Confidentiality Protection and Utility for Contingency Table Data: Algorithms and Links to Statistical Theory

Adrian Dobra,[1] Stephen E. Fienberg,[2,3] Alessandro Rinaldo,[2] and Yi Zhou[3]

[1] Department of Statistics and Center for Statistics and the Social Sciences, University of Washington
[2] Department of Statistics, Carnegie Mellon University
[3] Cylab and Machine Learning Department, Carnegie Mellon University

Abstract. One major strain of the statistical literature on disclosure limitation for contingency table data has focused on the the risk-utility tradeoff where utility has been measure either formally or informally in terms of information contained in marginal tables linked to a log-linear model analysis and risk has focused on disclosure potential of small cell counts, especially those equal to 1 or 2. Utility of margins for log-linear model analysis depends on estimability, e.g., existence of maximum likelihood estimates, and the ability to assess goodness-of-fit of models. One simple way to assess risk is to compute bounds for cell entries given a set of released marginals. Both of these methodologies become non-trivial to implement 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.

1 Introduction

The disclosure limitation literature for contingency table data is highly varied but over the past decade a substantial amount of it has focused on the the risk-utility tradeoff where risk has been measure 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 [15,16,23,26,28]). Among the ways considered for assessing risk have been the computation of bounds for cell entries, e.g., see [9,10,11,12,13,14], and counting of possible table realizations, e.g., see Fienberg and Slavkovic [28].

Recent advances in algebraic and polyhedral geometry have allowed us to gain greater insights into both the bounds problem and that of determining the existence of maximum likelihood estimates (MLEs) for log-linear models. Some of the recent literature on bounds referred to above has made direct use of algebraic geometry. Scaling up algebraic geometry methods to deal with large

tables remains a serious computational issue. The existence of MLEs is a major issue for large sparse contingency tables and it is now well-known that this is a direct function of the numbers and locations of the zero cell counts in a table. For details, see [31,1,20,25,39,40].

A key feature in both problems is the two-fold role of marginal tables, as a form of data release with significant implications for confidentiality and as minimal sufficient statistics for estimation. The formal links between these seemingly separate problems emanate from the common statistical and mathematical formalism of algebraic statistics. That the problems are related is somewhat surprising simply because zero cell values pose no direct disclosure limitation problem since they correspond to cells containing no respondents from a survey or a population dataset. But as we consider contingency tables of increasing dimensionality (i.e., numbers of cell) with a fixed total count, and then we condition on a collection of overlapping released marginal totals, we end up with very sparse tables whose entries become more constrained than one might naively expect.

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 for contingency table cell counts under log-linear models, and shows how these ideas can be made rigorous through the underlying mathematics of the same geometric/algebraic framework.

In the next section we illustrate that link in terms of 2×2 and $2 \times 2 \times 2$ tables and then we outline the technical mathematical details on which we draw. We illustrate the basic ideas using a series of numerical examples and data from two actual contingency tables, one with 64 cells and the other with approximately 60,000 cells. Others have written about bounds recently, see [4,5,36,37] but they have not examined in depth the proposal in Fienberg [21] as developed in detail in Dobra [9] and they have not discussed the link to existence of maximum likelihood estimates. We offer some clarifications and focus on the scaling up algorithms to deal with large sparse tables of practical interest.

2 Bounds for 2×2 and $2 \times 2 \times 2$ Tables and Their Generalizations

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

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

The extra lower bound component comes from the three upper bounds on the cells complementary to the (i, j) cell. These Fréchet bounds have been widely exploited in the disclosure limitation literature and have served as the basis for the development of statistical theory on copulas [38]. 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*, i.e., whose estimated expected values are expressible as explicit functions of marginal totals.

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 $n = n_{+++}$. Given the 2-way marginal totals, the bounds for the count in the (i, j, k) cell for $i = 1, 2$, $j = 1, 2$,and $k = 1, 2$, are

$$\begin{aligned} \min \ & (n_{ij+}, n_{i+k}, n_{+jk}, n_{ijk} + n_{\bar{i}\bar{j}\bar{k}}) \\ & \geq n_{ijk} \\ & \geq \max(n_{i++} - n_{i+k} - n_{ij+}, n_{+j+} - n_{ij+} - n_{+jk}, n_{++k} - n_{i+k} - n_{+jk}, 0) \end{aligned} \tag{2}$$

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) 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 $n_{ijk} + n_{\bar{i}\bar{j}\bar{k}}$. Again there is a link to maximum likelihood estimation but this time it is much more subtle. Under the no 2nd-order interaction model (e.g., see [1,31]), the minimal sufficient statistical marginals are the two-way totals. Further, Haberman [31] showed that the maximum likelihood estimates exist if and only if the minimal sufficient statistics are positive and, in addition, so are the pairs of cells (i, j, k) and $(\bar{i}, \bar{j}, \bar{k})$ for all possible values of i, j and k. Thus in particular, the MLEs do not exist when

$$n_{ijk} = n_{\bar{i}\bar{j}\bar{k}} = 0. \tag{3}$$

The extra component of the upper bound for this non-decomposable model seems inextricably bound up with the existence of MLEs and it is the generalization of this notion that the present article is built upon.

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 [9] then used a related construction with further refinements to develop what he referred to as the "generalized shuttle" algorithm, extending an idea in Buzzigoli and Giusti [2] that we can iterate between "naive" upper bounds and lower bounds in order to get sharp bounds. This algorithm has the nice property that it finds the sharp bounds in the decomposable case without extensive computation, and it can reduce the computation substantial in a number of other special cases. Nonetheless, it does not really scale well to situations involving large sparse contingency tables in part of the same reasons that the related non-integer bounds problem is known to be NP-hard, c.f., [4,5].

2.1 Non-sparse Example

We use the example of a real-life table to point out that nonexistence of the MLE, although generally to be expected in a sparse setting, can very well occur also in tables with large counts and very few empty cells.

Table 1 shows a $2 \times 2 \times 2 \times 3$ contingency table from [34], which describes the results of a clinical trial on the effectiveness on an analgesic drug, for patients of two statuses and centers. The sample size is large (193) with respect to the number of cells and, in fact, except for two zero counts, the cell counts are quite big.

With the goal of illustrating statistical disclosure limitation techniques and discussing the risk of disclosure associated to various marginal releases, Fienberg and Slavkovic [26] analyze two nested models which fit the data of Table 1:

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

The [CSR] margins has one zero entry, which causes the nonexistence of the MLE for both models. For reasons that will be explained in the discussion of the next example, since both models are decomposable the IPF algorithm does not provide any indications of nonexistence, except that some fitted values are zeros. Both of these models are decomposable and the IPF algorithm does not produce any indication of nonexistence, except that some fitted values are zeros.

Table 1. Results of clinical trial for the effectiveness of an analgesic drug. The variables are Center (C), Response (R), Status (S) and Treatment (T). Source: [34]. Using the software `MIM` (see [17]), the search for the best log-linear model for this data will result in three competing models: [CST][CSRT], [CST][CSR][TR] and [CST][CSR][CTR]. Close inspection will reveal that the MLE is not defined for any of these models because of the two zero entries, an anomaly that neither `MIM` nor `R` detected. In parentheses are the upper and lower bounds for the individual cell entries over all contingency tables with the same [CST][CSR][CTR] obtained using integer linear programming. See [26].

Center	Status	Response Treatment	Poor	Moderate	Excellent
1	1	Active	3 [0,6]	20 [9,28]	5 [0,13]
		Placebo	11 [8,14]	14 [6,25]	8 [0,13]
	2	Active	3 [0,6]	14 [6,25]	12 [4,17]
		Placebo	6 [3,9]	13 [2,21]	5 [0,13]
2	1	Active	12 [6,15]	12 [9,18]	**0** [**0**,**0**]
		Placebo	11 [8,17]	10 [4,13]	**0** [**0**,**0**]
	2	Active	9 [0,9]	3 [3,12]	4 [4,4]
		Placebo	6 [0,9]	9 [6,15]	3 [3,3]

3 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.

3.1 Technical Specifications and Geometrical Objects

Contingency tables are arrays of non-negative integers that arise from the cross-classification of a sample or a population of N objects based on a set of categorical variables of interest (see [1], [35]). We represent the contingency table $\mathbf{n}$ as a vector of non-negative integers, each indicating the number of times a given configuration of classifying criteria has been observed in the sample. We can fully specify log-linear models for the vector $\mathbf{p}$ of cell probabilities by a 0-1 design matrix A, in the sense that, for each $\mathbf{p}$ in the model, $\log \mathbf{p}$ belongs to the row span of A. The vector $\mathbf{t} = \mathrm{A}\mathbf{n}$ of marginal tables for the highest-order terms in the model gives a set of minimal sufficient statistics for the underlying parameters, e.g., see [1,31,35]. It is precisely because released margins are also minimal sufficient statistics that disclosure limitation techniques and inference for log-linear models are structurally linked and share much of the same statistical and mathematical formalism.

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. Their application has led to a series of theoretical results that offer novel, geometric representation of both the parameter and the sample spaces. We outline below the most relevant aspects of the algebraic statistics formalism, which essentially involves a representation of the interaction between parameter and sample space through geometric objects that can be explicitly described.

Parameter Space
In algebraic statistics we represent hierarchical log-linear models by certain polynomial maps. The parameter space, i.e., the set of probability distributions implied by such models, becomes a smooth hyper-surface of points satisfying polynomial equations, referred to as a *toric variety* [42]. Specifically, 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 vector in kernel(A), where $\mathbf{z}^+ = \{\max(\mathbf{z}(i), 0)\}$, $\mathbf{z}^- = \{-\min(\mathbf{z}(i), 0)\}$ and $\mathbf{p}^{\mathbf{z}} = \prod_i \mathbf{p}(i)^{\mathbf{z}(i)}$. For a 2×2 table and the model of independence the toric variety is the familiar surface of independence ([1]) and for general tables the toric variety is a hyper-surface.

The log-linear modeling paradigm hinges upon representing $\log \mathbf{p}$ as a point in the vector space spanned by the rows of A and thus constrain $\mathbf{p}$ to be strictly positive. In contrast, the algebraic geometry representation enjoys the crucial advantage of naturally providing an explicit representation of the closure of the parameter space, which consists of points in the simplex that belong to the

toric variety and have some zero coordinates. It is the possibility of identifying these points, both analytically via polynomials functions and geometrically, that allows for a full description of all possible patterns of sampling zeros leading to the nonexistence of the MLE.

Sample Space

Virtually all data-dependent objects encountered in the study of log-linear models are closed convex sets defined by linear inequalities. In particular, for a given log-linear model and a set of margins $\mathbf{t}$, consider the convex bounded set

$$P_{\mathbf{t}} = \{\mathbf{x} \text{ real, non-negative} : \mathbf{t} = A\mathbf{x}\}$$

of all real-valued non-negative tables having the same margins $\mathbf{t}$, computed using the design matrix A, where $P_{\mathbf{t}}$ is a high-dimensional polygon which can be explicitly described as a finite intersection of half-spaces (e.g., see [**?**]). The set of all integer points inside $P_{\mathbf{t}}$ is called the *fiber* of $\mathbf{t}$ and A. The fiber is the portion of the sample space associated with the same set of MSS margins and thus the support of the conditional distribution of tables given the margins, often called the *exact distribution*. Properties of the fiber are fundamental both for assessing the risk of disclosure and for conducting exact inference.

Three tasks of interest in disclosure limitation techniques are:

1. **Counting.** The simplest indication of the complexity of the fiber is its size, corresponding to the number of integer-valued tables with prescribed margins, i.e., the count of the lattice points in the polytope. Fienberg and Slavkovic [28] illustrate the use of `LattE` [6], a program for table enumeration, in the context of confidentiality. For the example in Table 1, there are 908 tables consistent with the margins [CST][CSR][CTR], 108,490 for the margins [CST][CSR][TR] and up to 65,419,200 for [CST][CSRT]. Counting is of particular interest to those who want to explore issues of confidentiality (e.g., see [28]).
2. **Sampling.** As the previous numbers show, the support of the conditional distribution can be quite big, in fact, so big that enumerating the points in the fiber may be an unrealistic task. Alternatively, one could characterize the fiber by means of *Markov moves*, i.e. integer valued vectors in kernel(A) that, added to the current table, will produce a new one with the same margins [8]. A *Markov basis* is a smallest set of moves that preserve connectedness in the fiber. Using Markov bases, it is possible to explore in a stochastic fashion the fiber and to estimate the conditional distribution of the tables given the margins. In most cases, Markov bases can only be computed with algebraic symbolic software, such as `4ti2` ([32]), based on algorithms that do not scale with the dimension of the problem and are not practical even for very small tables. For example, the Markov basis for a $4 \times 4 \times 4$ table and the model of no-second-order interaction consists of 145,512 moves, obtained with considerable computational efforts (see `http://math.harvard.edu/~seths/ccachallenge.html`).
3. **Optimizing.** An important task of great relevance for disclosure limitation techniques is integer linear programming over the fiber, e.g., see [41]. In

a limited sense, the technique solves the so-called "table entry data security problem," i.e., the computation of sharp lower and upper bounds for the individual table entries given a set of margins. These bounds allow for some assessment of the disclosure risk for a given marginal release. For example, the bounds for the entries in Table 1 show that the release of the [CST][CSR][CTR] margins will reveal with certainty four cell counts, namely the ones for which the upper and lower bounds coincide, and hence may be considered problematic [26]. Integer linear programming entails maximizing a linear function over the polyhedron $P_{\mathbf{t}}$, with the constraint that the solution has to be integral. Unlike linear programming, for which the integral constraint is removed, thus producing solutions that may be not sharp, it is computationally unfeasible with large tables (see, for example, [41]). Unfortunately, the linear programming solution is not guaranteed to be correct, as it may produce bounds that are fractional and not sharp (see, for example, Dobra et al. [11]).

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, called *marginal cone*, is an unbounded (here N is allowed to be any integer number) convex set consisting of all the linear combination of the columns of A with nonnegative coefficients, i.e.,

$$C_A = \{\mathbf{y} : \mathbf{y} = A\mathbf{x},\ \mathbf{x} \text{ real, non-negative}\}.$$

Since the margins are minimal sufficient statistics, the marginal cone provide the most efficient and parsimonious representation of the entire sample space. On the other hand, every point $\mathbf{t}$ in marginal cone C_A determines the polytope $P_{\mathbf{t}}$, which in turns contains the fiber, i.e., the portion of the sample space that is relevant for both statistical inference and disclosure limitation.

In summary, 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 hyper-surface representing the parameter space as a compact subset of the simplex.

3.2 Link between MLE and Bounds

The MLE $\widehat{\mathbf{p}}$, if it exists, is the unique point in the simplex such that $\log \widehat{\mathbf{p}}$ is in the row range of A and $A(N\widehat{\mathbf{p}}) = \mathbf{t}$ [31]. [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. Not only is this condition simple to interpret, but it also reduces the problem of detecting nonexistence of the MLE to a linear optimization program over a convex set. We 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 $\widehat{\mathbf{p}}^{e}$ [40]. The first condition implies $A(N\widehat{\mathbf{p}}^{e}) = \mathbf{t}$ 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*. 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 the MLE under the model of no-second-order interaction despite the margins being strictly positive. This table, along with others providing novel examples of "pathological" configurations of sampling zeros, were obtained using `polymake` (Gawrilow and Joswig [29]), a computational geometry software for the algorithmic treatment of convex polyhedra. 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.

Table 2. (a): Configurations of zero cells that cause nonexistence of the MLE 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 x indicate positive entries. Source: Fienberg and Rinaldo [25].

(a)

0	x	0
x	x	x
0	x	x

x	x	x
x	0	x
0	0	x

x	x	0
x	0	0
x	x	x

(b)

x	0	0
0	x	0
0	0	x

0	0	x
x	0	0
0	x	0

0	x	0
0	0	x
x	0	0

We use the geometric machinery to make the link between existence of the MLE and the computation of cells bounds explicit in the following Proposition:

Proposition 1. *For any lattice point* $\mathbf{t}$ *on the boundary of the marginal cone, let* $\widehat{\mathbf{p}}^{e}$ *be the extended MLE and let* $\mathcal{Z}_{\mathbf{t}} = \{i \colon \widehat{\mathbf{p}}^{e}(i) = 0\}$ *be the set of cells for which the extended MLE is zero. Then, each table* $\mathbf{n}$ *in the fiber 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 margin $\mathbf{t}$ and correspond to one of the many patterns of sampling zeros which invalidate the existence of the MLE.

The cells *not* in these configurations form a random set called the *facial set* (see [30,40]). The cells with positive entries in Tables 2 (a) and Table 5 are examples of facial sets. Proposition 1 then shows that the determination of the facial set associated to a given marginal table is key 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. This is, in fact, what emerges from the computational study conducted in
citeMLE:06, which suggests that the number of facial sets associated to a given hierarchical log-linear model may grow super-exponentially in the dimension of the table. 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 Fienberg and Rinaldo [**?**].

3.3 Open problems and their geometry

Sharp Bounds Linear programming relaxation methods for the problem of computing integers bounds for cell entries will often produce fractional and non-sharp bounds. See the example in Table 3. In recent years, researchers have made various attempts to quantify the maximal difference between the linear programming and integer linear programming solution, i.e., the *integer gap*. [44] construct pathological cases of contingency tables for dichotomous variables with exponentially large gaps. [33] give general algebraic conditions on the size of the integer gap. [43] uses the notion of compressed polytopes to derive necessary and sufficient conditions for a null integer gaps. These conditions include the well known case of decomposable models, for which Fréchet bounds are known to be sharp, but they are typically difficult to check for generic log-linear models, and they do not help with settings such as those involving reducible models [12] which may allow for a considerable simplifications.

The shuttle algorithm, originally put forward by Buzzugoli and Giusti [2], is a very efficient algorithm for computing integer bounds for cell entries that is based on the idea that the upper and lower bounds are interlinked, which means that bounds for some cells induce bounds for some other cells in the table. Dobra [9] generalized the shuttle algorithm by proposing 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. Because this algorithm is substituting 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. In the case of margins corresponding to decomposable graphs, the bounds have

		1	2	C
		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 3. 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: Hoşten and Sturmfels [33].

explicit representation (see [12]) and the branch and bound component is not needed, and when they correspond to reducible graphs this component effectively works on the reducible components!

Markov Bases Complexity and Disconnected Fiber By construction, Markov bases preserve connectedness in the fiber. This crucial feature implies, among other things, that Markov bases encode the maximal degree of geometric and combinatorial complexity for the fibers associated to a given log-linear model. De Loera and Ohn [7] indicate that the complexity of Markov bases has no bound and thus there is little hope for an efficient computation of Markov bases for problems of even moderate size, from the theoretical point of view. They also show in a constructive way that fibers can be largely (in fact, arbitrarily) disconnected, a fact that can be quantified by the degree of the Markov moves. A disconnected fiber implies that there may exist cells in the table for which the range of integer values that are compatible with the margins is not a finite sequence of integers, but contains instead gaps. In the presence of such gaps, it is apparent that the knowledge of sharp upper and lower integer bounds for the cell entries cannot be a definitive indication of the safety of a data release. The combinatorial and geometric assessment of the degree of disconnectedness of a given fiber is an open problem with important implications for disclosure limitation methodologies.

Table 4 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.

4 Examples

MLE Existence and Bounds. Table 5 shows a 4×4×4 table. There are 123 tables in the fiber. Table provides the cell bounds given the two-way marginals computed using the shuttle algorithm. Proposition 1 implies that the upper

(:,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 4. Margins of a 3×4×6 table with a gap in the entry range for the $(1, 1, 1)$ cell. Source: De Loera and Ohn [7].

bounds for the entries of the zero cells, which correspond to a set $\mathcal{Z}_{\mathbf{t}}$, is zero. Furthermore, it is easy to show 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.

(:,:,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 5. 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. Source: Fienberg and Rinaldo [25].

A Small Sparse Table. Edwards [18] reports on an analysis of genetics data in the form of a sparse 2^6 contingency table given in Table 7. 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 [3], included 37 loci for 81 offsprings, with 11 missing data—a very large sparse table.

(:,:,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 6. Sharp integer bounds of the 4×4×4 Table 5.

			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 7. Cell counts for the dataset analyzed by [18]. 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 (there is a total of 60 possible values for the margins). Using the program `polymake` ([29]), we found that the marginal cone for this model has 116,764 facets, each of them corresponding to a different pattern of sampling zeros causing nonexistence of the MLE , but only 60 of them produce null margins. Table 8 displays one facial set associated to one of these facets. The facet of the marginal cone specified by the unique null margins observed for the Table 7 has 11,432 facets.

			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 8. 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.

Table 9 shows the set $\mathcal{Z}_{\mathbf{t}}$ obtained when fixing the [ABCD][CDE][ABCEF] margins as the cells marked with a '0'. The '+' entries are cells for which the integer lower bound is positive, while '+0' cells indicate a 0 lower integer bound. The fiber in this case consists of 30 tables.

We note that Proposition 1 cannot be reversed, in the sense that null integer upper bounds for a set of cells does not imply the nonexistence of the MLE. In fact, Table 10 shows a set of sharp integer upper and lower bounds for a model for which the MLE exists! Despite the fact that there exist strictly positive real-valued tables in the fiber determined by the prescribed margins, there are cells, highlighted in red, 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 bigger than 1, while the smallest estimated mean value is 0.000691. 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 interlinked.

			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 9. Zero patterns when CDE, ABCD, ABCEF are fixed.

A Large Sparse Table. The 8-dimensional contingency table in Table 11 is a collapsed excerpt from a slightly altered version of the 1981 Australian population census, involving about 10 million individuals. The variables we consider are birthplace (BPL), sex (SEX), age (AGE), religious denomination (REL), marital status (MST), level of qualification (QAL), individual income (INC) and family income (FIN).

Structural zeros for some combinations of DUR and TIS, because these are impossible combinations, , but there are no reasons a priori to suspect that structural zeros may originate from the cross-classification of the remaining categories. Thus, we ignored the variables DUR and TIS and considered the resulting 9-way

															D
				1				2							E
			1		2			1		2					
			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 10. Exact upper and lower bounds for model obtained by fixing all positive 3-way margins.

	BPL	SEX	AGE	REL
#Categories	8	2	11	5
	MST	QAL	INC	FIN
#Categories	5	11	15	16

Table 11. Number of levels for the 8 variables extracted from the 1981 Australian population census.

table. Since even this reduced table is too large to be analyzed, we applied various collapsings over the levels of the variables $TALLY$, BPL and REL. Specifically, we aggregated 1) the levels of $TALLY$ into $TALLY \geq 3$ and $TALLY \leq 3$; 2) the levels of BPL into $Australia$, $England$, $OtherEurope$, $Asia$, $America$, $Africa$, $Ocean$, and $Other$; 3) the levels of REL into $Majority$ and $Minority$ where the label $Majority$ includes the original labels $CatholicRoman$, $CatholicNotRoman$, and $ChurchOfEngland$. The resulting 9-dimensional table is described in Table 11. Since it contains more than 13 millions of cells, a number of the same order of magnitude of the grand total. As more than half of the cells contains zero entries, this table offers an exemplary instance of large-dimensional sparse data sets, ideal as a test sets for many computations for cell bound and extended MLE.

We determined facial sets in a naive way by computing LP upper bounds for individual cells. (More efficient algorithms for the identification of facial sets are presented in Rinaldo [39] and are currently under development.) The rationale for this procedure is an immediate consequence of Proposition 1.

Corollary 1. *For any* $\mathbf{t}$ *in the marginal cone,* $\mathcal{Z}_{\mathbf{t}} = \{i\colon \sup_{\mathbf{x}\in P_{\mathbf{t}}} \mathbf{x}(i) = 0\}$.

Therefore, the real upper bound for an individual cell i is 0 if and only if $i \in \mathcal{Z}_{\mathbf{t}}$, a fact that can be checked by setting up a linear program, which can be run in parallel. Besides nonexistence of the MLE, this also implies $\max_{\mathbf{n}} \mathbf{x}(i) = 0$,

where the maximum is taken over all integer tables $\mathbf{n}$ inside the fiber. We note that $0 < \sup_{\mathbf{x} \in \mathrm{P_t}} \mathbf{x}(i) < 1$ would have the same implication.

We report on our results for the determination of the $\mathcal{Z}_\mathbf{t}$ sets for various sub-tables and different fixed margins. Specifically, we consider two 5-dimensional sub-tables 12 and 13 by collapsing over BPL, AGE and REL and over SEX, AGE and REL, respectively and the 6-dimensional sub-table 14 by collapsing over INC and FIN. The distributions of cell entries for these sub-tables, reported in Table 15, reveals a high degree of sparsity, whereby most of the entries are zeros or small counts.

Since existence of the MLE depends only on the position of the positive counts in the table and not on their values, we replaced all positive entries with 1. In order to perform linear programming optimization, we used `lp_solve` ([19]), a free linear/integer programming solver based on the revised simplex method and the branch-and-bound method for the integers, on a 128-processor linux Beowful cluster. For the 5-way Tables 12 and 13 we fixed the margins: [QAL, INC][SEX, MST, QAL][SEX, MST, INC][SEX, MST, FIN][SEX, INC, FIN] and [BPL, QAL][MST, QAL][QAL, INC][INC, FIN][BPL, MST, INC], respectively, and for the 6-way Tables 14 we fixed the margins [BPL, SEX, QAL][SEX, AGE, QAL][SEX, REL, QAL][SEX, MST, QAL][BPL, SEX, AGE, REL][BPL, SEX, REL, MST][SEX, AGE, REL, MST]. We chose these marginal configurations because they are all positive.

Table 16 summarizes these computations and gives the time cost. Since $|\mathcal{Z}_\mathbf{t}|$ is empty in all three cases the MLE exists and all estimated expected values are strictly positive. Given the high level of sparsity of these tables, however, we expect to observe a phenomenon similar to the one with describe earlier in our comment of Table 10, namely that many fitted values are very small and that there will be numerous instances of cells for which the upper and lower integer bound is zero.

	SEX	MST	QAL	INC	FIN
#Categories	2	5	11	15	16

Table 12. A 5-dimensional sub-table after collapsing the table over variable BPL, AGE and REL. 42% of cells are zero.

	BPL	MST	QAL	INC	FIN
#Categories	8	5	11	15	16

Table 13. A 5-dimensional sub-table after collapsing the table over variable SEX, AGE and REL. 65% of cells are zero.

	BPL	SEX	AGE	REL	MST	QAL
#Categories	8	2	11	5	5	11

Table 14. A 6-dimensional sub-table after collapsing the table over variable INC and FIN. 41% of cells are zero.

Cell count	No. Cells		
	Table 12	Table 13	Table 14
0	11613	69029	19771
1∼10	6722	20689	14107
11∼100	4168	10094	8965
101∼1000	2597	4360	4126
>1000	1300	1428	1431
Total cells	26400	105600	48400

Table 15. Distribution of the cell counts for the sub-tables 12, 13 and 14.

table	dimensionality	#cells	#equality constraints	time cost	$\lvert\mathcal{Z}_{\mathbf{t}}\rvert$
Table 12	5	26400	1065	14 minutes	0
Table 13	5	105600	1788	4 hours	0
Table 14	6	48400	2468	1 hour	0

Table 16. Summary of our computations for the Tables 12, 13 and 14.

5 Conclusions

In this paper, we follow up on ideas put forward by Fienberg [21] and elucidate some connections between the problem of computing cell bounds given a fixed set of margins and the existence of the maximum likelihood estimates for the cell counts under hierarchical log-linear models. We show that these two problems can be formulated using the same geometric framework of algebraic statistics and we describe the relevant geometric objects. We exemplify these results by presenting a variety of computations on simulated and real life examples of contingency tables of different dimensions and degrees of sparsity. In our calculations we relied a variety of software for symbolic algebra, computational geometry, linear and integer programming and the generalized shuttle algorithm.

ACKNOWLEDGMENTS

The research reported here was supported in part by NSF grants EIA9876619 and IIS0131884 to the National Institute of Statistical Sciences and by Army contract DAAD19-02-1-3-0389 to CyLab at Carnegie Mellon University. Fraser Jackson provided helpful suggests and other input.

References

1. Bishop, Y.M.M., Fienberg, S.E., and Holland, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice*, MIT Press, Cambridge, MA. Reprinted (2007), Springer-Verlag, New York.
2. Buzzigoli, L. and Giusti, A. (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. Christiansen, S. K. and Giese, H. (19991). "Genetic analysis of obligate barley powdery mildew fungus based on RFPL and virulence loci. " *Theoretical and Applied Genetics,* 79, 705–712.
4. Cox, L.H. (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.
5. Cox, L.H. (2003). "On properties of multi-dimensional statistical tables." *Journal of Statistical Planning and Inference*, 117, 251–273.
6. De Loera, A., Hemmecke, R., Tauzer, J. and Yoshida, R. (2003). "Effective lattice point counting in rational convex polytopes." Manuscript available at http://www.math.ucdavis.edu/~latte/pdf/lattE.pdf
7. De Loera, J.A. and Onn, S. (2006). "Markov bases of 3-way tables are arbitrarily complicated." *Journal of Symbolic Computation*, 41, 173–181.
8. Diaconis P. and Sturmfels, B. (1998). "Algebraic algorithms for sampling from conditional distribution." *Annals of Statistics*, 26, 363–397.
9. Dobra, A. (2002). *Statistical Tools for Disclosure Limitation in Multi-way Contingency Tables.* Ph.D. Dissertation, Department of Statistics, Carnegie Mellon University.
10. Dobra, A. (2003). "Markov bases for decomposable graphical models." *Bernoulli*, 9(6), 1–16.
11. Dobra, A., Erosheva, E.A., and Fienberg, S.E. (2003). "Disclosure limitation methods based on bounds for large contingency tables with application to disability data," in H. Bozdogan, ed., *Statistical Data Mining and Knowledge Discovery,* Chapman & Hall/CRC Press, New York, 93–116.
12. Dobra, A. and Fienberg, S.E. (2000). "Bounds for cell entries in contingency tables given marginal totals and decomposable graphs." *Proceedings of the National Academy of Sciences*, 97, 11885–11892.
13. Dobra, A. and Fienberg, S.E. (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.
14. Dobra, A. and Fienberg, S.E. (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.
15. Dobra, A., Fienberg, S.E., and Trottini, M. (2003). "Assessing the risk of disclosure of confidential categorical data." In J. Bernardo et al., eds., *Bayesian Statistics 7*, Oxford University Press, 125–144.

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

16. Duncan, G.T., Fienberg, S.E., Krishnan, R., Padman, R., and Roehrig, S.F. (2001). "Disclosure limitation methods and information loss for tabular data." In P. Doyle,

J. Lane, J. Theeuwes, and L. Zayatz, eds., *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies.* Elsevier, 135–66.
17. Edwards, D. (2000). *Introduction to Graphical Modelling*, 2nd edition, Springer-Verlag, New York.
18. Edwards, D. (1992). "Linkage analysis using log-linear models." *Computational Statistics and Data Analysis*, 10, 281–290.
19. Eikland K. and Notebaert, P. *lp_solve.* Manual and code available for download from http://lpsolve.sourceforge.net/5.5/.
20. Eriksson, N., Fienberg, S.E., and Rinaldo, A., and Sullivant, S. (2006). "Polyhedral conditions for the non-existence of the MLE for hierarchical log-linear models." *Journal of Symbolic Computation,* 41, 222–233.
21. Fienberg, S.E. (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. Fienberg, S.E. (2001). "Statistical perspectives on confidentiality and data access in public health." *Statistics in Medicine*, 20, 1347–1356.
23. Fienberg, S.E. (2004). "Datamining and disclosure limitation for categorical statistical databases." *Proceedings of Workshop on Privacy and Security Aspects of Data Mining, ICDM 2004.*
24. Fienberg, S.E. (2005). "Confidentiality and disclosure limitation." In *Encyclopedia of Social Measurement*, Academic Press, Vol. 1, 463–469.
25. Fienberg and Rinaldo (2007). "Three centuries of categorical data analysis: log-linear models and maximum likelihood estimation." *Journal of Statistical Planning and Inference*, 137, 3430–3445.
26. Fienberg, S.E. and Slavkovic, A.B. (2004a). "Making the release of confidential data from multi-way tables count." *Chance*, 17(3), 5–10.
27. Fienberg, S.E. and Slavkovic, A.B. (2004b). "Bounds for cell entries in two-way tables given conditional relative frequencies." In J. Domingo-Ferrer and V. Torra, eds., *Privacy in Statistical Databases: PSD 2004 Proceedings*, Lecture Notes in Computer Science, Volume 3050, Springer-Verlag, 30–43.
28. Fienberg, S.E. and Slavkovic, A.B. (2005). "Preserving the confidentiality of categorical databases when releasing information for association rules." *Data Mining and Knowledge Discovery,* 11, 155–180.
29. Gawrilow, E. and Joswig, M. (2005). "Geometric reasoning with polymake." Manuscript available at arXiv:math.CO/0507273
30. Geiger, D. and Meek, C. and Sturmfels, B. (2006) "On the toric algebra of graphical models." *Annals of Statistics*, 34, 1463–1492.
31. Haberman, S. J. (1974). *The Analysis of Frequency Data,* University of Chicago Press, Chicago, IL.
32. Hemmecke, R., Hemmecke, R., and Malokin, P. (2005). 4ti2 Version 1.2—Computation of Hilbert bases, Graver bases, toric Gröbner bases, and more. Manuscript available at http://www.4ti2.de
33. Hoşten, S. and Sturmfels, B. (2006). "Computing the integer programming gap." To appear in *Combinatorica.*
34. Koch, G.G. and Amara, J. and Atkinson, S. and Stanish, W. (1983). "Overview of categorical analysis methods." *SAS–SUGI*, 8, 785–795.
35. Lauritzen, S. L. (1996). *Graphical Models,* Oxford University Press, New York.
36. Lu, H., LI, Y., and Wu, X. (2006). "Disclosure Analysis for Two-Way Contingency Tables." In *Privacy in Statistical Databases (PSD'2006)*, Lecture Notes in Computer Science Vol. 4302, Springer-Verlag, Berlin, 57–67.

37. Lu, H. and LI, Y. (2007). "Practical Inference Control for Data Cubes." *IEEE Transactions on Dependable and Secure Computing (TDSC)*, accepted July, 2007.
38. Nelsen, R.B. (2006). *An Introduction to Copulas.* Springer-Verlag, New York.
39. Rinaldo, A. (2005). *Maximum Likelihood Estimation for Log-linear Models.* Ph.D. Dissertation, Department of Statistics, Carnegie Mellon University.
40. Rinaldo (2006). "On maximum likelihood estimation for log-linear models," submitted for publication.
41. Schrijver, A. (1998). *Theory of Integer and Linear Programming*, Wiley, New York.
42. Sturmfles, B. (1995). *Gröbner Bases and Convex Polytope*, American Mathematical Society, University Lecture Series, 8, Providence, RI.
43. Sullivant, S. (2005). "Compressed polytopes and statistical disclosure limitation." Manuscript available at http://math.berkeley.edu/~seths
44. Sullivant, S. (2005). "Small contingency tables with large gaps." *SIAM Journal of Discrete Mathematics,* 18(4), 787–79.
45. Ziegler, M. G. (1998). *Lectures on Polytopes*, Springer-Verlag, New York.