Partial Information Releases for Confidential Contingency Table Entries: Present and Future Research Efforts

Aleksandra B. Slavković
Department of Statistics
Pennsylvania State University
University Park PA 16802
sesa@stat.psu.edu

Abstract

Tabular data have been a staple product for disseminating information derived from the confidential microdata that fuel social science research and inform policy decisions. This paper outlines recent results on disclosure risk assessment associated with the release of high-dimensional contingency tables, and discusses some related research problems. The main focus is the *partial information release* strategy, through which agencies can release relevant marginal and conditional tables along with the sample size instead of a full contingency table. The most recent approaches in this area combine statistical and operations research methodologies with tools from computational algebra.

1 Introduction

Data privacy is an overarching concern in modern society, as government and non-government agencies alike collect, archive, and release increasing amounts of sensitive personal data. As the amount of data accumulates in the public realm and record-linkage methodologies improve, the threat to confidentiality and privacy magnifies. Certainly, many ethical, legal, and pragmatic considerations inhere in the issue of data privacy, and government agencies in particular are making efforts to critically evaluate both the type of data that are made publicly available from statistical databases and the format of the data product releases.

The confidential microdata that fuel social science research and inform policy decisions have long been disseminated in the derivative form of tabular data. Tabular data come in two formats: (1) *magnitude tables*, whose cells contain aggregates of non-negative reported values, such as income or sales volumes, and (2) *contingency tables*, whose cells contain frequency counts determined by cross-classifying a sample of n individuals from a large population of size N by their k attributes.

Agencies seek to report maximum information from such tables without releasing data that would allow individuals to be identified with a high degree of probability. Despite this proviso, though, the released data must be useful to the analyst, i.e., statistical inferences drawn from the selected data must be identical to statistical inferences drawn from the full k-way table. For an extensive literature on the analysis of contingency tables, see [3] and [2].

The goal of this paper is to outline recent results regarding disclosure risk assessment associated with the release of high-dimensional contingency tables, and to consider related open research problems. It should be noted that numerous approaches to investigating data privacy issues have been proposed in statistics- and computer science-related fields. Originating from the work done by national statistical offices, statistical disclosure limitation (SDL) or statistical disclosure control (SDC) methods aim to optimize the trade-off between data utility and disclosure risk by modifying either the microdata or the tables. Many SDL techniques are appropriately described as *data masking* or *matrix masking* both of which introduce bias and variance to data in order to minimize identity and attribute disclosure while trying to retain sufficient information for proper statistical inference.

Traditional approaches for tabular data include the use of (1) recoding (e.g., rounding and thresholding), (2) cell suppression, (3) data swapping, and (4) perturbation. The more modern approaches, which rely on sampling and simulation techniques, include the use of (1) synthetic data, (2) remote access servers, (3) secure computations, and (4) partial information releases. A full description of these techniques and their applications is beyond the scope of this paper, but see [26], [11], and [12] for more details.

This paper focuses on a *partial information release* strategy through which agencies can release relevant marginal and conditional tables along with sample size instead of releasing a full contingency table. Many categorical data summaries are in the form of marginal tables, but agencies also often release rates or percentages representing proportions of individuals who fall in a certain category given some other characteristic(s) (see [19], p. 7 for an example). Furthermore, the conditionals preserve association measures, such as odds and odds-ratios, that are relevant for data utility.

The next section outlines the current methodologies and issues associated with partial information releases from k-way contingency tables. These modern approaches have connections to algebraic geometry, combinatorial optimization, and multivariate analysis, such as graph theory and log-linear models. The last section of this paper considers some open research problems, including potential connections to methods proposed in the field of computer science.

2 Partial data releases for contingency tables

Many data summaries, e.g., marginal tables, conditional tables, odds-ratios, and relative-risk measures, can be compiled from contingency tables and subsequently released to the public. Assuming that data are reported without error, that marginals and conditionals are compatible, and that unweighted counts are used, the researchers thus far have focused on determining safe releases in terms of arbitrary sets of marginal and/or conditional tables coupled with sample size. The most recent approaches in this area are informed by a confluence of statistical methodologies, operations research methodologies, and tools from computational algebra (i.e., algebraic statistics). Four related problems are relevant to evaluating disclosure risk and data utility: (1) characterizing conditions under which released fragmentary data will uniquely identify the original table, and thus lead to full disclosure of the original table; (2) computing sharp bounds on cell counts with small entries; (3) counting the number of feasible tables consistent with released data; and (4) sampling, i.e., estimating the probability distributions of multi-way tables.

Notation and Definitions Let $X = (X_1, X_2, ..., X_k)$ be a discrete random vector with the probability function

$$p(x) = P(X = x) = P(X_1 = x_1, ..., X_k = x_k)$$

where $x = (x_1, ..., x_k)$. Each X_i is defined on a finite set of integers $[d_i] = \{1, 2, ..., d_i\}, d_i \geq 1, i = 1, ..., k$, with $\mathcal{D} = [d_1] \times ... \times [d_k]$. A k-way contigency table of counts, $\mathbf{n} = \mathbf{n}(i), i \in \mathcal{D}$, is a k-way dimensional *array* of non-negative integers, such that each cell entry $\mathbf{n}(i) = \#\{X = i\}$ represents the number of times the configuration i is observed in a series of independent realizations of $X_1, ..., X_k$ (see [3], [23]). The data of interest are counts in a k-way contingency table, $d_1 \times d_2 \times \cdots \times d_k$. Defined in this way, a table of counts is a point in a simplex of dimension equal to $\mathcal{D} - 1$, i.e., the number of cells-1. The values of X_i are lattice points in a convex polytope. The parameter sets lie in a related simplex. This sets up a link between the contingency tables and algebraic geometry. For more details on links between algebraic statistics, contingency table analysis, and disclosure limitation, see [7].

Consider a subset a of $K = \{1, ..., k\}$ and denoted by $\mathbf{n}_a$ and $\mathbf{p}_a$ the vectors of *marginal* counts (i.e., tables) 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 $\mathbf{n}_{a|b} = \mathbf{n}_{ab}/\mathbf{n}_b$. We define the array of *conditional probabilities* of a given b by $\mathbf{p}_{a|b} = \mathbf{p}_{ab}/\mathbf{p}_b$, where $\mathbf{p}_b > 0$. If $a \cup b = K$, $\mathbf{n}_{a|b}$ is referred to as a *full* conditional, otherwise it is referred to as a *small* or *partial* conditional.

Suppose we observe an arbitrary set of conditional and marginal tables, $\mathcal{T}$. Let A be an $r \times d$ matrix of integers that captures constraints imposed by $\mathcal{T}$ on a k-way table $\mathbf{n}$. In

applications with contingency tables, d is the number of cells in the table, and r is the number of parameters or constraints. For a given log-linear model, A is called the *design matrix*, or it can be referred to as a *constraint matrix* used in optimization problems. For a linear constraint $\mathbf{t} \in \mathbb{Z}_+^r$, let

$$\mathcal{F}_{\mathbf{t}} = A^{-1}[\mathbf{t}] := \{\mathbf{n} \in \mathbb{Z}_+^d : A\mathbf{n} = \mathbf{t}\} \tag{1}$$

be the set of all non-negative integer tables, called a *fiber* or a *reference set*, that satisfy the given linear constraints. For examples of A and t in this context, see [7], [33] and [25]. Properties of the fiber are fundamental for assessing the risk of disclosure (e.g., via optimization, enumeration, and/or sampling) and for making conditional or exact inferences (e.g., see [6], [15], [10],[9], [4], [24]).

Fibers are related to *Markov bases*. Consider a sublattice $\mathcal{L}_{\mathbf{t}}$ of $\mathbb{Z}^{\mathcal{D}}$ that depends on a collection $\mathcal{T}$, and a finite subset $\mathcal{B}_{\mathbf{t}}$ (e.g., a Markov basis is the smallest such subset) of $\mathcal{L}_{\mathbf{t}}$. Each element of $\mathcal{B}_{\mathbf{t}}$, $\mathbf{z}$ can be thought of as a contingency table with values in $\mathbb{Z}^{\mathcal{D}}$, and each is called a *move* that satisfies $A(\mathbf{n} + \mathbf{z}) = A\mathbf{n}$. The most important property of Markov bases, for our purposes, is that they *connect* all tables in the fiber; thus, they can be used for data swaps and for building a connected Markov chain. Helpful references for algebraic statistics, including the calculation and use of Markov and Gröbner bases, are [6], [29].

2.1 Optimization and Counting for Contingency Tables

Optimization, counting, and exhaustive enumeration in this setting 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.

Using the previously defined notation, a linear program (LP) consists of a linear objective function optimized subject to linear constraints:

$$\begin{aligned} &\textit{Minimize} && \mathbf{cn} \\ &\textit{subject to} && \mathbf{An} = \mathbf{t} \\ &&& \mathbf{Gn} = \mathbf{h} \\ &&& \mathbf{n} \geq \mathbf{0} \end{aligned} \tag{2}$$

where $\mathbf{c}$ is a row vector of length d, $\mathbf{n}$ is a column vector of length d representing the k-way array $\mathbf{n}$, and $\mathbf{G}$ and $\mathbf{h}$ encode additional assumptions needed for program feasibility. An integer program (IP) can be formulated as (2) with the additional constraints that all decision variables must be integers. We need to perform $2 \times d$ optimizations in order to calculate both the lower and upper bounds of each cell in the table. The narrower the intervals, especially for cells with low counts, the higher the risk of disclosure.

Releasing Marginal tables Calculating cell bounds for entries in contingency tables given marginal totals has a long history (e.g., see [17] and [18] for more details). Given an $I \times J$ table with a total sample size (n_{++}) and marginal totals (n_{i+} and n_{+j}), it is well known that *Fréchet bounds* for the ij^{th} cell are sharp and have the following form:

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

Now consider a situation in which instead of releasing a full k-way table, we release a set of lower-dimensional marginal totals. In principle, the bounds can be obtained by solving the corresponding LP problem, but in general these are NP-hard problems. Integer programs can be even more difficult to solve and expensive computationally. In part, these issues can be circumvented by using generalizations of Fréchet-type bounds that include explicit formulas for instances in which the released margins correspond to the minimal sufficient statistics of the decomposable log-linear models (e.g., [8], [**?**]).

Dobra and Fienberg [9] describe an updated version of the Generalized Shuttle Algorithm (GSA), which computes sharp integer bounds *and* exhaustively enumerates all feasible tables consistent with a set of linear constraints. The considered constraints comprise marginals, bounds, and structural and sampling zeros. The algorithm exploits the hierarchical structure of the categorical data and the multi-way tables in order to simplify the calculations. The authors claim that for very large sparse contingency tables (e.g., 2^{16}), GSA can produce exact sharp bounds or at least bounds that are very close to the actual bounds, and that GSA can often accomplish this in one quick step. This ability is in stark contrast to IP and LP, both of which must solve many separate optimization problems simultaneously (e.g., 2×2^{16}). The authors claim that any linear-type constraint can be evaluated using this algorithm. It remains to be explored if the algorithm would work for other constraints such as those imposed by observed conditional probabilities.

GSA is a multifaceted off-the-shelf method that can be used for assessing both the disclosure risk and utility associated with contingency table releases. It accounts correctly for the presence of zero counts and can calculate the exact p-values of goodness-of-fit tests in exact conditional inference; thus, it provides a valid assessment of utility. This characteristic is important because the presence of zero counts is tied to the non-existence of maximum likelihood estimates (e.g., see [7]). It has been shown empirically that the presence of counts of zero may increase the disclosure risk of certain cells by tightening their bounds, but a rigorous evaluation of the effect of zero cells on disclosure risk has yet to be undertaken.

Releasing Marginal and Conditional tables In comparison with calculating cell bounds given marginals, very little research has focused on examining bounds induced by a given sample size and observed conditional probabilities or to their combination with marginals. Slavković and Fienberg ([31], [20], [32]) characterize situations in a k-way table when the released collection $\mathcal{T}$ will uniquely identify the joint distribution, and thus produce full disclosure. Multiple realizations of the joint distribution for $\mathbf{n}$ suggest that more than

one table is capable of satisfying the constraints imposed by $\mathcal{T}$. For such cases, the same authors use LP, IP, and algebraic methods to calculate the bounds on cell entries. They also discuss the possible inadequacies of these methods for treating conditional constraints.

Smucker and Slavković's recent results([34], [33]) improve on the original LP/IP formulation given the observed conditional frequencies by allowing the counts in individual cells to be equal to zero. They also describe closed-form solutions for linear relaxation bounds, thus reducing or even eliminating the computing time required for optimization. Given an $I \times J$ table with a total sample size of n and observed conditional probabilities $n_{j|i}$, the bounds for the ij^{th} cell are

$$n_{j|i} \leq n_{ij} \leq (n - (I-1))n_{j|i}. \tag{3}$$

The proposed bounds generalize to k-way tables given either a full or partial conditional $\mathbf{n}_{a|b}$; they hold even if zero cell counts are observed. Similar to the marginal case, the zeros may reveal additional information about their complementary cells. This requires further careful investigation, particularly in regard to k-way tables.

The observed conditional proportions are typically rounded before publication. While the rounding can be seen as adding random noise to these releases, and thus offering "more" protection, it may lead to different bounds for the same table and to computing problems. 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. In principle, the agency can obtain sharp IP bounds by using the observed counts directly and then decide whether to release the partial information. Smucker and Slavković ([34], [33]) discuss these issues and propose two different formulations, one to be used when the data owner performs the calculation and the other when an intruder performs the calculation. More rigorous study is necessary, if we are to understand how these issues scale with high-dimensional tables and how they affect various measures of risk and utility.

In terms of disclosure risk, empirical examples have shown that sharp IP bounds given the sample size and full conditionals uniquely identify the counts in the original table. At present, we do not fully understand the underlying characteristics of a table that would produce such a unique specification and consequently full disclosure. One possibility lies in a trade-off between the sample size and the number of cells; however, the ratio between these two quantities in the investigated examples has not yet suggested a clear relationship.

Slavković, Zhu, and Petrović [30] describe bounds and enumeration algorithms that exploit the hierarchical structure of contingency tables and the links between conditional tables $\mathbf{n}_{a|b}$ and corresponding marginal tables $\mathbf{n}_{ab}$. In some instances, the bounds for small conditionals reduce to the bounds given by corresponding margins. This relationship indicates that GSA could be used with conditional tables, though how exactly to use GSA for that form and the problems that may be encountered by doing so requires investigation. The proposed calculations are also much faster than related algebraic tools (e.g., LattE) for counting the number of possible tables in $\mathcal{F}_{\mathbf{t}}$. Thus, they are likely to be easy to implement in readily available statistical packages, such as R and SAS.

Algebraic tools Neverthless, the algebraic approach does have some disadvantages: (1) it can be computationally infeasible to calculate Markov bases; and (2) for conditionals, Markov bases are extremely sensitive to the rounding of cell probabilities. Both of these disadvantages are the subject of active research; in fact, some recent improvements in computation have been achieved that mitigate their effects (e.g., see [21]). Slavković and Lee ([25],[24]) describe the calculation of Markov bases given fixed conditionals for two-way tables; they also propose a sampling algorithm and a way of creating synthetic tables (see Sec 2.2).

Bounds with gaps Two types of gaps are associated with bounds that can significantly affect both risk and utility measures. Solutions to IPs can be difficult to obtain; they can also be computationally expensive. These drawbacks are likely to mean that linear relaxation bounds are preferred as an approximation to the sharp IP bounds. Given the marginals, the first type of gap – the maximal gap between an IP solution and its linear relaxation – has been shown, in theory, to be exponentially large ([35], [22]). These results suggest that it could be misleading to assess disclosure risk by using linear relaxation bounds as an approximation to sharp integer bounds. Smucker and Slavković([34], [33]) show empirically that the same holds in the case of given conditional probabilities. Therefore, LP bounds may often not be a reasonable choice for detecting whether there is a "true" disclosure, except perhaps as a crude approximation in the event of time-prohibitive sharp IP calculations; note that a more precise mathematical definition of "reasonable" approximation is needed.

The second type of gap concerns the fiber $\mathcal{F}_{\mathbf{t}}$; i.e., for some cell entries the range of integer values is not a sequence of consecutive integers. In the presence of such gaps, even the sharp bounds for the cell entries do not constitute a definitive indication of the safety of a given data release. Onn [27] showed that there could be arbitrary gaps in the bounds on cell entries given the margins, which could further increase the risk of disclosure. Using the algebraic tools, the researchers have empirically shown similar results given conditional tables. For examples of such gaps and for further discussion of the implications of disclosure and utility, see [7] and references therein.

2.2 Sampling and Synthetic contingency tables

Synthetic data methods are defined by the use of Bayesian methodology and the release of multiple-imputed datasets instead of the original datasets. The probability of uniquely identifying an individual from such data is low, and Raghunathan et al. [13] describe methods for drawing valid inferences from such datasets. The U.S. Census Bureau is implementing and evaluating these and related methods; the agency is, in fact, building synthetic datasets [1].

Related synthetic methods for contingency tables utilize Markov bases and MCMC sampling algorithms. These methods replace the original tables by a random draw from the exact

distribution of the tables under a log-linear model whose sufficient statistics are released marginal totals (e.g, see [10] and [5]). The released table can be considered an instance of a fully synthetic or a partially synthetic dataset produced with a special type of data swapping. A synthetic table that preserves marginals is often an appropriate replacement, as marginals are minimal sufficient statistics for log-linear models. However, Lee and Slavković [25] show that such tables sometimes fail to produce reliable statistical inferences, as in the case of Mantel-Haenszel tests and some logistic regression models. This failure to yield reliable statistical inferences suggests that other statistics should be considered with the goal of expanding the range of data utility.

Lee and Slavković ([25], [24], [28]) also address the problem of generating a complete synthetic contingency table $\mathbf{n}$ that is consistent with the observed conditional probabilities from the reference set $\mathcal{F}_{\mathbf{t}}$. The algorithm has an algebraic and Bayesian flavor similar in spirit to the data augmentation algorithms described by [10]. Current results indicate that a viable alternative data release to marginal totals $\mathbf{n}_{ab}$ could be the corresponding small conditionals $\mathbf{n}_{a|b}$; they may have a smaller disclosure risk but maintain the data utility needed for valid statistical inference. The proposed algorithms work well for smaller tables, but it is foreseeable that scaling to large sparse tables may be problematic because the algorithm depends on calculating the Markov bases. A complementary research problem explores the structure and form of Markov bases given both the conditional and marginal tables (e.g., [30]). To estimate posterior distributions, unlike with margins, requires a choice of prior distribution. In addition, it is necessary to carefully analyze the sensitivity of the proposed technique to the choice of prior.

3 Conclusions and Research Directions

To date, statistical disclosure limitation methodologies for tables of counts have focused heavily on the release of unaltered marginal totals from such tables, and to some extent on the inferences that are possible from such releases. Many statistical agencies also release other forms of summary data from tables, such as tables of rates, observed conditional frequencies, and odds-ratios; however, less is understood about the impact of such information on disclosure risk and utility. These statistics are predominantly released as two-way and three-way tables with conditioning on a single variable.

This paper highlights some new results in regard to partial information releases particularly in relation to sample size and marginal and conditional tables tied to small and large k-way tables. GSA seems an efficient and multifaceted tool for assessing both the disclosure risk and utility inhering in the release of marginal tables. Whether GSA is flexible enough to account for the release of conditional proportions as well remains to be determined. The research on releasing conditionals has predominantly focused on assessing the bounds and distributions for smaller tables, and its results in this regard indicate a close connection between marginals and conditionals and point to a number of issues that require further

investigation:

- Understanding the underlying characteristics of a table capable of producing a unique specification of the original table **n** even when the partial information seems insufficient to uniquely identify the joint distribution.
- Understanding and characterizing the integer gap problem and gaps in the fiber, and how these affect disclosure risk and utility. Note that for large sparse contingency tables, calculation of sharp bounds is prohibitively time-consuming for some practical uses; thus the closed-form linear relaxation bounds could be a useful easy estimate for an agency considering releasing such summaries.
- Understanding the effects of zero counts, especially when releasing tables of rates. Zero cells in a table become zeros again when we condition on one or more of the variables with zero in the margins. Thus, the release of such conditionals reveals extra information about certain complementary cells and about the full cross-classification.
- Modeling the joint distribution of multivariate categorical data in the presence of partial information other than the marginal tables. Log-linear models are a very useful tool in the analysis of contingency tables, but the needed marginal tables may not be available if it is not considered safe to release them.
- Understanding the effects on both the disclosure risk and utility of releasing other types of partial information. Any useful statistic reveals some information about its related database. For multivariate categorical data, these statistics may be odds and odds-ratios, relative risk, specificity and sensitivity measures, for example.

For more open problems directly related to algebraic statistics, contingency tables, and disclosure limitation, see [7].

An additional ongoing problem for agencies is the presence of powerful data mining algorithms. There are close links between SDL techniques and privacy-preserving data mining (PPDM) techniques, but in general little is understood about how they interact in practice or how they affect risk and utility. Fienberg and Slavković [16], for example, discuss links between the release of marginal and conditional tables and the mining of association rules [36]. Furthermore, results in the recent SDL literature have focused on static database releases. Some argue that "selective revelation techniques that attempt to separate transactional data (e.g., ticket purchase, money transfers, etc.) from identity, and that reveal information incrementally may significantly reduce privacy the likelihood that privacy will be breached [14]. Since the bounds are calculated by accounting for the collective based on partially released information, the potential exists to dynamically update safe releases. In general, SDL techniques and practice in regard to official statistics have focused little on *automation* and *scaling*, both of which are the primary strengths of PPDM techniques.

Simulation/sampling methods are important since we can generate distributions of all possible tables given the margins and other statistics, and hence establish a probabilistic frame-

work for disclosure assessment and statistical inference. SDL literature has shown that risk of identification in such datasets is minimal. The utility is tied to the family of posterior predictive distributions from which data have been drawn. However, it is still unclear how these methods affect record linkage and privacy preserving data mining methods; i.e., how will information from various databases be combined?

Agencies are also faced with new challenges in negotiating modern databases, both in terms of the types of information these collect and their often large size. In addition to traditional types of information contained in census and medical studies, contemporary information repositories store social network data (e.g., Facebook data), product preferences (e.g., Amazon), web search data and other information that was not previously archived in digital format. Sensitive information can be leaked when an intruder has only modest partial knowledge about the data from external sources (e.g., [37]). In light of this, agencies need to consider more rigorous definitions of privacy. For example, how does the concept of differential privacy [13] fit with current statistical techniques and official statistics practices? Can these different yet complementary approaches be integrated to produce statistically sound techniques, yield broadly useful data, and yet preserve privacy in the face of realistic external information?

Acknowledgments

The research reported here was supported in part by NSF Grant SES-0532407 to the Department of Statistics, Pennsylvania State University.

References

[1] Abowd, J., Stinson, M., and Benedetto, G. (2006). Final Report to the Social Security Administration on the SIPP/SSA/IRS Public Use File Project. Technical report, US Census Bureau Longitudinal Employer-Household Dynamics Program.

[2] Agresti, A. (2002). *Categorical Data Analysis*. John Wiley & Sons, New York, 2nd edition edition.

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

[4] Caffo, B. and Booth, J. (2003). Monte Carlo Conditional Inference for Log-Linear and Logistic Models: A Survey of Current Methodology. *Statistical Methods in Medical Research*, 12(2):109–123.

[5] Chen, Y., Dinwoodie, I., and Sullivant, S. (2006). Sequential Importance Sampling for Multiway Tables. *Annals of Statistics*, 34(1):523–545.

[6] Diaconis, P. and Sturmfels, B. (1998). Algebraic algorithms for sampling from conditional distributions. *Annals of Statistics*, pages 363–397.

[7] Dobra, A., Fienberg, S., Rinaldo, A., Slavković, A., and Zhou, Y. (2008). Algebraic statistics and contingency table problems: Log-linear models, likelihood estimation and disclosure limitation,. In Putinar, M. and Sullivant, S., editors, *IMA Volumes in Mathematics and its Applications: Emerging Applications of Algebraic Geometry*, volume 149, pages 63–88. Springer Science+Business Media, Inc.

[8] Dobra, A. and Fienberg, S. E. (2001). Bounds for cell entries in contingency tables given marginal totals and decomposable graphs. *Statistical Journal of the United Nations Economic Commission for Europe*, 18(4):363–371.

[9] Dobra, A. and Fienberg, S. E. (2009). The generalized shuttle algorithm. In P. Gibilisco, E. Riccomagno, M. R. and Wynn, H., editors, *Algebraic and geometric methods in statistics*. Cambridge University Press, to appear.

[10] Dobra, A., Tebaldi, C., and West, M. (2006). Data augmentation in multi-way contingency tables with fixed marginal totals. *Journal of Statistical Planning and Inference*, 136(2):355–372.

[11] Domingo-Ferrer, J. and Franconi, L., editors (2006). *Privacy in Statistical Databases, CENEX-SDC Project International Conference, PSD 2006, Rome, Italy, December 13-15, 2006, Proceedings*, volume 4302 of *Lecture Notes in Computer Science*. Springer.

[12] Domingo-Ferrer, J. and Saygin, Y., editors (2008). *Privacy in Statistical Databases, UNESCO Chair in Data Privacy International Conference, PSD 2008, Istanbul, Turkey, September 24-26, 2008. Proceedings*, volume 5262 of *Lecture Notes in Computer Science*. Springer.

[13] Dwork, C. (2008). Differential privacy: A survey of results. *Lecture Notes in Computer Science*, 4978:1.

[14] Fienberg, S. (2006). Privacy and Confidentiality in an e-Commerce World: Data Mining, Data Warehousing, Matching and Disclosure Limitation. *Statistical Science*, 21(2):143–145.

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

[16] Fienberg, S. and Slavkovic, A. (2008). A survey of statistical approaches to preserving confidentiality of contingency table entries. In Aggarwal, C. C. and Yu, P. S., editors, *Privacy-Preserving Data Mining: Models and Algorithms*, pages 291–309. Springer.

[17] 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*, pages 115–129, Luxembourg. Eurostat.

[18] Fienberg, S. E. (2000). Contingency tables and log-linear models: Basic results and new developments. *Journal of the American Statistical Association*, 95(450):643–647.

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

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

[21] Hemmecke, R. and Hemmecke, R. (2003). 4ti2 version 1.1—computation of Hilbert bases, Graver bases, toric Gröbner bases, and more. Available at www.4ti2.de.

[22] Hoşten, S. and Sturmfels, B. (2007). Computing the integer programming gap. *Combinatorica*, 27(3):367–382.

[23] Lauritzen, S. (1996). *Graphical Models*. Oxford Science Publications.

[24] Lee, J. and Slavkovic, A. (2008). Sampling contingency tables preserving the observed conditional frequencies. Poster at ISBA 2008.

[25] Lee, J. and Slavković, A. (2008). Synthetic tabular data preserving observed conditional probabilities. In Domingo-Ferrer, J. and Saygin, Y., editors, *CD Proceedings, PSD2008*.

[26] Methodology, F. C. O. S. (2005). Statistical policy working paper 22 (revised 2005)—report on statistical disclosure limitation methodology. Technical report, Statistical and Science Policy, Office of Information and Regulatory Affairs, Office of Management and Budget, Washington, DC. http://www.fcsm.gov/working-papers/spwp22.html.

[27] Onn, S. (2006). Entry uniqueness in margined tables. In Domingo-Ferrer, J. and Franconi, L., editors, *Privacy in Statistical Databases – PSD 2006, Lecture Notes in Computer Science No. 4302*, pages 94–101. Springer-Verlag.

[28] Slavkovic, A. and Lee, J. (2009). Synthetic two-way contingency table preserving conditional frequencies. *Statistical Methodology, submitted*.

[29] Slavkovic, A. and Sullivant, S. (2006). The space of compatible full conditionals is a unimodular toric variety. *Journal of Symbolic Computation*, 41(2):196–209.

[30] Slavkovic, A., Zhu, X., and Petrovic, S. (2009). Mathematical aspects of the space of contingency tables given partial information. *Journal of Electronic Statistics, to be submitted*.

[31] Slavković, A. B. and Fienberg, S. E. (2004). Bounds for cell entries in two-way tables given conditional relative frequencies. In Domingo-Ferrer, J. and Torra, V., editors, *Privacy in Statistical Databases – PSD 2004, Lecture Notes in Computer Science No. 3050*, pages 30–43. Springer-Verlag.

[32] Slavkovic, A. B. and Fienberg, S. E. (2009). Algebraic geometry of 2×2 contingency tables. In P. Gibilisco, E. Riccomagno, M. R. and Wynn, H., editors, *Algebraic and geometric methods in statistics*. Cambridge University Press, to appear.

[33] Smucker, B. and Slavković, A. (2009). Calculating cell bounds in contingency tables based on conditional frequencies,. *Journal of Official Statistics*, to be submitted.

[34] Smucker, B. and Slavkovic, A. B. (2008). Cell bounds in two-way contingency tables based on conditional frequencies. In *Privacy in Statistical Databases*, pages 64–76.

[35] Sullivant, S. (2005). Small contingency tables with small gaps. *Siam J. Discrete Math*, 18(4):787–793.

[36] Verykios, V. and Gkoulalas-Divanis, A. (2008). A Survey of Association Rule Hiding Methods for Privacy. In Aggarwal, C. C. and Yu, P. S., editors, *Privacy-Preserving Data Mining: Models and Algorithms*, pages 267–286. Springer.

[37] Wong, R., Fu, A., Wang, K., and Pei, J. (2007). Minimality attack in privacy preserving data publishing. In *Proceedings of the 33rd international conference on Very large data bases*, pages 543–554. VLDB Endowment.