Journal of Privacy and Confidentiality (2013) **5**, Number 2, 133–156

An Intuitive Formulation and Solution of the Exact Cell-Bounding Problem for Contingency Tables of Conditional Frequencies

Stephen E. Wright*, and Byran J. Smucker†

1 Introduction

1.1 Overview

The overarching issue in statistical disclosure limitation is the compromise between data utility and data privacy. While data providers, such as government or nongovernment agencies, often give privacy guarantees to those whose data they collect, too much obfuscation dilutes the data's usefulness to the public and researchers. Researchers concerned with statistical disclosure limitation study this tradeoff in various contexts. The current paper addresses questions on the privacy side of the matter.

One common form of data release is the contingency table. Sometimes, there is concern that releasing a small cell count will allow private information to be deduced, and one way of reducing this risk is to release a summary of the table instead of cell counts directly. Traditionally, such releases have been in the form of marginal totals, which are minimal sufficient statistics for parameters in associated log-linear models [see 11, 8, 12]. Releasing such summaries allows for bounds on the underlying cell counts to be deduced, and if these bounds are too narrow for cells with small counts, a significant disclosure risk is incurred. This problem has been well-studied in the literature [e.g., 2, 6, 7].

An alternative to releasing marginals is to release observed conditional probabilities. These quantities, which preserve odds and odds ratios [20], represent proportions of cell counts having specified characteristics. There are also potential applications beyond analyses based on odds ratios. For instance, association rules in data mining seek relationships between variables in possibly sparse databases. This process involves the use of marginal and/or conditional information [13], and raises potential privacy issues that make the computation of bounds important. Also, Chen et al. [4] simulate from a distribution of contingency tables that assume given marginal totals and require cell bounds in the course of their sampling procedure. A similar procedure might be devised for which bounds given conditionals would be necessary or convenient (see, e.g., [24], whose Bayesian procedure includes priors that could be informed by cell bounds.

*Department of Statistics, Miami University of Ohio, Oxford, OH, USA, mailto:wrightse@MiamiOH.edu.

†Department of Statistics, Miami University of Ohio, Oxford, OH, USA, mailto:smuckerb@MiamiOH.edu.

As with marginals, releasing conditional probabilities incurs a possible disclosure risk and this problem has been studied as well. Slavković [23] and Fienberg and Slavković [13] first examined this issue, and more recently Smucker and Slavković [25] and Smucker et al. [26] studied integer programming formulations to determine bounds on cells given conditional probabilities and tabular sample size, as well as easier-to-solve linear programming relaxations. However, for large tables the integer programs may not solve in a reasonable amount of time, and the linear programming relaxations give bounds that are much wider than the sharp integer bounds. Indeed, in a review paper Slavković [20] suggests that "we do not fully understand the underlying characteristics of a table that would produce ... a unique specification [of the underlying cell counts] and consequently full disclosure."

In this paper, we address the problem of exactly identifying the sharpest bounds on the underlying cell counts based only on knowledge of the (unrounded) conditionals and the sample size. We reformulate the basic optimization problem as a linear knapsack problem, e.g., [18], that can be solved very quickly. The theoretical results that follow from this formulation allow integer bounds to be deduced in closed form for various easy-to-check cases. We consider tables of a variety of sizes, rearranging multi-way tables to two dimensions with one dimension denoting the response categories and the other the predictors. Our results also cover so-called "partial" conditionals that are created by aggregating out (i.e., marginalizing) some variables beforehand, in which case our bounds cover cells in the partial table of aggregated counts and thereby project onto the corresponding sums of cells in the full table. We also indicate briefly how our approach can be directly adapted to utilize information in the form of bounds on sums of cell counts within a given row that might be drawn from tables in other forms.

We note that the generalized shuttle algorithm of Dobra and Fienberg [7] appears at first glance to be another possible approach to finding the bounds. However, it cannot handle weighted linear constraints, which are an essential part of any mathematical formulation of the problem considered herein. On the other hand, it tackles the more challenging problem of bounds given multiway marginal information whereas we limit ourselves herein to two-way tables given conditionals and total sample size.

We set our work in the context of multiway contingency tables and the information that can be extracted about them from given a set of conditionals. Our present goal is relatively modest: we focus on the case in which a single set of conditionals is released. This is a problem that has not yet received a definitive treatment.

Within this context, the paper accomplishes several things. First, we list several mathematical consequences of the knapsack formulation that illuminate the problem structure, thereby providing simple and intuitive criteria for disclosure or non-disclosure of individual cells (and sometimes all cells) in the full or partial table. The earlier papers cited above have noted a lack of such fundamental insight. In particular, we address several conjectures/questions raised by Smucker et al. [26] concerning patterns they observed in the examples they studied. We illustrate our basic principles with numerous examples, worked in their entirety.

Second, the developments in this paper facilitate new understanding about the re-

lationship between released conditional probabilities and disclosure. Though this work focuses primarily on disclosure in the rather narrow sense of possible disclosures of individual characteristics associated with small counts in contingency tables, the insights gained from the mathematical and computational results allow us to comment more generally on the usefulness, or lack thereof, of releasing tables in the form of conditionals. See the discussion in Section 4.

Third, we demonstrate that even large cell-bounding problems with thousands of nonzero cells can be solved in seconds rather than hours. This has implications for disclosure risk because the apparent difficulty of exactly solving such large problems has been cited as evidence that tables of conditionals might not pose as much risk as their tight integer bounds would suggest [26]. Furthermore, table redesign, through aggregation over or within variables, is one of the key tools available to the data owner for protecting privacy [15]. The methodology given here can be used to facilitate such rapid table audits.

One issue that we do not address directly in the current paper is the fact that observed conditionals are published as rounded decimals rather than exact fractions. As noted by the referees, this can have serious effects on data utility with regard to inference involving small cell counts and (say) the ability to distinguish them from structural or observed zero counts. In the current paper, we again follow Smucker et al. [26] and note that the assumption of unrounded conditionals may be viewed as a worst-case scenario that aids the data owner in determining disclosure risk when designing a table for release. Subsequent rounding then provides additional protection against disclosure. Although some earlier works such as Smucker et al. [26] have made preliminary efforts in this direction by obtaining relaxed bounds, no method has been presented for calculating the sharpest possible bounds from rounded conditionals. A forthcoming paper focuses precisely on such calculations, extending the results presented here.

The paper is organized as follows. In §1.2 we describe two small examples that are used later in the paper as easily understood illustrations. Section 2 lays out the notation and assumptions, followed by a derivation of the general knapsack formulation. In §3 we present our main results, which deal with simplifying the knapsack problems and identifying cases with closed-form solutions. Those results are illustrated with real-world examples that appear fairly complex, but can be addressed quickly by hand with the ideas introduced in that section. We also include a brief discussion of the scalability of our results and methods. Some additional remarks are given in §4, including further comparisons with similar work in the literature and an attempt to put insights derived from our work into a broader context of disclosure and inference.

1.2 Motivation: Two Easy Examples

To illustrate basic ideas and results as they're introduced, we consider two 4×4 contingency tables. Table 1(a), a fictitious example taken from a report by the Federal Committee on Statistical Methodology [19], shows the number of delinquent children in a cross-tabulation of county and education level of the head of household. The corre-

sponding conditional probabilities are shown as fractions in Table 1(b). In this and all subsequent examples, we assume that such conditionals are to be released along with the total sample size. We wish to draw attention to several features of the conditionals in this example.

Table 1: 4×4 table with $N = 135$ shown in two forms: (a) observed counts for the number of delinquent children; (b) exact values of observed conditional probabilities, given education level.

(a) Observed counts

	Low	Medium	High	Very High	total
Alpha	15	1	3	1	20
Beta	20	10	10	15	55
Gamma	3	10	10	2	25
Delta	12	14	7	2	35

(b) Observed conditionals as fractions (each row sums to 1)

	Low	Medium	High	Very High
Alpha	$15/20 = 3/4$	$1/20$	$3/20$	$1/20$
Beta	$20/55 = 4/11$	$10/55 = 2/11$	$10/55 = 2/11$	$15/55 = 3/11$
Gamma	$3/25$	$10/25 = 2/5$	$10/25 = 2/5$	$2/25$
Delta	$12/35$	$14/35 = 2/5$	$7/35 = 1/5$	$2/35$

First, note that the *unreduced* fractions expose the actual cell counts, which are precisely the corresponding numerators. So the goal of a data snooper is essentially to identify the correct unreduced fractions in cells of interest. Second, for the *reduced* fractions (those in lowest terms), the denominators within each row are no longer all the same. More importantly, no reduced denominator in row 2 (Beta) corresponds to the actual total count in that row because the denominator 55 has been reduced to 11 in *all* columns. Because the lowest common denominators of the reduced fractions in each row (namely, 20, 11, 25, and 35) sum to 91 rather than $N = 135$, it takes the data snooper a bit of additional reasoning to decide which row(s) should be scaled up to expose their cell counts as numerators.

Here is an example of such reasoning. The goal is to determine how one might express $N = 135$ as a sum of positive multiples of the denominators 20, 11, 25, and 35. Taking just one copy of each row's common denominator leads to a base total of 91, which leaves $135 - 91 = 44$ to be accounted for. This remaining 44 must likewise be expressed as a sum of nonnegative integer multiples of 20, 11, 25, and 35. Clearly, $44 = 4(11)$ works and corresponds to unreduced denominators (by row) of 20, $11 + 44 = 55$, 25 and 35. The data snooper, unsure of the actual cell counts, now must decide whether this is the only possibility. It can quickly be seen that it is. For instance, adding 35 to 20, 11, or 25 exceeds the target of 44; this rules out using another 35 and incidentally proves

that row 4 is disclosed by its numerators. Similarly, one can work through the small number of cases involving only sums of multiples of 20, 11, and 25 to rule out any other possibilities. Thus, the counts in Table 1(a) are completely disclosed, which could be troubling because several of the counts are small. The purpose of this paper is to show that an equivalent analysis can be performed systematically, automatically, and quickly for tables of any size.

More generally, releasing fractional values for conditional probabilities potentially obscures some or all of the actual cell counts. To see this, consider the nearly identical table shown in Table 2, which was obtained from Table 1 by changing a single entry and the corresponding row sum. Reasoning again as in the preceding paragraph, we see that accounting for one copy of each row's common denominator leads to a base total of 90, leaving $130 - 90 = 40$. As before, we can rule out the use of more than one 25 or 35, but this time we can express the remaining target of 40 in several ways: $40 = 2(20) = 20 + 2(10) = 4(10)$. We conclude that the numerators in rows 3 and 4 disclose their corresponding cell counts, whereas the cell counts in rows 1 and 2 remain uncertain. For instance, the common denominator in row 1 might be 20, 40, or 60 corresponding to a cell count of 1, 2, or 3 in column 2.

Table 2: 4×4 table with $N = 130$, modified slightly from Table 1: only the entry in row 2 and column 2 has been changed, from 10 to 5.

(a) Observed counts

	Low	Medium	High	Very High	total
Alpha	15	1	3	1	20
Beta	20	5	10	15	50
Gamma	3	10	10	2	25
Delta	12	14	7	2	35

(b) Observed conditionals as fractions (each row sums to 1)

	Low	Medium	High	Very High
Alpha	$15/20 = 3/4$	$1/20$	$3/20$	$1/20$
Beta	$20/50 = 2/5$	$5/50 = 1/10$	$10/50 = 1/5$	$15/50 = 3/10$
Gamma	$3/25$	$10/25 = 2/5$	$10/25 = 2/5$	$2/25$
Delta	$12/35$	$14/35 = 2/5$	$7/35 = 1/5$	$2/35$

The central issue explored in the present work is efficient determination of the tightest possible bounds on each cell count, given only the reduced fractions and the total count N. Furthermore, the methods are easily adapted to account for *a priori* bounds on some cells or sums of cells, as might be gleaned from a related table (of marginals, say) or insider information.

The formulation introduced in §2 can also allow one to obtain the complete list of

possible values for a cell, as was done in the preceding example for Table 2. Corollary 3.9 captures the reasoning of that example for more general tables. This is relevant to disclosure: someone who knows that the cell count does not exceed 28 in row 1, column 1 of Table 2 can deduce that the cell count is actually 15, which thereby exposes the singletons in columns 2 and 4 of that row.

A final noteworthy aspect of disclosure limitation concerns rounding, as needed when published tables generally use decimal-place values instead of reduced fractions. One rounding of Table 1(b) is shown in Table 3. Note that the value $2/11 \approx 0.1818181818$ in row 2 is rounded up in one column and down in the next to preserve a row sum of unity. This is an example of *consistent* rounding (e.g., [5]) which has little effect on statistical analysis of the table but potentially aids in further obscuring the true cell counts. Publicly released conditionals would certainly be presented in rounded form, consistent or otherwise. Note that this potentially reduces data utility, particularly if it blurs distinctions between small counts and structural or observed zeros. A key aspect in any attempt to recover the cell counts from such reported decimal approximations is to identify the nearby fractions being approximated. For example, a quick computer calculation reveals that the only viable fractions within 0.001 of the values shown in Table 3 are the unreduced fractions corresponding to Table 1(b); hence, the 3-digit rounding likewise discloses all cell counts for this example. The methods described herein provide bounds on the cell counts for whatever fractions are considered, assuming only that they sum to unity within rows. Extensions to our methods can be made to account directly for uncertainty due to rounding in the released fractions, but there are additional subtleties that put the matter beyond the scope of the current work. Instead, we assume that the released conditional probabilities are exact and unrounded fractions, following earlier research on sharp bounds [25, 26]. This assumption is not particularly restrictive, and should be viewed as a sort of worst-case analysis for the data releasers, insofar as it considers the setting in which a data snooper has managed to obtain perfect information regarding the fractions.

Table 3: Consistently rounded decimal approximations of the observed conditional probabilities in Table 1(b).

	Low	Medium	High	Very High	total
Alpha	0.750	0.050	0.150	0.050	1.000
Beta	0.364	0.182	0.181	0.273	1.000
Gamma	0.120	0.400	0.400	0.080	1.000
Delta	0.343	0.400	0.200	0.057	1.000

2 A Minimal Formulation of the Cell-Bounding Problem

2.1 Notation and Assumptions

We consider a two-way contingency table with I rows and J columns, noting that a k-way table can be represented as a two-way table by specifying a partition of the vari-

ables into those whose levels comprise the rows and whose levels comprise the columns. Indeed, as illustrated by the examples in §3, every table of (full) conditionals can be viewed as a two-way table for the present purpose. The observed count in cell ij is denoted o_{ij}, with row counts $o_i = \sum_j o_{ij}$ and sample size $N = \sum_i o_i$. This paper examines the mathematical problem of recovering the cell counts o_{ij}, assuming that only the sample size and observed conditional probabilities $\hat{p}_{ij} = \hat{p}(j|i) = o_{ij}/o_i$ are known. More precisely, the goal is to identify the tightest bounds for integers n_{ij} satisfying $\hat{p}_{ij} = n_{ij}/\sum_j n_{ij}$ and $\sum_{i,j} n_{ij} = N$, where we assume that the fractions $\hat{p}_{ij}$ and sample size N are given. Note that we use o_{ij} to denote the actual observed count and n_{ij} to denote a guess for o_{ij}.

We treat each $\hat{p}_{ij}$ as an exact fraction in lowest terms. From the observed counts we calculate the following positive integers:

- each row's greatest common divisor $d_i = \gcd(o_{i1}, \ldots, o_{iJ})$;
- each cell's *reduced count* $r_{ij} = o_{ij}/d_i$;
- each row's *reduced sum* $r_i = o_i/d_i = \sum_j r_{ij}$;
- the table's *reduced total* $R = \sum_i r_i = \sum_{i,j} r_{ij}$.

It is evident that $\hat{p}_{ij} = r_{ij}/r_i$ and that r_i is the lowest common denominator of $\hat{p}_{i1}, \ldots, \hat{p}_{iJ}$. Consequently, the reduced counts r_{ij} are readily obtained from the fractions $\hat{p}_{ij}$ alone. We henceforth work with the reduced counts and their sums instead of the fractions $\hat{p}_{ij}$ or the observed counts o_{ij}.

To illustrate the above notation we refer again to Table 2(a), repeated here with reduced counts r_{ij} included in parentheses. Note that $d_2 = 5$, indicating that all reduced counts in row 2 differ from the corresponding observed counts by a factor of 5.

Table 4: Example of Table 2 ($N = 130$, $R = 90$), with each cell showing its observed count o_{ij} followed by its parenthesized reduced count $r_{ij} = o_{ij}/d_i$.

	Low		Medium		High		Very High		o_i	d_i	r_i
Alpha	15	(15)	1	(1)	3	(3)	1	(1)	20	1	20
Beta	20	(4)	5	(1)	10	(2)	15	(3)	50	5	10
Gamma	3	(3)	10	(10)	10	(10)	2	(2)	25	1	25
Delta	12	(12)	14	(14)	7	(7)	2	(2)	35	1	35
								total:	130		90

As noted above, our goal is the rapid identification of the tightest possible upper and lower bounds on the putative counts n_{ij}. Toward this end, two remarks are worthwhile regarding the definitions above. First, $d_i \geq 1$ implies $o_{ij} \geq r_{ij}$, and so our solutions can be constrained to satisfy $n_{ij} \geq r_{ij}$. In other words, because we know the reduced counts by way of the released fractions, they constitute a lower bound on each cell. Second, when $d_i = 1$ the actual counts o_{ij} in row i agree with the reduced counts r_{ij}, and

therefore the tightest lower bound on each cell ij in that row is precisely r_{ij}. In other words, an irreducible row is potentially disclosed in its entirety by the lower bounds calculated for it. The sequel provides simple criteria for determining disclosure risk in some other important cases.

One final bit of notation is needed. In Section 1.2, the calculations focused on choosing multiples of the reduced sums r_i needed to bridge the gap $N - R$ between the reduced total R and the sample size N. We denote that multiple by ν_i for row i and write its minimum and maximum attainable values as ν_i^- and ν_i^+. As we show in the next section, the interval of tightest possible bounds on the cell count n_{ij} is $[r_{ij}(\nu_i^- + 1),\, r_{ij}(\nu_i^+ + 1)]$.

2.2 Integer Programming and Knapsack Formulations

We describe two optimization formulations of the cell-bounding problem. The first one directly captures the problem as stated in the first paragraph of §2.1, and the second greatly streamlines the first to enhance efficiency and provide important insights.

We begin with a pair of integer linear programming problems for bounding the recovered cell count for a specified cell $\overline{ij}$:

$$\min / \max \quad n_{\overline{ij}} \quad \text{over all } n_{ij} \text{ subject to} \tag{1}$$

$$r_i n_{ij} = r_{ij} \sum_{j'} n_{ij'},\ \forall i, j, \tag{2}$$

$$\sum_i \sum_j n_{ij} = N, \tag{3}$$

$$n_{ij} \geq r_{ij},\ \forall i, j, \tag{4}$$

$$n_{ij} \text{ integer}, \forall i, j. \tag{5}$$

The minimization and maximization in (1) correspond, respectively, to finding the lower and upper bounds on $n_{\overline{ij}}$. The constraint (2) is equivalent to $n_{ij} / \sum_{j'} n_{ij'} = r_{ij}/r_i$, and therefore reformulates the requirement $\hat{p}_{ij} = n_{ij} / \sum_j n_{ij}$. Constraints (3)–(5) clearly represent the other required properties of n_{ij}.

The integer program (1)–(5) modifies that proposed by Smucker and Slavković [25] and Smucker et al. [26] in three ways. First, the earlier papers used o_i and o_{ij} rather than r_i and r_{ij} in Equation (2), suggesting that the actual observed counts were needed when in fact they aren't. Second, those papers omitted, for each i, one equation from (2) as being redundant, a simplification perhaps better left to the software used to solve the problem. Third, we replace the constraints $n_{ij} \geq 0$ and $\sum_j n_{ij} \geq 1$ used in those papers with the inequalities (4), which are algebraically much tighter. We note that although Smucker et al. [26] didn't use the reduced counts in their formulation, they did remark on the fact that their sharp upper and lower bounds were multiples of those reduced counts in the examples they considered. The theorem below shows that this must always be true.

The formulation (1)–(5) requires both minimization and maximization of each cell value n_{ij} in the data table. In other words, $2IJ$ optimization problems must be solved. In Section 1.2, however, cell bound calculations were obtained for entire rows at a time by considering how to rescale the lowest common denominators (i.e., the reduced row sums r_i) of the released fractions in each row. Specifically, knowledge of the reduced cell counts r_{ij} allows us to optimize only a single rescaling factor for each row to recover the value of the decision variable n_{ij} for the whole row. We now formalize that approach by simplifying the integer programming formulation given above. The key is to equate the variable ν_i introduced in the last paragraph of Section 2.1 with the quantity $(\sum_j n_{ij}/r_i) - 1$. We then consider these two much simpler problems:

$$\min / \max \quad \nu_{\bar{\imath}} \quad \text{over all integers } \nu_i \text{ subject to} \tag{6}$$

$$\sum_i r_i \nu_i = N - R, \tag{7}$$

$$\nu_i \geq 0, \ \forall i. \tag{8}$$

Let ν_i^- and ν_i^+ denote the minimum and maximum integer values of ν_i subject to (7)–(8).

Theorem 2.1. *The interval* $[r_{ij}(\nu_i^- + 1), r_{ij}(\nu_i^+ + 1)]$ *constitutes the tightest possible bounds on the cell count* n_{ij}. *In particular, the table of counts is fully disclosed if and only if the system (7)–(8) has exactly one integer solution* $(\nu_1, \ldots, \nu_I)$.

Note that this theorem appears in a somewhat different form in the paper by Slavković et al. [22]. It shows that we can replace the $2J$ problems (1)–(5) for cells $\overline{\imath j}$ in row $\bar{\imath}$ with an equivalent pair of simpler problems for the entire row $\bar{\imath}$. Again, we note that the formulation requires knowledge only of the fractional conditional probabilities and total count N. Because the theorem gives the tightest possible bounds for the given information, all examples in this paper yield the same sharp bounds as presented previously in Smucker et al. [26].

We have decreased the number of problems by a factor of J. Moreover, the problems (6)–(8) are much simpler than the original problems (1)–(5), because the number of variables per problem is lower by a factor of J and the number of equations has been decreased from $IJ + 1$ to 1. The optimization problems (6)–(8) are integer *knapsack* problems, which can be solved collectively in a tiny fraction of the time needed for the original formulations (1)–(5). In the next two subsections, we make use of the special structure of (6)–(8) to obtain simplifications that allow many problems to be solved by hand.

We close this section by noting that the knapsack formulation is easily adapted to accommodate *a priori* bounds one might have for a particular cell count or, more generally, a sum of cell counts within a row. To see how, consider such bounds expressed in the form $l \leq \sum_{j \in C} n_{ij} \leq u$, where C is some subset of the column indices. This is equivalent to $l \leq (\nu_i + 1) \sum_{j \in C} r_{ij} \leq u$, which can be solved for ν_i to yield

$$\frac{l}{\sum_{j \in C} r_{ij}} - 1 \leq \nu_i \leq \frac{u}{\sum_{j \in C} r_{ij}} - 1.$$

By the integrality of ν_i, we may round the lefthand side up and the righthand side down to provide integer bounds $l_i \leq \nu_i \leq u_i$, and then use these to replace the looser nonnegativity condition (8) in the knapsack formulation. Software packages for solving knapsack problems can handle such general bounds without additional difficulty.

3 Simplifications and Examples

This section lists mathematical consequences of the knapsack formulation, along with examples showing their uses. As noted earlier, k-way tables of conditionals are essentially two-way tables. The results in this section refer to ordinary *full* conditionals rather than partial conditionals, although the latter can be handled in a straightforward way. The bounds presented here are the tightest possible, and so they agree with the sharp bounds that have appeared elsewhere (e.g., [26]).

3.1 The Remainder and Conditions Guaranteeing Disclosure

We first use the knapsack formulation of the preceding section to formalize and generalize the simple ideas applied earlier to the small tables of Section 1.2. We draw on those examples as illustrations here and then look at more complicated instances. The emphasis for the moment is on directly addressing the remainder $N - R$. We begin by stating an obvious opportunity.

Proposition 3.1. *If $N - R$ is divisible by $r_{\bar{\imath}}$, then $\nu_{\bar{\imath}}^+ = (N - R)/r_{\bar{\imath}}$. In this case, $\nu_i^- = 0$ for all $i \neq \bar{\imath}$.*

In the context of Table 4 in Section 1.2, we see that $r_1 = 20$ and $r_2 = 10$ both evenly divide $N - R = 40$. Consequently, the minimum values are $\nu_i^- = 0$ for each i, with maximum values $\nu_1^+ = 40/20 = 2$ and $\nu_2^+ = 40/10 = 4$ for two of the variables. This leaves just the optimizations needed for ν_3^+ and ν_4^+, to be treated after the next result.

Full disclosure of all cell counts in the table means that each of the bounding problems of the preceding section have a unique feasible point. In general, determining whether such uniqueness holds could require solving many of the bounding problems. However, some circumstances lead to easily tested conditions, a few of which are listed below. These are particularly relevant to tables with small numbers of rows and moderate or large counts in each row. We focus mainly on disclosure of a specified row.

Proposition 3.2. *Suppose the (unrounded) fractions $\hat{p}_{ij}$ for row i are known and let $r_{\min} = \min_{i'}\{r_{i'} \mid r_{i'} > 0\}$. Then any of the following conditions suffices to disclose all the cell counts in row i:*

(a) the actual nonzero count for some cell in row i is known (possibly from a separate source of information);

(b) $r_i > N - R$;

(c) $N - R - r_{\min} < r_i < N - R$;

(d) $r_i < N - R$, *and for each* i' *either (1)* $r_i + r_{i'} > N - R$ *or (2)* $N - R - r_{\min} < r_i + r_{i'} < N - R$.

Note that to verify condition (d) in Proposition 3.2, one must check $i' = i$ as well as all $i' \neq i$. Parts (b)–(d) in Proposition 3.2 are the first three in an evident sequence of conditions forcing ν_i to its lower bound of 0. These three involve checking up to 1, I, or I^2 (respectively) possible combinations of row sums. At some point, of course, it is simply more efficient to solve the upper bounding problem for row i.

We return again to the example data of Table 4, which has $r_{\min} = 10$. Because row 4 gives $r_4 + r_{\min} = 45 > 40 = N - R$, we see by part (c) of Proposition 3.2 that the row is disclosed. In other words, from the known fractions $\hat{p}_{41} = 12/35$, $\hat{p}_{42} = 14/35$, $\hat{p}_{43} = 1/5$, and $\hat{p}_{44} = 2/35$, we can deduce that the actual counts are 12, 14, 7, and 2, respectively, because we know that ν_4 cannot be greater than 0. Similarly, we can use part (d) to show that row 3 is disclosed. Specifically, the condition $r_{i'} \leq \min\{r_3, N - R - r_3\} = 15$ only holds when $i' = 2$, for which we see that $N - R - r_{\min} = 30 < r_3 + r_2 = 35 < 40 = N - R$. We also recall that Proposition 3.1 gave us $\nu_i^- = 0$ for each $i \in \{1, 2, 3, 4\}$, along with $\nu_1^+ = 2$ and $\nu_2^+ = 4$. The arguments just presented provide the two remaining values: $\nu_3^+ = \nu_4^+ = 0$. The corresponding cell bounds $[r_{ij}(\nu_i^- + 1), r_{ij}(\nu_i^+ + 1)]$ are shown in Table 5. Three cells have small disclosed counts and two others have very low upper bounds on small counts, so the table is a likely disclosure risk. The risk might be reduced by aggregating columns and/or rows into fewer categories.

Table 5: Tightest possible cell bounds on observed counts calculated from the observed conditional probabilities of Table 2(b).

	Low	Medium	High	Very High	total
Alpha	[15,45]	[1,3]	[3,9]	[1,3]	[20,60]
Beta	[4,20]	[1,5]	[2,10]	[3,15]	[10,50]
Gamma	[3,3]	[10,10]	[10,10]	[2,2]	[25,25]
Delta	[12,12]	[14,14]	[7,7]	[2,2]	[35,35]

It is worth noting that if all but one of the rows in (a two-way representation of) a table is disclosed, then the entire table of counts is disclosed. Consider the three-way example of Table 6, which shows a dataset from the 1972 National Opinion Research Center General Society Survey regarding the attitude of white Christians toward abortion [14]. We see that each reduced row sum except $r_4 = 31$ exceeds the remainder $N - R = 31$, so part (b) of Proposition 3.2 guarantees disclosure of all rows except the third. Consequently, the third is disclosed as well (with $\nu_3 = 1$).

In fact, full table disclosure is commonplace for tables with a relatively small number of rows and a moderate or large number of responses. The main reason is indicated in the following result.

Table 6: $3^2 \times 3$ table of observed counts with $N = 1055$ and $N - R = 31$.

		Attitudes				
Religion	Education	Positive	Mixed	Negative	o_i	r_i
North Protestant	≤ 8	9	16	41	66	66
	9–12	85	52	105	242	242
	≥ 13	77	30	38	145	145
South Protestant	≤ 8	8	8	46	62	31
	9–12	35	29	54	118	118
	≥ 13	37	15	22	74	74
Catholic	≤ 8	11	14	38	63	63
	9–12	47	35	115	197	197
	≥ 13	25	21	42	88	88

Corollary 3.3. *Suppose the greatest common divisor for the cell counts o_{ij} in each row is 1, so that no row is reduced. If the actual fractions $\hat{p}_{ij}$ are known for all cells in the table, then the entire table is disclosed.*

We close this subsection with a first look at the more complicated four-way example shown in Table 7, which gives the number of patients in a clinical trial for an analgesic drug. The patients' recoveries varied in success according to the levels of three predictor variables [17]. By Proposition 3.1, we see that $\nu_5^+ = 17$ and $\nu_i^- = 0$ for $i \neq 5$ because $N - R = 34$ is divisible by $r_5 = 2$. On the other hand, part (c) of Proposition 3.2 implies that $\nu_2^+ = 0$ and part (d) implies that $\nu_3^+ = 0$. The remaining six maximums and one minimum can be ascertained by various *ad hoc* extensions of the preceding results, such as the comments following Proposition 3.2. However, the next section provides a more systematic approach to handling them.

Table 7: $2^3 \times 3$ table of observed counts with $N = 193$ and $N - R = 34$.

			Recovery				
Center	Status	Treatment	Poor	Modest	Excellent	o_i	r_i
1	1	1	3	20	5	28	28
		2	11	14	8	33	33
	2	1	3	14	12	29	29
		2	6	13	5	24	24
2	1	1	12	12	0	24	2
		2	11	10	0	21	21
	2	1	3	9	4	16	16
		2	6	9	3	18	6

3.2 Simplifications Based on Equality or Divisibility of Reduced Sums

We now present results that are particularly useful when the number of rows is large or some rows are sparsely populated, insofar as they suggest that full disclosure would seldom occur in such cases. In view of Theorem 2.1, the question hinges on the reduced sums that differ from their corresponding observed sums. Intuitively, one would expect many such differences in a table with sufficiently many rows (or a few sufficiently sparse rows). The key is that the reduced sums for many rows will then be equal to, or divisible by, combinations of the reduced sums for other rows. This provides flexibility in how the remaining count $N-R$ can be allocated among rows and allows us to focus on relatively small representative collections of rows. Moreover, the bounds can be calculated quickly and often in closed form. The flexibility afforded by divisibility of one reduced row sum by another is nicely illustrated by the following fact.

Theorem 3.4. *Suppose that $r_{\bar{\imath}}$ is divisible by r_i for some $i \neq \bar{\imath}$. Then the minimum value $\nu_{\bar{\imath}}^-$ is 0 and the maximum value $\nu_{\bar{\imath}}^+$ is the greatest integer less than or equal to $\nu_i^+ r_i / r_{\bar{\imath}}$.*

An important consequence of Theorem 3.4 is that almost all the knapsack optimizations become trivial if some $r_i = 1$. We record that fact here, but defer further discussion of this special case until the end of the section.

Corollary 3.5. *If $r_{\bar{\imath}} = 1$ then, for all i, the maximum value ν_i^+ is the greatest integer less than or equal to $(N-R)/r_i$; moreover, $\nu_i^- = 0$ for all $i \neq \bar{\imath}$. Consequently, the only optimization that remains is the calculation of $\nu_{\bar{\imath}}^-$.*

To see how Theorem 3.4 might be used, we look again at Table 7. Because $r_1 = 20$, $r_4 = 24$, $r_7 = 16$, and $r_8 = 6$ are all divisible by $r_5 = 2$, Theorem 3.4 says that $\nu_1^- = \nu_4^- = \nu_7^- = \nu_8^- = 0$. These minimums agree with the answers we obtained previously at the end of Section 3.1. But in contrast with Proposition 3.1, the approach taken here illustrates how the optimizations for one row might be referred entirely to some other row, *independent* of the value of $N - R$. Moreover, the corresponding maximums ν_i^+ are the greatest integers less than or equal to $\nu_5^+ r_2 / r_i$. These fractions, respectively, are

$$17(2)/20 = 1.7, \quad 17(2)/24 = 1.41\bar{6}, \quad 17(2)/16 = 2.125, \quad 17(2)/6 = 5.\bar{3},$$

which therefore yield $\nu_1^+ = 1$, $\nu_4^+ = 1$, $\nu_7^+ = 2$, and $\nu_8^+ = 5$.

Similar reasoning enables us to remove some variables completely from certain knapsack problems. For this purpose, it is helpful to introduce a little notation. Given row-index sets $\mathcal{L}, \mathcal{K} \subseteq \{1, \ldots, I\}$, we shall write $\mathcal{L} \preceq \mathcal{K}$ to indicate that $\mathcal{L} \subseteq \mathcal{K}$ and that there exist integers $\alpha_{i,k} \geq 0$ such that $r_k = \sum_{i \in \mathcal{L}} r_i \alpha_{i,k}$ for each $k \in \mathcal{K}$. It is readily verified that $\mathcal{L} \preceq \mathcal{M}$ if $\mathcal{L} \preceq \mathcal{K}$ and $\mathcal{K} \preceq \mathcal{M}$ both hold. An important special case is will be used in subsequent examples: provided that $\mathcal{L} \subseteq \mathcal{K}$, we see that $\mathcal{L} \preceq \mathcal{K}$ if each $r_{\bar{\imath}}$ with $\bar{\imath} \in \mathcal{K}$ is either divisible by some r_i with $i \in \mathcal{L}$ or divisible by a sum of r_i-values indexed by $\mathcal{L}$. The next two results show that this concept potentially allows us to remove variables.

Proposition 3.6. *If $\mathcal{L} \preceq \{1, \ldots, I\}$ and $\bar{\imath} \in \mathcal{L}$, then the maximum value $\nu_{\bar{\imath}}^+$ agrees with the optimal value in the (possibly smaller) knapsack problem*

$$\max\ \nu_{\bar{\imath}}\ \text{ over all integers } \nu_i \text{ s.t. } \sum_{i \in \mathcal{L}} r_i \nu_i = N - R, \quad \nu_i \geq 0,\ \forall i.$$

Proposition 3.7. *If $\mathcal{L} \preceq \{1, \ldots, I\} \setminus \{\bar{\imath}\}$, then the minimum value $\nu_{\bar{\imath}}^-$ agrees with the optimal value in the (possibly smaller) knapsack problem*

$$\min\ \nu_{\bar{\imath}}\ \text{ over all integers } \nu_i \text{ s.t. } r_{\bar{\imath}} \nu_{\bar{\imath}} + \sum_{i \in \mathcal{L}} r_i \nu_i = N - R, \quad \nu_i \geq 0,\ \forall i.$$

For the example data in Table 7, we can now find maximum ν-values for the third and sixth rows. For ν_6^+, we note that $r_5 = 2$ divides r_i for each $i \in \{1, 4, 7, 8\}$ and, moreover, that $r_3 = 29 = r_6 + 4r_5$. Thus we can choose $\mathcal{L} \in \{2, 5, 6\}$ and use Proposition 3.6 to omit $i \in \{1, 3, 4, 7, 8\}$ from the maximization. Since we earlier determined that $\nu_2 = 0$ for all feasible solutions, we just need to maximize ν_6 subject to $2\nu_5 + 21\nu_6 = 34$. Because we cannot obtain 34 by adding an even number to 21, we must have $\nu_6^+ = 0$.

For ν_3^+ we can also use Proposition 3.6 and take $\mathcal{L} \in \{2, 3, 5, 6\}$, again because $r_5 = 2$ divides r_i for each $i \in \{1, 4, 7, 8\}$. Since we have already determined that $\nu_2^+ = \nu_6^+ = 0$, the upper bound for the third row amounts to maximizing ν_3 subject to $29\nu_3 + 2\nu_5 = 34$, from which it is evident that $\nu_3^+ = 0$.

The sole remaining optimization for Table 7 is the lower bound for the fifth row, which has $r_5 = 2$. In contrast to the above maximizations, we potentially need several of the rows having even values of r_i because these can accommodate amounts that otherwise must be absorbed by ν_5. By Proposition 3.7, we can omit $i \in \{1, 4\}$ from consideration because $r_1 = 28 = r_7 + 2r_8$ and $r_4 = 24 = 4r_8$, so that $\mathcal{L} \in \{2, 3, 6, 7, 8\}$. Since $\nu_i^+ = 0$ for $i \in \{2, 3, 6\}$, they can be excluded from the minimization so that we must minimize ν_5 subject to $2\nu_5 + 16\nu_7 + 6\nu_8 = 34$. Taking $\nu_7 = 1$ and $\nu_8 = 3$, we can obtain $\nu_5^- = 0$.

The complete set of optimizations used in our discussion of Table 7 are summarized in Table 8. Note that cells in irreducible rows are exposed as their lower bounds, a fact that is guaranteed by the validity of the knapsack representation. Only one small cell count of 3 is disclosed, along with the two disclosive zeros in the third column.

We note that Propositions 3.6 and 3.7 above are immediate consequences of the following technical lemma, which show how to replace one set of row indices with a suitable smaller set.

Lemma 3.8. *Suppose that $\mathcal{L} \preceq \mathcal{K}$ and consider integers ν_i for $i \in \mathcal{K}$. Given any integers $\nu_i' \leq \nu_i$ for $i \in \mathcal{K} \setminus \mathcal{L}$, there exist integers $\nu_i' \geq \nu_i$ for $i \in \mathcal{L}$ so that*

$$\sum_{i \in \mathcal{K}} r_i \nu_i' = \sum_{i \in \mathcal{K}} r_i \nu_i.$$

In particular, if $\nu_i \geq 0$ for all $i \in \mathcal{K}$ then $\sum_{i \in \mathcal{L}} r_i \nu_i' = \sum_{i \in \mathcal{K}} r_i \nu_i$ for some choice of integers $\nu_i' \geq \nu_i$ for $i \in \mathcal{L}$.

Table 8: Optimization of row counts for Table 7.

r_i	ν_i^-	result used	ν_i^+	result used	actual counts and cell bounds $o_{ij} \in [r_{ij}(1+\nu_i^-), r_{ij}(1+\nu_i^+)]$		
28	0	(3.1),(3.4)	1	(3.4)	$3 \in [3,6]$	$20 \in [20,40]$	$5 \in [5,10]$
33	0	(3.1)	0	(3.2)	$11 \in [11,11]$	$14 \in [14,14]$	$8 \in [8,8]$
29	0	(3.1)	0	(3.6)	$3 \in [3,3]$	$14 \in [14,14]$	$12 \in [12,12]$
24	0	(3.1),(3.4)	1	(3.4)	$6 \in [6,12]$	$13 \in [13,26]$	$5 \in [5,10]$
2	0	(3.7)	17	(3.1)	$12 \in [1,18]$	$12 \in [1,18]$	$0 \in [0,0]$
21	0	(3.1)	0	(3.6)	$11 \in [11,11]$	$10 \in [10,10]$	$0 \in [0,0]$
16	0	(3.1),(3.4)	2	(3.4)	$3 \in [3,9]$	$9 \in [9,27]$	$4 \in [4,12]$
6	0	(3.1),(3.4)	5	(3.4)	$6 \in [2,12]$	$9 \in [3,18]$	$3 \in [1,6]$

With this lemma we can also extend the minimization conclusion of Theorem 3.4 to more general circumstances, as shown by the following result.

Corollary 3.9. *If $\mathcal{L} \preceq \mathcal{K}$ and $\bar{\imath} \in \mathcal{K} \setminus \mathcal{L}$, then $\nu_{\bar{\imath}}^- = 0$. Moreover, the knapsack problems* (6)–(8) *admit feasible solutions taking each integer value of $\nu_{\bar{\imath}}$ in the interval $[0, \nu_{\bar{\imath}}^+]$.*

In the context of Table 7, this result shows that $\nu_i^- = 0$ for $i \in \{1,2,3,4,7,8\}$ because the corresponding r_i for each of those is either divisible by $r_5 = 2$ or is a sum of $r_6 = 21$ and a multiple of $r_5 = 2$. This once again confirms six of the minimum values found by Proposition 3.1, but *without* reference to the specific value of $N - R$.

The second statement in Corollary 3.9 also has useful consequences, in that some methods for solving knapsack problems obtain the complete list of possible ν_i-values. If the ν_i-values in some row consisted (say) of unity plus multiples of 5, then that row would be disclosed to a data snooper who knows an upper bound that forces $\nu_i \le 4$. For a row $\bar{\imath}$ satisfying the assumptions of the corollary, however, we are assured that such a bound could not be tightened further without additional information about the table. When Corollary 3.9 does not apply, it still possible to identify all possible values for ν_i, but they may not be consecutive integers. The details are algorithm-specific and beyond the scope of the present work; a forthcoming paper of the authors presents such an algorithm. Algebraic table-enumeration concepts can also be used (e.g., [22]).

It is worth noting that there are many possible ways to extract an index set $\mathcal{L}$ from $\mathcal{K}$ for which $\mathcal{L} \preceq \mathcal{K}$. The two easiest consist simply of removing indices i' from $\mathcal{K}$ for which some smaller index i has $r_i = r_{i'}$, or for which $r_{i'}$ is divisible by a smaller r_i value. We can also remove i' if $r_{i'}$ is a sum of two smaller r_i-values, which we have seen can be helpful when analyzing small tables by hand. More generally, $r_{i'} = r_i\alpha_i + r_{\bar{\imath}}\alpha_{\bar{\imath}}$ for some choice of nonnegative integers α_i and $\alpha_{\bar{\imath}}$ if and only if there is an integer k satisfying $-\beta_{\bar{\imath}} r_{i'}/r_i \le k \le \beta_i r_{i'}/r_{\bar{\imath}}$, where β_i and $\beta_{\bar{\imath}}$ may be taken as any integers satisfying $\gcd(r_i, r_{\bar{\imath}}) = r_i\beta_i + r_{\bar{\imath}}\beta_{\bar{\imath}}$. Such β_i and $\beta_{\bar{\imath}}$ can be obtained by the standard Euclidean algorithm for calculating the greatest common denominator (e.g., [18]).

We end this section by noting that the maximization in Proposition 3.6 is trivial if some r_i divides all the others. This is the case when some row in the table has only a

single nonzero entry: its reduced sum is then 1 and therefore divides all other r_i-values. As noted in Corollary 3.5, the only problem therefore requiring any computational effort is the minimization knapsack of Proposition 3.7 for just that one row, as all the others can be solved in closed form. And even the one remaining problem is typically easy; indeed, it becomes trivial if there are two different rows with lone nonzero entries, because Theorem 3.4 then yields $\nu_i^- = 0$ for both rows.

3.3 Scalability

The methods and theoretical development described so far scale in a computationally efficient way to allow calculation of bounds for very large two-way tables, especially those with a binary response.

For instance, the data shown in Table 9 comprise a six-way dataset due to Edwards and Havranek [9], which records prognostic factors for coronary heart disease on 1841 Czech autoworkers. The two-way arrangement given here shows how one variable (F, whether the subject smokes) is related to combinations of the other five variables. Row 29 has a single nonzero entry and hence $r_{29} = 1$, which divides $N - R = 379$ and all of the other r_i-values. Consequently, we have $\nu_{29}^+ = 379$ and $\nu_i^- = 0$ for all $i \neq 29$ (by Proposition 3.1), and ν_i^+ is the greatest integer less than $379(1)/r_i$ for all $i \neq 29$ (by Theorem 3.4). The only remaining optimization is the calculation of ν_{29}^-, which is zero because $N - R = 379 = 4 + 375 = 2(2) + 75(5) = 2r_{32} + 75r_{28}$. We conclude that the sharpest bounds for all nonzero cells are quite wide. Note that a zero cell potentially constitutes a disclosure risk, insofar as it might reveal that a subject known to belong to a given row *must* satisfy the conditions of the complementary column; we refer to Hundepool et al. [15] for a discussion of disclosive zeros and their privacy implications.

Bounds for much larger two-way tables can also be calculated easily. Using the knapsack formulation (6)–(8), together with simplifications based on the results in the preceding two subsections, typically leads to speedups of one to two orders of magnitude over the integer programming formulation of Smucker et al. [26] and the improved integer programming formulation of §2.2. This enables the calculation, in a few seconds, of all bounds for tables with thousands of cells. Our tests have included two-way rearrangements of an eight-way table from the U.S. Census Bureau's 1993 Current Population Survey with $N = 48842$, and also of a 16-way table from the National Long Term Care Study (see, e.g., [10]) with $N = 21574$. For these tests, the multiway tables were again simplified by choosing some variables as responses and some variables as predictors. For instance, for one rearrangement of the eight-way table cited above, Smucker et al. [26] reported using over 5 hours to find all cell bounds with CPLEX 12.1 on a dedicated compute node (dual quad-core, 64-bit, 2.26 Ghz, 24 GB RAM, Linux). Subsequent improvements in CPLEX (and its MATLAB interface) have reduced that time to 3 minutes, whereas our knapsack formulation can be solved in only 7 seconds. Moreover, it can be handled in 20 milliseconds when we realize that the table has two rows with $r_i = 1$ (cf., Corollary 3.5), because the optimization can be handled in closed form. Even for large tables without that particular opportunity, solution times on the order of 0.1–2 seconds can be obtained.

Table 9: $2^5 \times 2$ table of counts for Czech autoworker data with $N = 1841$ and $N - R = 379$.

						F					
i	A	B	C	D	E	no	yes	o_i	r_i	cell bounds	
1	−	<3	<140	n	n	44	40	84	21	[11, 209]	[10, 190]
2					y	112	67	179	179	[112, 336]	[67, 201]
3				y	n	129	145	274	274	[129, 258]	[145, 290]
4					y	12	23	35	35	[12, 132]	[23, 253]
5			≥140	n	n	35	12	47	47	[35, 315]	[12, 108]
6					y	80	33	113	113	[80, 320]	[33, 132]
7				y	n	109	67	176	176	[109, 327]	[67, 201]
8					y	7	9	16	16	[7, 168]	[9, 216]
9		≥3	<140	n	n	23	32	55	55	[23, 161]	[32, 224]
10					y	70	66	136	68	[35, 210]	[33, 198]
11				y	n	50	80	130	13	[5, 150]	[8, 240]
12					y	7	13	20	20	[7, 133]	[13, 247]
13			≥140	n	n	24	25	49	49	[24, 192]	[25, 200]
14					y	73	57	130	130	[73, 219]	[57, 171]
15				y	n	51	63	114	38	[17, 170]	[21, 210]
16					y	7	16	23	23	[7, 119]	[16, 272]
17	+	<3	<140	n	n	5	7	12	12	[5, 160]	[7, 224]
18					y	21	9	30	10	[7, 266]	[3, 114]
19				y	n	9	17	26	26	[9, 135]	[17, 255]
20					y	1	4	5	5	[1, 76]	[4, 304]
21			≥140	n	n	4	3	7	7	[4, 220]	[3, 165]
22					y	11	8	19	19	[11, 220]	[8, 160]
23				y	n	14	17	31	31	[14, 182]	[17, 221]
24					y	5	2	7	7	[5, 275]	[2, 110]
25		≥3	<140	n	n	7	3	10	10	[7, 266]	[3, 114]
26					y	14	14	28	2	[1, 190]	[1, 190]
27				y	n	9	16	25	25	[9, 144]	[16, 256]
28					y	2	3	5	5	[2, 152]	[3, 228]
29			≥140	n	n	4	0	4	1	[1, 380]	[0, 0]
30					y	13	11	24	24	[13, 208]	[11, 176]
31				y	n	5	14	19	19	[5, 100]	[14, 280]
32					y	4	4	8	2	[1, 190]	[1, 190]

We conclude that the computational time required to find the tightest bounds is much less onerous than previously suggested. The authors have recently determined that even greater speed-ups are possible, and they are preparing an article showing how this allows a real-time interactive exploration of disclosure risk even for such large tables.

4 Discussion

4.1 Implications for Disclosure and Data Utility

The primary contribution of this paper is in the area of disclosure limitation for the specific case of a contingency table released in the form of conditional probabilities along with sample size. Our findings have two main implications.

First, disclosure risk is highest for relatively small, dense tables. Tables with few or no reducible rows are particularly susceptible to full disclosure of all cell counts. This may be of limited concern, however, unless some cell has a small count. Of course, such a situation implies that full utility is retained whereas the release of conditionals has provided no advantage regarding privacy. If the table is not fully disclosed, the conditional probabilities give bounds on the respective margins, which for two-way tables could inform inferential procedures that rely on those quantities.

Second, large sparsely populated tables have little risk of disclosure beyond the implications that might be drawn from cells or rows with zero counts. For instance, if there are rows with only one nonzero cell, this virtually guarantees that there will be no disclosure of that nonzero cell. In general, disclosure risk is lowered when there are more rows with small reduced sums r_i. This result is in contrast to the known situation for released marginal totals. Dobra et al. [8] indicate that for bounds based on released marginal totals, *small* counts in sparse tables are typically associated with tight bounds. Although the reduced sums are lower bounds for their corresponding margins and frequently equal the margins themselves, the presence of many small reduced sums almost certainly alleviates any risk in the context of conditionals and sample size alone.

On the other hand, sparse conditional tables that yield loose bounds will be essentially useless for inference that requires marginal totals. But even more strikingly, odds and odds ratios (which are preserved by conditionals) will also be of limited utility because the excessively wide bounds give inadequate information about the number of individuals upon which the statistics are based.

If conditionals fully disclose the table, inference can proceed undisturbed. Otherwise, additional information in the form of marginals or partial conditionals might enable inference [21] as well. But even for tables that are fully disclosed by the exact conditionals, rounding and perturbations introduced to ensure privacy [1, 3, 27] may substantially reduce the data utility.

4.2 Related and Ongoing Work

The results in this paper can be viewed as a worst-case scenario from the perspective of the data releaser because we assume that the exact fractions are released. In practice, conditional frequencies would be released in decimal form and this may result in an additional layer of data security. The analysis of bounds under released rounded conditionals is more difficult because decimal approximations may admit many possible exact conditional fractions; an investigation along these lines is underway. Still, the results in

this article give a baseline for the risk associated with releasing these data summaries. A larger question of interest regards the confidentiality issues surrounding the release of several sets of marginals and/or conditionals for a general multi-way table, and the degree to which the tools of the present paper might contribute to such a study.

As we prepared this paper for submission, we became aware of a recent unpublished manuscript [22] that examines the same problem as this paper through the lens of algebraic statistics. Their core result agrees with ours, namely, that the heart of the matter lies in the Diophantine equation that constitutes the principal constraint in our knapsack formulation. Both their paper and ours recognize that this provides a characterization of the case in which all cell counts are disclosed, but the papers largely diverge from there. Slavković et al. [22] use algebraic methods and number theory to count the number of possible tables for a given set of conditionals, and then suggest using a full enumeration of those tables to obtain cell bounds from unrounded conditionals. Using Markov bases and graph theory, they provide interesting results on how several tables might be combined to produce bounds and they also address some issues pertaining to inference. They demonstrate their methods on small examples with up to 12 cells. A quick method is outlined for approximate (relaxed) cell bounds from rounded conditionals, which appears to rely on prior knowledge of the unbounded conditionals or their denominators as fractions in lowest terms. In summary, their paper has the same starting point as ours and is aimed at addressing the problem structure, but they take a very different approach and focus on topics complementing those treated here. Karwa and Slavković [16] use similar methods to address conditional inference in contingency tables.

Acknowledgements

The authors thank Andrew J. Sage and the anonymous referees for many suggestions that improved the exposition of the ideas presented herein. We are also grateful to the editor, Stephen E. Fienberg, for guidance on focusing the paper. Miami University's "RedHawk" computing cluster was used for much of the preliminary work that led to this paper.

Appendix

In this appendix, we provide proofs of several results presented in the paper.

Proof of Theorem 2.1. Equation (2) amounts to $n_{ij} = r_{ij}(\nu_i + 1)$, which reflects the need for the cell counts in row i to be a fixed multiple of that row's reduced counts. It also allows us to eliminate the individual cell counts n_{ij} in favor of the row-scaling factors ν_i: simply rewrite the objective in (1) as $r_{\bar{\imath}\bar{\jmath}}(\nu_{\bar{\imath}}+1)$, the Equation (3) as $\sum_i r_i\nu_i = N - R$, and the inequalities (4) as $\nu_i \geq 0$ for all i. Moreover, we see that optimizing $r_{\bar{\imath}\bar{\jmath}}(\nu_{\bar{\imath}} + 1)$ is the same as optimizing $\nu_{\bar{\imath}}$, and then simply adding 1 and multiplying by $r_{\bar{\imath}\bar{\jmath}}$. To handle the integrality condition (5), note that it is equivalent to integrality of

$r_{ij}\nu_i = r_{ij}(\nu_i + 1) - r_{ij}$ for all i and j. We claim that for fixed i, this integrality is equivalent to integrality of ν_i. To see why, consider ν_i expressed as a fraction s/t in lowest terms. The integrality of $r_{ij}\nu_i = (r_{ij}s)/t$ implies that t divides r_{ij} for each j, and so t must divide $\gcd(r_{i1}, \ldots, r_{iJ}) = 1$. This implies that $t = 1$ and hence $\nu_i = s/t$ must be an integer. □

Proof of Proposition 3.2. Condition (a) suffices because the ratios of entries in row i are fixed by the given values of r_{ij}. Conditions (b)–(d) are sufficient because they each prevent the value of ν_i in the constraints (7)–(8) from being increased away from zero for any feasible solution, so that $\nu_i^+ = 0$. □

Proof of Corollary 3.3. No row can be reduced, so $R = N$ and we have $r_i > 0 = N - R$ for all i. Hence part (b) of Proposition 3.2 applies to all rows and the entire table is disclosed. □

Proof of Theorem 3.4. The knapsack constraint (7) is preserved when we replace values ν_i and $\nu_{\bar{\imath}}$ with values ν_i' and $\nu_{\bar{\imath}}'$ satisfying $\nu_i r_i + \nu_{\bar{\imath}} r_{\bar{\imath}} = \nu_i' r_i + \nu_{\bar{\imath}}' r_{\bar{\imath}}$. For convenience, we rewrite this as

$$\nu_i = \nu_i' + (\nu_{\bar{\imath}}' - \nu_{\bar{\imath}})r_{\bar{\imath}}/r_i \quad \text{and} \quad \nu_i' = \nu_i + (\nu_{\bar{\imath}} - \nu_{\bar{\imath}}')r_{\bar{\imath}}/r_i, \tag{9}$$

and use these to prove the following three claims:

1. There is a feasible point satisfying $\nu_{\bar{\imath}} = 0$, so the minimum value is $\nu_{\bar{\imath}}^- = 0$.
2. There is a feasible point satisfying $\nu_{\bar{\imath}} = k$, where k is the greatest integer less than or equal to $\nu_i^+ r_i/r_{\bar{\imath}}$.
3. Every feasible point must have $\nu_{\bar{\imath}} < k + 1$, so the maximum value is $\nu_{\bar{\imath}}^+ = k$.

First, consider a feasible solution whose i and $\bar{\imath}$ entries are given by ν_i' and $\nu_{\bar{\imath}}'$. By the left equation in (9), taking $\nu_{\bar{\imath}} = 0$ and $\nu_i = \nu_i' + (\nu_{\bar{\imath}}' - 0)r_{\bar{\imath}}/r_i \geq \nu_i'$ gives another feasible solution, which proves claim 1.

Next, note that k in the statement of claim 2 is precisely the integer satisfying

$$k \leq \nu_i^+ r_i/r_{\bar{\imath}} < k + 1. \tag{10}$$

The proof of the first claim shows that maximizing ν_i forces $\nu_{\bar{\imath}}$ to its lower bound of 0, giving a feasible solution whose i and $\bar{\imath}$ entries are $\nu_i' = \nu_i^+$ and $\nu_{\bar{\imath}}' = 0$. Taking $\nu_{\bar{\imath}} = k$ and $\nu_i = \nu_i^+ + (0 - k)r_{\bar{\imath}}/r_i$ in the left equation in (9), we see that the leftmost inequality in (10) is equivalent to $\nu_i \geq 0$. We therefore have a feasible solution with $\nu_{\bar{\imath}} = k$, which proves claim 2.

Finally, consider an arbitrary feasible solution with i and $\bar{\imath}$ entries given by ν_i and $\nu_{\bar{\imath}}$. By the right equation in (9), replacing these with $\nu_i' = \nu_i + (\nu_{\bar{\imath}} - 0)r_{\bar{\imath}}/r_i \geq \nu_i$ and

$\nu'_{\bar{\imath}} = 0$ gives a feasible solution. The maximality of ν_i^+ ensures that $\nu_i^+ \geq \nu'_i$, and the rightmost inequality in (10) then gives

$$k+1 > \nu_i^+ r_i/r_{\bar{\imath}} \geq \nu'_i r_i/r_{\bar{\imath}} = (\nu_i + \nu_{\bar{\imath}} r_{\bar{\imath}}/r_i)\, r_i/r_{\bar{\imath}} = \nu_i r_i/r_{\bar{\imath}} + \nu_{\bar{\imath}} \geq \nu_{\bar{\imath}}.$$

Hence $\nu_{\bar{\imath}} < k+1$, as needed for claim 3. This completes the proof of the theorem. □

Proof of Lemma 3.8. For $i \in \mathcal{L}$, let $\nu'_i = \nu_i + \sum_{k \in \mathcal{K}\setminus\mathcal{L}} \alpha_{k,i}(\nu_k - \nu'_k) \geq \nu_i$. Then

$$\begin{aligned}
\sum_{i\in\mathcal{K}} r_i\nu'_i &= \sum_{i\in\mathcal{L}} r_i \left[\nu_i + \sum_{k\in\mathcal{K}\setminus\mathcal{L}} \alpha_{i,k}(\nu_k - \nu'_k)\right] + \sum_{i\in\mathcal{K}\setminus\mathcal{L}} r_i\nu'_i \\
&= \sum_{i\in\mathcal{L}} r_i\nu_i + \sum_{i\in\mathcal{L}}\sum_{k\in\mathcal{K}\setminus\mathcal{L}} r_i\alpha_{i,k}(\nu_k - \nu'_k) + \sum_{i\in\mathcal{K}\setminus\mathcal{L}} r_i\nu'_i \\
&= \sum_{i\in\mathcal{L}} r_i\nu_i + \sum_{k\in\mathcal{K}\setminus\mathcal{L}} r_k(\nu_k - \nu'_k) + \sum_{i\in\mathcal{K}\setminus\mathcal{L}} r_i\nu'_i = \sum_{i\in\mathcal{K}} r_i\nu_i,
\end{aligned}$$

which completes the proof. □

Proof of Corollary 3.9. Consider a feasible solution having $\nu_{\bar{\imath}} = \nu_{\bar{\imath}}^+$. Let $\nu'_{\bar{\imath}}$ be any value from 0 through $\nu_{\bar{\imath}}^+$, let $\nu'_i = \nu_i$ for all $i \in \mathcal{K}\setminus\mathcal{L}$ with $i \neq \bar{\imath}$, and let $\nu'_i = \nu_i$ for all $i \notin \mathcal{K}$. Now apply the lemma to obtain values for ν'_i for $i \in \mathcal{L}$. All ν'_i-values are then nonnegative and the knapsack sum (7) is preserved. □

References

[1] Barak, B., Chaudhuri, K., Dwork, C., Kale, S., McSherry, F., and Talwar, K. (2007). Privacy, accuracy, and consistency too: A holistic solution to contingency table release. In *PODS '07 Proceedings of the 26th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*. 273–282.

[2] Buzzigoli, L. and Gusti, A. (1998). An algorithm to calculate the upper and lower bounds of the elements of an array given its marginals. In *Statistical Data Protection (SDP 1998) Proceedings*. Luxembourg: Eurostat. 131–147.

[3] Charest, A.-S. (2010). How can we analyze differentially-private synthetic datasets? *Journal of Privacy and Confidentiality*, 2(2):21–33.

[4] Chen, Y., Dinwoodie, I., and Sullivant, S. (2006). Sequential importance sampling for multiway tables. *Annals of Statistics*, 34(1):523–545.

[5] Cox, L. (1987). A constructive procedure for unbiased controlled rounding. *Journal of the American Statistical Association*, 82:520–524.

[6] — (2002). Bounds on entries in 3-dimensional contingency tables. In Domingo-Ferrer, J., ed., *Inference Control in Statistical Databases – From Theory to Practice*, vol. 2316 of *LNCS*. Springer-Verlag. 21–33.

[7] Dobra, A. and Fienberg, S. (2010). The generalized shuttle algorithm. In Gibilisco, P., Riccomagno, E., Rogantin, M., and Wynn, H. (eds), *Algebraic and Geometric Methods in Statistics*. Cambridge University Press. 135–156.

[8] Dobra, A., Fienberg, S., Rinaldo, A., Slavkovic, A., and Zhou, Y. (2009), Algebraic statistics and contingency table problems: Log-linear models, likelihood estimation and disclosure limitation, chap. in Putinar, M. and Sullivant, S. (eds), *Emerging Applications of Algebraic Geometry*, Springer Science+Business Media, Inc., vol. 149 of *The IMA Volumes in Mathematics and its Applications*. 63–88.

[9] Edwards, D. and Havranek, T. (1985). A fast procedure for model search in multidimensional contingency tables. *Biometrika*, 72:339–351.

[10] Erosheva, E., Fienberg, S., and Joutard, C. (2007). Describing disability through individual-level mixture models for multivariate binary data. *Annals of Applied Statistics*, 1:346–384.

[11] Fienberg, S., Makov, U., and Steele, R. (1998). Disclosure limitation using perturbation and related methods for categorical data. *Journal of Official Statistics*, 14(4):485–502.

[12] Fienberg, S. and Slavković, A. (2008). A survey of statistical approaches to preserving confidentiality of contingency table entries. In Aggarwal, C., Yu, P., and Elmagarmid, A. (eds), *Privacy-Preserving Data Mining*, vol. 34 of *Advances in Database Systems*. Springer. 291–312.

[13] Fienberg, S. E. and Slavković, 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.

[14] Haberman, S. J. (1978). *The Analysis of Qualitative Data*, vol. 1, 2. Orlando: Academic Press.

[15] Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E. S., Spicer, K., and de Wolf, P.-P. (2012), *Statistical Disclosure Control*. Chichester: John Wiley & Sons.

[16] Karwa, V. and Slavković, A. (2013). Conditional inference given partial information in contingency tables using Markov bases. *Wiley Interdisciplinary Reviews: Computational Statistics*, 5(3):207–218.

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

[18] Nemhauser, G. L. and Wolsey, L. A. (1988). *Integer and Combinatorial Optimization*. Wiley-Interscience.

[19] SDL Report (2005). Report on Statistical Disclosure Limitation Methodology. Federal Committee on Statistical Methodology, Statistical Policy Working Paper 22 (Version Two).

[20] Slavković, A. (2010). Partial information releases for confidential contingency table entries: Present and future research efforts. *Journal of Privacy and Confidentiality*, 1(2):253–264.

[21] Slavković, A. and Fienberg, S. (2009). Algebraic geometry of 2x2 contingency table, chap. in Gibilisco, P., Riccomagno, E., Rogantin, M. P., and Wynn, H. P. (eds), *Algebraic and Geometric Methods in Statistics*. Cambridge University Press. 63–81.

[22] Slavković, A., Zhu, X., and Petrović, S. (2013). Fibers of Multi-way Contingency Tables given Conditionals: Relation to Marginals, Cell Bounds and Markov Bases, Tech. report, The Pennsylvania State University.

[23] Slavković, A. B. (2004). Statistical Disclosure Limitation Beyond the Margins: Characterization of Joint Distributions for Contingency Tables, Ph.D. thesis, Carnegie Mellon University.

[24] Slavković, A. B. and Lee, J. (2010). Synthetic two-way contingency tables that preserve conditionals frequencies. *Statistical Methodology*, 7:225–239.

[25] Smucker, B. and Slavković, A. (2008). Cell bounds in two-way contingency tables based on conditional frequencies. In Domingo-Ferrer, J. and Saygin, Y. (eds), *PSD 2008*, vol 5262 of *LNCS*. Springer-Verlag Berlin Heidelberg. 64–76.

[26] Smucker, B., Slavković, A., and Zhu, X. (2012). Cell bounds in k-way tables given conditional frequencies. *Journal of Official Statistics*, 28(1):121–140.

[27] Yang, X., Fienberg, S., and Alessandro, R. (2012). Differential privacy for protecting multi-dimensional contingency table data: Extensions and applications. *Journal of Privacy and Confidentiality*, 4(1):101–125.