A Data Structure and Integer Programming Technique to Facilitate Cell Suppression Strategies

Colleen M. Sullivan and Errol G. Rowe*
Bureau of the Census, Washington, D.C. 20233

Keywords: Disclosure Avoidance, Cell Suppression, Tabular Data, Integer Programming

1. INTRODUCTION

The U.S. Bureau of the Census has the responsibility to collect data regarding economic sectors and to publish these data without violating confidentiality laws. Collected data contain sensitive data values that if directly published could identify an individual establishment's data. There are a number of methods available that prevent compromising the sensitive cells. These disclosure avoidance techniques include rounding, perturbation, and cell suppression, and are outlined in Cox, et al. (1986a).The Bureau's current practice is to protect any cell where n or fewer respondents make up k percent or more of a table cell's value (Zayatz, 1992). (The values of n and k are confidential.)

Since rounding and perturbation are unsatisfactory for economic aggregate magnitude data (Cox, et al. 1986b), the Economic Divisions have always chosen a cell suppression technique to protect published tabular data. Instead of the sensitive data value appearing in the publication, a "D" appears in its place. However, in most cases, the sensitive data values could be derived from non-sensitive data because most data items are published in additive tables. Therefore, additional data values must be suppressed. These additional suppressed data values are commonly referred to as complementary suppressions. The objective adhered to by the Census Bureau in applying complementary suppressions is to minimize the sum of the data values chosen as complementary suppressions. Minimizing the cost incurred through complementary suppressions produces a publishable table with maximum data utility; that is, the greatest amount of usable data is provided.

Furthermore, the Bureau uses complementary suppressions to ensure that a data user cannot estimate the value of a sensitive data cell within a predefined interval. That is, when choosing complementary suppressions for some primary suppression with true value X, we ensure that it cannot be estimated within a smaller interval than [X-L,X+U] where L is the amount of lower protection required by X, and U is the amount of upper protection required by X. Kelly, et al. (1991) discusses protection levels in greater detail.

In recent years, the Economic Divisions of the Bureau have employed a cell suppression technique that utilizes network flow methodology. The origin of using graph theory in the disclosure avoidance area lies in Cox (1980), and Gusfield (1984). More recently, Cox, et al. (1986a) has outlined this methodology, and a more complete history is given in Greenberg (1990). A general outline of the minimum cost network flow problem and related methodology appears in Bazaraa & Jarvis (1977), and Gondran & Minoux (1984).

The network flow system currently employed is implemented using the commercially available Minimum Cost Flow (MCF) program of Glover, Klingman and Mote. As described in its documentation, "MCF is a highly refined implementation of the upper bounded, revised primal simplex algorithm for linear programming." With this refined implementation, the primal simplex method can be performed directly on a network. Kennington and Helgason (1980) refer to this procedure as the "simplex on a graph" algorithm.

Although MCF is computationally fast, it often oversuppresses due to the structure of the objective function (See Section 3). The ideal technique for choosing complementary suppressions is the integer programming routine outlined in Section 2. This routine, however, is computationally impractical for census tables. This paper discusses a hybrid technique outlined in Section 4 (using MCF and integer programming) that lessens the oversuppression problem without adding substantial computation time (see Section 5).

2. THE IP FORMULATION

The cell suppression problem has a theoretical integer programming (IP) statement that produces an optimal solution (Greenberg, 1986). In this routine, there is an indicator variable, I_{ij}, that is restricted to be zero or one. Thus, we consider decisions in which just two outcomes are possible: we either assign a complementary suppression to a particular table cell or we do not. Consequently, the IP formulation minimizes the sum of the values chosen as complementary suppressions while maintaining the

confidentiality of the primary suppression within a specified tolerance level. The formulation follows:

$$\min_{I_{ij}} c = \sum_{i=1}^{m+1}\sum_{j=1}^{n+1} e_{ij}I_{ij}$$

subject to

$$f_{i,n+1} = \sum_{j=1}^{n} f_{ij} \quad i=1,m$$

$$f_{m+1,j} = \sum_{i=1}^{m} f_{ij} \quad j=1,n$$

$$g_{i,n+1} = \sum_{j=1}^{n} g_{ij} \quad i=1,m$$

$$g_{m+1,j} = \sum_{i=1}^{m} g_{ij} \quad j=1,n$$

$$f_{m+1,n+1} = \sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}$$

$$g_{m+1,n+1} = \sum_{i=1}^{m}\sum_{j=1}^{n} g_{ij}$$

$$f_{de} = e_{de} - l_{de}$$

$$g_{de} = e_{de} + u_{de}$$

$$e_{ij} - I_{ij}e_{ij} \le f_{ij} \le e_{ij} + I_{ij}e_{ij} \quad i=1,m+1 \text{ and } j=1,n+1$$

$$e_{ij} - I_{ij}e_{ij} \le g_{ij} \le e_{ij} + I_{ij}e_{ij} \quad i=1,m+1 \text{ and } j=1,n+1$$

$$I_{ij} = 0 \text{ or } 1 \quad i=1,m+1 \text{ and } j=1,n+1$$

$$I_{ij} = 1 \quad \forall (i,j) \in S$$

where e_{ij} is the value of the entry in row i column j of the table that requires complementary suppressions, l_{de} is the amount of lower protection required by the primary cell, u_{de} is the amount of upper protection required by the primary cell, f_{ij} is the amount of uncertainty e_{ij} contributes to achieving l_{de}, g_{ij} is the amount of uncertainty e_{ij} contributes to achieving u_{de}, m is the number of internal rows in the table, n is the number of internal columns in the table, S is the set of cells that are suppressed, I_{ij} is 1 if e_{ij} is suppressed, and 0 otherwise.

To illustrate, suppose Table 1 depicts 4 (m) products produced in 3 (n) counties.

	County 1	County 2	County3	Total
Product 1	146	444	213	803
Product 2	675	8	991	1674
Product 3	P 312	395	561	1268
Product 4	19	346	11	376
Total	1152	1193	1776	4121

Table 1

Suppose that the table entry in row 3, column 1 (e_{31}) is a primary suppression; i.e., e_{31} is considered too sensitive to be released. Further, suppose we want to protect e_{31} by an upper and lower protection of at least 46 units. That is, we want to prevent users from estimating the value of e_{31} any finer than the range $262 < P < 358$. If the above integer programming formulation is applied to this problem, the cells shown with a "C" in Table 2 are chosen as complementary suppressions.

	County 1	County 2	County 3	Total
Product 1	C 146	444	C 213	803
Product 2	675	8	991	1674
Product 3	P 312	395	C 561	1268
Product 4	19	346	11	376
Total	1152	1193	1776	4121

Table 2

This result is optimal. However, Kelly (1990) has shown that the cell suppression problem is NP-hard; that is, there is no known polynomial time algorithm to solve the problem with optimal results every time, and all known methods take exponential time. This implies that for large tables, as many census tables are, the IP formulation is an impractical choice.

3. THE MCF FORMULATION

Due to the unreasonable amount of time used by the IP formulation, the Bureau utilizes a heuristic known as the MCF program that employs network flow methodology. A key idea is to transform a two dimensional table, like Table 1, into a network diagram. The transformation from table to network is described in Sullivan and Zayatz (1991), and Rowe (1991).

Figure 1 presents a general network diagram. In this figure Δ_{ij} and δ_{ij} represent the amount of uncertainty that the table entry in row i, column j (e_{ij}) contributes to achieving an upper and lower protection required by the primary suppression.

To illustrate the MCF approach, again suppose that e_{31} is a primary suppression. Then finding a suppression pattern to protect e_{31} in the table corresponds to finding a set of closed paths (cycles) containing Δ_{31} in the network. All other cells, represented by arcs in the chosen cycles, would then be suppressed as complements in the table. Our objective is to choose the set of cycles through the network that suppresses the least amount of data

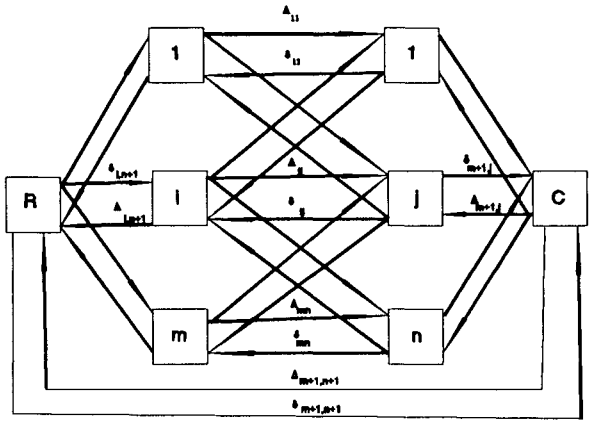

Figure 1. Network Diagram

value while protecting the sensitive cell. We do this by solving a minimal cost flow problem.

Each suppression problem can be viewed as a specialized linear programming problem of the general form:

$$\min_{\Delta_{ij},\delta_{ij}} z = \sum_{i=1}^{m+1}\sum_{j=1}^{n+1} a_{ij}(\Delta_{ij}+\delta_{ij})$$

subject to

$$\sum_{j=1}^{n+1}\Delta_{ij} - \sum_{k=1}^{n+1}\delta_{ki} = 0 \quad i=1,m+1$$

$$\sum_{i=1}^{m+1}\Delta_{ij} - \sum_{k=1}^{m+1}\delta_{kj} = 0 \quad j=1,n+1$$

$$0 \le \Delta_{ij},\delta_{ij} \le p_{ij} \quad i=1,m+1 \text{ and } j=1,n+1$$

$$0 \le \Delta_{de} \le c_{de}$$

$$\delta_{de} = 0$$

where a_{ij} is -99999 if e_{ij} is the primary suppression being protected, and e_{ij} otherwise; Δ_{ij} for i from 1 to m and j from 1 to n, and i equal to m+1 and j equal to n+1, $\delta_{i,n+1}$ for i from 1 to m, and $\delta_{m+1,j}$ for j from 1 to n is the amount of uncertainty e_{ij} contributes to achieving the required upper protection; δ_{ij} for i from 1 to n and j from 1 to n, and i equal to m+1, ,j equal to n+1, $\Delta_{i,n+1}$ for i from 1 to m, $\Delta_{m+1,j}$ for j from 1 to n is the amount of uncertainty e_{ij} contributes to achieving the required lower protection; p_{ij} is the maximum amount of uncertainty e_{ij} could contribute to achieving either the upper or lower protection; c_{de} is the amount of upper and lower protection required by the primary suppression.

Applying the MCF formulation to Table 1, (again protecting $e_{3,1}$ by an upper and lower protection of at least 46 units) the cells shown with a "C" in Table 3 are chosen as complementary suppressions.

Comparing the complementary suppressions chosen by MCF (shown in Table 3) with those chosen by the integer programming formulation (shown in Table 2),

	County 1	County 2	County 3	Total
Product 1	C 146	444	C 213	803
Product 2	675	8	991	1674
Product 3	P 312	C 395	C 561	1268
Product 4	C 19	C 346	C 11	376
Total	1152	1193	1776	4121

Table 3

we see that MCF is guilty of oversuppression. Since the objective function in the MCF formulation minimizes the sum of the products of the data values chosen as complementary suppressions and their corresponding uncertainty variables, the solution is suboptimal; that is, there is potential for oversuppression. For the above example, the total data value lost to complementary suppressions is 1691 for MCF, and only 920 for the IP formulation.

4. The HYBRID FORMULATION

Provided that the set of complementary suppressions generated by MCF is small enough, an IP routine can refine the MCF solution. The hybrid method performs the refinement operation using MCF and a variant of the IP formulation. We begin with the network shown in Figure 1. Applying MCF, it determines a complementary suppression scheme and produces a data structure tree that corresponds to the scheme. For instance, the suppression tree for the suppression pattern shown in Table 3 is given in Figure 2. The backedge (the arrow from 6 to 1) indicates the primary suppression and all other arcs represent cells in the suppression pattern.

First, we identify all closed paths in the tree using a recursive algorithm that performs a depth-first search on the nodes of the graph (Cormen, et al. (1990)). We augment the algorithm to determine the cost of using each path and the maximum amount of protection each path can provide the primary suppression. (The maximum protection each path can give is the smallest value in the path.) From Figure 2, we have three paths (excluding the backedge):

path A: 1-2-4-6 cost: 760 max prot: 19
path B: 1-3-4-6 cost: 591 max prot: 11
path C: 1-3-5-6 cost: 92 max prot: 146

Often the paths are not disjoint. As shown in Figure 2, path A and path B share the arc from 4 to 6, and paths B and C share the arc from 1 to 3.

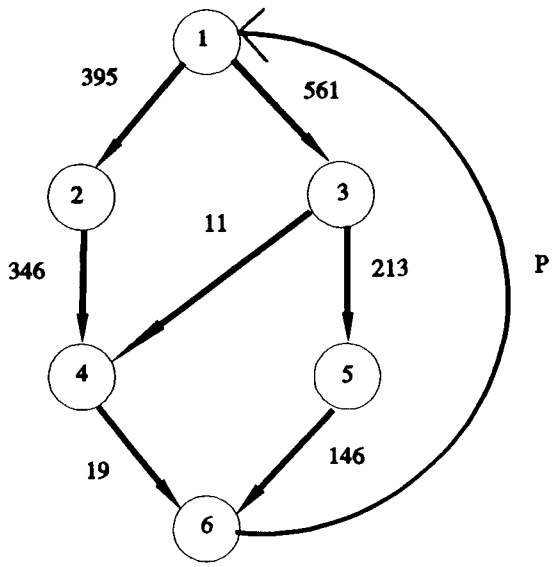

Figure 2. Suppression Tree

When two paths are not disjoint and the protection allowance of the arc they share is less than the protection needed for the primary suppression we call this a bottleneck. Bottlenecks are important since they limit the amount of protection a combination of paths can give to the primary. For example, in Figure 2, paths A and B used together supply nineteen units of protection, not thirty even though path A by itself supplies nineteen units of protection and path B by itself supplies eleven units of protection. This is because paths A and B use the common arc from 4 to 6 with a protection allowance of nineteen units.

Using the information obtained from the suppression tree, we are able to construct a reasonably small set of constraints to be used in the Balas 0-1 IP algorithm (see Syslo, et al. (1983)). The resulting IP formulation then attempts to refine the MCF solution by removing a subset of complementary suppressions. The formulation follows:

$$\min_{I_j} y = \sum_{j=1}^{n} c_j I_j$$

subject to

$$\sum_{j=1}^{n} a_j I_j \geq f \qquad (1)$$

$$\sum_{j \in B(i)} a_j I_j \leq b_i \qquad (2)$$

$$I_j = 0 \text{ or } 1$$

where n is the number of paths, c_j is the cost of suppressing path j, a_j is the maximum amount of protection provided by path j, f is the amount of protection required by the primary suppression, I_j is 1 if path j is to be suppressed, 0 otherwise, B(i) is the set of paths that share arc i, and b_i is the amount of protection supplied by the bottleneck arc.

In the above formulation, constraint (1) ensures that the cells in the paths chosen for suppression protect the primary suppression by the required amount. Constraint (2) accounts for the fact that the amount of protection given to the primary suppression by all paths that share an arc is not more than the protection allowance of the shared arc.

Applying the hybrid formulation to Figure 2, we have:

$$\min \quad y = 760\, I_1 + 591\, I_2 + 920\, I_3$$

subject to

$$19\, I_1 + 11\, I_2 + 146\, I_3 \geq 46$$

$$19\, I_1 + 11\, I_2 \leq 19$$

$$I_j = 0 \text{ or } 1, \text{ for all } j$$

Note the reason we do not include $11\, I_2 + 146\, I_3 \leq 561$ as a constraint in the above formulation is that the protection allowance of 561 shared by the two paths is greater than the protection needed for the primary suppression, and thus does not constitute a bottleneck.

Table 4 shows the results of applying the hybrid method to Table 2.

	County 1	County 2	County 3	Total
Product 1	C 146	444	C 213	803
Product 2	675	8	991	1674
Product 3	P 312	395	C 561	1268
Product 4	19	346	11	376
Total	1152	1193	1776	4121

Table 4

Comparing the complementary suppressions chosen by MCF (shown in Table 3) with those chosen by the hybrid method (shown in Table 4), we see that the hybrid method lessens the oversuppression problem caused by MCF. In fact, in this example, the complementary suppressions chosen by the hybrid method are the same as those chosen by the IP formulation (shown in Table 2).

As exemplified above, the hybrid method is frequently able to release some of the superfluous suppressions applied by MCF. Consequently, the sum of the data values suppressed by this hybrid

technique will lie between the sum of the suppressed values of an IP optimal solution and the corresponding sum of the MCF suboptimal solution.

5. COMPUTATIONAL COMPARISONS

The hybrid method would not be a worthwhile improvement if it added excessive computation time to the MCF method when used alone. However as described below, the hybrid technique takes only slightly more time, especially as compared to the IP formulation. Yet it substantially improves the total data value suppressed.

Tables 5 and 6 show the results of an experiment using each technique on square (n x n) tables with one primary suppression. For each n from 5 through 13, five tables of random numbers were generated, and the three techniques were applied to the table. Table 5 shows the average ratio of data suppressed by each technique as compared to the IP method.

n	MCF/IP	Hybrid/IP
5	1.407	1.035
6	1.503	1.218
7	1.374	1.015
8	1.910	1.513
9	1.576	1.269
10	1.777	1.100
11	1.642	1.446
12	1.319	1.245
13	1.733	1.169

Table 5

Table 5 shows that a significant number of MCF complementary suppressions were released using the hybrid technique. On average MCF suppressed 58 percent more data than necessary, while the hybrid suppressed 22 percent more than required by IP.

Table 6 shows the average CPU time (including disk time) used by the IP method.[1] The comparable averages for both the MCF method and the hybrid method were all less than one second. In fact, the MCF method and the hybrid method had average

n	IP
5	10.48 sec
6	24.8 sec
7	1.14 min
8	2.82 min
9	3.38 min
10	6.93 min
11	18.76 min
12	55.39 min
13	1.34 hrs

Table 6

CPU times less than 1 second for tables where $n \leq 50$. Since the IP method quickly became intractable for $n > 13$, comparable times are not available. However, times for IP where $n \geq 15$ can be estimated by first fitting an exponential function, $T(n) = a^n$, to the IP times for $n \leq 13$, and then substituting n into the fitted function.[2] The fitted function predicts that the IP method would require 32 days for $n=25$ and 2350 centuries for $n = 50$.

6. CONCLUSION

This paper showed that the MCF program currently used by the Bureau to apply complementary suppressions to economic data often oversuppresses. The IP routine, which is optimal, requires an exorbitant amount of computer time, and is thus impractical for census data. Thus, we have presented a hybrid technique using the suppression scheme produced by MCF, along with an IP routine to release superfluous suppressions. In an experiment, the hybrid routine oversuppressed 22 percent as compared to 58 percent by MCF, yet added no substantial computation time.

ENDNOTES

1. For all three methods, performance was measured on a Solbourne Series 5/605, using one of five 22 MIPS processors.

2. There is no guarantee that the time for the IP formulation is actually an exponential function. This is merely used to show that the IP formulation is impractical for census data.

REFERENCES

Bazaraa, M.S., and Jarvis, J.J. (1977), *Linear Programming and Network Flows*, New York: John Wiley and Sons, Inc.

Cormen, T.H., Leiserson, C.E., and Rivest, R.L. (1990), *Introduction to Algorithms*, McGraw-Hill Book Company.

Cox, L.H. (1980), "Suppression Methodology and Statistical Disclosure Control," *Journal of the American Statistical Association*, 75, 377-385.

Cox, L.H., Fagan, J.T., Greenberg, B.V., and Hemmig, R.J. (1986a), "Research at the Census Bureau into Disclosure Avoidance Techniques for Tabular Data," *Proceedings of the American Statistical Association, Survey Research Methods Section.*

Cox, L.H., McDonald, S., and Nelson, D. (1986b), "Confidentiality Issues at the United States Bureau of the Census," *Journal of Official Statistics*, 2, 135-160.

Gondran, M. and Minoux, M. (1984), *Graphs and Algorithms*, New York: John Wiley and Sons, Inc.

Greenberg, B.V. (1986), "Using Mathematical Programming to Find Complementary Suppression Cells in Tabular Data," Bureau of the Census, unpublished manuscript.

Greenberg, B.V. (1990), "Disclosure Avoidance Research at the Census Bureau," *Proceedings of the Bureau of the Census Sixth Annual Research Conference*, Bureau of the Census, Washington, D.C.

Gusfield, D. (1984), "A Graph Theoretic Approach to Statistical Data Security," Department of Computer Science, Yale University, New Haven.

Kelly, J.P. (1990), "Confidentiality Protection in Two and Three-Dimensional Tables," Ph.D. Dissertation, University of Maryland, College Park, Maryland 20742.

Kelly, J.P., Golden, B.L., and Assad, A.A. (1991), "Cell Suppression: Protection for Sensitive Tabular Data," Working Paper Series MS/S 91-014, College of Business and Management, University of Maryland, College Park, Maryland 20742.

Kennington, J.L., and Helgason, R.V. (1980), *Algorithms for Network Programming*, New York: John Wiley and Sons, Inc.

Rowe, E. (1991), "Some Considerations in the Use of Linear Networks to Suppress Tabular Data," *American Statistical Association, 1991 Proceedings of the Section on Survey Research Methods.*

Sullivan, C.M., and Zayatz, L. (1991), "A Network Flow Disclosure Avoidance system Applied to the Census of Agriculture," *American Statistical Association, 1991 Proceedings of the Section on Survey Research Methods.*

Syslo, M.M., Deo, N., and Kowalik, J.S. (1983), *Discrete Optimization Algorithms with PASCAL Programs*, Englewood Cliffs, New Jersey: Prentice-Hall, Inc.

University of Texas at Austin, "MCF PROGRAM DOCUMENTATION: Minimum Cost Flow Optimization Software," Center for Business Decision Analysis, College and Graduate School of Business Administration, Austin, Texas.

Zayatz, L.V. (1992), "Linear Programming Methodology Used for Disclosure Avoidance Purposes at the Census Bureau," to appear *American Statistical Association, 1992 Proceedings of the Section on Survey Research Methods.*

*This paper reports the general results of research undertaken by the Census Bureau staff. The views expressed are attributable to the authors and do not necessarily reflect those of the Census Bureau.