors $\mathbf{p}$ such that $\mathbf{p}^{\mathbf{z}^+} - \mathbf{p}^{\mathbf{z}^-} = 0$ for all integer valued vectors $\mathbf{z}$ in kernel(A), where $\mathbf{z}^+ = \max(\mathbf{z}, 0)$, $\mathbf{z}^- = -\min(\mathbf{z}, 0)$, the operations being carried elementwise, and $\mathbf{p}^{\mathbf{z}} = \prod_i p_i^{z_i}$. For a 2×2 table and the model of independence, the toric variety is the familiar surface of independence (see, e.g. [1]) depicted in Figure 1. For further details on the algebraic geometry of other aspects of 2×2 tables, see [37, 40, 3]. The advantage of the algebraic statistics representation for the parameter space over the traditional log-linear representation based on logarithms, is that the points on the toric variety are allowed to be on the relative boundary of the simplex. This implies that the toric variety includes not only the points $\mathbf{p}$ such that $\log \mathbf{p}$ belongs to the row span of A, but also points in its sequential closure.

Sample Space. The sample space indicates the set of possible observable contingency tables, namely the set of all non-negative integer-valued array

FIG. 1. *Surface of independence for the* 2×2 *table. The tetrahedron represents the set of all probability distributions* $\mathbf{p} = (p_{11}, p_{12}, p_{21}, p_{22})$ *for a* 2×2 *table, while the enclosed surface identifies the probability distributions satisfying the equation* $p_{11}p_{22} = p_{12}p_{21}$*, i.e. the toric variety for the model of independence.*



in $\mathbb{R}^{\mathcal{D}}$ with entries summing to n. Virtually all data-dependent objects encountered in the study of log-linear models are polyhedra (see, e.g., [44]).

In particular, for a given log-linear model and a set of margins $\mathbf{t}$, consider the polytope

$$\mathrm{P}_{\mathbf{t}} = \{\mathbf{x} \in \mathbb{R}^d \colon \mathbf{t} = \mathrm{A}\mathbf{x}, \mathbf{x} \geq 0\}$$

of all real-valued non-negative tables having the same margins $\mathbf{t}$, computed using the design matrix A. The set of all integer points inside $\mathrm{P}_{\mathbf{t}}$ is the *fiber* of $\mathbf{t}$. Formally, if $\mathbf{n}$ is any table such that $\mathbf{t} = \mathrm{A}\mathbf{n}$, then the fiber of $\mathbf{t}$ is $\mathcal{F}(\mathbf{n}) \equiv \{\mathbf{v} \in \mathbb{N}^d \colon \mathbf{n} - \mathbf{v} \in \mathrm{kernel}(\mathrm{A})\}$. Thus the fiber is the portion of the sample space associated with the same set of margins. Since margins are also minimal sufficient statistics, $\mathcal{F}(\mathbf{n})$ consists precisely of all the possible tables that would provide the same information on the unknown underlying vector $\mathbf{p}$ as the observed table. These tables form the support of the conditional distribution given the margins, often called the *exact distribution*, because it does not depend on the model parameters. Properties of the fiber are fundamental both for assessing the risk of disclosure and for conducting "exact" inference, e.g., see [10, 22].

Fibers are also related to *Markov bases*. Let $\mathcal{B} \subseteq \mathrm{kernel}(\mathrm{A}) \cap \mathbb{Z}^d$ and, for each table $\mathbf{n}$, let $\mathcal{F}(\mathbf{n})_{\mathcal{B}}$ be the undirected graph whose nodes are the elements of $\mathcal{F}(\mathbf{n})$ and in which two nodes $\mathbf{v}$ and $\mathbf{v}'$ are connected if and only if either $\mathbf{v} - \mathbf{v}' \in \mathcal{B}$ or $\mathbf{v}' - \mathbf{v} \in \mathcal{B}$. $\mathcal{B}$ is said to be a Markov basis if $\mathcal{F}(\mathbf{n})_{\mathcal{B}}$ is connected for each $\mathbf{n}$. Markov bases can be obtained as the minimal generators of the toric ideal $\langle \mathbf{p}^{\mathbf{z}^+} - \mathbf{p}^{\mathbf{z}^-}, \mathbf{z} \in \mathrm{kernel}(\mathrm{A}) \cap \mathbb{Z}^d \rangle$ (see, e.g. [10]). Although Markov bases are by no means unique, for a given design matrix A all Markov bases which are minimal with respect to inclusion have the same cardinality, so that it is customary to talk about "the" Markov basis. When the Markov basis $\mathcal{B}$ is available, one can walk around all the points in the fiber without leaving $\mathrm{P}_{\mathbf{t}}$: for any $\mathbf{n}_1, \mathbf{n}_2 \in \mathcal{F}(\mathbf{n})$, there is a sequence of Markov moves $(\mathbf{z}_1, \ldots, \mathbf{z}_L) \in \mathcal{B}$ such that $\mathbf{n}_1 = \mathbf{n}_2 + \sum_{i=1}^{L} \mathbf{z}_i$, with

$\mathbf{n}_1 + \sum_{i=1}^{j} \mathbf{z}_i \geq 0$ for all $1 \leq j \leq L$. In fact, this is possible if only if $\mathcal{B}$ is a Markov basis. Therefore, Markov bases fully characterize the set of fibers associated to all possible observable tables.

Another polyhedral object relevant to log-linear modeling is the convex hull of all the possible margins $\mathbf{t}$ that could be observed for a given design matrix A. This object, a polyhedral cone called the *marginal cone*, is an unbounded (here n is allowed to be any integer number) convex set consisting of all the linear combinations of the columns of A with nonnegative coefficients, i.e.,

$$\mathrm{C_A} = \{\mathbf{y} \in \mathbb{R}^m \colon \mathbf{y} = \mathrm{A}\mathbf{x},\ \mathbf{x} \in \mathbb{R}^d, \mathbf{x} \geq 0\}.$$

As a result, the marginal cone comprises the set of all possible observable sufficient statistics and also their expectations, which are the points $\mathrm{A}(n\mathbf{p})$ with $\mathbf{p}$ ranging over the simplex. For example, for the $I \times J \times K$ table and the model of no-2nd-order interaction, the minimal sufficient statistics are the three sets of two-dimensional marginal sums, $\{n_{ij+}\}$, $\{n_{i+k}\}$, and $\{n_{+jk}\}$, so that $\mathbf{t}$ is an integer valued random vector of dimension $IJ + IK + JK$.

To summarize, we use the design matrix A and the marginal tables $\mathbf{t}$ to obtain geometric representations of the parameter and sample space for log-linear models. On one hand, A determines a system of polynomial equations that encode the dependencies among the random variables in the table. The solution set of these equations is the variety representing the parameter space as a compact subset of the simplex. On the other hand, every point $\mathbf{t}$ in the marginal cone $\mathrm{C_A}$ determines the polytope $\mathrm{P}_{\mathbf{t}}$, which in turn contains the fiber, i.e., the portion of the sample space that is relevant for both statistical inference and disclosure limitation.

2.2. Maximum Likelihood Estimation. The method of *maximum likelihood* (ML) is a standard approach to parameter estimation which chooses as the estimate of $\mathbf{p}$ the point in the simplex which maximizes the probability of the observed data $\mathbf{n}$ as a function of the parameter $\mathbf{p}$.

The maximum likelihood estimate (MLE) $\widehat{\mathbf{p}}$ is said to exist when $\widehat{\mathbf{p}}$ lies on the interior of the simplex. In this case, $\widehat{\mathbf{p}}$ is the unique point such that $\log \widehat{\mathbf{p}}$ is in the row range of A and satisfies the marginal constraints $\mathrm{A}\widehat{\mathbf{p}} = \frac{1}{n}\mathbf{t}$ (see, e.g., Haberman [30]). In particular, for decomposable log-linear models, the entries of the MLE are rational functions of the sample size n and the entries of the marginal vector $\mathbf{t}$, and we can compute them easily. Furthermore, the MLE of a decomposable model exists if and only if $\mathbf{t} > 0$. In contrast, for non-decomposable models, there is no closed-form expression for the MLE, which we can only evaluate numerically, and positivity of the observed marginals is only a necessary condition.

More generally, Eriksson et al. [20] show that existence of the MLE is equivalent to requiring that the marginal table $\mathbf{t}$ belongs to the interior of the marginal cone $\mathrm{C_A}$. Not only is this condition simple to interpret, but it

TABLE 2

(a): *Configurations of zero cells that cause nonexistence of the MLE (a facial set) for the model of no-second-order interaction without producing null margins.* **(b):** *Example of a table with many sampling zeros but for which the MLE for the model of no-second-order interaction is well defined. Cells with entries + indicate positive entries. Source: Fienberg and Rinaldo [24].*

(a)

0	+	0
+	+	+
0	+	+

+	+	+
+	0	+
0	0	+

+	+	0
+	0	0
+	+	+

(b)

+	0	0
0	+	0
0	0	+

0	0	+
+	0	0
0	+	0

0	+	0
0	0	+
+	0	0

also reduces the problem of detecting nonexistence of the MLE to a linear optimization program over a convex set. Furthermore, we can use the same geometric formalism to characterize cases in which the MLE does not exist, a circumstance that occurs whenever $\mathbf{t}$ lies on the boundary of C_A. In fact, for any point $\mathbf{t}$ in the marginal cone, the polytope $P_{\mathbf{t}}$ containing the fiber is never empty and it intersects always the toric variety describing the model implied by A at one point $\widehat{\mathbf{p}}^e$ [35]. The first condition implies $A\widehat{\mathbf{p}}^e = \frac{1}{n}\mathbf{t}$ and the second that $\widehat{\mathbf{p}}^e$ is in the closure of the log-linear parameter space. If $\mathbf{t}$ is in the interior of C_A, then these are precisely the defining conditions for the MLE, hence $\widehat{\mathbf{p}}^e = \widehat{\mathbf{p}}$. If $\mathbf{t}$ is instead a point on the boundary of C_A, $\widehat{\mathbf{p}}^e$ will have some zero coordinates and will be the MLE of a restricted log-linear model at the boundary of the parameter space, an *extended MLE*. Notice that an extended MLE does not possess a representation as the logarithm of a point in the simplex belonging to the row range of A. Nonetheless, it is a well defined point on the toric variety. The extended MLE realizes, both statistically and geometrically, the connection between the sample space and the parameter space.

For example, the pattern of zero cells in Table 2(a) leads to the nonexistence of the MLE under the model of no-second-order interaction even though the entries in the margins are strictly positive. We obtained this table, along with others providing novel examples of "pathological" configurations of sampling zeros, using `polymake` (Gawrilow and Joswig [28]). The example in Table 2(b) is sparser than the one in Table 2(a) but the MLE exists in the former case and not in the latter.

2.3. Bounds for Cell Counts. Public/government agencies collect high quality, multi-dimensional census and survey data and they generate databases that they do not make fully accessible to the public. These categorical data are often represented in tabular form as large and sparse

contingency tables with small counts. The release of partial information from such databases is of public utility and typically consists of publishing marginal and conditional tables. Users can cumulate the released information and translate it in upper and lower bounds for cell counts. If these bounds are close for a particular cell, then an intruder could learn the corresponding count and this might compromise the confidentiality offered to individual respondents.

Individuals or establishments that have an uncommon combinations of attributes will show up in the contingency table in cells with small counts of "1" or "2". A count of "1" might correspond to a population unique whose identity might be at risk unless the table contains a significant number of sample uniques that are not population uniques. In this case the "true" population uniques are concealed by the counts of "1" associated with "false" uniques. Counts of "2" can lead to similar violations if the intruder is one of the two persons and, for other small counts, we have the notion of inferential or probabilistic disclosure, i.e. the possibility of determining with a high degree of certainty individuals in the database. Such small counts also have the highest disclosure risk because their upper and lower bounds are close to the true value even for modest amounts of released information. This is especially true for large sparse categorical databases (e.g., census data) in which almost all counts are zero. These counts translate into small counts in the released marginals, which in turn lead to tight upper and lower bounds.

Consider a 2×2 contingency table with cell counts n_{ij} and row and column totals, n_{i+} and n_{+j} respectively, adding to the total n_{++}. If we are given the row and column totals, then the well-known Fréchet bounds for the individual cell counts are:

$$\min(n_{i+}, n_{+j}) \geq n_{ij} \geq \max(n_{i+} + n_{+j} - n, 0) \text{ for } i = 1, 2, j = 1, 2. \quad (2.1)$$

The extra lower bound component comes from the upper bounds on the cells complementary to (i, j). These bounds have been widely exploited in the disclosure limitation literature and have served as the basis for the development of statistical theory on copulas [33]. The link to statistical theory comes from recognizing that the minimum component $n_{i+} + n_{+j} - n$ corresponds to the MLE of the expected cell value under independence, $n_{i+}n_{+j}/n$. The bounds are also directly applicable to $I \times J$ tables and essentially a related argument can be used to derive exact sharp bounds for multi-way tables whenever the marginal totals that are fixed correspond to the minimal sufficient statistics of a log-linear model that is *decomposable*.

Next we consider a $2 \times 2 \times 2$ table with cell counts n_{ijk}, and two way marginal totals n_{ij+}, n_{i+j}, and n_{+jk}, adding to the grand total n_{+++}. Given the 2-way margins, the bounds for the count in the (i, j, k) cell for

$i = 1, 2$, $j = 1, 2$,and $k = 1, 2$, are

$$\begin{aligned}&\min(n_{ij+}, n_{i+k}, n_{+jk}, n_{ijk} + n_{\bar{i}\bar{j}\bar{k}})\\ &\quad \geq n_{ijk}\\ &\quad \geq \max(n_{i++} - n_{i+k} - n_{ij+}, n_{+j+} - n_{ij+} - n_{+jk}, n_{++k} - n_{i+k} - n_{+jk}, 0)\end{aligned} \tag{2.2}$$

where $(\bar{i}, \bar{j}, \bar{k})$ is the complementary cell to (i, j, k) found by replacing 1 by 2 and 2 by 1, respectively. Equation (2.3) consists of a combination of Fréchet bounds for each of the rows, columns, and layers of the full table plus an extra upper bound component $n_{ijk} + n_{\bar{i}\bar{j}\bar{k}}$.

Fienberg [21] suggested how to use this basic construction to get bounds for an $I \times J \times K$ table by considering all possible collapsed $2 \times 2 \times 2$ versions (based on all possible permutations of the subscripts). Dobra [11] refined this construction and developed a "generalized shuttle" algorithm, extending an idea in Buzzigoli and Giusti [2] in order to obtain sharp bound by iterating between "naive" upper bound and lower bounds. This algorithm finds the sharp bounds for decomposable models without extensive computation, which is reduced in other special cases, e.g., see Dobra and Fienberg [16]. Nonetheless, it does not scale well for large sparse tables, c.f. the fact that non-integer bound problems are NP-hard (see [6, 7]).

2.4. Link between Maximum Likelihood Estimates and Cell Bounds. We now use the algebraic geometric machinery to make the link between existence of the MLE and the computation of cell bounds explicit through the following result.

PROPOSITION 2.1. *For any lattice point* $\mathbf{t}$ *on the boundary of the marginal cone, let* $\widehat{\mathbf{p}}^{\mathrm{e}}$ *be the extended MLE and let* $\mathcal{Z}_{\mathbf{t}} = \{i \colon \widehat{\mathbf{p}}^{\mathrm{e}}_i = 0\}$ *be the set of cells for which the extended MLE is zero. Then, each table* $\mathbf{n}$ *in the fiber* $\mathrm{P}_{\mathbf{t}}$ *is such that* $\mathbf{n}_i = 0$ *for all* $i \in \mathcal{Z}_{\mathbf{t}}$.

The set $\mathcal{Z}_{\mathbf{t}}$ is uniquely determined by the margins $\mathbf{t}$ and corresponds to one of the many patterns of sampling zeros which invalidate the existence of the MLE. For the $2 \times 2 \times 2$ table and the model of no-2nd-order interaction, "impermissible" patterns of zeros occur in pairs of cells where $i + j + k$ is odd for one and even for the other. These cells can either be in adjacent cells adding to a margin zero in one of the two-way margins or they can be of the form (i, j, k) and $(\bar{i}, \bar{j}, \bar{k})$ for all possible values of i, j and k, c.f., Haberman [30]. Thus in particular, the MLE does not exist when $n_{ijk} + n_{\bar{i}\bar{j}\bar{k}} = 0$. The extra component of the upper bound for this non-decomposable model in equation (2.3) is thus inextricably bound up with the existence of MLEs.

The cells *not* belonging to $\mathcal{Z}_{\mathbf{t}}$ form a (random, as it depends on the random quantity $\mathbf{t}$) *facial set* (see [29, 35]). The cells with positive entries in Tables 2 (a) and Table 3 are examples of facial sets. Proposition 2.1 then shows that the determination of the facial set associated with a given

(:,:,1)=				(:,:,2)=				(:,:,3)=				(:,:,4)=			
0	0	0	5	0	0	1	1	0	1	2	2	4	2	3	3
4	5	5	1	0	0	6	0	0	5	5	0	2	2	2	0
1	5	0	1	5	3	2	2	0	4	0	0	2	2	0	0
1	0	0	1	5	0	2	2	3	2	4	3	2	0	0	0

TABLE 3

A 4×4×4 table with a pattern of zeros corresponding to a non empty $\mathcal{Z}_{\mathbf{t}}$ and, therefore, to a nonexistent MLE for the model of no-second-order interaction. Source: Fienberg and Rinaldo [24].

marginal table is crucial, not only for computing the extended MLE, but also for calculating individual cell bounds, as it implies that one only needs to consider the cells in the facial set for performing the tasks of counting, sampling and optimizing over the fiber.

The determination of the facial sets is an instance of what in computational geometry is known as the face-enumeration problem: the computations of all the faces of a given polyhedron. Unfortunately, the number of solutions of this problem is often affected by a combinatorial explosion. As a result, complete enumeration of all the facial sets is impractical. A much more efficient solution consists in finding just the facial set corresponding to the observed margins $\mathbf{t}$, using the methods developed in [23].

Table 3 shows a 4×4×4 table for which the MLE for the model of no-2nd-order interaction is nonexistent. The zero entries correspond to facial set for the minimal sufficient statistics, which as we noted above are the sets of all two-way margins. There are 123 tables in the fiber. Table provides the cell bounds given the two-way marginals computed using the shuttle algorithm [11]. Proposition 2.1 implies that the upper bounds for the entries of the zero cells, which correspond to a set $\mathcal{Z}_{\mathbf{t}}$, is zero. The integer bounds for this table are shown in Table 4. Notice that the entry range for each cell is an interval of integer points, i.e., the fiber is connected, and thus the knowledge of cell bounds is very informative for assessing the risk of disclosure.

3. More on Disclosure Limitation: From Margins to Margins and Conditionals. Because data from both marginal and conditional tables are widely reported as summary data from multi-way contingency tables, we need to understand how they differ from the sets of marginals in terms of the information they convey about the entries in the tables. For example, we want to see whether or not sets of marginal and conditional distributions for a contingency table are sufficient to uniquely identify the full joint distribution. When this is not the case we can protect against disclosure further by replacing a marginal table by constituent marginal and conditional components.

At first blush, one might think that there would also be a direct role for

(:,:,1)=				(:,:,2)=			
[0, 0]	[0, 0]	[0, 0]	[5, 5]	[0, 0]	[0, 0]	[0, 2]	[0, 2]
[2, 6]	[3, 7]	[5, 5]	[1, 1]	[0, 0]	[0, 0]	[6, 6]	[0, 0]
[0, 4]	[3, 7]	[0, 0]	[0, 2]	[4, 6]	[3, 3]	[2, 2]	[1, 3]
[0, 2]	[0, 0]	[0, 0]	[0, 2]	[4, 6]	[0, 0]	[1, 3]	[0, 4]
(:,:,3)=				(:,:,4)=			
[0, 0]	[0, 3]	[0, 4]	[1, 3]	[4, 4]	[0, 3]	[2, 5]	[3, 3]
[0, 0]	[3, 6]	[4, 7]	[0, 0]	[0, 4]	[0, 6]	[0, 3]	[0, 0]
[0, 0]	[4, 4]	[0, 0]	[0, 0]	[0, 4]	[0, 4]	[0, 0]	[0, 0]
[3, 3]	[2, 2]	[3, 5]	[2, 4]	[2, 2]	[0, 0]	[0, 0]	[0, 0]

TABLE 4
Sharp integer bounds of the 4×4×4 Table 3.

marginals and conditionals in the estimation of Bayes net models described in algebraic terms by Garcia et al. [27], especially when all variables are categorical. In such settings, we replace an omnibus log-linear model by a series of linear log-odds or *logit* models corresponding to a factorization of the joint probabilities into a product of conditional distributions and there is an interesting issue of whether we can use any reduction via minimal sufficient statistics to exploit the ideas that follow.

To extend the ideas from the preceding section, we consider a subset a of $K = \{1, ..., k\}$ and denote by $\mathbf{n}_a$ and $\mathbf{p}_a$ the vectors of marginal counts and probabilities for the variables in a, respectively, of dimension $d_a = \prod_{i \in a} d_i$. If a and b are two disjoint subsets of K, we denote by $\mathbf{n}_{ab}$ and $\mathbf{p}_{ab}$ the corresponding marginal quantities for the variables in $a \cup b$. Provided that the entries of $\mathbf{n}_b$ are strictly positive, we define the array of *observed conditional proportions* of a given b by $\mathbf{n}_{a|b} = \mathbf{n}_{ab}/\mathbf{n}_b$, and the array of *conditional probabilities* of a given b by $\mathbf{p}_{a|b} = \mathbf{p}_{ab}/\mathbf{p}_b$, where $\mathbf{p}_b > 0$.

When $k = 2$, so that $a = \{1\}$, $b = \{2\}$, by any of the following sets of distributions uniquely identifies the joint distribution: (1) $\mathbf{p}_{a|b}$ and $\mathbf{p}_{b|a}$, (2) $\mathbf{p}_{a|b}$ and $\mathbf{p}_b$, or (3) $\mathbf{p}_{b|a}$ and $\mathbf{p}_a$. Cell entries can be zero as long as we do not condition on an event of zero probability. Sometimes the sets $\{\mathbf{p}_{a|b}, \mathbf{p}_a\}$ and $\{\mathbf{p}_{b|a}, \mathbf{p}_b\}$ uniquely identify the joint distribution. The following result, due to Slavkovic [37] and described in Slavkovic and Fienberg [39], characterizes this situation for a generalization to a k-way table.

THEOREM 3.1. *(Slavkovic(2004)) Consider a k-way contingency table and a pair of matrices* $\mathcal{T} = \{\mathbf{p}_{a|b}, \mathbf{p}_a\}$, *where* $a, b \subset \{1, \ldots, k\}$ *and* $a \cap b = \emptyset$. *If the matrix of conditional probabilities has full rank, and* $d_a \geq d_b$, *then* $\mathcal{T}$ *uniquely identifies the marginal table of probabilities* $\mathbf{p}_{ab}$.

Often, there are multiple realizations of the joint distribution for $\mathbf{n}$, i.e., there is more than one table that satisfies the constraints imposed by them. Slavkovic [37], and Slavkovic and Fienberg [39] describe the calculation of

bounds given an arbitrary collection of marginals and conditionals. They use linear programming (LP) and integer programming (IP) methods and discuss potential inadequacies in treating conditional constraints via LP. These results rely on the fact that any k-way table satisfying 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, an intruder intent on learning about individuals represented in a table of counts knows with a high degree of certainty that there is only a small number of them possessing the characteristics corresponding to that cell. This may pose a risk of disclosure of the identity of these individuals. For example, equation (2.1) gives the bounds when all that is released are the two one-way marginals in a two-way table. When we have a single marginal or a single conditional, 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 provided we know the sample size N. When we are working with released marginals we know N, but when we work only with conditionals this is an extra piece of information that needs to be provided to rescale the proportions to infer possible values for tables of counts.

When the conditions of Theorem 3.1 are not satisfied, we can obtain bounds for cell entries, and in some two-way cases there are closed-form solutions. Slavkovic [37] and Fienberg and Slavkovic [39] derive such closed-form solutions for $2 \times J$ tables. Then the corresponding marginal and conditional cell probabilities are denoted as $p_{i+} = \sum_j p_{ij}$ for $\mathbf{p}_a, i \in a$, $p_{+j} = \sum_i p_{ij}$ for $\mathbf{p}_b, j \in b$, and $p_{i|j} = p_{ij}/p_{+j}$ for $\mathbf{p}_{a|b}$, respectively. The closed form solutions for the p_{ij}'s rely on solving a linear programming problem via the simplex method and are given in the following theorem.

THEOREM 3.2. *Consider a $2 \times J$ contingency table and a pair of matrices $\mathcal{T} = \{\mathbf{p}_{a|b}, \mathbf{p}_a\}, i \in a, j \in b$. Let*

$$UB_1 = p_{i|j} \frac{p_{i+} - \max\limits_{r \neq j}\{p_{i|r}\}}{p_{i|j} - \max\limits_{r \neq j}\{p_{i|r}\}},$$

and

$$UB_2 = p_{i|j} \frac{p_{i+} - \min\limits_{r \neq j}\{p_{i|r}\}}{p_{i|j} - \min\limits_{r \neq j}\{p_{i|r}\}}.$$

Then there are sharp upper bounds (UB) and lower bounds (LB) on the cell probabilities, p_{ij} given by

$$UB = \begin{cases} UB_1 & \text{if} \quad p_{i+} \geq p_{i|j} \\ \\ UB_2 & \text{if} \quad p_{i+} < p_{i|j}, \end{cases} \tag{3.1}$$

and

$$LB = \begin{cases} \max\{0, UB_2\} \ s.t. \ UB_2 \leq UB & if \quad p_{i+} \geq p_{i|j} \\ \\ \max\{0, UB_1\} \ s.t. \ UB_1 \leq UB & if \quad p_{i+} < p_{i|j}. \end{cases} \tag{3.2}$$

Given a set of low dimensional tables with nicely rounded conditional probability values, these bounds will be sharp. For higher dimensions, linear approximations of the bounds may be far from the true sharp bounds for the table of counts, and thus 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. Thus if the observed counts are considered sensitive, the database owner is the only one who can produce the sharp IP bounds in the case of the conditionals, e.g., see Slavkovic and Smucker [38].

Using algebraic tools for determining Gröbner and Markov bases, 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 program approach fails, and (2) we can use it to describe all possible tables satisfying costraints imposed by information given by $\mathcal{T}$. In particular, a set of minimal Markov bases allows us to build a connected Markov chain and perform a random walk over all the points in the fiber that have the same fixed marginals and/or conditionals. Thus we can either enumerate or sample from the space of tables via Sequential Importance Sampling (SIS) or Markov Chain Monte Carlo (MCMC) sampling, e.g., see Chen et al. [5]. Some disadvantages of the algebraic approach are that (1) calculation of Markov bases are computationally infeasible even for tables of small dimension, and (2) for conditionals, Markov bases are extremely sensitive to rounding of cell probabilities. Slavkovic [37] provides a description of calculation and structure of Markov bases given fixed conditionals for two-way tables. In this setting, the design matrix A does not rely on a log-linear model, but is a $m \times d$ *constraint matrix* A where d is the number of cells and m the number of linear constraints induced by $\mathcal{T}$ and the row of ones induced by the fixed sample size N; the corresponding m dimensional vector $\mathbf{t}^{-1} = \begin{bmatrix} N & \mathbf{0} \end{bmatrix}$. The reported results in the examples below rely on this methodology.

4. Two Examples. Here we illustrate several of the results described in the preceding sections in the context of two examples of sparse contingency tables. The examples illustrate the limits of current computational approaches. To simplify the description of log-linear models we use the common short-hand notation for marginals, referring to the variables comprising them. For example, in a 3-way table involving variables A, B, and C, we denote the 3 2-way marginals as [AB], [AC], and [BC].

4.1. Example: Genetics Data. Edwards [19] reports on an analysis of genetics data in the form of a sparse 2^6 contingency table given in Table 5.

The six dichotomous categorical variables, labeled with the letters A-F, record the parental alleles corresponding to six loci along a chromosome strand of a barely powder mildew fungus, for a total of 70 offspring. The original data set, described in [4], included 37 loci for 81 offsprings, with 11 missing data—a rather sparse table.

			1				2				D
			1		2		1		2		E
			1	2	1	2	1	2	1	2	F
1	1	1	0	0	0	0	3	0	1	0	
		2	0	1	0	0	0	1	0	0	
	2	1	1	0	1	0	7	1	4	0	
		2	0	0	0	2	1	3	0	11	
2	1	1	16	1	4	0	1	0	0	0	
		2	1	4	1	4	0	0	0	1	
	2	1	0	0	0	0	0	0	0	0	
		2	0	0	0	0	0	0	0	0	
A	B	C									

TABLE 5

Cell counts for the dataset analyzed by Edwards [19]. Data publicly available at `http://www.hypergraph.dk/`.

For the model implied by fixing all the 2-way margins, the MLE is nonexistent because there is one null entry in the [AB] margins. Using `polymake` [28], we found that the marginal cone for this model has 116,764 facets, each corresponding to a different pattern of sampling zeros causing nonexistence of the MLE , but only 60 of them correspond to null margins. Table 6 displays the facial set associated to one of these facets. The facet of the marginal cone containing in its relative interior the null margins observed for the Table 5 is, in turn, a polyhedral cone with 11,432 facets.

Table 7 shows the set $\mathcal{Z}_{\mathbf{t}}$ obtained when we release three marginals: [ABCD][CDE][ABCEF]. The cells marked with a '0' correspond to values constrained to be zero, the '+' entries are cells for which the integer lower bound is positive, and the '+0' cells indicate a zero lower integer bound. The fiber in this case consists of 30 tables.

Proposition 2.1 is about LP and not ILP. In fact, null integer upper bounds for a set of cells do not imply that the MLE does not exist. In fact, Table 8 shows a set of sharp integer upper and lower bounds for a model for which the MLE exists—espite the fact that there exist strictly positive real-valued tables in the fiber determined by the prescribed margins, there are cells, highlighted in boldface, for which no positive integer entries can occur. Although the MLE is well defined, many estimated cell mean values are rather small: 28 out of 64 values were less than 0.01 and only 14 were

			1				2				D
			1		2		1		2		E
			1	2	1	2	1	2	1	2	F
1	1	1	0	+	+	+	0	0	+	0	
		2	0	0	+	+	0	0	0	0	
	2	1	0	0	0	0	0	0	0	0	
		2	0	0	+	0	0	0	0	0	
2	1	1	0	+	0	0	+	+	+	0	
		2	0	+	+	+	0	0	+	0	
	2	1	0	0	0	0	+	0	0	0	
		2	+	+	+	0	+	0	+	0	
A	B	C									

TABLE 6

Example of a 2^6 sparse table with a nonexistent MLE for the model specified by fixing all 2-way margins. The '+' signs indicate cells in a facial set corresponding to one facet of the marginal cone.

bigger than 1. For such small estimates, which correspond mostly to the cells for which the upper and lower integer bound is zero, the standard error is clearly very large. In fact, it is reasonable to expect that cells for which the maximal integer entries compatible with the fixed margins are zero will correspond to cell estimates with large standard errors. In this sense, cell bounds and maximum likelihood inference are strongly linked.

			1				2				D
			1		2		1		2		E
			1	2	1	2	1	2	1	2	F
1	1	1	0	0	0	0	+	0	+	0	
		2	0	+	0	0	0	+	0	0	
	2	1	+0	+0	+	0	+	+0	+	0	
		2	+0	+0	0	+	+0	+	0	+	
2	1	1	+	+0	+	0	+0	+0	+0	0	
		2	+0	+	+0	+	+0	+0	+0	+0	
	2	1	0	0	0	0	0	0	0	0	
		2	0	0	0	0	0	0	0	0	
A	B	C									

TABLE 7

Zero patterns when the margins [CDE], [ABCD], [ABCEF] are fixed.

4.2. Example 2: Data from the 1993 U.S. Current Population Survey. Table 9 describes data extracted from the 1993 Current Population Survey. Versions of these data have been used previously to illustrate

			1				2				D
			1		2		1		2		E
			1	2	1	2	1	2	1	2	F
1	1	1	[0,1]	**[0,0]**	[0,2]	**[0,0]**	[1,4]	[0,1]	[0,2]	[0,1]	
		2	**[0,0]**	[0,2]	**[0,0]**	[0,2]	[0,1]	[0,2]	[0,1]	[0,1]	
	2	1	[0,1]	**[0,0]**	[0,2]	**[0,0]**	[6,9]	[0,1]	[1,4]	[0,1]	
		2	**[0,0]**	[0,1]	**[0,0]**	[0,2]	[0,1]	[1,4]	[0,1]	[9,12]	
2	1	1	[15,18]	[0,1]	[0,4]	[0,1]	[0,1]	**[0,0]**	[0,1]	**[0,0]**	
		2	[0,1]	[2,5]	[1,2]	[1,5]	**[0,0]**	[0,1]	**[0,0]**	[0,1]	
	2	1	[0,1]	**[0,0]**	[0,2]	[0,1]	[0,1]	**[0,0]**	[0,1]	**[0,0]**	
		2	**[0,0]**	[0,1]	[0,1]	[0,2]	**[0,0]**	[0,1]	**[0,0]**	[0,1]	
A	B	C									

TABLE 8

Exact upper and lower bounds for the model obtained by fixing all positive 3-way margins.

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.

Variable	Label	Categories
Age (in years)	A	$< 25, 25 - 55, > 55$
Employer Type (*Empolyment*)	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	$< \$50K, \$50K+$

TABLE 9

Description of variables in CPS data extract.

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 C and G from 5 and 3 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

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

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. The marginal table [ABCFG] has 1 cell with a count of "5", while the margin [ACDGH] has a low count of "4" and two cells with a count of "8"; e.g., see Table 10. Even without any further analysis, most agencies would not release such margins. Because we are fitting a decomposable model, this initial exploratory analysis reveals that there will be at least one cell with a tight sharp upper bound of size "4". Now 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 8-way tables.

		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

TABLE 10
Marginal table [ACDGH] from 8-way CPS table.

Model 1 is decomposable and thus there are closed-form solutions for the bounds given the margins. Almost all lower bounds are 0. As expected from the analysis above, the smallest upper bound is 4. 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 1						Model 2					
Difference Cell count	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

TABLE 11

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

model we directly identified 12 out of 126 sensitive cells.

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 11. Thus releasing the margins corresponding to Model 1 poses a substantial risk of disclosure.

Model 2 is a non-decomposable log-linear model and it requires an iterative algorithm for parameter estimation and extensive calculations for computing the cell bounds. This model has 6 marginals as sufficient statistics. The 5 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], but 5-way margin [ACDGH] is still problematic.

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, which 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 an 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 reduce 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 11. 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

C	F	A G	1	2	3
1	1	1	1128	740	1893
		2	552	502	2271
	2	1	1416	1329	4381
		2	1117	1186	5677
2	1	1	1462	525	3069
		2	677	334	3039
	2	1	1268	1386	7442
		2	812	867	5769

TABLE 12
Upper bounds for [ACFG]

algebraic techniques and software such as LattE [8] or sampling techniques such as MCMC and SIS [5]. Due to the 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.

Based on Model 1, the 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 are only three small cell counts, two of size "2" and one of size "3", 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.

First, we recall that given only one margin, the lower bounds are all zero and the upper bound corresponds to the values of the observed margin. For example, given [ACFG], the smallest upper bound is 502 for the cell (211112), but for the small counts of "2" the upper bounds are 677 for (121112) and 1117 for (111122). Table 12 includes all of these upper bound values. We can carry out a similar analysis for the [ADEFG] margin.

For the decomposable model of interest above, the sharp bounds are easy to calculate. The upper bound for a particular cell is a minimum of the relevant marginal counts [ACFG] and [ADEFG]. The lower bound is the maximum between the 0 count and the value equal to the ([ACFG] +

[ADEFG]-[AFG]), where the marginal [AFG] is a separator in the decomposable model. The smallest bound for the whole table is on the sensitive cells (121112) with the count of 2; the bound is [0,15]. If we consider releasing the corresponding conditionals, e.g., [C|AFG] and[DE|AFG], we would obtain the same sharp bounds! In fact, any conditional that corresponds to the same marginal table and involves all variables of the marginal table will produce the same sharp bounds, e.g., [AFG|C] would have the same IP bounds as [ACFG]. The same argument holds for [DE|AFG] and [ADEFG]. Moreover, since the model is decomposable we can consider the pieces separately.

The LP relaxation bounds are typically much wider for the conditionals than for the corresponding margins, however, and the space of tables is different and often larger. For this example, due to computational complexity we were unable the obtain the counts via LattE.

5. Some Open Statistical Problems and Their Geometry. We present below a list of open problems that are pertinent to the topics introduced in this article. We purposely formulate them in rather general terms, as all of the problems pose challenges that are both of theoretical and computational nature, and we believe are relevant to the mathematical and statistical audience jointly.

5.1. Patterns for Non-Existence of MLEs.

PROBLEM 5.1. *Suppose that d_1 is allowed to grow, while keeping k and $d_2, \ldots, d_k$ fixed. What is the smallest integer d such that the number different patterns of zeros that lead to the non-existence of MLEs is constant for $d_1 > d$?*

Eriksson et al. [20] posed a related conjecture, and wondered whether some finite complexity properties of the facial structure of the marginal cone is related to the finite complexity properties of Markov bases proved in Santos and Sturmfels [36].

Carrying the algebraic statistical results on existence of MLEs to large sparse contingency tables in a fashion that allows relatively easy computational verification has proven to be difficult. Thus we pose the following challenge:

PROBLEM 5.2. *Given a marginal cone $\mathrm{C_A}$ and a vector of observed margins $\mathbf{t} = \mathrm{A}\mathbf{n}$, design an algorithm for finding the facial set associated with $\mathbf{t}$ that is computationally feasible for very large, sparse tables.*

5.2. Sharp Bounds.

Linear programming relaxation methods for the problem of computing integers bounds for cell entries will often produce fractional and non-sharp bounds, e.g., for Table 13.

In recent years, researchers became aware of the seriousness of the *integer gap* problem defined as the maximum difference between the real and integer bounds—see Hosten and Sturmfels [31] and Sullivant [43]. A

relevant example is presented in Table 13. The necessary and sufficient conditions for null integer gaps given in Sullivant [42] are the geometric counterpart to similar results by Dobra and Fienberg [13] already existing in the statistical literature.

		1		2		C
		1	2	1	2	D
1	1	0	1	1	0	
	2	1	0	0	0	
2	1	1	0	0	0	
	2	0	0	0	1	
A	B					

TABLE 13

An example of a table with integer gap of 1.67 for the entry $(1,1,1,1)$ *with fixed 2-way margins. For that cell the integer upper bounds is* 0*. Incidentally, we note that the MLE is defined and that the fiber contains one table only. Source: Hosten and Sturmfels [31].*

The generalized shuttle algorithm propose by Dobra [11] is based on a succeeding branch-and-bound approach to enumerate all feasible tables, thus adjusting the shuttle bounds to be sharpest, and implemented a parallel version of the enumerating procedure which permits efficient computation for large tables. Dobra and Fienberg [16] provide further details and applications. Because this algorithm substitutes for the traversal of all lattice points in the convex polytope, and this involves aspects of the *exact distribution* without the probabilities, it is not surprising that there are links with the issues of maximum likelihood estimation. When the margins correspond to decomposable graphs, the bounds have explicit representation (see [13]) and the branch and bound component is not needed. When they correspond to reducible graphs this component effectively works on the reducible components!

PROBLEM 5.3. *Can we formalize the algebraic geometric links for the bounds problem in a form that scales to large sparse tables?*

5.3. Markov Bases Complexity and Gaps in the Fiber. By a fiber with gaps we mean a fiber in which, for some of the cells entries, the range of integer values that are compatible with the given margins is not a finite sequence of integers, but instead contains gaps. In the presence of such gaps, knowledge of sharp upper and lower integer bounds for the cell entries cannot be a definitive indication of the safety of a data release. By construction, Markov bases preserve connectedness in the fiber and thus they encode the maximal degree of geometric and combinatorial complexity for all the fibers associated to a given log-linear model. De Loera and Onn [9] show that the complexity of Markov bases has no bound and thus there is little hope for an efficient computation of Markov bases for problems

$(:,j,k) =$						$(i,:,k) =$						$(i,j,:) =$			
2	1	2	0	2	0	2	1	2	3	0	0	2	2	2	2
1	0	2	0	0	2	2	1	0	0	2	1	3	1	1	1
1	0	0	2	2	0	0	0	2	1	2	3	2	2	2	2
0	1	0	2	0	2										

TABLE 14

Margins of a 3×4×6 table with a gap in the entry range for the $(1,1,1)$ *cell. Source: De Loera and Onn [9].*

of even moderate size, from the theoretical point of view. They also show in a constructive way that fibers can have large (in fact, arbitrarily large) gaps, a fact that can be quantified by the degree of the Markov moves.

PROBLEM 5.4. *What combinatorial and geometric tools allow us to assess and quantify gaps in a given fiber?*

These open problems have important implications for disclosure limitation methodologies. Table 14 gives an example of an integer gap for a 3×4×6 with fixed 2-way margins. The fiber contains only 2 feasible tables and the range entry for the first cell is $\{0,2\}$, thus exhibit a gap, since a value of 1 cannot be observed. In principle, it is possible to generate examples of tables with arbitrarily disconnected fiber.

Markov bases are *data independent*, in the sense that they prescribe all the moves required to guarantee connectedness for *any* fiber. However, there are instances which depend on the observed table $\mathbf{n}$, when in fact some (potentially many) of the moves are not needed: for example when the observed fiber contains gaps and when the observed margins lie on the boundary of the marginal cone $\mathrm{C_A}$.

PROBLEM 5.5. *Is it possible to reduce the computational burden of calculating Markov bases by computing only the moves that are relevant to the observed fiber* $\mathrm{P_t}$*?*

5.4. Bounds for Released Margins and Conditionals. The degree of Markov moves for given conditionals is arbitrary in a sense that it depends on the values of conditional probabilities, that is it depends on the smallest common divisor of the actual cell counts for a given conditional. In the disclosure limitation context, the database owner who knows the original cell counts can calculate the sharp LP and IP bounds, and the Markov bases for given conditionals only subject to the computational limitations of current optimization and algebraic software. In practice, however, the conditional values are reported as real numbers and depending on the rounding point the LP/IP bounds, the moves and thus the fibers generated for a given table will differ.

PROBLEM 5.6. *Characterize the difference between these bases, the fibers, and bounds due to rounding of the observed conditional probabilities.*

We have observed that the gap in the bounds for the cell counts, and thus the degree of gap of a given fiber, is more pronounced with conditionals than with the corresponding marginals. While the sharp bounds on the cells maybe the same, the fibers differ in their content and size resulting in different conditional distributions on the space of tables. This has important implications for exact inference, and disclosure limitation methods as certain conditionals may release less information than the corresponding margin. Consider a k-way contingency table, and two fibers; one for a matrix of conditional values $\mathbf{p}_{a|b}$ and a second for the corresponding margin $\mathbf{p}_{ab}$ where $a, b \subset \{1, ..., k\}$. The size of the first fiber will be greater than equal to the size of the second fiber. Also the Markov bases for $\mathbf{p}_{a|b}$ will include all the elements of the moves from fixed margin $\mathbf{p}_{ab}$ plus some additional ones. These observations lead to the following challenge:

PROBLEM 5.7. *Characterize the difference of two fibers, one for a conditional probability array and the other for the corresponding margin, and thus simplify the calculation of Markov bases for the conditionals by using the knowledge of the moves of the corresponding margin.*

This is related to the characterization of Theorem 3.2 when $\mathbf{p}_{a|b}$ and $\mathbf{p}_a$ do not uniquely identify the marginal table $\mathbf{p}_{ab}$.

REFERENCES

[1] Y.M.M. BISHOP, S.E. FIENBERG, AND P.W. HOLLAND (1975). *Discrete Multivariate Analysis: Theory and Practice*, MIT Press, Cambridge, MA. Reprinted (2007), Springer-Verlag, New York.

[2] L. BUZZIGOLI AND A. GIUSTI (1999). An algorithm to calculate the lower and upper bounds of the elements of an array given its marginals, in *Proceedings of the Conference on Statistical Data Protection*. Luxemburg: Eurostat, pp. 131–147.

[3] E. CARLINI AND F. RAPALLO (2005). *The geometry of statistical models for two-way contingency tables with fixed odds ratios,* Rendiconti dell'Istituto di Matematica dell'Università di Trieste, 37, 71–84.

[4] S.K. CHRISTIANSEN AND H. GIESE (19991). *Genetic analysis of obligate barley powdery mildew fungus based on RFPL and virulence loci,* Theoretical and Applied Genetics, 79, 705–712.

[5] Y. CHEN, I.H. DINWOODIE, AND S. SULLIVANT (2006). *Sequential importance sampling for multiway tables,* Annals of Statistics, 34, 523–545.

[6] L.H. COX (2002). *Bounds on entries in 3-dimensional contingency tables subject to given marginal totals,* In J. Domingo-Ferrer (Ed.), *Inference Control in Statistical Databases*, Springer-Verlag LNCS 2316, pp. 21–33.

[7] L.H. COX (2003). *On properties of multi-dimensional statistical tables,* Journal of Statistical Planning and Inference, 117, 251–273.

[8] J.A. DE LOERA, R. HEMMECKE, J. TAUZER AND R. YOSHIDA (2004). *Effective lattice point counting in rational convex polytopes,* Journal of Symbolic Computation, 38, 1273–1302.

[9] J.A. DE LOERA AND S. ONN (2006). *Markov bases of 3-way tables are arbitrarily complicated,* Journal of Symbolic Computation, 41, 173–181.

[10] P. DIACONIS AND B. STURMFELS (1998). *Algebraic algorithms for sampling from conditional distribution,* Annals of Statistics, 26, 363–397.

[11] A. Dobra (2002). *Statistical Tools for Disclosure Limitation in Multi-way Contingency Tables.* Ph.D. Dissertation, Department of Statistics, Carnegie Mellon University.

[12] A. Dobra (2003). *Markov bases for decomposable graphical models,* Bernoulli, 9(6), 1–16.

[13] A. Dobra and S.E. Fienberg (2000). *Bounds for cell entries in contingency tables given marginal totals and decomposable graphs,* Proceedings of the National Academy of Sciences, 97, 11885–11892.

[14] A. Dobra and S.E. Fienberg (2001). "Bounds for cell entries in contingency tables induced by fixed marginal totals with applications to disclosure limitation," Statistical Journal of the United Nations ECE, 18, 363–371.

[15] A. Dobra and S.E. Fienberg (2003). *Bounding entries in multi-way contingency tables given a set of marginal totals,* in Y. Haitovsky, H.R. Lerche, and Y. Ritov, eds., *Foundations of Statistical Inference: Proceedings of the Shoresh Conference 2000*, Physica-Verlag, 3–16.

[16] A. Dobra and S.E. Fienberg (2008). *The generalized shuttle algorithm,* in P. Gibilisco, Eva Riccomagno, Maria-Piera Rogantin (eds.) *Algebraic and Geometric Methods in Probability and Statistics*, Cambridge University Press, to appear.

[17] A. Dobra, S.E. Fienberg, and M. Trottini (2003). *Assessing the risk of disclosure of confidential categorical data*, in J. Bernardo et al., eds., *Bayesian Statistics 7*, Oxford University Press, 125–144.

[18] P. Doyle, J. Lane, J. Theeuwes, and L. Zayatz (eds.) (2001). *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies.* Elsevier.

[19] D. Edwards (1992). *Linkage analysis using log-linear models,* Computational Statistics and Data Analysis, 10, 281–290.

[20] N. Eriksson, S.E. Fienberg, A. Rinaldo, , and S. Sullivant (2006). *Polyhedral conditions for the non-existence of the MLE for hierarchical log-linear models,* Journal of Symbolic Computation, 41, 222–233.

[21] S.E. Fienberg (1999). *Fréchet and Bonferroni bounds for multi-way tables of counts With applications to disclosure limitation,* In *Statistical Data Protection, Proceedings of the Conference, Lisbon*, Eurostat, pp. 115–131.

[22] S.E. Fienberg, U.E. Makov, M.M. Meyer and R.J. Steele (2001). "Computing the exact distribution for a multi-way contingency table conditional on its marginal totals," in A.K.M.E. Saleh, ed., *Data Analysis from Statistical Foundations: A Festschrift in Honor of the 75th Birthday of D. A. S. Fraser*, Nova Science Publishers, Huntington, NY, 145–165.

[23] S.E. Fienberg and A. Rinaldo (2006). *Computing maximum likelihood estimates in log-linear models,* Technical Report 835, Department of Statistics, Carnegie Mellon University.

[24] S.E. Fienberg and A. Rinaldo (2007). *Three centuries of categorical data analysis: log-linear models and maximum likelihood estimation,* Journal of Statistical Planning and Inference, 137, 3430–3445.

[25] S.E. Fienberg and A.B. Slavkovic (2004a). *Making the release of confidential data from multi-way tables count,* Chance, 17(3), 5-10.

[26] S.E Fienberg and A.B. Slavkovic (2005). *Preserving the confidentiality of categorical databases when releasing information for association rules,* Data Mining and Knowledge Discovery, 11, 155–180.

[27] L. Garcia, M. Stillman, and B. Sturmfels (2005). *Algebraic geometry for Bayesian networks,* Journal of Symbolic Computation, 39, 331–355.

[28] E. Gawrilow and M. Joswig (2005). *Geometric reasoning with polymake,* Manuscript available at `arXiv:math.CO/0507273`

[29] D. Geiger C. Meek and B. Sturmfels (2006). *On the toric algebra of graphical models,* Annals of Statistics, 34, 1463–1492.

[30] S.J. Haberman (1974). *The Analysis of Frequency Data,* University of Chicago

Press, Chicago, Illinois.
[31] S. Hosten and B. Sturmfels (2006). *Computing the integer programming gap,* Combinatorica, 27, 367–382.
[32] S.L. Lauritzen (1996). *Graphical Models,* Oxford University Press, New York.
[33] R.B. Nelsen (2006). *An Introduction to Copulas.* Springer-Verlag, New York.
[34] A. Rinaldo (2005). *Maximum Likelihood Estimation for Log-linear Models.* Ph.D. Dissertation, Department of Statistics, Carnegie Mellon University.
[35] A. Rinaldo (2006). *On maximum likelihood estimation for log-linear models,* submitted for publication.
[36] F. Santos and B. Sturmfels (2003). *Higher Lawrence configurations,* J. Combin. Theory Ser. A, 103, 151–164.
[37] A.B. Slavkovic (2004). *Statistical Disclosure Limitation Beyond the Margins: Characterization of Joint Distributions for Contingency Tables.* Ph.D. Dissertation, Department of Statistics, Carnegie Mellon University.
[38] A.B. Slavkovic and B. Smucker (2007). *Calculating Cell Bounds in Contingency Tables Based on Conditional Frequencies,* Technical Report, Department of Statistics, Penn State University.
[39] A.B. Slavkovic 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.
[40] A.B. Slavkovic and S.E. Fienberg (2008). *The algebraic geometry of* 2×2 *contingency tables,* forthcoming.
[41] B. Sturmfels (1995). *Gröbner Bases and Convex Polytope*, American Mathematical Society, University Lecture Series, 8.
[42] S. Sullivant (2006). *Compressed polytopes and statistical disclosure limitation,* Tohoku Mathematical Journal, 58(3), 433-445.
[43] S. Sullivant (2005). *Small contingency tables with large gaps,* SIAM Journal of Discrete Mathematics, 18(4), 787–793.
[44] G.M. Ziegler (1998). *Lectures on Polytopes*, Springer-Verlag, New York.