



reach a solution with proven optimality on 2-dimensional tables with up to 500 rows and 500 columns. The problem is solved in a few minutes on a standard PC. Fischetti and Salazar-Gonzales (2000) extended their work to other tabular data including k-dimensional table with  $k > 2$ , hierarchical tables, linked tables etc., using branch-and-cut based procedures. Alternatively, instead of completely suppressing table cells, Salazar (2001); Fischetti and Salazar (2003) proposed a “partial cell suppression” method that will publish a subset of table cells with variable estimation intervals. Though FIPP and CSP shares the same MILP model, unfortunately, so far we think all of the above mentioned secondary cell selection methods do not apply directly to selecting protecting cells (PCs) that are to be published in FIs, neither optimally nor heuristically. The reason is that these models can not accommodate the knowledge of the FI bounds.

### 3. Selection Improvement Algorithm

In previous research, Cohen and Li (2005) have proposed an iterative “selection-improvement” algorithm, which improves cell selection upon each previous step until all primary cells are sufficiently protected. The iterative selection-improvement algorithm has two stages at each iteration, (1) selecting primary and secondary PCs and replacing them with FIs; and (2) conducting an audit on the publication table with the newly selected PCs in FIs. If the audit finds any primary cell is still at risk, the algorithm re-iterates by selecting more PCs and conducting another audit until all primary cells are protected. The initial set of PCs is the set of cells selected through one of the CSP methods. In case the iterations fail at the end, i.e. no candidate PCs available for selection while there are still unprotected cells, the method defaults back to the usual CSP solutions targeting only the remaining exposed cells. The steps involved in the selection improvement algorithm follow:

- Step 1. Identify primary and secondary cells in a table via a CSP method and publish them in pre-defined FIs.
- Step 2. Apply linear constrained optimization to identify those primary cells with disclosure risks.
- Step 3. For those primary cells at risk, select additional cells that have not been selected previously from the publication table and publish them in FIs. Three specific methods are proposed for this research and will be briefly described in following paragraph and sections. This is the ‘selection step’.
- Step 4. Apply linear constrained optimization again to check if any primary cell in the original table is still at risk. If yes, return to step 3; otherwise EXIT the algorithm,

the table is successfully protected. This is the “audit step”.

- Step 5. If the step 2 – 4 iteration fails to protect every primary cells, i.e. no further unsuppressed cells available for selection while there are still disclosed primary cells, use any solution method to CSP, i.e. completely suppress these exposure primary and corresponding secondary cells.

There are several alternative methods can be used to select additional PCs in Step 3. We can randomly select cells that are within the same row or column of the exposed primary cells, or we can select through more complex MILP models and mathematical programming techniques. We would like to minimize either the number of cells to be selected or the total value of the selected cells. We studied the following three methods in the selection step: the Systematic, Single-Source Shortest Path (SSSP) and the Random Selection methods in Cohen and Li (2005):

1. Systematic Method. To minimize values published in fixed intervals, this method selects the smallest cell among all cells that form additive relationship with two selected exposure cells that need further protection that has not been suppressed during the previous iteration(s). This cell is published as a pre-defined FI. Default to Random Selection Method (see 3 next) at the end if this method fails.
2. Single-Source Shortest Path (SSPS) Method. This method models the table as a network similar to Travelling Salesman’s Problem (TSP), treat all primary exposure cells on a table as destinations of a travelling map. The method aims to find the shortest path through these destinations, to minimize the total cell values expressed in FIs. To make this TSP solvable for all tables, the method fixes the order of the destinations or vertices on the table network. The method only needs to find the shortest path connecting the order-fixed set of vertices to form a closed “loop” with minimized path. Publish all cells that are not already selected in previous iterations on the chosen loop in FIs. Default to Random Selection Method if this method fails at the end.
3. Random Selection Method. This method randomly selects a cell among all cells that form additive relationship with the primary exposure cells. The candidate cells are cells that are either in the same row or column as the primary cell. If all cells forming additive

relationships are already selected during previous iteration(s), or it by itself is the only decent from the higher hierarchy, go one hierarchy step higher until additional protecting cells can be found through additive relationships. Randomly select protecting cells among the candidates, publish these and all cells along the hierarchical searching path as FIs.

Table 1a displays a portion of a current publication table currently a user sees in BLS publications. In this table the cells marked with “x” are suppressed cells due to primary and secondary suppressions. Table 1b shows the results of the selection-improvement algorithm.

#### 4. Noise Model

For estimation based upon a noise model, Evans, Zayatz and Slanta (1998) proposed a disclosure limitation technique for establishment magnitude tabular data based upon a noise model. The noise model distorts every data element by some minimum amount. For a given report all data are always disturbed in the same direction (increased or decreased.)

Specifically, each record is perturbed by introducing p % noise onto each establishment’s values. To perturb an establishment’s data by p % we multiply its data by a number that is close to 1.0 +/- p %. The exact multiplier is chosen from a distribution centred at 1.0 +/- p %. Both the value of p and the distribution used should be considered agency confidential. The same identical distribution must be used at both ends. Multipliers should be assigned randomly, both whether elements are going to be inflated or deflated and the exact multiplier value chosen from the distributions centred at 1.0 +/- p %. Under this model it can be shown that the estimates are unbiased.

An example from the Evan/Zayatz/Slanta paper shows how the model would work:

Company	Establishment	Direction	Multiplier
Company A		1.1	
	A1		1.12
	A2		1.09
	A3		1.1
Company B		0.9	
	B1		0.89
	B2		0.93
Company C		1.1	
	C1		1.08

It should be noted that it is BLS practice to evaluate disclosure limitation at the company level. Thus, all establishments in a sensitive cell from a single company must all be perturbed in the same direction. Note that for certainty units the weight becomes the multiplier and for non-self representing units the weight is [multiplier + (weight – 1)].

#### 5. Measures of Data Utility

The tradeoffs between data utility and disclosure have been extensively studied by Duncan et Al (1999). Basically, minimizing disclosure risk increases data loss. Data utility can be defined as a measure of the value of information to a legitimate data user. In this section we will propose to measure the amount of data loss due to confidentiality protection using FI or the noise model. Information in tabular data is clearly lost with cell suppression, FI and estimates produced under the noise model. Some data recovery is possible by applying linear program techniques to get bounds on suppressed cells or tighter bounds on FI given the data released. Data loss of FIs as a disclosure avoidance protection will be compared to cell suppression which we assume to be the gold standard for data protection.

The aim is to measure the amount of information loss that is to be accepted by the data user accessing the published tables compared to the actual estimates not released.

A sophisticated data user knows that bounds can be placed on suppressed cells via linear programming using the information released. Similarly, tighter bounds on FIs than published by a statistical agency can be determined using the data released. For each cell protected by cell suppression or FI we will determine the minimum and maximum value possible for a cell given the table structure produced. The information loss statistic will be computed. Various aggregations will be

NAICS code	Counties of a U.S. State								
	Total	County 1	County 2	County 3	County 4	County 5	County 6	County 7	etc.
...	...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...
451	13940	113	1758	2691	111	X	241	64	
4511	9070	82	1121	1699	x	X	166	x	
45111	4187	26	703	773	89	-	51	51	
451110	4187	26	703	773	89	-	51	51	
45112	2648	x	274	451	x	X	x	-	
451120	2648	x	274	451	x	X	x	-	
45113	1237	x	110	302	-	X	x	x	
451130	1237	x	110	302	-	X	x	x	
45114	998	x	35	173	-	-	38	x	
451140	998	x	35	173	-	-	38	x	
4512	4870	31	637	992	x	-	75	x	
45121	3415	x	504	444	x	-	x	x	
451211	3193	x	x	438	x	-	x	x	
451212	222	x	x	6	-	-	-	x	
45122	1455	x	133	548	x	-	x	-	
451220	1455	x	133	548	x	-	x	-	
...	...	...	...	...	...	...	...	...	
...	...	...	...	...	...	...	...	...	
Total	1166388	15589	98129	190226	7524	5018	22485	12171	etc.

"x" are nondisclosable data due to primary and secondary suppressions

**Table 1a.** A sample evaluation data set as published perturbed for confidentiality

NAICS code	Counties of a U.S. State								
	Total	County 1	County 2	County 3	County 4	County 5	County 6	County 7	etc.
...	...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...
451	13940	113	1758	2691	111	0-19	241	64	
4511	9070	82	1121	1699	20-99	0-19	166	20-99	
45111	4187	26	703	773	89	-	51	20-99	
451110	4187	26	703	773	89	-	51	20-99	
45112	2648	0-19	274	250-499	0-19	0-19	0-19	-	
451120	2648	0-19	274	250-499	0-19	0-19	0-19	-	
45113	1237	0-19	110	302	-	0-19	20-99	0-19	
451130	1237	0-19	110	302	-	0-19	20-99	0-19	
45114	998	20-99	20-99	173	-	-	38	0-19	
451140	998	20-99	20-99	173	-	-	38	0-19	
4512	4870	31	637	992	0-19	-	75	0-19	
45121	3415	20-99	504	444	0-19	-	20-99	0-19	
451211	3193	20-99	250-499	438	0-19	-	20-99	0-19	
451212	222	0-19	20-99	6	-	-	-	0-19	
45122	1455	0-19	133	548	0-19	-	20-99	-	
451220	1455	0-19	133	548	0-19	-	20-99	-	
...	...	...	...	...	...	...	...	...	
...	...	...	...	...	...	...	...	...	
Total	1166388	15589	98129	190226	7524	5018	22485	12171	etc.

**Table 1b.** The same section of the evaluation data set as it is published under FIPP method

tabulated to compare information loss for FI data compared to the gold standard for tabular disclosure protection: cell suppression.

For the noise model, no additional techniques will improve the estimates of the cell values. However, information loss can still be calculated.

Some notation:

$X_m^{ij}$ : mid-point between published upper and lower bound,

$X_m'^{ij}$ : optimized mid-point. Mid-point between  $F_l$  and  $F_u$ , where in the case of fixed interval publication:

$F_l = \max$  (feasibility lower bound, published lower bound), and

$F_u = \min$  (feasibility upper bound, published upper bound);

And in the case of complete suppression:

$F_l$  = feasibility lower bound, and

$F_u$  = feasibility upper bound;

$X_m''^{ij}$ : non-optimized mid-point. Mid-point between  $F_l$  and  $F_u$ , where in the case of fixed interval publication:

$X_n^{ij}$ : Cell value aggregated from using micro data perturbed by the noise model

$X_o^{ij}$ : actual value of cell  $ij$ .

Then, for a two-dimensional table with I rows and J columns, let  $PIL_{ij}$  denotes the percent information loss per cell attributed to the  $ij^{\text{th}}$  publication cell in the table, defined as:

A) Minimal information loss (%) obtained by a sophisticated data user by applying linear programming techniques to obtain the tightest bounds on cells altered for disclosure reasons

a. Complete suppression

$$PIL_{ij} = \begin{cases} |X_m'^{ij} - X_o^{ij}| / X_o^{ij}, \\ \text{where } i, j \in \text{set of CS cells} \\ 0, \text{ otherwise} \end{cases}$$

b. Fixed interval suppression

$$PIL_{ij} = \begin{cases} |X_m^{ij} - X_o^{ij}| / X_o^{ij}, \\ \text{where } i, j \in \text{set of FI cells} \\ 0, \text{ otherwise} \end{cases}$$

B) User information loss (%) experienced by a user without significant technical expertise.

a. Complete suppression

$$PIL_{ij} = \begin{cases} 1, \text{ for } i, j \in \text{set of CS cells} \\ 0, \text{ otherwise} \end{cases}$$

b. Fixed interval suppression

$$PIL_{ij} = \begin{cases} |X_m''^{ij} - X_o^{ij}| / X_o^{ij}, \\ \text{where } i, j \in \text{set of FI cells} \\ 0, \text{ otherwise} \end{cases}$$

C) Information loss for the noise model:

$$PIL_{ij} = |X_n^{ij} - X_o^{ij}| / X_o^{ij}, \text{ for all } i, j$$

For the complete suppression case user information loss, we assume the unsophisticated data user can not estimate any cell value. However, with algebraic manipulation, non-unique cell values can be estimated in a non-optimal way even by an unsophisticated data user.

## 6. Analysis

BLS QCEW data used for this study was for the State of Maryland 1<sup>st</sup> quarter 2004 in manufacturing, retail trade, transportation and warehousing and selected services sectors (Table 2).

The initial analysis of information loss was over all cells and protected cells. The  $PIL_{ij}$  s are averaged over all cells of desired study types to obtain the average percent information loss over these types of cells (Table 3a). Similarly  $PIL_{ij}$  s are averaged over only protected publications cells to obtain average percent information loss for protected publication cells (Table 3b). These two tables show different aspects of the complete suppression and interval publication methods. Analyses of protected cells are of most interest as this is where we get information loss for FI. Fixed intervals provide only a slight improvement in information loss

while the increase in data utility for estimates produced under the noise model is significant.

Define absolute information loss at cell  $(i, j)$  for FI as:

$$IL_{ij} = \begin{cases} |X_m^{ij} - X_o^{ij}|, \\ \text{where } i, j \in \text{set of CS cells} \\ 0, \text{ otherwise} \end{cases}$$

Define absolute information loss at cell  $(i, j)$  under the noise model as:

$$IL_{ij} = |X_n^{ij} - X_o^{ij}|, \text{ for all } i, j$$

Note that the cell  $(i, j)$  information loss  $IL_{ij}$  is defined similarly to  $PIL_{ij}$ , except it is not normalized to the actual cell value. Total information loss in level of employment is calculated by adding absolute information loss due to each protected cell. The  $IL_{ij}$ s are then aggregated over respective types of cells to obtain Table 4.

To calculate total information loss at hierarchical levels for FI, let:

$$IL'_{ij} = \begin{cases} X_m^{ij} - X_o^{ij}, \\ \text{where } i, j \in \text{set of CS cells} \\ 0, \text{ otherwise} \end{cases}$$

To calculate total information loss at hierarchical levels under the noise model, let:

Notice that absolute values are not taken for  $IL'_{ij}$  in this case. To calculate the percent information loss at one NAICS digit level, first calculated the total information loss at each cell. Then, aggregate the relative value over the total employment. In table 5 we examine the total information loss in level of employment at each NAICS hierarchical levels the effect of aggregation on information loss.

It should be noted that the results in tables 3 through 5 for minimal information loss for FI are conservative since each LP is solved independently.

## 7. Conclusions

We used both relative information loss and absolute IL in the analysis. The relative IL, which is the percentage of IL in a publication relative to the actual cell value,

provides the value delivered to data user that lost regardless of the magnitude of the cell value. On the contrary, the absolute IL provides the user the quantity that is removed from the final publication table due to confidentiality protection reasons. Relative IL are more meaningful to users who treats small and large segments equally, for example, a city planner who observes employment growth of all industry sectors; absolute IL may be favoured on the other hand by users want to know the levels, for example, an observer of nation employment growth.

Overall, regardless of confidentiality protection method used, there is slightly less information loss under the minimal information loss scenario than under the user information scenario for FI. Estimates produced under the noise model provide more data utility than FI over complete suppression. Similarly, as previous research has indicated, regardless of the user scenario, complete suppression method has less data utility than either the fixed interval publication method or estimates produced under the noise model.

When IL is averaged over all cells, marginal cells tend to lose less information than the interior cells. This is caused by the fact that there are more interior cells being protected in percentage than the marginal cells. This is not true if averaged over only protected cells. Marginal cells lose more in average among protected cells. However, by looking at Table 5, where IL are listed at each aggregation levels, as aggregation level goes up, less IL happens. This indicates most IL is actually occurring on the lower aggregates.

The level and relative information loss are about the same for CS method between primary and secondary cells. But FI method tends to lose less than the second and primary cell suppression combined.

In summary, sophisticated users, including purposeful attackers will drive out more information from protected tables, under either protection method; FI method will provide more information to data users than the CS method; and at lower aggregates and interior cell level, this is the location where most information loss actually occurred.

**Table 2: Industrial distribution of actual test data sample**

2-digit NAICS	Industry	Total Number of Establishments	Total Employment
31-32	Manufacturing	1,008	54,244
44-54	Retail Trade	4,356	332,011
48-49	Transportation and Warehousing	864	65,224
51	Information	451	37,650
52	Finance and Insurance	1,764	137,511
53	Real Estate and Rental and Leasing	1,078	80,104
54	Professional, Scientific and Technical Services	3,871	289,659
62	Health Care and Social Assistance	3,151	169,985
	<b>Total</b>	<b>16,527</b>	<b>1,166,388</b>

Ref: QCEW State of Maryland 1<sup>st</sup> quarter 2004.

**Table 3a: Average percent information loss per cell among different cell types**

Cell type	Minimal information loss (%)			User information loss (%)		
	Complete Suppression	Fixed interval publication	Random noise	Complete Suppression	Fixed interval publication	Random noise
Primary suppression	89.5	35.2	10.9 <sup>3</sup>	100	52.5	10.9 <sup>3</sup>
Secondary suppression	78.5			100		
Interior cells <sup>1</sup>	68.3	9.07	3.61	72.2	20.5	3.61
Marginal cells <sup>2</sup>	29.6	7.55	0.19	31.8	10.9	0.19
All cells	51.2	8.18	0.1	55.6	17.5	0.1

1: Cells that are not sums of other cells

2: Cells that are sums of at least one other cell.

3: Only primary cells are counted for random noise method

**Table 3b: Average percent information loss per cell among *protected* publication cells**

Cell type	Minimal information loss (%)			User information loss (%)		
	Complete Suppression	Fixed interval publication	Random noise	Complete Suppression	Fixed interval publication	Random noise
Primary suppression	92.5	75.2	10.9 <sup>3</sup>	100	81.2	10.9 <sup>3</sup>
Secondary suppression	89.2			100		
Interior cells <sup>1</sup>	88.4	74.8	6.93	100	80.1	6.93
Marginal cells <sup>2</sup>	94.6	75.9	1.55	100	82.2	1.55
All cells	92.5	75.2	0.68	100	81.2	0.68

1: Cells that are not sums of other cells

2: Cells that are sums of at least one other cell.

3: Only primary cells are counted for random noise method

**Table 4: Total information loss in level of employment in different types of publication cells**

Cell type	Minimal information loss			User information loss		
	Complete Suppression	Fixed interval publication	Random noise	Complete Suppression	Fixed interval publication	Random noise
Primary suppression	60,535	86621	NA	71,317	95355	NA
Secondary suppression	41,204			55,962		
Interior cells	79,894	66729	10284	92,335	74547	10284
Marginal cells	21,845	19892	2734	34,944	20808	2734
All cells	101,739	86621	13365	127,279	95355	13365

**Table 5: Total information loss in level of employment at each NAICS hierarchical levels**

NAICS hierarchical levels	Minimal information loss (%)			User information loss (%)		
	Complete Suppression	Fixed interval publication	Random noise	Complete Suppression	Fixed interval publication	Random noise
Six-digit	7.85	6.99	4.57	8.07	7.11	4.57
Five-digit	3.95	3.05	0.022	3.21	3.29	0.022
Four-digit	0.199	0.182	0.014	0.212	0.195	0.014
Three-digit	0.015	0.001	0.011	0.021	0.001	0.011
Two-digit	0	0	0.0059	0	0	0.0059

## 8. References

- Castro, J. (2001). "Using Modeling Languages for the Complementary Suppression Problem Through Network Flow Models." Joint ECE/Eurostat Work Session on Statistical Data Confidentiality.
- Castro, J. and N. Nabona (1996). "An Implementation of Linear and Nonlinear Multi-commodity Network Flows." European Journal of Operational Research 92: 37-53.
- Cohen, Stephen H. and Timothy Li (2005), "Using Fixed Intervals to Protect Sensitive Cells Instead of Cell Suppression", Monographs of Official Statistics, United Nations Economic Commission for Europe, November 2005
- Cox, L. H. (1980). "Suppression Methodology and Statistical Disclosure Control." Journal of the American Statistical Association 75: 377-385.
- Cox, L. H. (1995). "Network Models for Complementary Cell Suppressions." Journal of the American Statistical Association 90: 1453-1462.
- Duncan, George T.; Feinberg, Stephen; Krishnan, R.; Padman, R.; Roehrig, S. (2001); "Disclosure Limitation Methods and Information Loss for Tabular Data"; Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies; eds. P. Doyle, J. Lane, J. Theeuwes and L. Zayatz; North-Holland, 135-166
- Evans, Timothy; Zayatz, Laura, and Slanta, John (1998); "Using Noise for Disclosure Limitation of Establishment Tabular Data"; Journal of Official Statistics, Vol 14, No. 4.
- Fischetti, M. and J. J. Salazar-Gonzales (2000). "Models and Algorithms for Optimizing Cell Suppression Problems in Tabular Data with Linear Constraints." Journal of the American Statistical Association 95: 916-928.
- Fischetti, M. and J. J. Salazar (1999). "Models and Algorithms for the 2-Dimensional Cell Suppression Problem in Statistical Disclosure Control." Mathematical Programming 84: 283-312.
- Fischetti, M. and J. J. Salazar (2003). "Partial Cell Suppression: A New Methodology for Statistical Disclosure Control." Statistics and Computing 13: 13-21.
- Giessing, S. (2001). Nonperturbative Disclosure Control Methods for Tabular Data. Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies. Doyle, Lane, Theeuwes and Zayatz, North-Holland.
- Kelly, J. P. (1990). Confidentiality Protection in Two and Three-Dimensional Tables. College Park, Maryland, University of Maryland, College Park, Maryland. Ph.D. Thesis.
- Repsilber, R. D. (1994). Preservation of Confidentiality in Aggregated Data. Second International Seminar on Statistical Confidentiality. Luxembourg.
- Salazar, J. J. (2001). "Improving Cell Suppression in Statistical Disclosure Control." Joint ECE/Eurostat Work Session on Statistical Data Confidentiality.

Note that the views expressed in this paper are those of the authors and do not necessarily represent the policy of the Bureau of Labor Statistics.