Journal of Privacy and Confidentiality (2009) **1**, Number 2, pp. 235–251

Vulnerability of Complementary Cell Suppression to Intruder Attack

Lawrence H. Cox[*]

Abstract. Complementary cell suppression was the first and remains a popular method for disclosure limitation of magnitude data such as economic censuses data. We show that, when not solved in a rigorous mathematical way, suppression can fail to protect data, sometimes fatally. When solved properly as a mathematical programming problem, suppression is guaranteed to meet certain conditions related to protecting individual data, but we demonstrate that other vulnerabilities exist. Suppression sacrifices both confidential and nonconfidential data, forcing potentially significant degradation in data quality and usability. These effects are often compounded because mathematical relationships induced by suppression tend to produce "over-protected" solutions. To mitigate these effects, it has been suggested that the data releaser provide exact interval estimates of suppressed cell values. We demonstrate for two standard data sensitivity measures that, even when safe, exact intervals further threaten data security, in some situations completely.

Keywords: magnitude data, sensitivity measure, exact intervals, statistical disclosure, p-percent rule, p/q-ambiguity rule, controlled tabular adjustment

1 Introduction

Measurements on variables such as blood pressure, income, number of retail outlets, or number of children are obtained for a population or sample of subjects. It is often convenient to organize the subjects into *categories* to facilitate data presentation or analysis. Categories may be defined by a single characteristic (gender, NAICS code, country of origin, age in ten year bands, or employment class = 0, 1–100, 101–200, ...) or by multiple characteristics (NAICS code by US state, NAICS code by employment class, or age by gender by highest degree held). Categories are typically mutually exclusive, so that each subject is assigned to precisely one category. Some categories may be empty. Tabular data are formed for each category by aggregating measurement data across all subjects within the category. When the data simply count the number of subjects in the category, the data are *categorical data*. Otherwise, the data are *magnitude data* such as total value of shipments, total retail sales, or total number of children. Tabular data are often presented in statistical tables of appropriate dimension(s), and each category corresponds to a unique cell of the table. Data based on averages or percentages can be presented in tables, but here we are not concerned with such tables. Likewise, here all data are nonnegative.

[*] National Center for Health Statistics, Centers for Disease Control and Prevention, Hyattsville, MD, mailto:ldc@cdc.gov

Over time, organizations that release statistical data, such as national statistical offices, develop their own (or, possibly, shared) notions of statistical disclosure, on which agency confidentiality policies and procedures are based. In order that disclosure be defined unambiguously and limited effectively, these concepts must be quantified. For tabular data, this may be accomplished by means of a linear functional over the measurement data called a *sensitivity measure*, discussed in Section 2. The sensitivity measure identifies *sensitive* (disclosure) cells to be those cells achieving a positive value for the sensitivity measure. Moreover, the sensitivity measure can be used to measure disclosure and, based on this measurement, to compute a lower bound on how much the value x of a particular sensitive cell would have to be increased (to $x + r$) to reach a hypothetical cell that is nonsensitive. Defining a corresponding lower bound, e.g., equal to $x - r$, defines the cell's *protection interval* $(x - r, x + r)$—equal to the set of all *unsafe estimates* of x. Statistical disclosure limitation (**SDL**) of tabular data is *complete* if and only if only the tightest (*exact*) interval estimate of each sensitive cell value x computable from the released tabulations is *safe*, viz., strictly contains the protection interval. The theory of sensitivity measures is developed fully in [1].

Complementary cell suppression (**CCS**) is a methodology for statistical disclosure limitation in tabular data. CCS replaces the value of each sensitive cell by a symbol (**D** for "disclosure"). Suppressing only these *primary suppressions* is usually not sufficient to ensure complete SDL, and consequently additional nonsensitive cells, called *complementary suppressions*, must also have their values replaced by **D**. CCS methodology is focused on assuring good choices for the complementary cells, viz., a collection of cells that assures that no protection interval is breached (complete SDL), while suppressing as little useful data as possible. These concepts are summarized in Section 2 and fully developed in [2, 3].

In this paper, we examine the data protection capabilities of complementary cell suppression. There are two widely used rules to define disclosure—the p-percent rule and the p/q-ambiguity rule. Treating magnitude data as continuous data, sophisticated users can compute exact interval estimates of suppressed cell values using linear programming. A fundamental question arises: shouldn't the data releaser make the same information available to all users to assist in their understanding, interpretation, and analysis of suppressed magnitude data? In Section 2, we provide SDL preliminaries on sensitivity measures and CCS, and discuss exact intervals. In Section 3, we examine CCS mathematically, and demonstrate how a suppression pattern protects sensitive data. In Section 4, we present vulnerabilities of CCS: first, for CCS performed in a heuristic, as opposed to rigorous, manner; second, under the p-percent rule; third, for exact intervals under the p/q-ambiguity rule; and finally, for reporting intervals, exact or not, under fairly general assumptions about intruder knowledge. Section 5 contains concluding comments.

2 Statistical Disclosure Limitation Preliminaries

2.1 Sensitivity Measures

Complementary cell suppression has been applied to major, important data collections of economic and other magnitude data in the US, Canada, and the European Union, in some cases for decades, based on software developed at the US Census Bureau, the US National Center for Health Statistics, Statistics Canada, and the EU CASC Project. Suppression has, for the most part, been used exclusively for SDL purposes in such applications. For this reason, we discuss suppression in terms of magnitude data, but remark that our conclusions are equally valid for contingency (count) data.

A simple, widely used disclosure rule for magnitude data is the *p-percent rule* which, in simplified form, states: a tabulation cell **X** is sensitive if, after subtracting the second largest contribution from the cell value, the remainder is within p-percent of the largest contribution. This rule is designed to prevent narrow estimation of any contribution to a cell value by a second contributor or third party. It follows from simple algebra that protecting the largest contributor from the second largest assures that each contributor is protected from each of the others and from an outside third party. Note that magnitude data SDL focuses on contributors as intruders, and that the risk of such *insider disclosure* exceeds that from the outside. p-percent rules have been used in US Economic and Canadian Agriculture Censuses.

X denotes a tabulation cell, and its cell value is x. Denote respondent contributions to x by x_i, ordered from largest to smallest, so that $x = \sum_i x_i$; $x_1 \geq x_2 \geq .... x_i \geq ....$ Express p as a decimal; e.g., 20% = 0.20. Disclosure under the p-percent rule is expressed by the sensitivity measure: $S_p(X) = px_1 - \sum_{i \geq 3} x_i > 0$.

An enhancement of the p-percent rule that incorporates prior information available to an intruder is the *p/q-ambiguity rule***:** the releaser assumes that an intruder can estimate any contribution to within q-percent, $1 \geq q >> p$. Express q as a decimal. Disclosure under the p/q-ambiguity rule is expressed by the sensitivity measure: $S_{p/q}(X) = (p/q)x_1 - \sum_{i \geq 3} x_i > 0$. When $q = 1$, the p-percent and p/q-ambiguity rules are identical; otherwise, it is evident that the p/q-ambiguity rule is *stricter* than the p-percent rule—it identifies as sensitive all p-sensitive cells, possibly more. The p/q-rule has been used in the Canadian Annual Census of Manufactures.

A sensitivity measure is a continuous function. If, as above, it is normalized with the final coefficient = –1, then its value r provides an exact lower bound on the distance from a sensitive cell value x to hypothetical larger cell values that are nonsensitive. We refer to r as the *upper protection limit.* Typically, $-r$ is selected as the lower protection limit, and the open interval $(x - r, x + r)$ is the *protection interval.* See [1] for complete details.

2.2 Complementary Cell Suppression

Complementary cell suppression is a very difficult theoretical and computational problem. CCS can be accomplished using mathematical programming, as follows.

Represent the tabular structure as $\mathbf{Ay} = \mathbf{t}$**.** Entries of **A** are 0 or 1. Original data are $\mathbf{a} = (a_1, \ldots, a_n)$, so that $\mathbf{Aa} = \mathbf{t}$. Sensitive cell values are denoted $a_{d(i)}$, i = 1, . . . , s, and their protection limits $r_{d(i)}$, $0 \leq r_{d(i)} \leq a_{d(i)}$, with r_k= 0 otherwise. Upper and lower protection limits here are equal, as is typical in practice, but in general can be unequal. A mathematical programming model for CCS is given by (1). See also [4].

$$\begin{gathered}
\min \sum_k c_k z_k \\
i = 1, \ldots, s; j = 1, 2; k = 1, \ldots, n : \\
Ay_i, j = t \\
l_k \leq y_{i,1,k} \leq a_k - r_k z_k \\
u_k \geq y_{i,2,k} \geq a_k + r_k z_k \\
z_j = 0, 1; z_{d(i)} = 1
\end{gathered} \tag{1}$$

The first constraint of (1) preserves the tabular structure. The second and third enforce the sensitivity measure. $M \geq 1$ is a suitable constant. Nonnegative lower (l_k) and upper bounds (u_k) on feasible cell values represent intruder prior knowledge, e.g., percentages under the p/q-rule or pseudo-bounds (0 and infinity) if prior knowledge is not assumed. The objective function is selected to preserve or optimize the releaser's notion of data quality: to minimize number of cells suppressed, set $c_k = 1$; to minimize total value suppressed, c_k= a_k; and, to minimize Berg entropy (a compromise between number and total value suppressed), c_k = log (1+ a_k). Note that these are geometrical, not necessarily statistical, measures of quality. Connecting these notions with, e.g., distributional metrics such as Chi-square or Kullback-Liebler distance, would be beneficial.

2.3 Releasing Exact Intervals for Suppressed Values

Magnitude data are treated as continuous data, and therefore exact interval estimates of suppressed cell values y_k can be obtained via linear programming: min y_k (respectively, max y_k) subject to $\mathbf{Ay} = \mathbf{t}$**.** The exact interval for a_k is [min y_k, max y_k]. The model constraints of (1) assure that exact intervals contain protection intervals.

For the p-percent rule, sophisticated users can compute exact intervals, albeit at some effort. It has been argued that the releaser should provide exact intervals in lieu of suppressed data under the p-percent rule. Similarly, for the p/q-rule, as the releaser assumes that an intruder can estimate data within q-percent, shouldn't the releaser provide these estimates for use by legitimate analysts? Doing so would improve analysts' ability to manipulate disclosure-limited tabular data and perhaps analytical precision as well. The question of releasing intervals has been asked since the 1970s, reemerging recently as *partial cell suppression* [5, 6].

If exact interval estimates of suppressed values are released, the user could impute interval midpoints for suppressed data and analyze the "midpoint tables." Unsophisticated users are likely to do this because it is simple. The resulting tables may fail to be additive, but additivity could be regained through *controlled tabular adjustment* [7] or application of iterative proportional fitting (IPF) to impute suppressed values [8].

3 Mathematical Properties of CCS

3.1 Mathematics Underlying CCS

Complementary cell suppression replaces all sensitive and selected nonsensitive values, which are fixed, by symbols, which can be treated as variables. Recall the definition: a CCS solution is complete if the exact interval for each variable corresponding to sensitive values x contain the cell's protection interval $(x - r,\ x + r)$. Table 1 provides a working example.

Table 1: 4x5 Working Example.

T	O	T	A	L	T
					O
	X(10)		**B(5)**		T
					A
	C(7)		**A(8)**		L

X denotes a sensitive cell, and **A, B, C** denote **X**'s complementary suppressions. For our analysis, we extract the essence of Table 1 and provide hypothetical values for the reduced marginal totals, represented in Table 2.

Table 2: Essentials of Table 1.

17	13	30
x=10	b=5	15
c=7	a=8	15

As it contains only four suppressions, Table 2 is the simplest possible example in two or more dimensions. Nevertheless, Table 2 is a building block for the following reasons. Complete suppression patterns in two dimensions can be decomposed into patterns containing 4 or 6 or 8 . . . or 2k cells (*even cycles*) and the mathematical analysis of Section

4 applies mutatis mutandis. This analysis is simply easier to see and present in the simplest, 4-element, case. In higher dimensions or linked tables, each two dimensional slice of a complete pattern is a complete two dimensional pattern and consequently amenable to this analysis.

Let $r = 2$. Then **X** is protected if and only if any interval derivable for x contains $(x - r, x + r) = (10 - 2, 10 + 2) = (8, 12)$. This condition holds if **X** is in an *alternating cycle* of suppressed cells and if the cycle permits a *flow* of $r = 2$ units from x = 10 in both + and − directions. The alternating cycle is given by

17	13	30
X (10)+/−	**B** (5)−/+	15
C (7)−/+	**A** (8)+/−	15

In the + direction, we can move up to 5 units into **X**—more would force $b < 0$.

17	13	30
X (15)	**B** (0)	15
C (2)	**A** (13)	15

In the − direction, we can move up to 8 units out of **X**—more would force $a < 0$.

17	13	30
X (2)	**B** (13)	15
C (15)	**A** (0)	15

Verification that Table 2 protects **X** is demonstrated by exact intervals below. As we can move $r = 2$ units in either direction, **X** is protected.

17	13	30
X [2, 15]	**B** [0, 13]	15
C [2, 15]	**A** [0, 13]	15

CCS is data dependent—the exact intervals above are much broader than the protection interval, whereas the same pattern fails to protect the table below.

17	6	23
X (10)+/−	**B** (5)−/+	15
C (7)−/+	**A** (1)+/−	8

3.2 CCS, Cycles and Protection

Movement of up to 5 (respectively, 8) units through sensitive cell **X** may be represented by the alternating cycle below.

17	13	30
X (10)+/–	**B** (5)–/+	15
C (7)–/+	**A** (8)+/–	15

Equivalently,

17	13	30
x+/–	**b**–/+	15
c–/+	**a**+/–	15

Cells marked with +/– have the *same parity* as *x;* those with –/+ have *opposite parity* to *x.* In general,

- maximum increase to x = minimum value with opposite parity (here, b = 5)
- maximum decrease to x = minimum value with same parity (here, a = 8)
- exact interval for $x = [x - a, x + b]$ (here, = [2, 15])
- *width* of exact interval = $(b + a)$ (here, = 13)
- *radius* of exact interval = $(b + a)/2$ (here, = 6.5)
- interval *midpoint* = $x + (b - a)/2$ (here, = 8.5)
- *bias* in midpoint estimate of $x = (b - a)/2$ (here, = –1.5)

CCS is based on creating cycles that

- contain the sensitive cells **X**
- collectively permit increase/decrease of x to at least $(x - r(X), x + r(X))$
- minimize information loss measured by linear cost function $\sum_k c_k z_k$

Typically (but not necessarily)

- a sensitive cell will be used as a complement for another if possible

- complementary cells are large enough to accommodate protection limits *r(X)*
- but as small as possible to minimize information loss
- an alternative is to select many small cells as complements

Multi-dimensional and linked tables exhibit much more complex cycle structure, including cycles of odd length and not square-free [9], but as mentioned each two-dimensional slice of such systems must comprise complete alternating cycles. If a sensitive cell is contained in multiple cycles, these are analyzed sequentially. Thus, we focus on alternating two dimensional cycles.

4 Confidentiality Characteristics of CCS

4.1 CCS Based on Heuristics Is Vulnerable

If complementary cell suppression is performed using a mathematical model that incorporates protection constraints explicitly, such as (1), exact intervals for suppressed sensitive cells must be nonsensitive and disclosure limitation is complete. Model (1) is a mixed integer linear program, which can be difficult or impossible to solve computationally by direct means such as branch and bound, except for small problems. Recent research has focused on solving medium to large CCS problems using branch and cut and specialized techniques [4]. Unfortunately, many organizations continue to solve CCS problems "by hand" or using computer programs based on "by hand" reasoning. These programs are faster than humans, but in the absence of CCS methodology, offer little improvement in terms of protection or data quality. We give two examples based on a typical disclosure rule for counts that defines the unsafe protection interval to be the interval (0, 5) (*5-threshold rule*).

Table 3: 3x3 Table With Internal Entries Suppressed.

$$\begin{array}{cc} \begin{pmatrix} * & * & * \\ * & * & * \\ * & * & * \end{pmatrix} & \begin{pmatrix} 11 \\ 5 \\ 5 \end{pmatrix} \\ (11 \quad 5 \quad 5\ ) & (21) \end{array} \qquad \begin{array}{cc} \begin{pmatrix} * & * & * \\ * & * & * \\ * & * & * \end{pmatrix} & \begin{pmatrix} 5 \\ 11 \\ 5 \end{pmatrix} \\ (5 \quad 11 \quad 5) & (21) \end{array} \qquad \begin{array}{cc} \begin{pmatrix} * & * & * \\ * & * & * \\ * & * & * \end{pmatrix} & \begin{pmatrix} 5 \\ 5 \\ 11 \end{pmatrix} \\ (5 \quad 5 \quad 11) & (21) \end{array}$$

$$\begin{pmatrix} 1 & 10 & 10 \\ 10 & 1 & 10 \\ 10 & 10 & 1 \end{pmatrix}$$

Table 3 is a 3x3x3 contingency table with all internal entries suppressed. Release of Table 3 is equivalent to release of all 2-dimensional table cells (“line” marginal totals) in place of the 3-dimensional cells (internal entries, marked *). Table 3 is not a realistic confidentiality example because it contains published marginal totals with value = 1, failing the 5-threshold rule. We ignore this issue momentarily, returning to it in the next paragraph. Meantime, compute 2-dimensional Frechet lower bounds for cells (1,1,1), (2,2,2) and (3,3,3) within planes k = 1, 2, 3, respectively. Each of these lower bounds = 1. Each of these three cells is constrained by a marginal total (vertical) = 1, and consequently these cells cannot achieve value > 1. Hence, each has value = 1, which has been revealed and is sensitive—disclosure limitation in this hypothetical example would have been entirely unsuccessful.

We return to the issue of realism. Replace Table 3 by a table comprising 5 copies of Table 3 stacked vertically, viz., a 3x3x15 table with two sets of planar marginals unchanged and the third (vertical) set with values five times those of Table 1 (table not drawn here). This is a realistic example (no marginals < 5) for which 15 cells are revealed to have value = 1, so that disclosure limitation fails completely.

A second example illustrating the failure of non-mathematical CCS methods to protect data is given by Table 4, a two-dimensional table with suppressions.

Table 4: 4x5 Table with Suppressions.

18	**21**	**18**	**23**	**80**
D_{11}	D_{12}	D_{13}	9	**20**
6	D_{22}	D_{23}	6	**20**
D_{31}	5	5	D_{34}	**15**
D_{41}	5	6	D_{44}	**25**

CCS in Table 4 may appear successful, as each suppressed cell is contained in a row and a column containing one or two additional suppressions and as corresponding sums are ≥ 5. But, in fact, $\mathrm{D}_{11} = 1$ can be deduced: add the first two rows: $D_{11} + D_{12} + D_{13} + D_{23} + D_{33} = 19$; add the second and third columns: $D_{12} + D_{13} + D_{23} + D_{33} = 18$; subtract the latter from the former, to obtain $\mathrm{D}_{11} = 1$. Disclosure limitation has failed.

Both examples illustrate that CCS should be done based on a verifiable mathematical model and NOT “by hand” or by software based, in essence, on “by hand” reasoning. For continuous data, these flaws can be detected by linear programming. For contingency data, efficient methods for computing exact intervals are available only for specialized classes of tables [10, 11]. A heuristic iterative min-max method—a *shuttle algorithm* [12]—to refine inexact intervals and on iterative midpoint refinement to construct consistent tables, is presented in [13].

4.2 Vulnerability of the p-Percent Rule

The data releaser may opt to release exact intervals [l, u] for the suppressed cells. Even if the releaser does not do so, the sophisticated user can compute these intervals independently, at least for continuous data. So, it suffices to assume that exact intervals are available. We return to our working example. Again, we remind the reader that all situations are not as simple as this 4-element two-dimensional cycle; but, that all situations do comprise two-dimensional cycles that the intruder can analyze in precisely the same manner as we now proceed to do.

17	13	30
x+/−	**b**−/+	15
c−/+	**a**+/−	15

Assume for concreteness that $b \leq c$ and $a \leq x$. By virtue of the polyhedral geometry of linear constraint systems, the intruder can determine the following.

- $\mathrm{l(x)} = x - a$: a of same parity as x, and $\mathrm{l(a)} = 0$
- $\mathrm{u(x)} = x + b$: b of opposite parity to x, and $\mathrm{l(b)} = 0$
- intruder knows the width of the exact interval $= a + b$
- if intruder can determine a or b or $b - a$ or b/a, then x is revealed

Consequently, protection on a cycle hinges on the intruder's ability to determine a single quantity. If $(b - a)/(2x)$ is small, then the midpoint estimate is precise. Similarly, if **A**, **B** are not historically sensitive, then the intruder can examine historical data to estimate a or b or $b - a$ or b/a, directly, or via regression, and consequently estimate x.

Often, contributor counts are released, so the intruder knows precisely the one contributor cells. If **X** involves two contributors, then by subtracting its own value, the second contributor can obtain even sharper bounds than those to follow.

If **X** involves only one contributor, then the intruder can deduce

$$\mathrm{l(x)} \leq (1 - \mathrm{p})\mathrm{x}$$
$$(1 + \mathrm{p})\mathrm{x} \leq \mathrm{u(x)}$$

Consequently,

$$\mathrm{l(x)}/(1 - \mathrm{p}) \leq \mathrm{x} \leq \mathrm{u(x)}/(1 + \mathrm{p}) \tag{2}$$

which is sharper than $\mathrm{l(x)} \leq \mathrm{x} \leq \mathrm{u(x)}$. In addition,

$$\mathrm{u(x)/l(x) \geq (1+p)/(1-p)}$$

so that

$$\mathrm{(u(x) - l(x)/(u(x) + l(x)) \geq p}$$

By virtue of (2), exact intervals can be "shrunk" if p is known. It is often discussed as to whether the releaser should make the value of p public to enhance analyzability of the data. It would appear that the answer to that question is a resounding "NO". The next question is whether the releaser should release the minimal safe interval (x – r(X), x + r(X)). Again the answer is "NO" because in doing so, x = midpoint of (x – r(X), x+r(X)) is divulged, as are *r = r(X) = (x + r(X)) – (x)* and *p = r/x.* Under *sliding protection* [14], which requires only that the width of the protection interval be at least *2r(X)*, the second question is moot.

If releasing p erodes protection, how well protected is the value of this parameter? For each one or two contributor cell **X**

$$\begin{aligned} \mathrm{p \leq p(X)} &= \mathrm{(u(x) - l(x))/(u(x) + l(x)} \qquad (3)\\ &= \mathrm{((u(x) - l(x))/2)/((u(x) + l(x))/2)}\\ &= \text{(radius/midpoint) of the protection interval for } \mathbf{X} \end{aligned}$$

These inequalities provide (many) upper bounds for p. In the context of a national census, or a set of different censuses or censuses conducted over multiple years, and a great many one contributor cells, many upper bounds (3) for p become available. The smallest, *p' = p(X')*, could be very precise.

A lower bound *p"* on p can be substituted into (2) to sharpen estimation of sensitive single contributor values x. The intruder thus obtains a tighter interval than the exact interval $\mathrm{l \leq x \leq u}$:

$$l \leq l/(1-p'') \leq l/(1-p) \leq x \leq u/(1+p) \leq u/(1+p'') \leq u \qquad (4)$$

A lower bound can be obtained via trial and error as follows.

- begin with any solution (e.g., adjusted midpoint or IPF)

- choose *p"* and protect **X** to within p"-percent

- if the current cycle is not selected, then *p > p"*

- do this for each one contributor cell **X**

- the largest *p"* is a lower bound for p

4.3 Vulnerability of Exact Intervals Under the p/q-Ambiguity Rule

Enhancements to the p/q–ambiguity related to weighting and imputation are examined in [15]. In this section, we discover fundamental weaknesses related to release of exact intervals under a p/q-rule. **X** denotes a sensitive cell under a p/q-rule, and is suppressed with complementary suppressions **A**, **B**, **C**. Assume that exact intervals are released in place of suppressions.

Table 5: Alternating Cycle for Cell Suppression

X $[l_X, u_X]$......+/−	**B** $[l_B, u_B]$ −/+
C $[l_C, u_C]$ −/+	**A** $[l_A, u_A]$ +/−

Assume l_A, l_C, $l_X \geq l_B$ (other cases analogous). Thus, *a, c, x* $\geq$ *b*. From the polyhedral geometry of linear constraints, the intruder can deduce

$$u_X - l_X = u_B - l_B = u_A - l_A = u_C - l_C = \ 2q\ \min\{a, b, c, x\} = 2qb$$

$$\begin{array}{ll} l_B \ = (1-q)b & l_C = c - qb \\ u_B = (1+q)b & u_C = c + qb \\ l_A \ = a - qb & l_X = x - qb \\ u_A = a + qb & u_X = x + qb \end{array}$$

Should the releaser release the value of *q*? The answer is definitely “NO” because these equations would reveal *a, b, c* and *x*. Indeed, it makes no difference whether or not the releaser reveals *q*, as *q* is in fact knowable. For q < 1,

$$u_B/l_B = (1+q)/(1-q)$$

$$q = (u_B - l_B)/(u_B + l_B)$$

resulting in

$$\begin{aligned} b &= l_B/(1-q) \\ a &= l_A + qb \\ c &= l_C + qb \\ x &= l_X + qb \end{aligned}$$

Consequently, release of exact intervals for a p/q-rule results in completely failed disclosure limitation when the suppression pattern corresponds to an alternating cycle. The general case is examined in the next section.

4.4 Vulnerability of Intervals Based on Symmetric Intruder Knowledge

The p/q-ambiguity rule assumes that intruder knowledge is uniform across all contributions and cells, in that the intruder can obtain lower and upper bounds for any value with q-percent accuracy, for fixed q < 1 (expressed as a decimal). A natural generalization is to permit *symmetric intruder knowledge* which may vary from cell to cell, viz., for cells A, B, C, X, the releaser assumes that the intruder can obtain lower or upper bounds for cell values or its contributions accurate to within fractions $0 \leq \mathrm{q_A}$, $\mathrm{q_B}\, \mathrm{q_C}\, \mathrm{q_X} < 1$, respectively. In this scenario, the releaser incorporates the corresponding constraints into the mathematical model (1) for cell suppression, solves the model for a final suppression pattern, and reports an exact interval for each suppressed cell value. Unfortunately, this scheme is also vulnerable, because along a suppression pattern based on an alternating cycle such as Table 5, we have:

$$
\begin{aligned}
a &= u_A - \tfrac{u_B - l_B}{2} \\
b &= u_B - \tfrac{u_B - l_B}{2} \\
c &= u_C - \tfrac{u_B - l_B}{2} \\
x &= u_X - \tfrac{u_B - l_B}{2}
\end{aligned}
\tag{5}
$$

The reason why this scheme fails is given by the following proposition.

Proposition 4.1 Exact intervals $[l_X,\, u_X]$ for a suppression pattern based on an alternating cycle that are computed assuming symmetric intruder knowledge are symmetric around true values x, viz., $x = \frac{u_X + l_X}{2}$, and thus true values are revealed.

Proof Under symmetric intruder knowledge, the exact lower bound l_X for any suppressed cell **X** is at least $(1 - q_X)x$, and the exact upper bound u_X is at most $(1 + q_X)x$. Along an alternating cycle, exact lower bounds are given by

$$
l_X = x - \min\{q_X x,\, \min\{q_B b :\ b\, of\, opposite\, parity\, to\, X\},\, \min\{q_A a :\ a\, of\, same\, parity\, as\, X\}\}
$$

This is because downward deviations of x correspond to both upward deviations of cell values opposite in parity to **X** and to downward deviations of cells of the same parity and, under symmetric intruder knowledge, these deviations cannot exceed the corresponding fraction of the true value. Similarly,

$$u_X = x + \min\{q_X x, \min\{q_B b : \mathit{b\,of\,opposite\,parity\,to\,X}\}, \min\{q_A a : \mathit{a\,of\,same\,parity\,as\,X}\}\}$$

Consequently, $[l_X, u_X]$ is symmetric about the true value x. *Q.E.D.*

It is straightforward to generalize Proposition 4.1 to suppression patterns in two-way tables and further to tables of network type [10]. Indeed, the same result holds in complete generality, as we now demonstrate.

Let **A'y = t'** denote the *system of suppression equations*, derived from the original system **Ay = t** by replacing unsuppressed entries by their true values. Let $l_k \le y_k \le u_k$ denote constraints corresponding to an assumption of symmetric intruder knowledge and let **l**, **u** denote the corresponding vectors. S is the number of suppressed cells. Define the linear program

$$L: \quad \begin{array}{ll} \max\ (0) & \\ A' \begin{pmatrix} y_1 \\ \vdots \\ y_S \end{pmatrix} = t' & \\ l_k \le y_k \le u_k & k = 1, \ldots, S \end{array} \tag{6}$$

Theorem 4.1 Exact intervals for a suppression pattern computed assuming *symmetric intruder knowledge* $l_k \le y_k \le u_k$, $a_k - l_k = u_k - a_k$, are symmetric around true values a_k , regardless of the underlying tabular structure.

Proof Denote the vector of true values of suppressed entries by $\mathbf{x^{(0)}}$, so that $\mathbf{x^{(0)}}$ satisfies L. Consider any suppressed true value ${x^{(0)}}_K$, 1 ≤ K ≤ S. A quantity ${x^{(0)}}_K + u$ is a feasible upper bound for ${x^{(0)}}_K$ if u ≥ 0 and if there exists a vector **w** with $w_K = u$ satisfying L, viz., $A'(x^{(0)} + w) = t'$, $l_k \le x_k^{(0)} + w_k \le u_k$ $(k = 1, \ldots, S)$, $A'(w) = 0$. Similarly, ${x^{(0)}}_K$ is a feasible lower bound for ${x^{(0)}}_K$ if l ≥ 0 and there exists a vector **v** with $v_K = l$ satisfying L, viz., $A'(x^{(0)} - v) = t'$, $l_k \le x_k^{(0)} - v_k \le u_k$ $(k = 1, \ldots, S)$, $A'(-v) = 0$. Let $\mathbf{v^{(0)}}$ and $\mathbf{w^{(0)}}$ denote vectors satisfying these respective conditions whose respective Kth coordinates $v_K^{(0)}$, $w_K^{(0)}$ are maximal. As $x^{(0)} - l = u - x^{(0)}$, then $l \le x^{(0)} - w^{(0)} \le u$. Thus, $A'(x^{(0)} - w^{(0)}) = t'$, $l_k \le x_k^{(0)} - w_k^{(0)} \le u_k$ $(k = 1, \ldots, S)$, $A'(-w^{(0)}) = 0$, and $x_K^{(0)} - w_K^{(0)}$ is a feasible lower bound for ${x^{(0)}}_K$. Hence, $v_K^{(0)} \ge w_K^{(0)}$. Identical reasoning yields $w_K^{(0)} \ge v_K^{(0)}$. Consequently, $v_K^{(0)} = w_K^{(0)}$ and the Kth interval is symmetric. *Q.E.D.*

Thus, if the releaser wishes to provide bounds for suppressed entries, these bounds must be nonsymmetric. If the original disclosure rule gives rise to nonsymmetric protection intervals, then $x_k^{(0)} - l_k$, $u_k - x_k^{(0)}$ are not necessarily equal, and interval midpoints are not necessarily equal to true values. If, in addition, no functional relationship exists between l_k, u_k, then the releaser might be able to provide exact intervals for true values. Otherwise, the releaser might adopt the following procedure or an adaptation for the case $x_k^{(0)} - l_k = u_k - x_k^{(0)}$.

Procedure: Randomly select values l'_k, $l \leq l'_k \leq a_k - r_k$, $k = 1, ..., S$ from S independent uniform distributions. If $l'_k - l_k \geq a_k - r_k - l'_k$, define $n_\kappa = a_k + r_k + (a_k - r_k - l'_k)$ and $\sigma_\kappa = \frac{1}{4}(a_k - r_k - l'_k)$; otherwise $n_\kappa = u_k - (l'_k - l_k)$ and $\sigma_\kappa = \frac{1}{4}(l'_k - l_k)$. Randomly select values u'_k, $u_k \geq u'_k \geq a_k + r_k$, $k = 1, ..., S$ from S independent truncated normal distributions $N(n_k, \sigma_k^2)$, $k = 1, ..., S$, truncated to intervals$[n_\kappa - 4\sigma_\kappa, n_\kappa + 4\sigma_\kappa)]$. In (1), replace symmetric bounds l_k, u_k by general bounds l'_k, u'_k.

A triangular or other symmetric distribution can be substituted for the truncated normal. The procedure resolves the symmetry problem (deterministic), and the choice of distribution (stochastic) assures that the expected value of each interval midpoint equals the true value. This property is useful, e.g., for data analysis at higher levels of aggregation. Conversely, if this property is unnecessary, a different choice of distributions can be made.

This procedure resolves the problem raised by Theorem 4.1. But, there is a caveat. To be effective in moving interval midpoints off true values, any procedure must in some cases narrow the original interval, perhaps considerably. Doing so opens the releaser to the vulnerability discussed in Section 4.2.

5 Concluding Comments

We have shown that complementary cell suppression has negative effects on data quality. Attempts to mitigate these effects that have been suggested elsewhere include

- release the parameter p of a p-percent rule
- release exact intervals in place of suppressions
- release the parameter q of a p/q-rule

We have shown that these alternatives may seriously threaten confidentiality. We have demonstrated means by which an intruder may compromise the security of suppressed data. The extent of these threats in practice needs to be examined, potential remedies explored, and alternatives, such as controlled tabular adjustment [7], considered.

We have shown how polyhedral geometry associated with suppression forces maximal masking of suppressed sensitive cell values. Exact intervals that are overly broad result when relative positions and sizes of values in the table combine to force selection of complementary suppressions that exceed the protection limits. The objective function is designed to minimize such *overprotection*, but it remains possible that the achieved minimum appears overly suppressive, e.g., in Table 1, $r(\mathrm{X}) = 2$ units of protection is required but 5 (or 8) units are actually provided. Providing safe but not minimally safe (exact) intervals is one way to mitigate these effects, but the problem of symmetry must be avoided. We provide a procedure by which this may be accomplished, but caution that any such procedure has the potential to reveal values for parameters of the underlying disclosure limitation rule, leading potentially to disclosure of true values of suppressed data. The security of any suppression-based disclosure limitation scheme needs to be vetted through analysis of vulnerabilities similar to that presented here.

Author's Statement This work solely represents the findings and opinions of the author and should not be interpreted as representing the policies or practices of the Centers for Disease Control and Prevention or any other organization or group. Partial results of this research were presented in [16].

6 References

1. Cox, L. H. Linear sensitivity measures in statistical disclosure control. *Journal of Statistical Planning and Inference* **5** (1981) 153–164.

2. Cox, L. H. Suppression methodology and statistical disclosure control. *Journal of the American Statistical Association* **75** (1980) 377–385.

3. Cox, L. H. Network models for complementary cell suppression. *Journal of the American Statistical Association* **90** (1995) 1453–1462.

4. Fischetti, M. and J. J. Salazar. Solving the cell suppression problem on tabular data with linear constraints. *Management Science* **47/7** (2001) 1008–1026.

5. Fischetti, M. and J. J. Salazar. Partial cell suppression: A new methodology for statistical disclosure control. *Statistics and Computing* **13** (2003) 13–21.

6. Salazar, J. J. A unified mathematical programming framework for different statistical disclosure limitation methods. *Operations Research* **53** (2005) 819–829.

7. Cox, L. H., J. P. Kelly and R. Patil. Balancing quality and confidentiality for multivariate tabular data. In *Privacy in Statistical Databases 2004, volume 3050 of Lecture Notes in Computer Science.* (J. Domingo-Ferrer and V. Torra, eds.), Berlin: Springer-Verlag (2004) 87–98.

8. Cox, L. H , J. G. Orelien and B. V. Shah. A method for preserving statistical distributions subject to controlled tabular adjustment. In *Privacy in Statistical Databases 2006, volume 4302 of Lecture Notes in Computer Science* (J. Domingo-Ferrer and L. Franconi, eds.), Heidelberg: Springer (2006) 1–11.

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

10. Cox, L. H. Contingency tables of network type: Models, Markov basis and applications. *Statistica Sinica* **17** (2007) 1371–1393.

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

12. Buzzigoli, L. and A. Giusti. An algorithm to calculate the lower and upper bounds of the elements of an array given its marginals. Statistical data protection: Proceedings of the conference, Lisbon, March 25–27, 1998, Luxembourg: European Communities (1999) 131–147.

13. Isserman, A. M and J. Westervelt. 1.5 million missing numbers: Overcoming employment suppression in *County Business Patterns* data. *International Regional Science Review* **29** (2008) 311–335.

14. Kelly, J. P., B. L. Golden and A. A. Assad. Cell suppression: Disclosure protection for sensitive tabular data. *Networks* **22** (1992) 397–417.

15. O'Malley, M. and L. R. Ernst. Practical considerations in applying the p/q-rule for primary disclosure suppression. *2007 Proceedings of the American Statistical Association, Survey Research Methods Section*, [CD-ROM}, Alexandria, VA: American Statistical Association (2007).

16. Cox, L. H. A data quality and data confidentiality assessment of complementary cell suppression. In *Privacy in Statistical Databases 2008, volume 262 of Lecture Notes in Computer Science* (J. Domingo-Ferrer and Y. Saygin, eds.), Heidelberg: Springer (2008) 13–23.