Markov Bases of Three-way Tables are Arbitrarily Complicated

Jesús A. De Loera [a], Shmuel Onn [b]

[a] *University of California at Davis, One Shields Ave. Davis, CA 95616, USA*

[b] *Technion- Israel Institute of Technology, 32000 Haifa, Israel*

Abstract

We show the following two universality statements on the entry-ranges and Markov bases of spaces of 3-way contingency tables with fixed 2-margins:

(1) For any finite set D of nonnegative integers, there are r, c, and 2-margins for $(r, c, 3)$-tables such that the set of values occurring in a fixed entry in all possible tables with these margins is D.

(2) For any integer n-vector d, there are r, c such that any Markov basis for $(r, c, 3)$-tables with fixed 2-margins must contain an element whose restriction to some n entries is d.

In particular, the degree and support of elements in the minimal Markov bases when r and c vary can be arbitrarily large, in striking contrast with 1-margined tables in any dimension and any format and with 2-margined (r, c, h)-tables with both c, h fixed.

These results have implications to confidential statistical data disclosure control. Specifically, they demonstrate that the entry-range of 2-margined 3-tables can contain arbitrary gaps, suggesting that even if the smallest and largest possible values of an entry are far apart, the disclosure of such margins may be insecure. Thus, the behavior of sensitive data under disclosure of aggregated data is far from what has been so far believed. Our results therefore call for the reexamination of aggregation and disclosure practices and for further research on the issues exposed herein.

Preprint submitted to The Journal of Symbolic Computation

Our constructions also provides a powerful automatic tool in constructing concrete examples, such as the possibly smallest 2-margins for $(6, 4, 3)$-tables with entry-range containing a gap.

Key words: Contingency tables, Markov bases, statistical data security, sampling and random generation, confidentiality of statistical data, transportation polytopes

1 Introduction

In this article we apply our recently discovered universality of 3-way transportation polytopes (De Loera and Onn 2004b) to two problems about the space of all contingency tables with fixed prescribed collections of margins.

The first problem concerns the *entry-range problem* - the set of all integer values that can occur in a fixed entry of any table with fixed margins. It is motivated by the practical problem of *confidential statistical data disclosure* facing agencies, such as the U.S. Census Bureau, wishing to maximize public access to information on their database while protecting confidentiality of individuals whose data is in the base (see for example (Cox 2002; Cox 2003; Duncan et al. 2001; Gusfield 1988; Irving and Jerrum 1994; Mehta and Patel 1983) and references therein). A common practice (Duncan et al. 2001), taken by the Bureau (American Factfinder), is to allow access to margins of tables according to some aggregation model, but not to the individual entries themselves. The security of an entry is closely related to the range of values it can attain in any table with the fixed released collection of margins: if the number of values that can occur is small, then the entry may be exposed, whereas if it is large, then the entry may be assumed secure. The common assumption by users of contingency tables and practitioners of data disclosure has been that the entry-range is always an interval. Thus, under this assumption, if the lower and upper bounds on an entry-range (computable by linear programming methods) are far apart, then the entry is safe. However, our results now show that, even for simple aggregation models, the entry-range can contain arbitrarily large gaps. Thus, the behavior of sensitive data under disclosure of aggregated data is far from what has been so far believed. Our results therefore call for the reexamination of aggregation and disclosure practices and for further research on the issues exposed herein.

The second problem concerns the structure of the *Markov basis* of an aggregation model - the set of moves that connects any pair of tables in the model that have the same set of margins. It is motivated by the problem of *sampling* the space of tables with fixed collection of margins according to the model, and estimating various statistics on that space. This topic has been stud-

ied by many authors, see for example (Aoki and Takemura 2003; Diaconis and Gangolli 1995; Diaconis and Sturmfels 1998; Dobra 2003; Hoşten and Sullivant 2002) and references therein. It is well-known that for any 1-margin models (e.g. 2-way tables with all 1-margins known) the elements of the Markov bases are simple vectors with $0, 1, -1$ entries. They have even been proved to be useful in random generation of contingency tables (see (Cryan et al. 2003; Cryan et al. 2002) and references therein). In drastic contrast, we show here that Markov bases of 3-way tables with given 2-margins are forced to contain entries of arbitrarily large size.

The case of 2-*margins of* 3-*way tables* this turns out to be, in a sense, the threshold case between simple models (where the entry-range is always an interval and the degree and support of the Markov basis elements are bounded) and complex models; we will discuss this below. Specifically, we consider *slim* 3-tables, by which we mean tables of format $(r, c, 3)$, that is, having r rows and c columns where r, c are variable, and fixed depth 3; note that this is the smallest depth for which the tables are genuinely 3-dimensional: tables of format (r, c, h) of depth $h \leq 2$ with fixed 2-margins are equivalent to 2-dimensional tables. We show the following statements (the precise formal definitions for the terms appearing in the statements are provided in following sections).

Theorem 1.1 Universality of entry-range: *For any finite set $D \subset \mathbb{N}$ of nonnegative integers, there are $r, c,$ and 2-margins for $(r, c, 3)$-tables such that the set of values occurring in a fixed entry in all possible tables with these margins is precisely D.*

Theorem 1.2 Universality of Markov bases: *For any nonnegative integer vector $d \in \mathbb{N}^n$, there are r and c such that any Markov basis for the model of $(r, c, 3)$-tables with fixed 2-margins must contain an element whose restriction to some n entries is precisely d. In particular, the degree and support of elements in the minimal Markov bases when r and c vary can be arbitrarily large.*

Example 1.3 Complex 2-margined 3-tables: The following (possibly smallest) collection of 2-margins of $(6, 4, 3)$-tables has entry-range (set of values occurring as the first entry $x_{1,1,1}$ in all possible tables with these margins) $D = \{0, 2\}$, and hence has a gap. Further, any Markov basis for 2-margined $(6, 4, 3)$-tables must contain an element of degree at least $d = 2$ and hence

is not square-free.

$$\begin{pmatrix} 2 & 1 & 2 & 0 & 2 & 0 \\ 1 & 0 & 2 & 0 & 0 & 2 \\ 1 & 0 & 0 & 2 & 2 & 0 \\ 0 & 1 & 0 & 2 & 0 & 2 \end{pmatrix}, \quad \begin{pmatrix} 2 & 1 & 2 & 3 & 0 & 0 \\ 2 & 1 & 0 & 0 & 2 & 1 \\ 0 & 0 & 2 & 1 & 2 & 3 \end{pmatrix}, \quad \begin{pmatrix} 2 & 3 & 2 \\ 2 & 1 & 2 \\ 2 & 1 & 2 \\ 2 & 1 & 2 \end{pmatrix} \quad .$$

The margins for this example, and moreover, margins that realize any desired entry-range $D \subset \mathbb{N}$, and an element realizing any desired $d \in \mathbb{N}^n$ that any minimal Markov basis must contain, can be automatically constructed following the proofs of Theorem 1.1 and Theorem 1.2. In fact, these procedures have been implemented in a computer and will be soon available online over the Internet (De Loera and Onn 2004c).

The above universality results for the class of 2-margined 3-tables of format (r, c, h) for any *fixed* $h \geq 3$ but variable r, c are compatible with the computational intractability of the problem of deciding the existence of any table of that format with given margins, established in (De Loera and Onn 2004a). In contrast, if both c, h are assumed fixed and only r is variable then the problem is polynomial time solvable (for *unary presented* margins), see (De Loera and Onn 2004a). This suggests that the class of models with both c, h fixed and only r variable is not universal; and indeed, it was recently shown in (Aoki and Takemura 2003; Hoşten and Sullivant 2003; Santos and Sturmfels 2003) that in this case the degree and support of any element of a minimal Markov basis can be bounded in terms of c and h only. An intriguing remaining open problem posed in (De Loera and Onn 2004a) regards the complexity when both c, h are fixed and only r is variable of deciding the existence of any table with given *binary presented* margins.

The proofs of Theorems 1.2 and 1.3 make use of the following result established recently in (De Loera and Onn 2004b). Roughly speaking it states that any rational convex polytope is in fact a 3-way transportation polytope:

Proposition 1.4 (De Loera - Onn (De Loera and Onn 2004b)) *Any polytope* $P = \{y \in \mathbb{R}^n_+ : Ay = b\}$ *with integer* $m \times n$ *matrix* A *and integer* m*-vector* b *is polynomial-time repre-*

sentable as a 3-way transportation polytope

$$T \quad = \quad \{\, x \in \mathbb{R}_+^{r\times c\times 3} \;:\; \sum_i x_{i,j,k} = w_{j,k}\,,\; \sum_j x_{i,j,k} = v_{i,k}\,,\; \sum_k x_{i,j,k} = u_{i,j} \,\}\,,$$

with $r = O(m^2(n+L)^2)$ *rows and* $c = O(m(n+L))$ *columns, where* $L := \sum_{j=1}^{n} \max_{i=1}^{m} \lfloor \log_2 |a_{i,j}| \rfloor$.

Here $\mathbb{R}_+$ denotes the set of nonnegative reals. A polytope $P \subset \mathbb{R}^p$ is *representable* as a polytope $Q \subset \mathbb{R}^q$ if there is an injection $\sigma : \{1,\ldots,p\} \longrightarrow \{1,\ldots,q\}$ such that the coordinate-erasing projection

$$\pi : \mathbb{R}^q \longrightarrow \mathbb{R}^p \;:\; x = (x_1,\ldots,x_q) \mapsto \pi(x) = (x_{\sigma(1)},\ldots,x_{\sigma(p)})$$

provides a bijection between Q and P and between the sets of integer points $Q \cap \mathbb{Z}^q$ and $P \cap \mathbb{Z}^p$.

Before proceeding to prove the above theorems and related statements, we set some terminology. A *d-table of size* $n = (n_1,\ldots,n_d)$ is an array of nonnegative integers $x = (x_{i_1,\ldots,i_d})$, $1 \le i_j \le n_j$. For any $0 \le k \le d$ and any k-subset $J \subseteq \{1,\ldots,d\}$, the *k-margin* of x corresponding to J is the k-table $x^J := (x^J_{i_j : j\in J}) := (\sum_{i_j : j\notin J} x_{i_1,\ldots,i_d})$ obtained by summing the entries over all indices *not in* J. For instance, the 2-margins of a 3-table $x = (x_{i_1,i_2,i_3})$ are its *line-sums* x^{12}, x^{13}, x^{23} such as $x^{13} = (x^{13}_{i_1,i_3}) = (\sum_{i_2} x_{i_1,i_2,i_3})$, and its 1-margins are its *plane-sums* x^1, x^2, x^3 such as $x^2 = (x^2_{i_2}) = (\sum_{i_1,i_3} x_{i_1,i_2,i_3})$.

An *aggregation model* is a triple $M = (d, J, n)$, where J is a family of subsets of $\{1,\ldots,d\}$ none containing the other and $n = (n_1,\ldots,n_d)$ is a tuple of positive integers. The model dictates the collection of margins for d-tables of size n to be specified. Our results concern the models $(3, \{12, 13, 23\}, (r, c, 3))$, that is, slim, $(r, c, 3)$-tables, with all three of their 2-margins specified. Note that we do not assume that the data is governed any *statistical* model. The reader is referred to (Dobra et al. 2003) for an introduction to the problems that arise when one does assume, for instance, that the data is a result of i.i.d. draws from any particular statistical model.

Finally, for an aggregation model $M = (d, J, n)$ and a specified margin collection $u = (u^J : J \in J)$ under the model M, the corresponding set of *contingency tables* with collection of margins u is

$$C(M;u) \quad := \quad \{\, x \in \mathbb{N}^{n_1\times\cdots\times n_d} \;:\; x^J = u^J\,,\; J \in J \,\}\,.$$

2 Markov Bases

We start with the proof that Markov Bases of 3-way tables are universal as stated in Theorem 1.2. Fix any model $M = (d, J, n)$. A *Markov basis* for M is a set of integer arrays $B(M) \subseteq \mathbb{Z}^{n_1 \times \cdots \times n_d}$ that *connects* every pair of tables having the same margins in the model. More precisely, every $m \in B(M)$ has zero margins ($m^J = 0 \,:\, J \in J$), and for any $x, y \in \mathbb{N}^{n_1 \times \cdots \times n_d}$ with ($x^J = y^J \,:\, J \in J$), there is a sequence of elements $m^1, \ldots, m^k$ in $B(M)$, possibly with repetitions, such that $y = x + \sum_{j=1}^{k} m^j$ and $x + \sum_{j=1}^{i} m^j$ is nonnegative for $i = 1, \ldots, k$.

Proof of Theorem 1.2. We need to show that for any n and any nonnegative integer vector $d \in \mathbb{N}^n$ there are r and c such that any Markov basis $B(M)$ for the model $M := (3, \{12, 13, 23\}, (r, c, 3))$ must contain an element m with the following property: the restriction of m to some n of its table entries, indexed by some n triples $\sigma(1), \ldots, \sigma(n) \in [r] \times [c] \times [3]$, coincides with d, that is, $d_i = m_{\sigma(i)}$ for $i = 1, \ldots, n$. Consider the polytope

$$P \quad := \quad \{y \in \mathbb{R}^{n+2}_{+} \,:\, y_0 + y_{n+1} = 1\,, \quad d_j \cdot y_0 - y_j = 0\,, \quad j = 1, \ldots, n\}\,.$$

By Proposition 1.4, there are r, c and $(u_{i,j}), (v_{i,k}), (w_{j,k})$ such the corresponding transportation polytope T represents P. Let $\sigma : \{0, 1, \ldots, n+1\} \longrightarrow [r] \times [c] \times [3]$ be the injection giving that representation, which in particular embeds y_i as $x_{\sigma(i)}$ for $i = 1, \ldots, n$. The right-hand-side data for T naturally gives a 2-margin collection $u = (u^{12}, u^{13}, u^{23})$ by $(u^{12}_{i,j}) := (u_{i,j})$, $(u^{13}_{i,k}) := (v_{i,k})$ and $(u^{23}_{j,k}) := (w_{j,k})$.

Clearly, the set $C(M; u)$ of contingency tables is the set of integer points in T, and by Proposition 1.4, T is integer equivalent to P. Now, P contains precisely two integer points, $y^1 := (0, 0, \ldots, 0, 1)$ and $y^2 := (1, d_1, \ldots, d_n, 0)$. Let x^1, x^2 be the corresponding tables in $C(M; u)$. Any Markov basis $B(M)$ of M must connect these tables, and since they are the only ones in $C(M; u)$, it must be that $m := x^2 - x^1$ is in $B(M)$. But then indeed $d_i = m_{\sigma(i)}$ for $i = 1, \ldots, n$ as desired. □

As pointed out by one of the referees, the above proof can be easily modified to extend Theorem 1.2 to arbitrary integer (and not necessarily nonnegative) vectors, giving the following nice

corollary.

Corollary 2.1 *For any integer vector $d \in \mathbb{Z}^n$, there are r and c such that any Markov basis for the model of $(r, c, 3)$-tables with fixed 2-margins must contain an element whose restriction to some of its entries is precisely d.*

Proof. Let $d^+, d^- \in \mathbb{N}^n$ be the positive and negative parts of d defined as usual by $d_i^+ = \max\{0, d_i\}$ and $d_i^- = -\min\{0, d_i\}$. Then the corollary is obtained by simply repeating the above proof of Theorem 1.2 starting from the polytope

$$P \quad := \quad \{(y, z) \in \mathbb{R}_+^{2(n+1)} \ : \ y_0 + z_0 = 1\,, \ \ d_j^+ \cdot y_0 - y_j = 0\,, \ \ d_j^- \cdot z_0 - z_j = 0\,, \ \ j = 1, \ldots, n\,\}\,. \quad \square$$

As usual, this can be lifted to the language of toric ideals in the corresponding algebra $\mathbb{C}[X] = \mathbb{C}[X_{i_1,\ldots,i_d}]$ of complex polynomials with variables indexed by table entries. Each table $x = (x_{i_1,\ldots,i_d})$ lifts to a monomial $X^x := \prod_{i_1,\ldots,i_d} X_{i_1,\ldots,i_d}^{x_{i_1,\ldots,i_d}}$. The model M gives rise to a *model toric ideal* I_M generated by all binomials coming from pairs of tables x, y with same margins in the model, that is,

$$I_M \quad := \quad \text{ideal}\{\, X^x - X^y \ : \ x^J = y^J \text{ for all } J \in \mathcal{J}\,\}\,.$$

It was shown in (Diaconis and Sturmfels 1998) that a set G of binomials in the toric ideal I_M generates it if and only if the corresponding set of integer arrays $B := \{x - y : X^x - X^y \in G\}$ is a Markov basis for the model M. This provides a fundamental link between commutative algebra and aggregation model theory; in particular, a finite Markov basis always exists and is computable by Gröbner bases methods. Corollary 2.1 has the following interesting implication on the complexity of "slim 3-way" toric ideals.

Corollary 2.2 Universality of model toric ideals: *For any vector $d \in \mathbb{Z}^n$, there are r, c such that any minimal generating set of the ideal I_M of the model $M = (3, \{12, 13, 23\}, (r, c, 3))$ must contain a binomial satisfying the following: one of its monomials restricted to a suitable subset of variables has multi-degree d^+ and its other monomial restricted to a suitable subset of variables has multi-degree d^-.*

3 Entry-range

Next, we consider entry-ranges. Permuting coordinates, we may always consider the first entry $x_{\mathbf{1}}$, where $\mathbf{1} := (1, \ldots, 1)$. The *entry-range* of a collection of margins u under model M is the set $R(M;u) := \{x_{\mathbf{1}} : x \in C(M;u)\} \subset \mathbb{N}$ of values $x_{\mathbf{1}}$ can attain in any table with these margins.

We start with the following characterization of the entry-ranges of 1-margin models in any dimension and of any format, showing that they are always intervals and hence presumably secure whenever the lower and upper bounds on the entry-range are far apart.

Proposition 3.1 *For any 1-margin model $M = (d, \{1, 2, \ldots, d\}, (n_1, \ldots, n_d))$ and any collection of margins $u = (u^1, \ldots, u^d)$ under M, the entry-range is an interval, that is, for some $a, b \in \mathbb{N}$*

$$R(M;u) \quad = \quad [a,b] \quad := \quad \{r \in \mathbb{N} : a \leq r \leq b\} \quad .$$

Proof. It is well known that any such 1-margin model M admits a $\{0, \pm 1\}$-valued Markov basis $B(M)$. Suppose indirectly that there is a collection of margins u under M for which the entry-range is not an interval. Thus, there are nonnegative integers a and $b \geq a + 2$ such that there are tables $x, y \in C(M;u)$ with $x_{\mathbf{1}} = a$ and $y_{\mathbf{1}} = b$, but no table $z \in C(M;u)$ with $z_{\mathbf{1}} = a + 1$. But then any table $z = x + \sum_{j=1}^{k} m^j$ reachable from x by an admissible sequence of elements $m^1, \ldots, m^k$ in $B(M)$ must satisfy $z_{\mathbf{1}} \leq a$, so x, y are not connected by the Markov basis $B(M)$, a contradiction. □

Our universality result Theorem 1.1 stands in contrast with the situation of Proposition 3.1 for 1-margined models and with recent attempts by statisticians to better understand entry behavior of slim 3-tables (see e.g. (Cox 2002; Cox 2003; Duncan et al. 2001)), and implies that entry-ranges of 2-margined slim 3-table models consist of all finite sets of nonnegative integers. This shows in particular that the entry-range can contain arbitrarily large gaps and so, even if the lower and upper bounds on the entry-range are far apart, the entry may be vulnerable. Thus, the behavior of sensitive data under disclosure of aggregated data is far from what has been so far believed, and this result calls for the reexamination of aggregation and disclosure practices and for further research on the subject.

Here is the proof, making use again of Proposition 1.4.

Proof of Theorem 1.1. We need to show that for any finite subset $D \subset \mathbb{N}$ there are r, c and 2-margins $u = (u^{12}, u^{13}, u^{23})$ for the model $M := (3, \{12, 13, 23\}, (r, c, 3))$ such that the corresponding entry-range $R(M; u)$ equals D. Let then $D = \{d_1, \ldots, d_n\} \subset \mathbb{N}$ be any such set. Consider the polytope

$$P \quad := \quad \{y \in \mathbb{R}^{n+1}_+ \ : \ y_0 - \sum_{j=1}^{n} d_j \cdot y_j = 0 \, , \ \sum_{j=1}^{n} y_j = 1 \} \ .$$

By Proposition 1.4, there are r, c and $(u_{i,j}), (v_{i,k}), (w_{j,k})$ such the corresponding transportation polytope T represents P. By suitable permutation of coordinates, we can assume that the injection σ giving that representation satisfies $\sigma(0) = (1, 1, 1)$, embedding y_0 as $x_{\mathbf{1}} = x_{1,1,1}$. Again, the right-hand-side data for T naturally gives a 2-margin collection $u = (u^{12}, u^{13}, u^{23})$ by $(u^{12}_{i,j}) := (u_{i,j})$, $(u^{13}_{i,k}) := (v_{i,k})$ and $(u^{23}_{j,k}) := (w_{j,k})$. Again, the set $C(M; u)$ of contingency tables is the set of integer points in T, and by Proposition 1.4, T is integer equivalent to P. The entry-range is therefore, as desired,

$$\begin{aligned} R(M; u) &= \{x_{\mathbf{1}} \ : \ x \in C(M; u)\} \\ &= \{x_{\mathbf{1}} \ : \ x \in T \cap \mathbb{Z}^{r \times c \times 3}\} \ = \ \{y_0 \ : \ y \in P \cap \mathbb{Z}^{n+1}\} \ = \ D \ . \quad \square \end{aligned}$$

An automatic universal generator and Example 1.3 revisited: The procedures described in the proofs of Theorems 1.1 and 1.2 have been implemented and will be soon available online, see (De Loera and Onn 2004c). For instance, the following 2-margins for $(16, 11, 3)$-tables giving entry-range $D = \{0, 2\}$ can be produced that way starting with the polytope in three variables

$P = \{y \geq 0 : y_0 - 2y_1 = 0,\ y_1 + y_2 = 1\}$,

$$\begin{pmatrix}
2 & 0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 \\
2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 & 2 \\
2 & 2 & 2 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 2 & 2 & 2 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 2 & 2 & 2 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 2 & 2 & 2 \\
2 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 2 & 0 & 0 & 0 \\
0 & 2 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 2 & 0 & 0 \\
0 & 0 & 2 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 2 & 0 \\
0 & 0 & 0 & 2 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 2
\end{pmatrix},$$

$$\begin{pmatrix}
2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 \\
4 & 0 & 0 & 0 & 0 & 2 & 2 & 0 & 0 & 2 & 2 & 0 & 0 & 0 & 0 & 4 \\
2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2
\end{pmatrix}, \quad \begin{pmatrix}
4 & 1 & 3 & 6 & 6 & 6 & 6 & 0 & 0 & 0 & 0 \\
2 & 3 & 3 & 0 & 0 & 0 & 0 & 2 & 2 & 2 & 2 \\
0 & 0 & 0 & 2 & 2 & 2 & 2 & 6 & 6 & 6 & 6
\end{pmatrix} \quad ;$$

with a suitable human short cut (which has to do with certain choices of parameters in the construction underlying Proposition 1.4 and which could also be coded with suitable extra effort), it is possible to get it down to $(6, 4, 3)$-tables as in Example 1.3.

We conclude with the notion of spectrum. The *spectrum* of a model $M = (d, J, n)$ is the family of all entry-ranges of collections of margins under the model,

$$\mathrm{Spec}(M) := \{R(M; u) \,:\, u = (u^J : J \in J) \text{ some margin collection under } M\}\,.$$

The spectrum of a class C of models is the union $\mathrm{Spec}(C) := \bigcup_{M \in C} \mathrm{Spec}(M)$ of spectra of its models.

With this terminology, Theorem 1.1 says that the spectrum of the class of 2-margined slim 3-table models consists of all finite subsets of $\mathbb{N}$ and hence the mere computation of lower and upper bounds on the entry-range does not provide sufficient information for deciding whether the disclosure of margins is secure or not: thus, further study is necessary in order to quantify the meaning of safe disclosure in such complex classes of aggregation models. On the other

hand, Proposition 3.1 says that the spectrum of the class of 1-margined models of any format consists of intervals only and hence that class is simple and its security is presumably directly related to the difference between the upper and lower bounds on the entry-range. Finally, the results of (Hoşten and Sullivant 2003) describe a broad class of hierarchical models for which the complexity of the spectrum is somewhere in between, indicating some ambiguity regarding the security of these models. We pose as an interesting research direction the classification of spectra of classes of models and the determination of the universal ones.

4 Concluding Remarks

As mentioned before, the results of this paper, in particular those in Section 3, indicate that the behavior of sensitive data under disclosure of aggregated data is far from what has been so far believed. This calls for the reexamination of aggregation and disclosure practices and for further research on the issues exposed herein.

In particular, we conclude with the following interesting questions raised by one of the referees. As pointed out by the referee, the constructions in this paper involve tables with zero margins. Can the constructions be strengthened to table spaces with no zero margins? Is it possible to determine whether the universality behavior is frequent or sparse? Is there a class of real data where this behavior occurs?

Also, as pointed out by the referee, from a practical standpoint, the sample size of a survey or the number of people who are taken in a census is some fixed amount. How does the complexity of various aggregation models, in particular 2-margined 3-tables, change if the grand total of all entries is fixed and the size of the table gets large?

Acknowledgments

The first author is grateful for support received by the NSF through grant DMS-0309694, the 2003-2008 UC Davis Chancellor fellow award, and an Alexander von Humboldt fellowship. The

second author was supported in part by a grant from ISF - the Israel Science Foundation, by AIM - the American Institute of Mathematics, by the Technion President Fund, and by the Fund for the Promotion of Research at the Technion. He thanks the American Institute of Mathematics and the organizers of the AIM Research Workshop on Computational Algebraic Statistics (December 2003, Palo Alto) for inviting him to this workshop, which provided the stimulation and inspiration for the research described herein. Both authors thank the AIM for its support, and the referees for very helpful suggestions that improved the presentation of this paper and for suggesting Corollary 2.1 which strengthens Theorem 1.2.

References

[Aoki and Takemura 2003] Aoki, S., Takemura, A.: Minimal basis for connected Markov chain over $3 \times 3 \times K$ contingency tables with fixed two-dimensional marginals. Austr. New Zeal. J. Stat. **45** (2003) 229–249

[Cox 2002] Cox, L.H.: Bounds on entries in 3-dimensional contingency tables. Inference Control in Statistical Databases - From Theory to Practice, Lec. Not. Comp. Sci. **2316** 21–33, Springer, New York, 2002

[Cox 2003] Cox, L.H.: On properties of multi-dimensional statistical tables. J. Stat. Plan. Infer. **117** (2003) 251–273

[Cryan et al. 2002] Cryan, M., Dyer, M., Goldberg, L.A., Jerrum, M., and Martin R. Rapidly Mixing Markov Chains for Sampling Contingency Tables with a Constant Number of Rows, Proceedings of the 43rd Annual Symposium on Foundations of Computer Science (FOCS 02), 711–720, 2002.

[Cryan et al. 2003] Cryan, M., Dyer, M., Müller, H., Stougie, L.: Random walks on the vertices of transportation polytopes with constant number of sources. Proc. 14th Ann. ACM-SIAM Symp. Disc. Alg. (Baltimore, MD) 330–339, ACM, New York, 2003

[De Loera and Onn 2004a] De Loera, J., Onn, S.: The complexity of three-way statistical tables. SIAM J. Comp. **33** (2004) 819–836

[De Loera and Onn 2004b] De Loera, J., Onn, S.: All rational polytopes are transportation polytopes and all polytopal integer sets are contingency tables. Proc. 10th Ann. Math. Prog. Soc. Symp. Integ. Prog. Combin. Optim., (Columbia University, New York), Lec. Not. Comp. Sci., Springer, New York, 3064:338–351, 2004

[De Loera and Onn 2004c] De Loera, J., Onn, S.: Universal Generator (2004), available online at
http://www.math.ucdavis.edu/∼deloera, http://ie.technion.ac.il/∼onn

[Diaconis and Gangolli 1995] Diaconis, P., Gangolli, A.: Rectangular arrays with fixed margins. In: Discrete Probability and Algorithms (Minneapolis, MN, 1993), IMA Vol. Math. App. **72** 15–41, Springer, New York, 1995

[Diaconis and Sturmfels 1998] Diaconis, P., Sturmfels, B.: Algebraic algorithms for sampling from conditional distributions. Ann. Statist. **26** (1998) 363–397

[Dobra 2003] Dobra, A.: Markov bases for decomposable graphical models. Bernoulli **9** (2003) 1093–1108

[Dobra et al. 2003] Dobra, A., Fienberg, S.E., Torttini M.: Assessing the risk of disclosure of confidential categorical data. In: Bayesian Statistics 7 (Bernardo, J.M., Bayarri, M.J., Berger, J.O., Dawid, A.P., Heckerman, D., Smith, A.F.M., West, M. eds.), Oxford University Press 125–144, 2003

[Duncan et al. 2001] Duncan, G.T., Fienberg, S.E., Krishnan, R., Padman, R., Roehrig, S.F.: Disclosure limitation methods and information loss for tabular data. In: Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies (Doyle, P., Land, J.I., Theeuwes, J.M., Zayatz, L.V. eds.), North-Holland, 2001

[Gusfield 1988] Gusfield, D.: A graph theoretic approach to statistical data security. SIAM J. Comp. **17** (1988) 552–571

[Hoşten and Sullivant 2002] Hoşten, S., Sullivant, S.: Gröbner bases and polyhedral geometry of reducible and cyclic models. J. Combin. Theory Ser. A **100** (2002) 277–301

[Hoşten and Sullivant 2003] Hoşten, S., Sullivant, S.: Finiteness theorems for Markov bases of hierarchical models, draft, (2003)

[Irving and Jerrum 1994] Irving, R., Jerrum, M.R.: Three-dimensional statistical data security problems. SIAM J. Comp. **23** (1994) 170–184

[Mehta and Patel 1983] Mehta, C.R., Patel, N.R.: A network algorithm for performing Fisher's exact test in $r \times c$ contingency tables. J. Amer. Stat. Assoc. **78** (1983) 427–434

[Santos and Sturmfels 2003] Santos, F., Sturmfels, B.: Higher Lawrence configurations. J. Combin. Theory Ser. A **103** (2003) 151–164

[American Factfinder] U.S. Census Bureau, *American FactFinder*, http://factfinder.census.gov/servlet/BasicFactsServlet