BUREAU OF THE CENSUS
STATISTICAL RESEARCH DIVISION
RESEARCH REPORT SERIES
No. RR-92/10

THE FUNDAMENTAL PRINCIPLES OF A NETWORK FLOW DISCLOSURE AVOIDANCE SYSTEM

by

Colleen M. Sullivan
U.S. Bureau of the Census
Statistical Research Division
Washington, D.C. 20233

Report issued: October 5, 1992

The Fundamental Principles of A Network Flow Disclosure Avoidance System

Colleen M. Sullivan

1. INTRODUCTION

The U.S. Bureau of the Census has the responsibility to collect data regarding economic sectors and to publish this data without violating confidentiality laws. As discussed in Sullivan (1992), collected data often contain sensitive data values, commonly called primary suppressions, that if directly published could identify an individual or establishment's data. The Bureau uses 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, a data user could still derive the sensitive data values from non-sensitive data because most data items are published in additive tables. Therefore, additional data values must be suppressed. We call these additional suppressed data values 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. The Bureau currently uses network flow methodology to choose the set of complementary suppressions to protect the sensitive cells.

This paper discusses the network flow methodology used to apply complementary suppressions to economic tabular data. We begin in Section 2 with a description of the elemental principles of the network flow method. In Section 3 we present a system of two dimensional tables with "appendages", and we provide concluding remarks in Section 4. A reading section is provided for those interested in more detail.

2. NETWORK METHODOLOGY

The goal of a disclosure avoidance system is to choose a group of complementary suppressions that protects the sensitive data values with a minimal reduction in the information provided by the published table. The Economic Disclosure Avoidance System is based on network flow methodology.

We begin by considering the following two-dimensional table:

	State	MSA 1	MSA 2	Non-MSA
SIC Total	a 173,536	e 14,566	f 45,105	g 113,865
SIC 1	b 84,842	h 5,413	i 18,177	j 61,252
SIC 2	c 43,588	k 1,377	l 20,146	m 22,065
SIC 3	d 45,106	n 7,776	o 6,782	p 30,548

Table 1. A Two Dimensional Table

A key step in using network flow methodology is to take a two-dimensional table and transform it into what we call a network flow diagram. The diagram, shown in Figure 1, consists of a series of points connected by lines. The points are called transshipment nodes and the lines are called arcs.

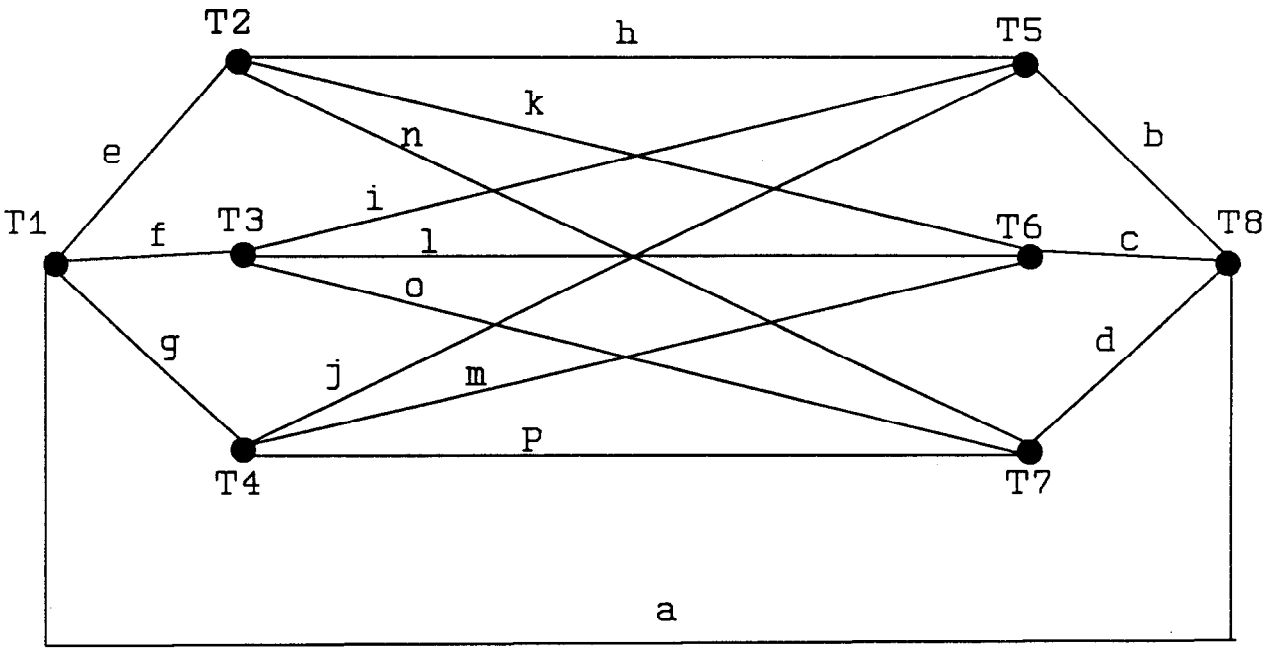


Figure 1. Network Flow Diagram Associated with Table 1

Each arc represents a cell from the table; there are sixteen cells in the table and sixteen arcs in the diagram. In Figure 1, *arc b* represents *cell b*, *arc c* represents *cell c*, and so on. Each transshipment node symbolizes an additive relationship from the table. Figure 1 shows that *arc e* enters the transshipment node labeled T2 and that *arcs h, k* and *n* exit. This represents the fact that *cells h, k* , and *n* sum to *cell e* as shown in the second column of Table 1.

Network Flow Diagram:

- **Arcs represent table cells**
- **Transshipment Nodes represent table relationships**

We use the network flow diagram instead of the table when choosing complementary suppressions for a sensitive data value. Finding a group of cells to protect a sensitive cell in the table corresponds to finding a collection of arcs in the diagram that form a closed path (or paths) and that contain the arc representing the sensitive cell. We call the closed path(s) of arcs a suppression pattern. (Note one closed path is called a cycle and a suppression pattern may contain several cycles.) All cells in the table corresponding to arcs in the suppression pattern are then suppressed.

Consider Table 2 where *cell i* is sensitive and is denoted as (S).

	State	MSA 1	MSA 2	Non-MSA
SIC Total	a 173,536	e 14,566	f 45,105	g 113,865
SIC 1	b 84,842	h 5,413	i (S)	j 61,252
SIC 2	c 43,588	k 1,377	l 20,146	m 22,065
SIC 3	d 45,106	n 7,776	o 6,782	p 30,548

Table 2. A Sensitive Data Value

We must choose other cells in the table to be suppressed for if we only suppressed *cell i*, a data user could subtract *cell l* and *cell o* from *cell f* and find that *cell i* must be 18,177.

The arc that corresponds to *cell i* is denoted as arc S and shown with a bold line in Figure 2.

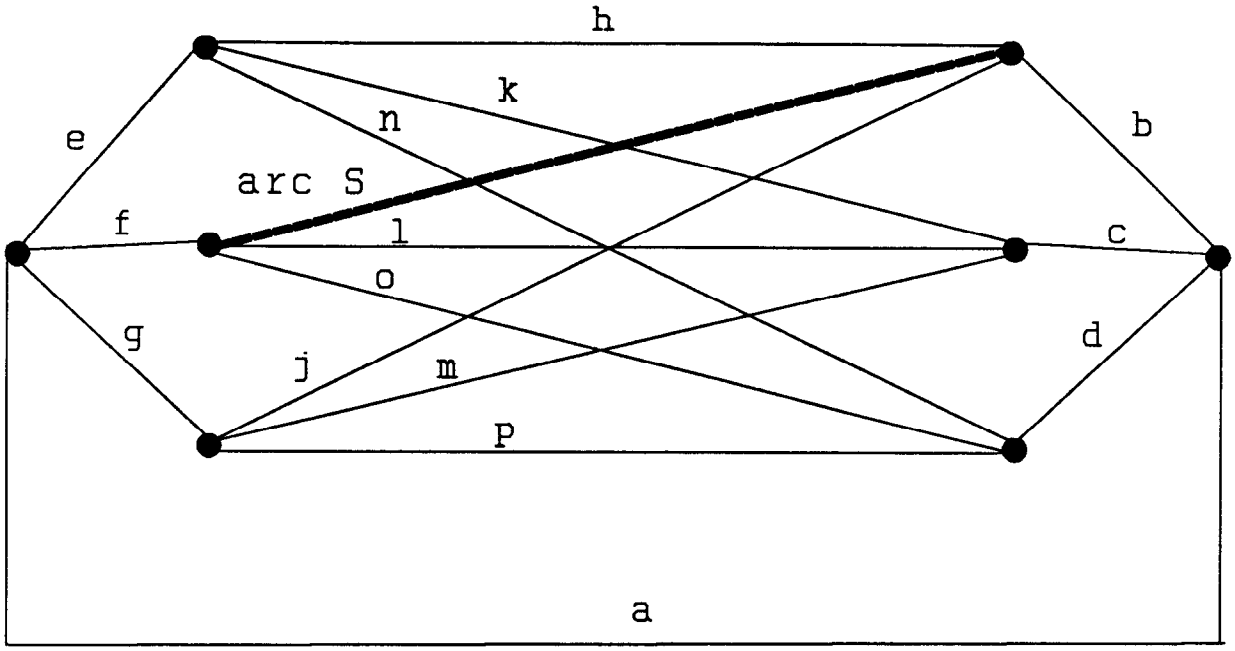


Figure 2. Network Flow Diagram Associated with Table 2

To protect the sensitive cell in Table 2, we want to find a cycle(s) in the diagram that includes arc S. Figure 3 shows a single cycle that includes arc S.

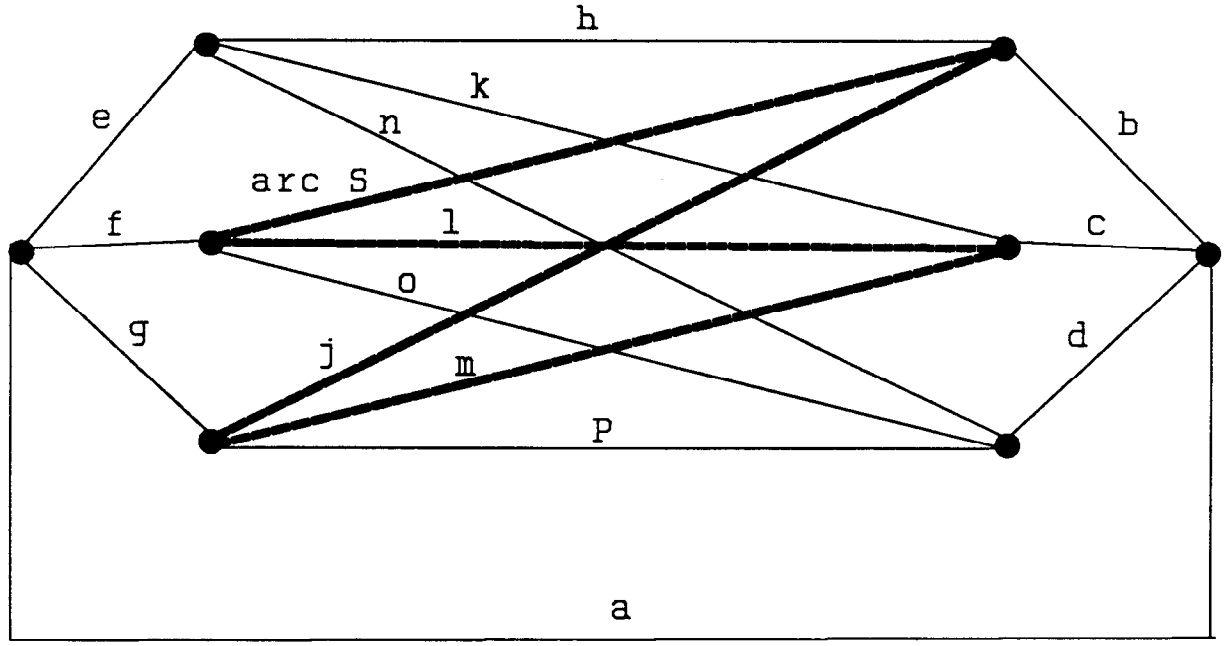


Figure 3. A Suppression Pattern That Contains A Single Cycle

The bold arcs in Figure 3 correspond to the cells denoted as (C) and (S) in Table 3.

	State	MSA 1	MSA 2	Non-MSA
SIC Total	a 173 536	e 14 566	f 45 105	g 113 865
SIC 1	b 84 842	h 5 413	i (S)	j (C)
SIC 2	c 43 588	k 1 377	l (C)	m (C)
SIC 3	d 45 106	n 7 776	o 6 782	p 30 548

Table 3. Suppression Pattern Corresponding to Figure 3

Now that we have suppressed *cells j, l* and *m* as complementary suppressions, a data user cannot find the exact value of *cell i.*

It is convenient to envision the formation of the suppression pattern as a process that sends (flows) a set of units around a cycle (or cycles) of arcs in the network. The number of units required to flow through the cycle(s) is decided by either the n-k or p% primary suppression rule. We will use the p% rule in all further examples.

The p% primary suppression rule follows:

Let T = the total value of a given cell,
L = the value of the largest contributor to the cell,
S = the value of the second largest contributor to the cell, and
p = the percentage of protection required.

Then R = T - L - S is the total value of the remaining contributors to the cell.

The p% rule says that a cell must be suppressed if R<(p/100)L. This implies that the minimum amount of protection needed by the sensitive cell is (p/100)L−R; that is, at least (p/100)L−R units must flow through the cycle(s) to protect the sensitive cell.

We consider the required protection of the sensitive arc to be the minimum amount of protection needed by the primary suppression; that is, the required protection is the number of units that flow through the sensitive arc, and in fact, the network. All other arcs must be assigned a capacity, which is the maximum number of units that can flow through a given arc. (Note the capacity of the sensitive arc is set to the required protection.)

Required Protection:

The minimum amount of protection required by the primary suppression as invoked by either the n-k or p% rule.

For example, suppose the required protection of the sensitive arc (determined by (p/100)L-R) is 2363 units. Also, suppose that all other arcs are given a capacity equal to half their corresponding cell value. Then *arc j* has a capacity of 30,626 units, *arc l* has a capacity of 10,073 units and *arc m* has a capacity of 11,032 units. Therefore, we can flow 2363 units around the cycle shown in Figure 3 and satisfy the required protection of the sensitive arc.

Now suppose the required protection of the sensitive arc is 13,463 units. Again suppose that all other arcs are given a capacity equal to half their corresponding cell values. Then we cannot use the cycle shown in Figure 3 alone to protect *cell i* by 13,463 units since the capacity of *arc l* is less than 13,463, as is the capacity of *arc m*. Therefore, additional arcs must be chosen to protect the sensitive arc. The cycle shown in Figure 3 gives the sensitive arc 10,073 units of protection since that is the smallest capacity of any of the arcs in the cycle. However, we still need 13463-10073=3390 units to meet the required protection of the sensitive arc. Therefore, if we flow (or send) 3390 additional units through *cells o, p* and *j* we have protected the primary by 13,463 units and have not violated any of the arc capacities. The suppression pattern consisting of two cycles is shown with bold arcs in Figure 4.

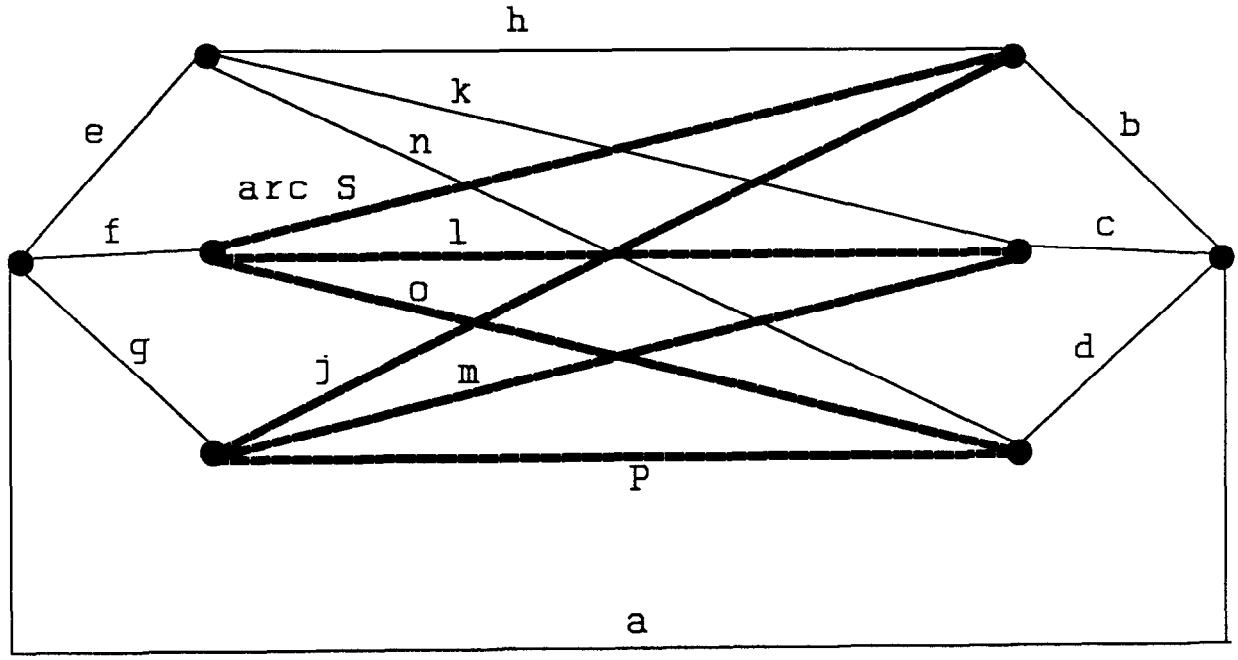


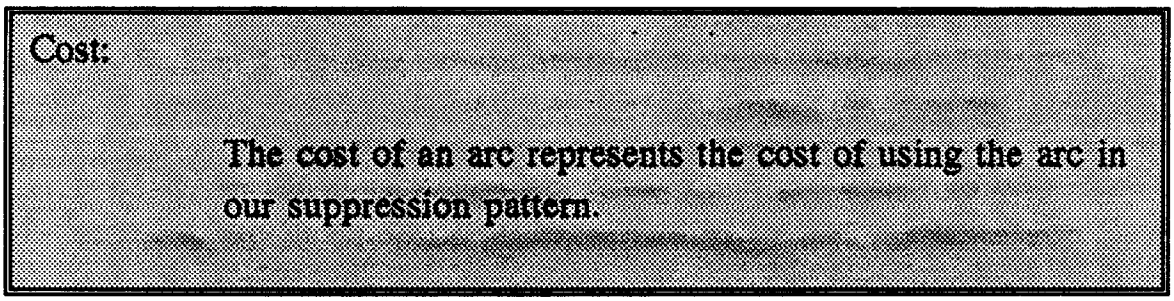
Figure 4. A Suppression Pattern Consisting of Two Cycles

For the simplest cases, all arcs (other than the sensitive arc) are given a capacity equal to their corresponding cell value. (Capacity assignment is a complex issue and will not be discussed in this paper.) The capacity of an arc represents the maximum amount of protection the corresponding table cell can give to the sensitive cell. Our objective is to choose the cycle(s) through the network that protects the sensitive cell by the required protection while suppressing the least amount of data value.

Capacity:

The capacity of an arc represents the maximum amount of protection the corresponding table cell can give to the sensitive cell.

To suppress the least amount of data, we start by giving each arc a cost that represents the cost of using the arc in our suppression pattern. Usually, the corresponding data value from the table is assigned as the cost. However, different cost functions based on the preference of suppressing various cells can be used. If we assign the cost of the sensitive arc to be a very large negative number, like -10^9, then the minimum cost suppression pattern will contain the sensitive arc; that is, the sensitive arc will be included in the suppression pattern since it lowers the cost.



If the corresponding data value from the table is assigned as the cost, then the cycle shown in Figure 5 has a cost of $(19{,}971-10^9)$ and is the minimum cost flow suppression pattern for the sensitive data value shown in Table 2. If the cell value also represents the capacity, then the sensitive data value is protected by 5413 units since this cell has the smallest capacity of all cells in the suppression pattern and is therefore the maximum protection given to the sensitive cell. If the amount of protection required by the primary suppression ((p/100)L−R) is less than 5413, then we have protected the sensitive cell and have chosen the minimum cost flow suppression pattern.

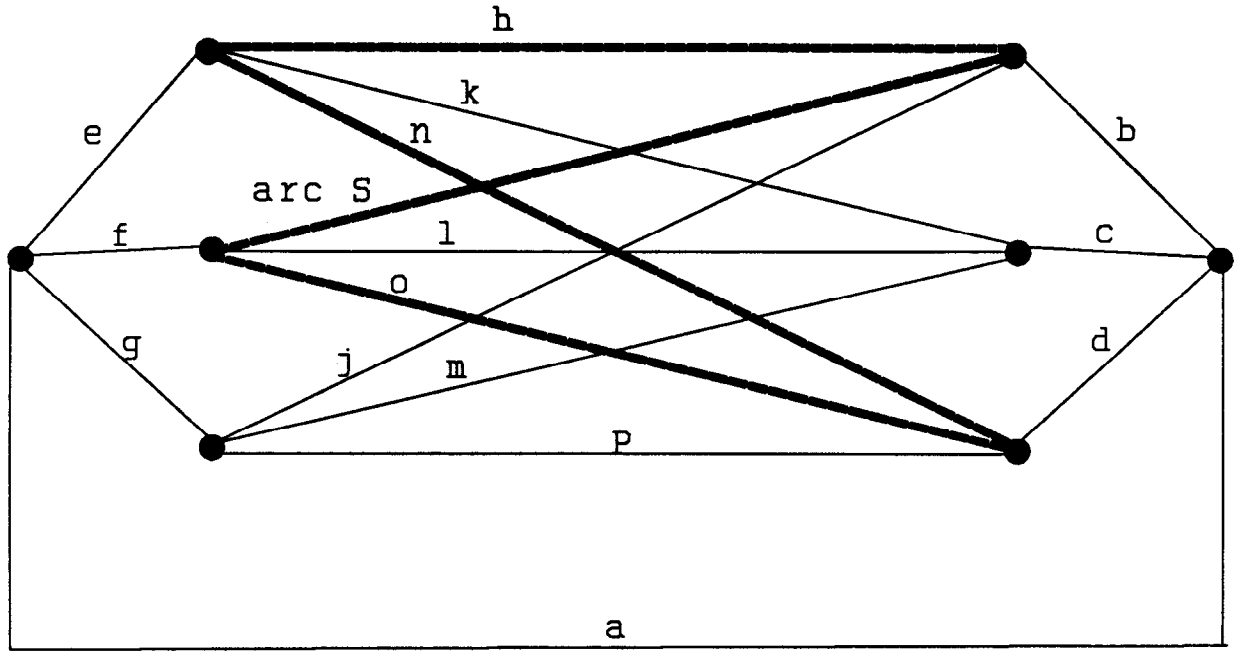


Figure 5. A Minimum Cost Flow Suppression Pattern

The complementary suppressions corresponding to Figure 5 are shown in Table 4.

	State	MSA 1	MSA 2	Non-MSA
SIC Total	a 173 536	e 14 566	f 45 105	g 113 865
SIC 1	b 84 842	h (C)	i (S)	j 61 252
SIC 2	c 43 588	k 1 377	l 20 146	m 22 065
SIC 3	d 45 106	n (C)	o (C)	p 30 548

Table 4. Suppression Pattern Corresponding to Figure 5

3. NETWORK WITH APPENDAGES

The basic network flow methodology presented above is straightforward for a single two-dimensional table. However, almost all economic data are contained in a "system of two-dimensional tables" due to the hierarchical structure of Standard Industrial Classification (SIC) codes. To illustrate, suppose besides our previous sales value table (Table 1), we also provide a more detailed break down of SIC 1, which is shown in Table 5.

	State	MSA 1	MSA 2	Non-MSA
SIC 1	b 84 842	h 5 413	i 18 177	j 61 252
SIC 11	q 52 388	r 2 500	t 7 249	u 42 639
SIC 12	v 32 454	w 2 913	x 10 928	y 18 613

Table 5. Detailed Break Down of SIC 1

Table 5, along with Table 1, is called a system of two-dimensional tables with an appendage. We call the first table a root table and the second table an appendage table. As long as the system is hierarchical, we can translate it into a single network flow diagram. This lets us process all tables of the system as if they were one. For example, Table 6 is a general root table that has an appendage from the SIC 1 Row.

	State	MSA 1	MSA 2	Non-MSA
SIC Total	a	e	f	g
SIC 1	b	h	i	j
SIC 11	q	r	s	t
SIC 12	u	v	w	x
SIC 2	c	k	l	m
SIC 3	d	n	o	p

Table 6. A Root and Appendage Table Presented as One Table

Notice that Table 6 represents the two dimensional system of tables with an appendage as one table and that it contains the following row relationships:

$$\text{SIC Total} = \text{SIC 1} + \text{SIC 2} + \text{SIC 3}$$
$$\text{SIC 1} = \text{SIC 11} + \text{SIC 12}$$

Figure 6 shows the network flow diagram that represents the system shown in Table 6. The highlighted segment of the diagram represents the appendage portion.

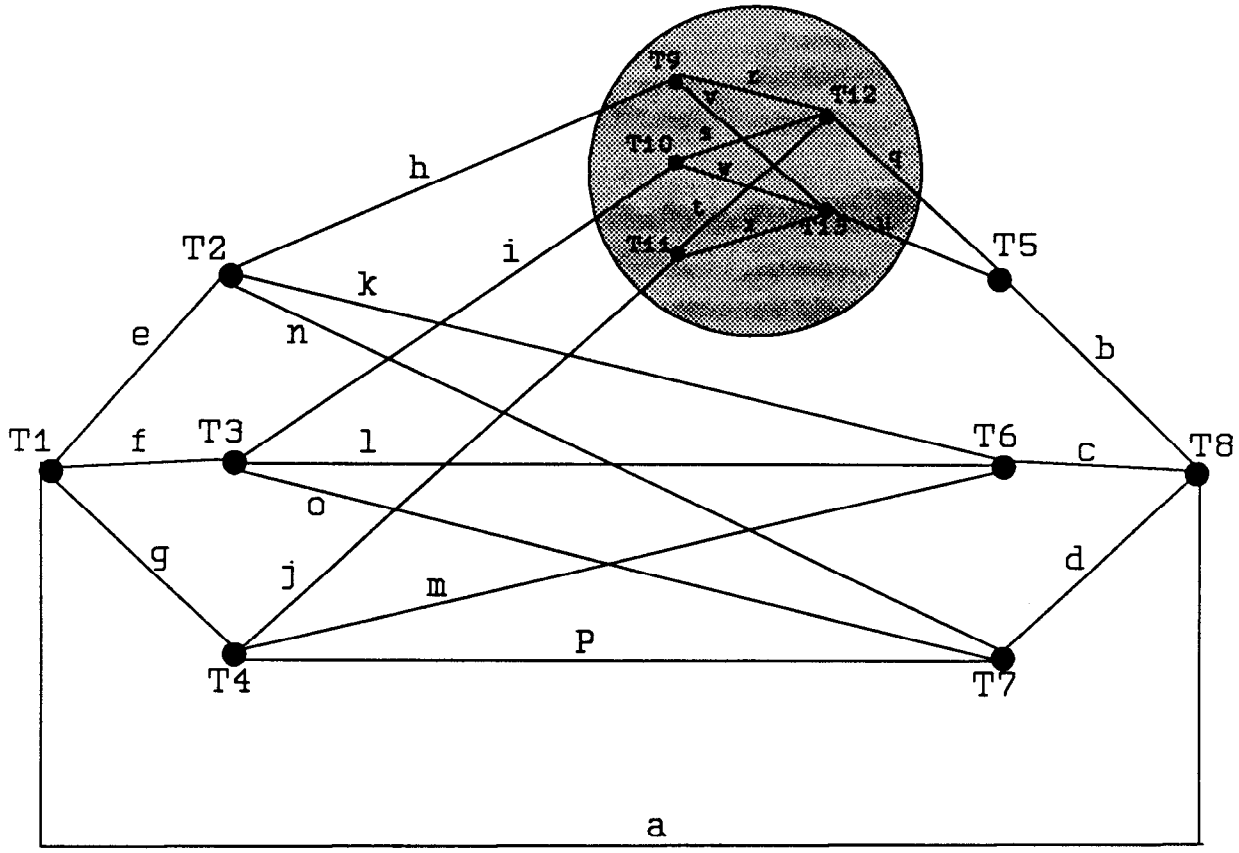


Figure 6. A Network Diagram with An Appendage

Again, each arc represents a cell from the table; there are twenty-four cells in the table and twenty-four arcs in the diagram. As before, each transshipment node symbolizes an additive relationship from the table. Figure 6 shows that *arc h* enters the transshipment node labeled T9 and that *arcs r* and *v* exit. This represents the fact that *cells r* and *v* sum to *cell h* as shown in Table 6.

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 6. *Arcs h, i,* and *j* in Figure 6 no longer directly enter T5 as they did in Figure 1. Instead, they flow to T9, T10, and T11, respectively, are split into more detailed arcs, and eventually enter T5.

A single network also can be designed for a system of two dimensional tables with more than one appendage. That is, each row can be broken down into more detailed relationships and, as long as it remains hierarchical, a single network diagram can be created.

At this point we can protect a sensitive data value the same way as before, by finding a minimum cost flow suppression pattern. As in the case without appendages, our objective is to find a cycle(s) through the network that suppresses the least amount of data value while protecting the sensitive cell. Suppose that *cell k* in Table 6 is sensitive. Figure 7 shows a single cycle that protects the corresponding *arc k.*

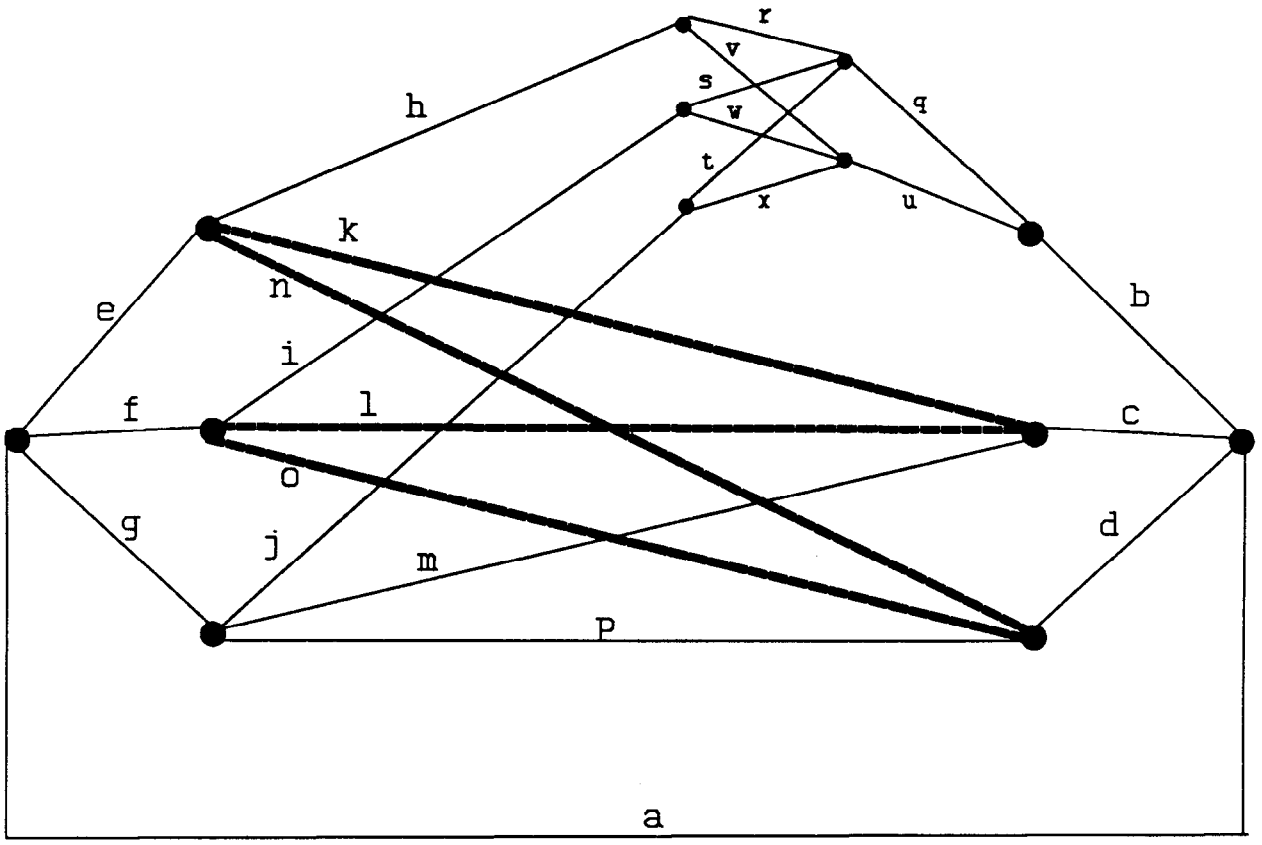


Figure 7. A Suppression Pattern with a Single Cycle

Table 7 shows the corresponding complementary suppression pattern.

	State	MSA 1	MSA 2	Non-MSA
SIC Total	a	e	f	g
SIC 1	b	h	i	j
SIC 11	q	r	s	t
SIC 12	u	v	w	x
SIC 2	c	k (S)	l (C)	m
SIC 3	d	n (C)	o (C)	p

Table 7. Suppression Pattern Corresponding to Figure 7

Notice that none of the arcs contained in the appendage portion of the network diagram were used in the suppression pattern. This is not always the case. If using the appendage arcs would have produced a lower cost suppression pattern, they would have been used in the pattern. Also, when a cell appearing in both the root table and the appendage table is sensitive, then interior cells of the appendage table (arcs in the highlighted portion of Figure 6) must be used in the suppression pattern.

Again consider Table 6 and suppose *cell i* is sensitive. Note *cell i* appears in both Table 1 (a root table) and Table 5 (an appendage table). Figure 8 shows a suppression pattern that protects the sensitive arc.

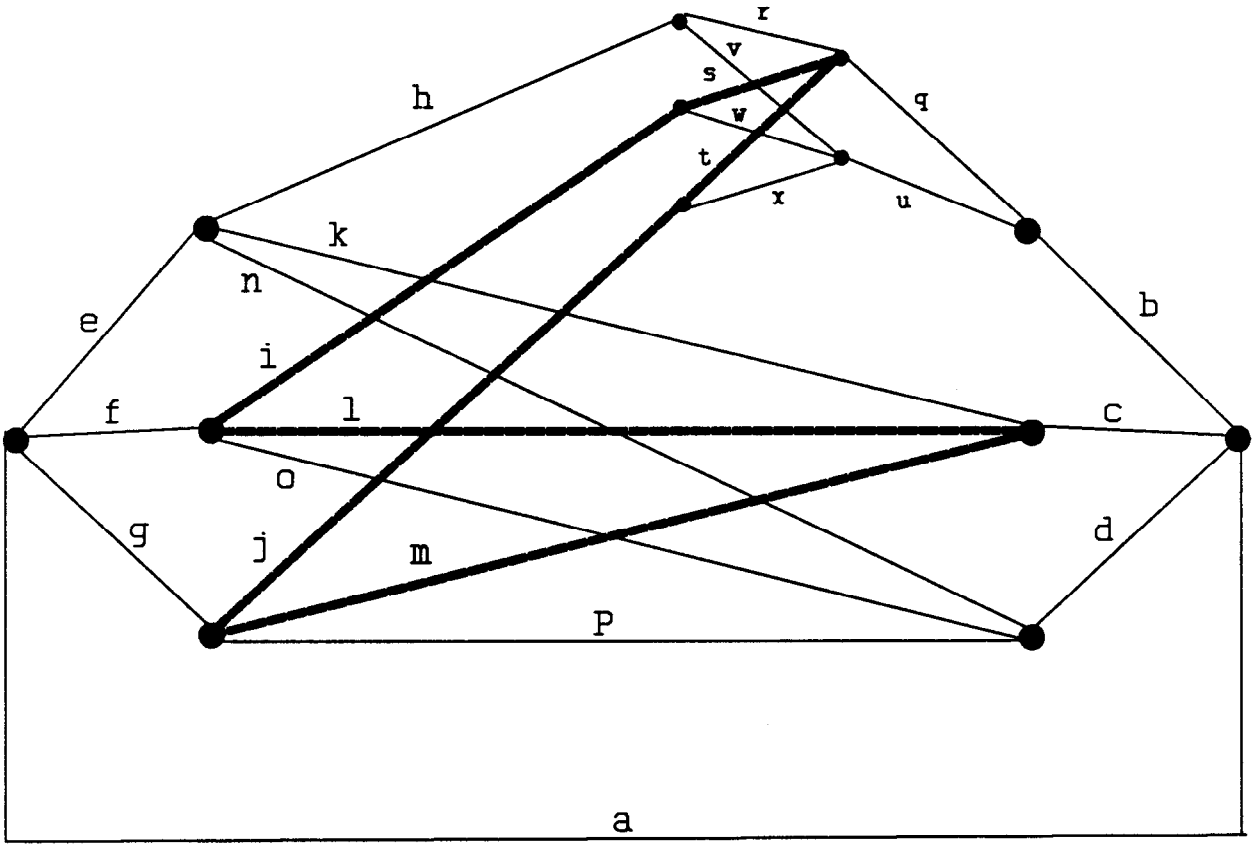


Figure 8. A Suppression Pattern Containing Appendage Arcs

The corresponding complementary suppression pattern is shown in Table 8.

	State	MSA 1	MSA 2	Non-MSA
SIC Total	a	e	f	g
SIC 1	b	h	i (S)	j (C)
SIC 11	q	r	s (C)	t (C)
SIC 12	u	v	w	x
SIC 2	c	k	l (C)	m (C)
SIC 3	d	n	o	p

Table 8. Suppression Pattern Corresponding to Figure 8

Notice that *cells s* and *t* are contained in the suppression pattern. If only *cells i, j, l* and *m* were suppressed, a data user could add the values of *cells s* and *w* and learn the exact value of *cell i*, and add the values of *cells t* and *x* and learn the exact value of *cell j*. However, since a single network diagram contains arcs representing both the root table cells and the appendage table cells, we need not perform any additional work to ensure that all necessary cells are contained in the suppression pattern. The single network diagram ensures this for us.

4. SUMMARY

We have shown how a two dimensional table can be transformed into a network flow diagram. We then use the network diagram to find complementary suppressions for corresponding sensitive table cells. The required protection of the sensitive cell determines how many units will flow through the network. Each arc is assigned a cost and capacity. The cost represents the cost of using a particular cell in a suppression pattern and the capacity represents the number of units of protection a given cell provides the sensitive cell. We then find a minimum cost flow cycle(s) through the network. This cycle(s) corresponds to complementary suppressions in our table.

We have also shown how a two dimensional system of tables with appendages translates to a single network diagram. Once the corresponding network diagram has been formed, we protect sensitive data in the same way as we did for the basic two dimensional tables.

REFERENCES

Sullivan, C.M. (1992), "An Overview of Disclosure Principles," SRD Research Report Series, No. RR-92/09, Bureau of the Census, Statistical Research Division, Washington, D.C. 20233.

FURTHER READINGS

Further discussion on the general topic of disclosure can be found in the following papers:

Cox, L.H. (1975), "Disclosure Analysis and Cell Suppression ," *Proceedings of the American Statistical Association, Social Statistics Section.*

Cox, L.H., Fagan, J.T., Greenberg, B.V., and Hemmig, R.J. (1986), "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. (1986), "Confidentiality Issues at the United States Bureau of the Census," *Journal of Official Statistics*, 2,135-160.

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. 20233

Further discussion on network flow methodology can be found in the following papers:

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. (1986), "Research at the Census Bureau into Disclosure Avoidance Techniques for Tabular Data," *Proceedings of the American Statistical Association, Survey Research Methods Section.*

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

Kelly, J.P., Golden, B.L., and Assad, A.A. (1992), "Cell Suppression: Disclosure Protection for Sensitive Tabular Data," *Networks*, 22, 397-417.

Sullivan, C.M. and Rowe, E.G. (1992), "A Data Structure and Integer Programming Technique to Facilitate Cell Suppression Strategies," *American Statistical Association, 1992 Proceedings of the Section on Survey Research Methods*, to appear.

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.*

Sullivan, C.M., and Zayatz, L. (1992), "A Disclosure Avoidance System Using Network Methodology for the Census of Agriculture," SRD Census Confidential Research Report Series, No. CCRR-92/02, Bureau of the Census, Statistical Research Division, Washington, D.C. 20233.

A linear programming approach can be used instead of network flow methodology. This alternative method is discussed in the following papers:

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

Zayatz, L. (1992), "Using Linear Programming Methodology for Disclosure Avoidance Purposes," to appear in *Proceedings of International Seminar on Statistical Confidentiality* held in Dublin, Ireland.

ACKNOWLEDGEMENTS

I gratefully acknowledge the help of those who reviewed and provided motivation for this paper, especially Alan Saalfeld, Robert Jewett and Laura Zayatz. Special thanks are also due to Dennis Shoemaker for comments on an earlier version.