A NETWORK FLOW DISCLOSURE AVOIDANCE SYSTEM APPLIED TO THE CENSUS OF AGRICULTURE

Colleen M. Sullivan and Laura Zayatz*
U.S. Bureau of the Census, Washington, D.C. 20233

KEY WORDS: Disclosure Avoidance, Cell Suppression, Tabular Data, Network Flow System, Linked Networks

1. INTRODUCTION

The U.S. Bureau of the Census, Agriculture Division, has the responsibility to collect data regarding the agricultural sector and to publish this data without violating confidentiality laws. Collected data contains sensitive data values, commonly referred to as primary suppressions, that if directly published could identify an individual or farm operation. There are a number of methods available which prevent compromising the primary suppressions. These disclosure avoidance techniques include rounding, perturbation, and cell suppression, and are outlined in the article by Cox, et al. (1986a).

Since rounding and perturbation are unsatisfactory for aggregate magnitude data (Cox, et al, 1986b), the Economic and Agriculture 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 still 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 in applying complementary suppressions is to ensure the protection of the sensitive data value at minimum cost. Note that this requires assigning a cost of suppression to each data cell. Commonly, the original data value that would have appeared in the publication is assigned as the cost. 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.

In recent years, the Bureau has conducted research on a cell suppression technique which 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. 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).

Prior to the 1978 Census of Agriculture, analysts in the division performed cell suppression by hand using a technique occasionally referred to as the "nearest-smallest method". For an outline of this method see Zayatz, et al (forthcoming). The cell suppression procedure was first automated for the 1978 Census of Agriculture by programming a portion of the hand procedure. However, a major portion of the complementary procedure was still performed manually. Minor revisions were made to the existing automated cell suppression procedure for the 1982 Census of Agriculture and the remainder of the hand procedure was automated. This was the first time the entire disclosure avoidance procedure was automated. After the 1982 Census of Agriculture, the disclosure procedure was reviewed and recommendations for improvements were made. These improvements were implemented for the 1987 Census of Agriculture. However, since the automated cell suppression procedure was not based upon any statistical or mathematical methodology, it was not always reliable. Frequently, oversuppression occurred which decreased the amount of usable data published. Also, undersuppression occurred which required analyst intervention to fully protect all sensitive data values.

For the 1992 Census of Agriculture, research was conducted on the cell suppression technique using the network flow system of applying complementary suppressions. However, the network flow system used by other divisions of the Bureau could only accommodate a single two dimensional table. Almost all agricultural data (as well as most of the data in other economic areas) are contained in a system of two dimensional tables. In addition, although Business Division and Industry Division have strictly hierarchical data structures, Agriculture Division does not. Further contributing to the complexity of agricultural data are systems of three dimensional tables. Because of these problems, the existing network flow system was not optimal for agricultural data, thereby requiring customization.

This paper discusses the formulation of the customized network methodology and the limitations encountered with the customized version when applied to agricultural data. In Section 2 we describe the fundamentals of the network flow system of applying complementary suppressions. A system of two dimensional tables with "appendages" is presented in Section 3. In Section 4 we discuss a heuristic that will link networks to accommodate three dimensional tables with appendages. We present the main limitations in Section 5 and provide concluding remarks in Section 6.

2. NETWORK FLOW METHODOLOGY

Network flow methodology is a convenient way to choose the group of complementary suppressions that protects the sensitive data value at minimum cost.

2.1 A Network Diagram

A key idea is to transform a two-dimensional table into a network flow diagram. The two dimensional tables are fundamental to the network flow system.

Consider the following two dimensional table:

A_{11} A_{12} A_{13} | A_{10}
A_{21} A_{22} A_{23} | A_{20}
A_{31} A_{32} A_{33} | A_{30}

A_{01} A_{02} A_{03} | A_{00}

Table 1

The network diagram consists of a set of points and arrows where the points are connected by the arrows. The points are referred to as transshipment nodes and the arrows are referred to as arcs. The arcs transport (or direct) data or units from one transshipment node to another. The transportation of units through an arc constitutes a flow through that arc. Since arcs connect various points, a closed path of arcs can be formed. This closed path is referred to as a cycle.

Figure 1 is the network diagram associated with Table 1.

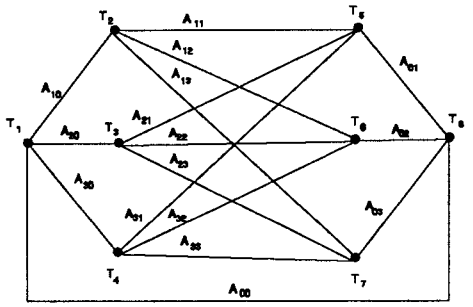

Figure 1

The transshipment nodes are labeled T_1 through T_8 in Figure 1. Each transshipment node symbolizes a relationship from the associated table. For example, arc A_{10} is entering transshipment node T_2 and arcs A_{11}, A_{12}, and A_{13} are exiting. This symbolizes the relationship $A_{10} = A_{11} + A_{12} + A_{13}$ which is the first row of Table 1.

Each cell from the table has two associated arcs in the network. For clarity, we have only drawn one line for each cell in this figure. However, each line in the figure represents two arcs in the actual network, one going from left to right and the other going from right to left. The arcs labeled A_{10}, A_{20}, and A_{30} represent the row totals, the arcs labeled A_{11}, A_{12}, A_{13}, A_{21}, A_{22}, A_{23}, A_{31}, A_{32}, and A_{33} represent the interior cell values, the arcs labeled A_{01}, A_{02}, and A_{03} represent the column totals, and the arc labeled A_{00} represents the overall table total. Hence each cell from Table 1 is accounted for in the network.

Each arc in the network has an associated capacity and cost. The capacity of an arc is the maximum number of units that can flow through that arc. The cost of an arc is the cost of flowing one unit through that arc.

For agricultural data, the capacity of an arc is assigned to be the corresponding farm count and the cost of flowing one unit through an arc is assigned to be the corresponding data value. The data values are obtained from Table 1 and each data value has an associated farm count (not shown in a table). (Note that only data values are suppressed while the farm counts associated with them are published.) For instance, A_{ij} represents a data value and has an associated farm count, say m_{ij}. Then we can flow a maximum of m_{ij} units across the arc corresponding to the data value A_{ij}, resulting in a maximum cost of $m_{ij}A_{ij}$. When a network flow system is actually implemented to apply complementary suppressions, the costs and capacities differ for sensitive data values and previously applied complementary suppressions. This is discussed further in Section 2.2 and Section 2.3.

2.2 Applying Complementary Suppressions

The network flow system is used to protect the sensitive cell by choosing other cells contained in the table for suppression. Finding a suppression pattern to protect a sensitive cell, S, in the table corresponds to finding a cycle in the network which contains one of the arcs representing the sensitive cell, S. It is convenient to think of this as sending one unit around a cycle of arcs in the network. All other cells represented by arcs in the chosen cycle would then be suppressed as complements. Our objective is to choose the cycle through the network which suppresses the least amount of data value while protecting the sensitive cell; that is, we find the minimum cost flow.

The network flow system is implemented using the Minimum Cost Flow (MCF) program discussed in Glover and Klingman (1982). The MCF Program, as the name implies, finds the minimum cost cycle available in the network provided by the user. The user supplies a cost and capacity for each arc following the convention outlined in Section 2.1. However, the cost of one of the arcs corresponding to the sensitive data value we wish to protect is assigned a very negative integer (e.g., -10^9). (The other arc representing the sensitive cell is temporarily given a capacity of zero, thereby essentially removing it from the network.) By assigning the cost as a sufficiently negative value, the flow is forced through the sensitive cell's arc. (Recall the nonsensitive arcs are assigned positive costs.) In addition, the capacity of the arc corresponding to the sensitive data value we wish to protect is changed to one. This allows only one unit to flow through the suppression cycle.

After these assignments are made, all possible cycles through the network are examined, and the cycle with the minimum cost flow is chosen.

To illustrate how complementary suppressions are chosen to protect a single sensitive cell in a table, consider Table 1, with A_{32} now representing a sensitive cell and referred to as S. Suppose the minimum cost cycle chosen to protect the sensitive arc of A_{32} in Figure 1 includes arcs A_{12}, A_{13} and A_{33}. This cycle is shown in Figure 2 with bold arrows indicating the direction of flow. Each arc in this cycle corresponds to a complementary suppression in Table 1.

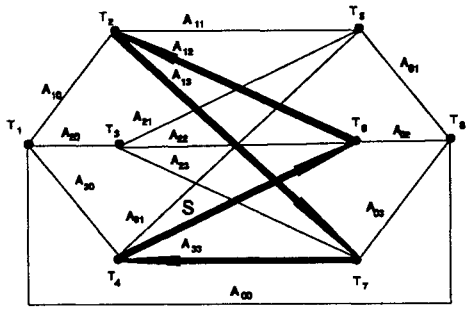

Figure 2

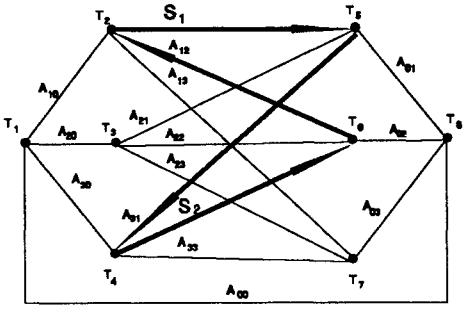

Figure 3

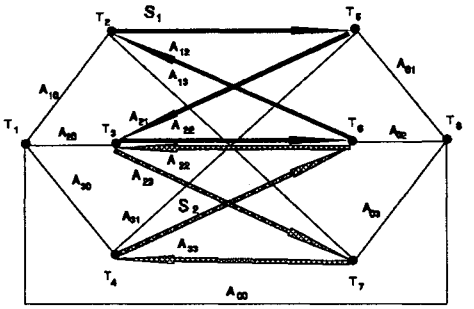

Figure 4

Often more than one sensitive cell exists in a table. When this occurs, a cycle must be chosen for each one. However, network-flow methodology can examine only one sensitive cell at a time; that is, it cannot process all simultaneously. To utilize network flow methodology with multiple sensitive cells, the sensitive cells are sorted into descending order based on their true cell value. The MCF program begins by assigning an extremely negative integer cost (e.g., -10^9) to one of the arcs corresponding to the largest sensitive cell. At this point, all other arcs representing sensitive cells are given a cost of zero. Then the system finds the cycle with minimum cost. By assigning a sufficiently negative integer as the cost of the sensitive arc being processed, we ensure that the minimum cost flow (cycle) chosen will contain that arc. Next, the arc representing the second largest sensitive cell in question is assigned a very negative integer as the cost and all other arcs corresponding to sensitive cells, including the first one, and all arcs representing previously chosen complementary suppressions are given a cost of zero. Then, a minimum cost flow cycle is chosen containing the second sensitive cell. All other sensitive cells are evaluated in this manner until a cycle has been chosen containing each sensitive cell. It is important to note that a cycle containing one sensitive cell may contain other sensitive cells or it may not. This solely depends on the cost of the non-sensitive data values.

As an illustration, consider Table 1 when both A_{32} and A_{11} are sensitive cells. Assume the value represented by A_{11} is greater than the value represented by A_{32}. Since A_{11} will be processed first, denote it as the first sensitive cell, S_1, and denote A_{32} as the second sensitive cell, S_2.

Then, depending on the non-sensitive data values, we can have a cycle, such as the one shown in Figure 3, that contains both S1 and S2. We could also have an alternative cycle, such as the one shown in Figure 4, that contains a separate cycle for each sensitive cell.

3. A SYSTEM OF TABLES WITH APPENDAGES

The basic network-flow methodology presented in Section 2, and used by economic divisions of the Bureau for their 1987 censuses, can only accommodate a single two dimensional table. Almost all agricultural data (as well as most of the data in other economic areas) are contained in a system of two dimensional tables. Often an interior column (or row) is a total column (or total row) in another table. This results from the level of detail provided to the public. Suppose that column one of Table 1 is broken down such that the original table along with a table containing the added relationships is represented with Tables 2a and 2b. Notice the third column (column of row totals) of Table 2b is the same as the first interior column of Table 2a.

$A_{11}\ A_{12}\ A_{13} \mid A_{10}$	$\tilde{A}_{11}\ \tilde{A}_{12} \mid A_{11}$
$A_{21}\ A_{22}\ A_{23} \mid A_{20}$	$\tilde{A}_{21}\ \tilde{A}_{22} \mid A_{21}$
$A_{31}\ A_{32}\ A_{33} \mid A_{30}$	$\tilde{A}_{31}\ \tilde{A}_{32} \mid A_{31}$
$A_{01}\ A_{02}\ A_{03} \mid A_{00}$	$\tilde{A}_{01}\ \tilde{A}_{02} \mid A_{01}$
Table 2a	Table 2b

We refer to Table 2a as a "root" table and Table 2b as an "appendage" table. The root table is a table that does not have its column of row totals (far right-hand column) appearing as an interior column of another table or as a column of row totals anywhere else in the network. An appendage table is a table which contains an interior

column of either the root table or a previous appendage table as its column of row totals. It is important to note that each column of a table need not be broken down into appendage tables. Also, a single network can contain as many appendage tables as desired. However, certain restrictions do exist: (1) a column of row totals must be unique to a single table in the network; that is, each column of data can appear as a column of row totals only once in any single network, and (2) a row and column which intersect cannot both be broken down in the same network. If either of these two situations exist, a single network cannot be created to accommodate all relationships between table cells.

For the 1987 censuses, complementary suppressions were applied to one table at a time even if they were contained within a system of tables. But if the system is hierarchical we can translate it into a single network flow diagram. This method, designed for Agriculture Division, lets us process all tables of a system as if they were one.

3.1 Creating the Network

Again, each cell in Tables 2a and 2b has two associated arcs in the network even if a cell appears in both the root table and the appendage table. The network associated with Tables 2a and 2b is presented in Figure 5.

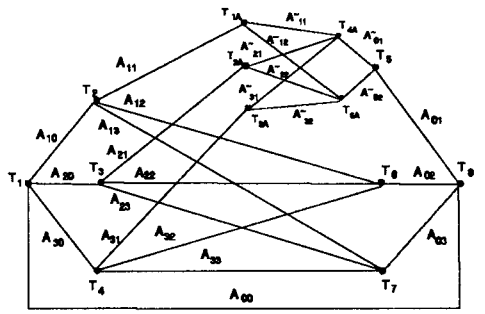

Figure 5

The arcs labeled A_{10}, A_{20}, and A_{30} represent the row totals of Table 2a. The arcs labeled A_{11}, A_{21} and A_{31} represent interior cells of Table 2a and row totals of Table 2b, while the arcs labeled A_{12}, A_{13}, A_{22}, A_{23}, A_{32} and A_{33} represent interior cell values of Table 2a only. The arcs labeled, $\tilde{A}_{11}$, $\tilde{A}_{12}$, $\tilde{A}_{21}$, $\tilde{A}_{22}$, $\tilde{A}_{31}$, and $\tilde{A}_{32}$ represent the interior values of Table 2b and the arcs labeled $\tilde{A}_{01}$ and $\tilde{A}_{02}$ represent the column totals of Table 2b. The arc labeled A_{01} represents a column total of Table 2a and the overall table total of Table 2b, while the arcs labeled A_{02} and A_{03} represent column totals of Table 2a only. The arc labeled A_{00} represents the overall table total of Table 2a.

Besides the transshipment nodes associated with the simple table (represented in Figures 1, 2 and 3), there are transshipment nodes associated with Table 2b. These are labeled T_{1A}, T_{2A}, T_{3A}, T_{4A}, and T_{5A} in Figure 5. Each new transshipment node symbolizes a relationship from Table 2b. For example, arc A_{11} is entering transshipment node T_{1A} and arcs $\tilde{A}_{11}$ and $\tilde{A}_{12}$ are exiting. This indicates $A_{11} = \tilde{A}_{11} + \tilde{A}_{12}$ as shown in the first row of Table 2b.

The difference between a network associated with a basic table and one associated with a table having appendages can be seen by comparing Figure 1 and Figure 5. Arcs A_{11}, A_{21}, and A_{31} in Figure 5 no longer directly enter T_5 as they did in Figure 1. Instead, they flow to T_{1A}, T_{2A}, and T_{3A}, respectively, are split into more detailed arcs, and eventually, enter T_5.

3.2 Applying Complementary Suppressions

To illustrate how network-flow methodology is used to protect sensitive data (an initial suppression) contained in a system of two dimensional tables with an appendage, consider Tables 2a and 2b, with A_{31} now representing a sensitive cell.

As in the case without appendages, our objective is to find a cycle through the network that suppresses the least amount of data value while protecting the sensitive cell; that is, we find the minimum cost flow.

At this point we can protect the sensitive data value as outlined in Section 2. Figure 6 shows such a cycle that protects A_{31}.

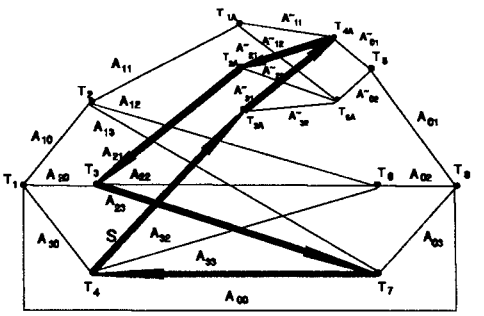

Figure 6

4. THREE DIMENSIONAL TABLES WITH APPENDAGES

Agricultural data consists of three dimensional tables with appendages. The first two dimensions (rows and columns) account for geography and detail of a given data item. The third dimension accounts for three sales categories of agricultural data: ALL FARMS, FARMS WITH SALES OF $10,000 OR MORE, and FARMS WITH SALES LESS THAN $10,000. These sales categories form the following relationship:

All Farms = Farms with Sales Less Than $10,000
+ Farms with Sales of $10,000 or more

However, typical network-flow methodology cannot be used to apply complementary suppressions to three dimensional tables because a single network cannot be

created to accommodate all of the relationships among the arcs associated with the cells of a three dimensional table. We did, however, design a heuristic to link networks to accommodate three dimensional tables with appendages. To illustrate the formation of this heuristic, begin by considering a simple three dimensional table.

Suppose the three dimensional table is composed of three levels and each level consists of a two dimensional table. Also, suppose level A = level B + level C, where level A is associated with ALL FARMS, level B is associated with FARMS WITH SALES OF $10,000 OR MORE, and level C is associated with FARMS WITH SALES LESS THAN $10,000. A heuristic can be created by viewing each level as a separate two dimensional table, creating a network for each two dimensional table, and linking the networks together.

However, for agricultural data, we can eliminate level C since no data are published for FARMS WITH SALES LESS THAN $10,000. But we must account for the relationship between data for ALL FARMS and its corresponding data for FARMS WITH SALES OF $10,000 OR MORE since both of these sales categories are published. For instance, suppose a value in level A is suppressed, say A_{11}. Also, suppose that A_{11} is equal to its corresponding value in level B, B_{11}. Then we must ensure that B_{11} is suppressed in level B because a data user would be able to see that the farm counts for the two items are the same and thus infer that the two values are the same. This link is accomplished by "carrying over" a suppression to the corresponding value in level B and treating it as a primary suppression. We added routines which helped reduce the number of times we had to carry suppressions between levels and thereby reduced the number of times we had to reexamine the other level.

Although the heuristic described was to link networks to accommodate three dimensional tables, it can easily be extended to one that will link networks to accommodate three dimensional tables with appendages. The heuristic can also be extended to include the third level that was eliminated for agricultural data. We have examined other heuristics but all so far involve viewing the three dimensions as separate two dimensional tables with appendages and linking them together.

5. LIMITATIONS

We consider the current MCF program that applies complementary suppressions to tabular data to be unsatisfactory for agricultural data. Following is a summary of the limitations.

5.1 Non-Hierarchical Data Structures

Although Business Division and Industry Division have strictly hierarchical data structures, Agriculture Division does not. For instance, Business and Industry Division publish data by standard Industrial Classification (SIC) codes. These codes result in a strictly hierarchical structure of the data. Any given SIC is only broken down one way. In addition to hierarchical data structures, agricultural data forms non-hierarchical structures where several different data items sum to the same total.

For example, the data total for land in farms is broken down into owned land in farms plus rented land in farms; land irrigated plus nonirrigated land; in addition to five other relationships which all sum to land in farms.

Each two-dimensional table associated with these seven relationships must be processed separately and then linked together for the total land in farms. We were unable to design a way of using the network system to do this to our satisfaction.

5.2 Multidimensional Tables

Another problem, as outlined in Section 4, is that the three dimensional structure of agricultural data requires reprocessing networks and carrying suppressions between them. We found this causes repetition and results that we were not always willing to accept.

5.3 Multiple Sensitive Data Values

It is quite common to have more than one sensitive data value in a system of tables. However, the network system only has the ability to process one primary suppression at a time. For example, consider the following system of tables with three primary suppressions and the complementary suppressions returned from the MCF program.

95 C	2259	6730 P	23758 I	32842
554	4325	9449	22766 I	37094
1067	11308	16902	25462 I	54739
1716 C	17892	33081 P	71986 I	124675

53 C	42 P I	95 C
306 C	248 C I	554
357	710 I	1067
716 C	1000 I	1716 C

The values 53, 306, 716, and 248 in the second table of the system sum to 1323. If the cell containing the value 1000 was chosen we would not have needed to suppress the cell values 53, 306, 716 or 248. This type of less than optimal result occurred many times in agricultural data. The problem is that the MCF methodology minimizes the cost flow of suppressing the three sensitive values separately rather than looking for a set of flows which protects the sensitive cells jointly at the minimum cost.

In an attempt to improve the results when there were multiple sensitive cells, various cost adjustments were tested which involved adjusting the cost of arcs in the network. However, after a significant amount of testing, none of the cost adjustment procedures we tested improved the results for all cases.

5.4 Unpublished Data Items

Often, the detailed break down of frequencies and categories contained in agricultural data are not published for county data. They are usually only published for the state total. Also, data for FARMS WITH SALES OF $10,000 OR MORE is seldom published, and when it is, it is usually only for the state total and a category total. The detailed break downs are not published. In the past, complementary suppression was performed on all data values, whether or not it appeared in a published table. Performing complementary suppression on all published and unpublished data values can cause a great deal of published data to be suppressed. This situation occurs often when an unpublished value for FARMS WITH SALES OF $10,000 OR MORE must be suppressed and is equal to its corresponding data for ALL FARMS which is published. Then the published data value must be suppressed and protected by suppressing other published data. However, for the 1992 Census of Agriculture, unpublished data will be eliminated from the disclosure analysis procedure. This in turn, eliminates the two dimensional structure which is fundamental to the network-flow methodology because the interior values in the two-dimensional tables are not published. Therefore, network-flow methodology cannot be used to apply complementary suppressions to agricultural data.

6. CONCLUSION

We were able to adapt the network-flow methodology for agricultural data to some extent. However, we feel that in its present form, it is still unsuitable. Often the structure of agricultural data is not strictly hierarchical. Rather several relationships sum to the same total. Each of these relationships must be processed separately and then linked together. This leads to reprocessing networks and undesirable results. Also, the links required for multidimensional tables requires reprocessing networks which again leads to undesirable results. In addition, there often is more than one sensitive data value in a single network. Since the MCF program minimizes the cost of the suppression cycle for each sensitive data value separately, the overall total value suppressed for all sensitive data values is far from minimal. Finally, by eliminating unpublished data from the disclosure analysis procedure, the two dimensional structure, which is fundamental to the network flow system, is eliminated.

REFERENCES

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

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.

Glover, F., and Klingman, D. (1982), "Recent Developments in Computer Implementation Technology for Network Flow Algorithms," INFOR, 20, 433-452.

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

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.

Zayatz, L., Greenberg, B., Sullivan, C., and Ward, J. "Methodology for Applying Complementary Suppressions to Data from The Census of Agriculture," *Statistical Research Division Report Series*, Bureau of the Census, Washington, D.C. (forthcoming)

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