

# TOTAL PROTECTION OF ANALYTIC INVARIANT INFORMATION IN CROSS TABULATED TABLES\*

MING-YANG KAO<sup>†</sup>

**Abstract.** To protect sensitive information in a cross tabulated table, it is a common practice to suppress some of the cells in the table. An *analytic invariant* is a power series in terms of the suppressed cells that has a unique feasible value and a convergence radius equal to  $+\infty$ . Intuitively, the information contained in an invariant is not protected even though the values of the suppressed cells are not disclosed. This paper gives an optimal linear-time algorithm for testing whether there exist nontrivial analytic invariants in terms of the suppressed cells in a given set of suppressed cells. This paper also presents NP-completeness results and an almost linear-time algorithm for the problem of suppressing the minimum number of cells in addition to the sensitive ones so that the resulting table does not leak analytic invariant information about a given set of suppressed cells.

**Key words.** statistical tables, data security, analytic invariants, mathematical analysis, mixed graph connectivity, graph augmentation.

**AMS subject classifications.** 68Q22, 62A99, 05C99, 54C30

**1. Introduction.** Cross tabulated tables are used in a wide variety of documents to organize and exhibit information, often with the values of some cells suppressed in order to conceal sensitive information. Concerned with the effectiveness of the practice of cell suppression [12], statisticians have raised two fundamental issues and developed computational heuristics to various related problems [5, 7, 8, 9, 10, 11, 28, 29, 30, 31]. The *detection* issue is whether an adversary can deduce significant information about the suppressed cells from the published data of a table. The *protection* issue is how a table maker can suppress a small number of cells in addition to the sensitive ones so that the resulting table does not leak significant information.

This paper investigates the complexity of how to protect a broad class of information contained in a two-dimensional table that publishes (1) the values of all cells except a set of sensitive ones, which are *suppressed*, and (2) an upper bound and a lower bound for each cell, and (3) all row sums and column sums of the complete set of cells. The cells may have real or integer values. They may have different bounds, and the bounds may be finite or infinite. The upper bound of a cell should be strictly greater than its lower bound; otherwise, the value of that cell is immediately known even if that cell is suppressed. The cells that are not suppressed also have upper and lower bounds. These bounds are necessary because some of the unsuppressed cells may later be suppressed to protect the information in the sensitive cells. (See Figures 1.1 and 1.2 for an example of a complete table and its published version.)

An *unbounded feasible assignment* to a table is an assignment of values to the suppressed cells such that each row or column adds up to its published sum. An *bounded* feasible assignment is an unbounded one that also obeys the bounds of the suppressed cells. An *analytic function* of a table is a power series of the suppressed cells, each regarded as a variable, such that the convergence radius is  $\infty$  [1, 4, 21, 22, 26, 27]. An *analytic invariant* is an analytic function that has a unique value at all the bounded feasible assignments. If an analytic invariant is formed by a

---

\*A preliminary version of this work appeared in Proc. 11th Annual Symposium on Theoretical Aspects of Computer Science, Caen, France, February 24–26, 1994, pp. 723–734.

<sup>†</sup>Department of Computer Science, Duke University, Durham, NC 27708. Supported in part by NSF grants MCS-8116678, DCR-8405478, and CCR-9101385. Part of this work was done while the author was at the Department of Computer Science, Yale University, New Haven, Connecticut 06520.

linear combination of the suppressed cells, then it is called a *linear invariant* [17, 19]. Similarly, a suppressed cell is called an *invariant cell* [14, 15] if it is an invariant by itself. For instance, in the published table in Figure 1.2, let  $X_{p,q}$  be the cell at row  $p$  and column  $q$ .  $X_{6,i}$  is an invariant because it is the only suppressed cell in row 6.  $X_{2,c}$  and  $X_{3,c}$  are invariant cells because their values are between 0 and 9.5, their sum is 19, and both cells are forced to have the same unique value 9.5. Consequently,  $(X_{3,c} \cdot X_{2,c} + 0.5 \cdot X_{2,c} - 95)^2 \cdot X_{1,b} + \sin(X_{2,c} \cdot X_{2,a} - 9.5 \cdot X_{2,a})$  is also an invariant.

Intuitively, the information contained in an analytic invariant is unprotected because its value can be uniquely deduced from the published data. In this paper, a set of suppressed cells is *totally protected* if there exists no analytic invariant in terms of the suppressed cells in the given set, except the trivial invariant that contains no nonzero terms. As the analytic power series form a very broad family of mathematical functions, total protection conceals from the adversary a very large class of information. This paper gives a very simple algorithm for testing whether a given set of suppressed cells is totally protected. When a graph representation, called the *suppressed graph*, of a table is given as input, this algorithm runs in optimal  $O(m + n)$  time, where  $m$  is the number of suppressed cells and  $n$  is the total number of rows and columns. This paper also considers the problem of computing and suppressing the minimum number of additional cells so that a given set of original suppressed cells becomes totally protected. This problem is shown to be NP-complete. For a large class of tables, this optimal suppression problem can be solved in  $O((m + n) \cdot \alpha(n, m + n))$  time, where  $\alpha$  is an Ackerman's inverse function and its value is practically a small constant [2, 3, 6, 16]. Moreover, for this class of tables, every optimal set of cells for additional suppression forms a spanning forest of some sort. As a consequence, at most  $n - 1$  additional cells need to be suppressed to achieve the total protection of a given set of original suppressed cells. As the size of a table may grow quadratically in  $n$ , the suppression of  $n - 1$  additional cells is a negligible price to pay for total protection for a reasonably large table.

Previously, four other levels of data security have been considered that protect information contained, respectively, in individual suppressed cells [14, 15], in a row or column as a whole, in a set of  $k$  rows or  $k$  columns as a whole, and in a table as a whole [18]. These four levels of data security and total protection differ in two major aspects. First, these four levels of data security primarily protect information expressible as linear invariants, whereas total protection protects the much broader class of analytic invariant information. Second, these four levels of data security emphasize protecting regular regions of a table, whereas total protection protects any given set of suppressed cells and is more flexible. These four levels of data security and total protection share some interesting similarities. As total protection corresponds to spanning forests in suppressed graphs, these four levels of data security are equivalent to some forms of 2-edge connectivity [14, 15], 2-vertex connectivity,  $k$ -vertex connectivity and graph completeness [18]. In this paper, the NP-completeness results and efficient algorithms for total protection rely heavily on its graph characterizations. Similarly, the equivalence characterizations of these four levels of data security have been key in obtaining efficient algorithms [14, 15, 18] and NP-completeness proofs [18] for various detection and protection problems.

Section 2 discusses basic concepts. Section 3 formally defines the notion of total protection and gives a linear-time algorithm to test for this notion. Sections 4 and 5 give NP-completeness results and efficient algorithms for optimal suppression problems of total protection. Section 6 concludes this paper with discussions.

**2. Basics of two-dimensional tables.** This section discusses basic relationships between tables and graphs.

A *mixed* graph is one that may contain both undirected and directed edges. A *traversable* cycle or path in a mixed graph is one that can be traversed along the directions of its edges. A *direction-blind* cycle or path is one that can be traversed if the directions of its edges are disregarded. The word direction-blind is often omitted for brevity. A mixed graph is *connected* (respectively, *strongly connected*) if each pair of vertices are contained in a direction-blind path (respectively, traversable cycle). A *connected component* (respectively, *strongly connected component*) of a mixed graph is a maximal subgraph that is connected (respectively, strongly connected). A set of edges in a mixed graph is an *edge cut* if its removal disconnects one or more connected components of that graph. An edge cut is a *minimal* one if it has no proper subset that is also an edge cut.

From this point onwards, let  $\mathcal{T}$  be a table, and let  $\mathcal{H}' = (A, B, E')$  and  $\mathcal{H} = (A, B, E)$  be the bipartite mixed graphs constructed below.  $\mathcal{H}'$  and  $\mathcal{H}$  are called the *total graph* and the *suppressed graph* of  $\mathcal{T}$ , respectively [15]. For each row (respectively, column) of  $\mathcal{T}$ , there is a unique vertex in  $A$  (respectively,  $B$ ). This vertex is called a *row* (respectively, *column*) vertex. For each cell  $X_{i,j}$  at row  $i$  and column  $j$  in  $\mathcal{T}$ , there is a unique edge  $e$  in  $E$  between the vertices of row  $i$  and column  $j$ . If the value of  $X_{i,j}$  is strictly between its bounds, then  $e$  is undirected. Otherwise, if the value is equal to the lower (respectively, upper) bound, then  $e$  is directed towards its column (respectively, row) endpoint. Note that  $\mathcal{H}'$  is a *complete* bipartite mixed graph, i.e., there is exactly one edge between each pair of vertices from the two vertex sets of the graph. The graph  $\mathcal{H}$  is the subgraph of  $\mathcal{H}'$  whose edge set consists of only those corresponding to the suppressed cells of  $\mathcal{T}$ . Figure 2.1 illustrates a table and its suppressed graph. For convenience, a row or column of  $\mathcal{T}$  will be regarded as a vertex in  $\mathcal{H}$  and a cell as an edge, and vice versa.

**THEOREM 2.1** ([15]). *A suppressed cell of  $\mathcal{T}$  is an invariant cell if and only if it is not in an edge-simple traversable cycle of  $\mathcal{H}$ .*

The *effective area* of an analytic function  $F$  of  $\mathcal{T}$ , denoted by  $EA(F)$ , is the set of variables in the nonzero terms of  $F$ . The function  $F$  is called *nonzero* if  $EA(F) \neq \emptyset$ . Note that because the convergence radius of  $F$  is  $\infty$ ,  $EA(F)$  is independent of the point at which  $F$  is expanded into a power series.

**THEOREM 2.2** ([17]). *For every minimal edge cut  $Y$  of a strongly connected component of  $\mathcal{H}$ ,  $\mathcal{T}$  has a linear invariant  $F$  with  $EA(F) = Y$ .*

The *bounded kernel* (respectively, *unbounded kernel*) of  $\mathcal{T}$ , denoted by  $BK(\mathcal{T})$  (respectively,  $UK(\mathcal{T})$ ), is the real vector space consisting of all linear combinations of  $x - y$ , where  $x$  and  $y$  are arbitrary bounded (respectively, unbounded) feasible assignments of  $\mathcal{T}$ .

Because  $\mathcal{H}$  is bipartite, every cycle of  $\mathcal{H}$  is of even length. Thus, the edges of an edge-simple direction-blind cycle of  $\mathcal{H}$  can be alternately labeled with  $+1$  and  $-1$ . Such a labeling is called a *direction-blind labeling*. A direction-blindly labeled cycle is regarded as an assignment to the suppressed cells of  $\mathcal{T}$ . If the corresponding edge of a suppressed cell is in the given cycle, then the value assigned to that cell is the label of that edge; otherwise, the value is 0. Note that this assignment needs not be an unbounded feasible assignment of  $\mathcal{T}$ .

**THEOREM 2.3** ([19]).

1.  $UK(\mathcal{T}) = BK(\mathcal{T})$  if every connected component of  $\mathcal{H}$  is strongly connected.
2. Every direction-blindly labeled cycle of  $\mathcal{H}$  is a vector in  $UK(\mathcal{T})$ .

**3. Total protection.** A set  $Q$  of suppressed cells of  $\mathcal{T}$  is *totally protected* in  $\mathcal{T}$  if there is no nonzero analytic invariant  $F$  of  $\mathcal{T}$  with  $EA(F) \subseteq Q$ . The goal of total protection can be better understood by considering  $Q$  as the set of suppressed cells that contain sensitive data. The total protection of  $Q$  means that no precise analytic information about these data, not even their row and column sums, can be deduced from the published data of  $\mathcal{T}$ . As analytic power series form a very large class of functions in mathematical sciences, this notion of protection requires a large class of information about  $Q$  to be concealed from the adversary.

The next lemma and theorem characterize the notion of total protection in graph concepts.

**LEMMA 3.1.** *If  $F$  is a nonzero analytic invariant of  $\mathcal{T}$  such that the edges in  $EA(F)$  are contained in the strongly connected components of  $\mathcal{H}$ , then for some strongly connected component  $D$  of  $\mathcal{H}$ ,  $EA(F) \cap D$  is an edge cut of  $D$ .*

*Remark.* The converse of this lemma is not true; for a counter example, consider the linear combination  $X_{1,a} + 2 \cdot X_{1,b}$  for the table in Figure 2.1. Also, if  $F$  is a nonzero linear invariant, then for every strongly connected component  $D$  of  $\mathcal{H}$ , the set  $D \cap EA(F)$  is either empty or is an edge cut of  $D$  [17].

*Proof.* Let  $\mathcal{T}_s$  be the table constructed from  $\mathcal{T}$  by also publishing the suppressed cells that are not in the strongly connected components of  $\mathcal{H}$ . By Theorem 2.1,  $F$  remains a nonzero analytic function of  $\mathcal{T}_s$ . Also, the connected components of the suppressed graph  $\mathcal{H}_s$  of  $\mathcal{T}_s$  are the strongly connected components of  $\mathcal{H}$ . Thus, to prove the lemma, it suffices to prove it for  $\mathcal{T}_s$ ,  $\mathcal{H}_s$ , and  $F$ .

Let  $x_0$  be a fixed bounded feasible assignment of  $\mathcal{T}_s$ . Let  $K = \{x - x_0 \mid x \text{ is a bounded feasible assignment of } \mathcal{T}_s\}$ . Since  $F$  is an analytic invariant of  $\mathcal{T}_s$ , the function  $G(x) = F(x) - F(x_0)$  is an analytic invariant of  $\mathcal{T}_s$  with  $EA(G) = EA(F)$  and its value is zero over  $x_0 + K$ . Because  $K$  contains a nonempty open subset of  $BK(\mathcal{T}_s)$ ,  $G$  is zero over  $x_0 + BK(\mathcal{T}_s)$ . By Theorem 2.3(1) and the strong connectivity of the connected components of  $\mathcal{H}_s$ ,  $BK(\mathcal{T}_s) = UK(\mathcal{T}_s)$  and  $G$  is zero over  $x_0 + UK(\mathcal{T}_s)$ . Thus, it suffices to show that if  $D - EA(F)$  is connected for all connected components  $D$  of  $\mathcal{H}_s$ , then  $G(x_0 + z_0) \neq 0$  for some  $z_0 \in UK(\mathcal{T}_s)$ . To construct  $z_0$ , let  $EA(G) = \{e_1, \dots, e_k\}$ . Let  $D_i$  be the connected component of  $\mathcal{H}_s$  that contains  $e_i$ . By the connectivity of  $D_i - EA(F)$ , there is a vertex-simple path  $P_i$  in  $D_i - EA(F)$  between the endpoints of  $e_i$ . Let  $C_i$  be the vertex-simple cycle formed by  $e_i$  and  $P_i$ . Next, direction-blindly label  $C_i$  with  $e_i$  labeled  $+1$ . Since  $G$  is a nonzero power series,  $G(x_0 + y_0) \neq 0$  for some vector  $y_0$ . Note that  $y_0$  is not necessarily in  $UK(\mathcal{T}_s)$ . So, let  $z_0 = \sum_{i=1}^k h_i \cdot C_i$ , where  $h_i$  is the component of  $y_0$  at variable  $e_i$ . Then, by Theorem 2.3(2),  $z_0 \in UK(\mathcal{T}_s)$ . Because  $P_i$  is in  $\mathcal{H}_s - EA(F)$ ,  $e_i$  appears only in the term  $C_i$  in  $\sum_{i=1}^k h_i \cdot C_i$ . Thus  $z_0$  and  $y_0$  have the same component values at the variables in  $EA(G)$ . Since the variables not in  $EA(G)$  do not appear in any expansion of  $G$ ,  $G(x_0 + z_0) = G(x_0 + y_0) \neq 0$ , proving the lemma.  $\square$

**THEOREM 3.2.** *A set  $Q$  of suppressed cells is totally protected in  $\mathcal{T}$  if and only if the two statements below are both true:*

1. *The edges in  $Q$  are contained in the strongly connected components of  $\mathcal{H}$ .*
2. *For each strongly connected component  $D$  of  $\mathcal{H}$ , the graph  $D - Q$  is connected.*

*Proof.* It is equivalent to show that  $Q$  is not totally protected if and only if  $Q$  contains some edges not in the strongly connected components of  $\mathcal{H}$  or for some strongly connected component  $D$  of  $\mathcal{H}$ , the graph  $D - Q$  is not connected. The  $\Rightarrow$  direction follows from Lemma 3.1. As for the  $\Leftarrow$  direction, if  $Q$  contains some edges not in the strongly connected components of  $\mathcal{H}$ , then by Theorem 2.1,  $Q$  contains

some invariant cells of  $\mathcal{T}$  and thus cannot be totally protected. If for some strongly connected component  $D$  of  $\mathcal{H}$ , the graph  $D - Q$  is not connected, then some subset  $Y$  of  $Q$  is a minimal edge cut of  $D$ . By Theorem 2.2,  $\mathcal{T}$  has a linear invariant  $F$  with  $EA(F) = Y$  and thus  $Q$  is not totally protected.  $\square$

This paper investigates the following two problems concerning how to achieve total protection.

PROBLEM 1 (Protection Test).

- Input: The suppressed graph  $\mathcal{H}$  and a set  $Q$  of suppressed cells of a table  $\mathcal{T}$ .
- Output: Is  $Q$  totally protected in  $\mathcal{T}$ ?

THEOREM 3.3. *Problem 1 can be solved in linear time in the size of  $\mathcal{H}$ .*

*Proof.* This problem can be solved within the desired time bound by means of Theorem 3.2 and linear-time algorithms for computing connected components and strongly connected components [2, 3, 6, 16].  $\square$

PROBLEM 2 (Optimal Suppression).

- Input: A table  $\mathcal{T}$ , a subset  $Q$  of  $E$ , and an integer  $p \geq 0$ , where  $E$  is the set of all suppressed cells in  $\mathcal{T}$ .
- Output: Is there a set  $P$  consisting of at most  $p$  published cells of  $\mathcal{T}$  such that  $Q$  is totally protected in the table  $\overline{\mathcal{T}}$  formed by  $\mathcal{T}$  with the cells in  $P$  also suppressed?

This problem is clearly in NP. Section 4 shows that this problem with  $Q = E$  is NP-complete. In contrast, Section 5 proves that if the total graph of  $\mathcal{T}$  is undirected, then this problem with general  $Q$  can be solved in almost linear time.

**4. NP-completeness of optimal suppression.** Throughout this section, the total graph of  $\mathcal{T}$  may or may not be undirected.

THEOREM 4.1. *Problem 2 with  $Q = E$  is NP-complete.*

To prove this theorem, the idea is to first transform Problem 2 with  $Q = E$  to the following graph problem and then prove the NP-completeness of the graph problem.

PROBLEM 3.

- Input: A complete bipartite mixed graph  $\mathcal{H}' = (A, B, E')$ , a subgraph  $\mathcal{H} = (A, B, E)$ , and an integer  $p \geq 0$ .
- Output: Does any set  $P$  of at most  $p$  edges in  $E' - E$  hold the following two properties?

Property N1: Every connected component of  $(A, B, E \cup P)$  is strongly connected.

Property N2: The vertices of each connected component of  $\mathcal{H}$  are connected in  $(A, B, P)$ , i.e., contained in a connected component in  $(A, B, P)$ .

LEMMA 4.2. *Problem 2 with  $Q = E$  and Problem 3 can be reduced to each other in linear time.*

*Proof.* Given an instance  $\mathcal{T}$  and  $p$  of Problem 2 with  $Q = E$ , the desired instance of Problem 3 is the total graph  $\mathcal{H}' = (A, B, E')$  and the suppressed graph  $\mathcal{H} = (A, B, E)$  of  $\mathcal{T}$ , and  $p$  itself. This transformation can easily be computed in linear time. There are two directions to show that it reduces Problem 2 to Problem 3. Assume that  $P$  is a desired set for Problem 3. By Property N1, Statement 1 in Theorem 3.2 is true. Also, every strongly connected component of  $(A, B, E \cup P)$  is a union of edge-disjoint connected components in  $\mathcal{H}$  and  $(A, B, P)$ . Therefore, by Property N2, Statement 2 of Theorem 3.2 holds. As a result,  $P$  itself is a desired set for Problem 2. On the other hand, assume that  $P$  is a desired set for Problem 2. Let  $P'$  be the set of all edges in  $P$  that are also in the strongly connected components of  $(A, B, E \cup P)$ . By Statement 1 of Theorem 3.2 and the total protection of  $E$  in  $\overline{\mathcal{T}}$ , the connected components of

$(A, B, E \cup P')$  are the strongly connected components of  $(A, B, E \cup P)$ . Thus,  $P'$  holds Property N1. Next, because a connected component of  $\mathcal{H}$  is included in a strongly connected component of  $(A, B, E \cup P')$ , by Statement 2 of Theorem 3.2,  $P'$  also holds Property N2 and thus is a desired set for Problem 3.

Given an instance  $\mathcal{H}'$ ,  $\mathcal{H}$ , and  $p$  of Problem 3, the desired instance of Problem 2 with  $Q = E$  is  $p$  itself and the table defined as follows. For each vertex in  $A$  (respectively,  $B$ ), there is a row (respectively, column). The upper and lower bounds for each cell are 2 and 0. For each edge  $e$  in  $E'$ , its corresponding cell is at the row and column corresponding to its endpoints. The value of that cell is 1 (respectively, 0 and 2) if  $e$  is undirected (respectively, directed from  $A$  to  $B$ , or directed from  $B$  to  $A$ ). For each edge  $e$  in  $\mathcal{H}$ , its corresponding cell is suppressed. Note that the total and suppressed graphs of this table are  $\mathcal{H}'$  and  $\mathcal{H}$  themselves. Thus, the remaining proof details for this reduction are essentially the same as for the other reduction.  $\square$

Both Problem 2 with  $Q = E$  and Problem 3 are clearly in NP. To prove their completeness in NP, by Lemma 4.2 it suffices to reduce the following NP-complete problem to Problem 3.

PROBLEM 4 (Hitting Set [13]).

- Input: A finite set  $S$ , a nonempty family  $W$  of subsets of  $S$ , and an integer  $h \geq 0$ .
- Output: Is there a subset  $S'$  of  $S$  such that  $|S'| \leq h$  and  $S'$  contains at least one element in each set in  $W$ ?

Given an instance  $S = \{s_1, \dots, s_q\}$ ,  $W = \{S_1, \dots, S_r\}$ ,  $h$  of Problem 4, an instance  $\mathcal{H}' = (A, B, E')$ ,  $\mathcal{H} = (A, B, E)$ ,  $p$  of Problem 3 is constructed as follows:

- Rule 1: Let  $A = \{a_0, a_1, \dots, a_q\}$ . The vertices  $a_1, \dots, a_q$  correspond to  $s_1, \dots, s_q$ , but  $a_0$  corresponds to no  $s_i$ .
- Rule 2: Let  $B = \{b_0, b_1, \dots, b_r\}$ . The vertices  $b_1, \dots, b_r$  correspond to  $S_1, \dots, S_r$  of  $S$ , but  $b_0$  corresponds to no  $S_j$ .
- Rule 3: Let  $E'$  be the union of the following sets of edges:
  1.  $\{b_0 \rightarrow a_0\}$ .
  2.  $\{a_0 \rightarrow b_j \mid \forall j \text{ with } 1 \leq j \leq r\}$ .
  3.  $\{a_i \rightarrow b_0 \mid \forall i \text{ with } 1 \leq i \leq q\}$ .
  4.  $\{b_j \rightarrow a_i \mid \forall s_i \text{ and } S_j \text{ with } s_i \in S_j\}$ .
  5.  $\{a_i \rightarrow b_j \mid \forall s_i \text{ and } S_j \text{ with } s_i \notin S_j\}$ .
- Rule 4: Let  $E = \{a_0 \rightarrow b_1, \dots, a_0 \rightarrow b_r\}$ .
- Rule 5: Let  $p = h + r + 1$ .

The above construction can easily be computed in polynomial time. The next two lemmas show that it is indeed a desired reduction.

LEMMA 4.3. *If some set  $S' \subseteq S$  with  $|S'| \leq h$  contains at least one element in each  $S_j$ , then there is a set  $P \subseteq E' - E$  consisting of at most  $p$  edges that holds Properties N1 and N2.*

*Proof.* For each  $S_j$ , let  $s_{i_j}$  be an element in  $S' \cap S_j$ ; by the assumption of this lemma, these elements exist. Next, let  $P_1 = \{b_1 \rightarrow a_{i_1}, \dots, b_r \rightarrow a_{i_r}\}$  and  $P_2 = \{a_{i_1} \rightarrow b_0, \dots, a_{i_r} \rightarrow b_0\}$ ; by Rule 3, these two sets exist. Now, let  $P = P_1 \cup P_2 \cup \{b_0 \rightarrow a_0\}$ . Note that  $P \subseteq E' - E$ . Since  $P_1$  consists of  $r$  edges and  $P_2$  consists of at most  $|S'|$  edges,  $P$  has at most  $p$  edges.  $P$  holds Property N1 because  $E \cup P$  consists of the edges in the traversable cycles  $b_0 \rightarrow a_0, a_0 \rightarrow b_j, b_j \rightarrow a_{i_j}, a_{i_j} \rightarrow b_0$ . Property N2 of  $P$  follows from the fact that  $P$  connects  $\{a_0, b_1, \dots, b_r\}$ , which forms the only connected component of  $\mathcal{H}$  with more than one vertex.  $\square$

LEMMA 4.4. *If some set  $P \subseteq E' - E$  consisting of at most  $p$  edges holds Properties*

*N1 and N2, then there exists a set  $S' \subseteq S$  with  $|S'| \leq h$  that contains at least one element in each  $S_j$ .*

*Proof.* By Property N1,  $P$  must contain some edge  $b_j \rightarrow a_{i_j}$  for each  $j$  with  $1 \leq j \leq r$ . By Rule 3(4),  $s_{i_j} \in S_j$ . Now let  $S' = \{s_{i_1}, \dots, s_{i_r}\}$ . To calculate the size of  $S'$ , note that by Property N1,  $P$  must also contain  $b_0 \rightarrow a_0$  and at least one edge leaving  $a_{i_j}$  for each  $j$ . Thus  $|P| \geq |S'| + r + 1$ . Then  $|S'| \leq h$  because  $|P| \leq p = r + h + 1$ .  $\square$

The above lemma completes the proof of Theorem 4.1.

**5. Optimal suppression in almost linear time.** Under the assumption that the total graph of  $\mathcal{T}$  is undirected, this section considers the following optimization version of Problem 2.

PROBLEM 5 (Optimal Suppression).

- Input: The suppressed graph  $\mathcal{H} = (A, B, E)$  of a table  $\mathcal{T}$  and a subset  $Q$  of  $E$ .
- Output: A set  $P$  consisting of the smallest number of published cells in  $\mathcal{T}$  such that  $Q$  is totally protected in the table  $\overline{\mathcal{T}}$  formed by  $\mathcal{T}$  with the cells in  $P$  also suppressed.

For all positive integers  $n$  and  $m$ , let  $\alpha$  denote the best known function such that  $m + n$  unions and finds of disjoint subsets of an  $n$ -element set can be performed in  $O((m + n) \cdot \alpha(n, m + n))$  time [2, 3, 6, 16].

THEOREM 5.1. *Problem 5 can be solved in  $O((m + n) \cdot \alpha(n, m + n))$  time, where  $m$  is the number of suppressed cells and  $n$  is the total number of rows and columns in  $\mathcal{T}$ .*

To prove Theorem 5.1, Problem 5 is first converted to the next problem.

PROBLEM 6.

- Input: An undirected bipartite graph  $\mathcal{H} = (A, B, E)$  and a subset  $Q$  of  $E$ .
- Output: A forest  $P$  formed by the smallest number of undirected edges between  $A$  and  $B$  but not in  $E$  such that the vertices of each connected component of  $(A, B, Q)$  are connected in  $(A, B, (E - Q) \cup P)$ , i.e., contained in a connected component of  $(A, B, (E - Q) \cup P)$ .

LEMMA 5.2. *Problems 5 and 6 can be reduced to each other in linear time.*

*Proof.* The proof uses arguments similar to those in the proof of Lemma 4.2. The strong connectivity properties in Problem 3 and Theorem 3.2 can be ignored because this section assumes that the total graph of  $\mathcal{T}$  is undirected. The forest structure of  $P$  follows from its minimality.  $\square$

Note that because  $Q \subseteq E$ , the vertices of each connected component of  $(A, B, Q)$  are connected in  $(A, B, (E - Q) \cup P)$  if and only if the vertices of each connected component of  $\mathcal{H}$  are connected in  $(A, B, (E - Q) \cup P)$ . Using this equivalence, the next stage of the proof of Theorem 5.1 further reduces Problem 6 to another graph problem with the steps below:

- M1. Compute the connected components  $D_1, \dots, D_r$  of  $\mathcal{H}$ .
- M2. For each  $D_i$ , compute a maximal forest  $K_i$  over the vertices of  $D_i$  using only the edges in  $E - Q$ .
- M3. For each  $D_i$ , extend  $K_i$  to a maximal forest  $L_i$  over the vertices of  $D_i$  using additional edges only from the complement graph  $D_i^c$  of  $D_i$ .
- M4. Construct a graph  $\hat{\mathcal{H}}$  from  $\mathcal{H}$  by contracting each tree in each  $L_i$  into a single vertex.
- M5. For each  $D_i$ , compute its contracted version  $\hat{D}_i$  in  $\hat{\mathcal{H}}$ .

M6. Divide the vertices of  $\hat{\mathcal{H}}$  into three sets,  $V_A, V_B, V_{AB}$ , where a vertex in  $V_A$  (respectively,  $V_B$ ) consists of a single vertex from  $A$  (respectively,  $B$ ), and a vertex in  $V_{AB}$  contains at least two vertices (thus with at least one from each of  $A$  and  $B$ ).

A set of undirected edges between vertices in  $V_A, V_B, V_{AB}$  is called *semi-tripartite* if every edge in that set is between two of the three sets or is between two vertices in  $V_{AB}$ . Note that the set of edges in  $\hat{\mathcal{H}}$  is semi-tripartite.

PROBLEM 7.

- Input: Three disjoint finite sets  $V_A, V_B, V_{AB}$ , and a partition  $\hat{D}_1, \dots, \hat{D}_r$  of  $V_A \cup V_B \cup V_{AB}$ .
- Output: A semi-tripartite set  $\hat{P}$  consisting of the smallest number of edges such that no edge in  $\hat{P}$  connects two vertices in the same  $D_i$  and the vertices in each  $D_i$  are connected in the graph formed by is  $\hat{P}$ .

LEMMA 5.3. *Problem 6 can be reduced to Problem 7 in  $O((m+n) \cdot \alpha(n, m+n))$  time, where  $m$  is the number of edges and  $n$  is the number of vertices in  $\mathcal{H}$ .*

*Proof.* The key idea is that an optimal  $P$  for Problem 6 can be obtained by connecting the vertices of each  $D_i$  first with edges in  $E - Q$ , which can be used for free, next with edges in  $D_i^c$ , and then with edges outside  $D_i \cup D_i^c$ . Let  $P'$  be a set of  $|\hat{P}|$  edges in the complement of  $\mathcal{H}$  that becomes  $\hat{P}$  after Step M4. Then,  $P' \cup (L_1 - K_1) \cup \dots \cup (L_r - K_r)$  is a desired output  $P$  for Problem 6, showing that Steps M1–M6 can indeed reduce Problem 6 to Problem 7. Step M3 is the only step that requires more than linear time. It is important to avoid directly computing  $D_i^c$  at Step M3. Computing these complement graphs takes  $\Theta(|A| \cdot |B|)$  time if some  $D_i$  contains a constant fraction of the vertices in  $\mathcal{H}$ . In such a case, if  $\mathcal{H}$  is sparse, then the time spent on computing  $D_i^c$  alone is far greater than the desired complexity. Instead of this naive approach, Step M3 uses efficient techniques recently developed for complement graph problems [20] and takes the desired  $O((m+n) \cdot \alpha(n, m+n))$  time.  $\square$

The last stage of the proof of Theorem 5.1 is to give a linear-time algorithm for Problem 7. A component  $\hat{D}_i$  is *good* if it has at least two vertices with at least one from  $V_{AB}$ ; it is *bad* if it has at least two vertices with none from  $V_{AB}$  (and thus with at least one from each of  $V_A$  and  $V_B$ ). The goal is to use as few edges as possible to connect the vertices in each of these components. Let  $w_g$  and  $w_b$  be the numbers of good and bad components, respectively. There are three cases based on the value of  $w_g$ .

*Case 1:*  $w_g = 0$ . If  $w_b = 0$ , then let  $\hat{P} = \emptyset$  because no  $\hat{D}_i$  needs to be connected. If  $w_b > 0$  and  $|V_{AB}| > 0$ , then include in  $\hat{P}$  an edge between each vertex in the bad components and an arbitrary vertex in  $V_{AB}$ . If  $w_b > 0$  and  $|V_{AB}| = 0$ , then there does not exist a desired  $\hat{P}$  and the given instance of Problem 7 has no solution.

*Case 2:*  $w_g = 1$ . Let  $\hat{D}_j$  be the unique good component.

If  $w_b > 0$ , then find a bad component  $\hat{D}_k$ , and three vertices  $u \in V_{AB} \cap \hat{D}_j$ ,  $v_1 \in V_A \cap \hat{D}_k$ ,  $v_2 \in V_B \cap \hat{D}_k$ . Next, include in  $\hat{P}$  an edge between  $v_2$  and each vertex in  $(\hat{D}_j \cap (V_A \cup V_{AB})) - \{u\}$ , an edge between  $v_1$  and each vertex in  $\hat{D}_j \cap V_B$ , and an edge between  $u$  and each vertex in the bad components.

If  $w_b = 0$  and  $V_{AB} - \hat{D}_j \neq \emptyset$ , then include in  $\hat{P}$  an edge between every vertex in  $\hat{D}_j$  and an arbitrary vertex in  $V_{AB} - \hat{D}_j$ .

If  $w_b = 0$  and  $V_{AB} - \hat{D}_j = \emptyset$ , then there are sixteen subcases depending on whether  $V_A \cap \hat{D}_j = \emptyset$ ,  $V_A - \hat{D}_j = \emptyset$ ,  $V_B \cap \hat{D}_j = \emptyset$ ,  $V_B - \hat{D}_j = \emptyset$ . If  $V_A \cap \hat{D}_j \neq \emptyset$ ,



$V_A - \hat{D}_j \neq \emptyset$ ,  $V_B \cap \hat{D}_j \neq \emptyset$ ,  $V_B - \hat{D}_j \neq \emptyset$ , then include in  $\hat{P}$  an edge between each vertex in  $V_A \cap \hat{D}_j$  and a vertex  $v_2 \in V_B - \hat{D}_j$ , an edge between each vertex in  $V_B \cap \hat{D}_j$  and a vertex  $v_1 \in V_A - \hat{D}_j$ , and an edge between  $v_1$  and each vertex in  $V_{AB} \cup \{v_2\}$ . The other fifteen subcases are handled similarly.

*Case 3:*  $w_g \geq 2$ . Let  $d$  be the total number of vertices in the good and bad components. Let  $w'$  be the number of connected components in  $\hat{P}$  that contain the vertices of at least one good or bad  $\hat{D}_i$ ; let  $d'$  be the number of vertices in these connected components of  $\hat{P}$  that are not in any good or bad  $\hat{D}_i$ . By its minimality,  $\hat{P}$  forms a forest and  $|\hat{P}| = d' + d - w'$ . The techniques for Cases 1 and 2 can be used to show that there exists an optimal  $\hat{P}$  with  $d' = 0$ . Thus, to minimize  $|\hat{P}|$  is to maximize  $w'$ . Because two bad components cannot be connected by edges between them alone, the strategy for maximizing  $w'$  is to pair a good component with a bad one, whenever possible, and include in  $\hat{P}$  edges between them to connect their vertices into a tree. After this step, if there remain unconnected bad components but no unconnected good ones, then add to  $P$  an edge between each vertex in the remaining bad components and an arbitrary vertex in the intersection of  $V_{AB}$  and a good component. On the other hand, if there remain good components but no bad ones, then pair up these good components similarly. After this step, if there remains a good component, then add to  $\hat{P}$  an edge between each vertex in this last good component and an arbitrary vertex in the intersection of  $V_{AB}$  and another good component. (As a result, if  $w_g \leq w_b$ , then  $|\hat{P}| = d - w_g$ ; otherwise,  $|\hat{P}| = d - \lfloor \frac{w_g + w_b}{2} \rfloor$ .)

The above discussion yields a linear-time algorithm for Problem 7 in a straightforward manner. This finishes the proof of Theorem 5.1.

**6. Discussions.** Lemma 5.2 has several significant implications. Since  $P$  is a forest, it has at most  $n - 1$  edges. Thus, for a table with an undirected total graph, no more than  $n - 1$  additional cells need to be suppressed to achieve total protection. This is a small number compared to the size of the table, which may grow quadratically in  $n$ . Moreover, when  $\mathcal{H}$  is connected and  $E = Q$ ,  $(A, B, P)$  is a spanning tree. In this case, many well-studied tree-related computational concepts and tools, such as minimum-cost spanning trees, can be applied to consider other optimal suppression problems for total protection.

**Acknowledgements.** The author is deeply grateful to Dan Gusfield for his constant encouragement and help. The author wishes to thank an anonymous referee for very helpful and thorough comments. The referee has also pointed out that some very interesting materials related to Theorems 2.2 and 2.3 have been developed in the context of protecting sums of suppressed cells [23, 24, 25].

## REFERENCES

- [1] L. AHLFORS, *Complex Analysis*, McGraw-Hill, New York, NY, 1979.
- [2] A. V. AHO, J. E. HOPCROFT, AND J. D. ULLMAN, *The Design and Analysis of Computer Algorithms*, Addison-Wesley, Reading, MA, 1974.
- [3] ———, *Data Structures and Algorithms*, Addison-Wesley, Reading, MA, 1983.
- [4] T. M. APOSTOL, *Mathematical Analysis*, Addison-Wesley, Reading, MA, 1974.
- [5] G. J. BRACKSTONE, L. CHAPMAN, AND G. SANDE, *Protecting the confidentiality of individual statistical records in Canada*, in Proceedings of the Conference of the European Statisticians 31st Plenary Session, Geneva, 1983.
- [6] T. H. CORMEN, C. L. LEISERSON, AND R. L. RIVEST, *Introduction to Algorithms*, MIT Press, Cambridge, MA, 1991.

- [7] L. H. COX, *Disclosure analysis and cell suppression*, in Proceedings of the American Statistical Association, Social Statistics Section, 1975, pp. 380–382.
- [8] ———, *Suppression methodology in statistics disclosure*, in Proceedings of the American Statistical Association, Social Statistics Section, 1977, pp. 750–755.
- [9] ———, *Automated statistical disclosure control*, in Proceedings of the American Statistical Association, Survey Research Method Section, 1978, pp. 177–182.
- [10] ———, *Suppression methodology and statistical disclosure control*, Journal of the American Statistical Association, Theory and Method Section, 75 (1980), pp. 377–385.
- [11] L. H. COX AND G. SANDE, *Techniques for preserving statistical confidentiality*, in Proceedings of the 42<sup>nd</sup> Session of the International Statistical Institute, the International Association of Survey Statisticians, 1979.
- [12] D. DENNING, *Cryptography and Data Security*, Addison-Wesley, Reading, MA, 1982.
- [13] M. GAREY AND D. JOHNSON, *Computers and Intractability: A Guide to the Theory of NP-Completeness*, Freeman, New York, NY, 1979.
- [14] D. GUSFIELD, *Optimal mixed graph augmentation*, SIAM J. Comput., 16 (1987), pp. 599–612.
- [15] ———, *A graph theoretic approach to statistical data security*, SIAM J. Comput., 17 (1988), pp. 552–571.
- [16] E. HOROWITZ AND S. SAHNI, *Fundamentals of Data Structures*, Computer Science Press, New York, NY, 1976.
- [17] M. Y. KAO, *Minimal linear invariants*, in Lecture Notes in Computer Science 1023: Algorithms, Concurrency and Knowledge, 1995 Asian Computing Science Conference, K. Kanchanasut and J. J. Levy, eds., Springer-Verlag, New York, NY, 1995, pp. 23–33.
- [18] ———, *Data security equals graph connectivity*, SIAM Journal on Discrete Mathematics, 9 (1996), pp. 87–100.
- [19] M. Y. KAO AND D. GUSFIELD, *Efficient detection and protection of information in cross tabulated tables I: Linear invariant test*, SIAM Journal on Discrete Mathematics, 6 (1993), pp. 460–476.
- [20] M. Y. KAO AND S. H. TENG, *Simple and efficient compression schemes for dense and complement graphs*, in Proceedings of the 5th International Symposium on Algorithms and Computation, 1994, pp. 201–210.
- [21] S. LANG, *Complex Analysis*, Springer-Verlag, New York, NY, 1985.
- [22] L. H. LOOMIS AND S. STERNBERG, *Advanced Calculus*, Addison-Wesley, Reading, MA, 1968.
- [23] F. M. MALVESTUTO, *A universal-scheme approach to statistical databases containing homogeneous summary tables*, ACM Transactions on Database Systems, 18 (1993), pp. 679–708.
- [24] F. M. MALVESTUTO AND M. MOSCARINI, *Query evaluability in statistical databases*, IEEE Transactions on Knowledge and Data Engineering, 2 (1990), pp. 425–430.
- [25] F. M. MALVESTUTO, M. MOSCARINI, AND M. RAFANELLI, *Suppressing marginal cells to protect sensitive information in a two-dimensional statistical table*, in Proceedings of ACM Symposium on Principles of Database Systems, 1991, pp. 252–258.
- [26] H. L. ROYDEN, *Real Analysis*, Macmillan, New York, NY, 1988.
- [27] W. RUDIN, *Principles of Mathematical Analysis*, McGraw-Hill, New York, NY, 1975.
- [28] G. SANDE, *Towards automated disclosure analysis for establishment based statistics*, tech. report, Statistics Canada, 1977.
- [29] ———, *A theorem concerning elementary aggregations in simple tables*, tech. report, Statistics Canada, 1978.
- [30] ———, *Automated cell suppression to preserve confidentiality of business statistics*, Statistical Journal of the United Nations, 2 (1984), pp. 33–41.
- [31] ———, *Confidentiality and polyhedra, an analysis of suppressed entries on cross tabulations*, tech. report, Statistics Canada, unknown date.

row column index	a	b	c	d	e	f	g	h	i	row sum
1	9.5	4.5	1.5	7	1.5	1.5	5.5	2	3	36.0
2	4.5	9.5	9.5	4.5	4.5	9.5	9.5	9.5	4.5	65.5
3	6	1.5	9.5	0	9.5	6	5.5	2	5.5	45.5
4	2	1.5	4	7	1.5	4.5	9.5	5.5	2	37.5
5	1.5	5.5	4	6	5.5	0	0	4.5	9.5	36.5
6	2	3	3	4	6	5.5	2	2	9.5	37.0
column sum	25.5	25.5	31.5	28.5	28.5	27.0	32.0	25.5	34.0	

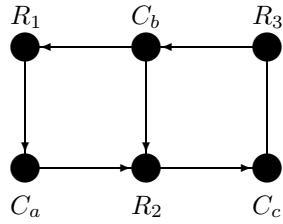
FIG. 1.1. *A Complete Table.*

row column index	a	b	c	d	e	f	g	h	i	row sum
1			1.5	7	1.5	1.5	5.5	2	3	36.0
2										65.5
3	6	1.5				6	5.5	2	5.5	45.5
4	2	1.5	4	7	1.5			5.5	2	37.5
5	1.5	5.5	4	6	5.5					36.5
6	2	3	3	4	6	5.5	2	2		37.0
column sum	25.5	25.5	31.5	28.5	28.5	27.0	32.0	25.5	34.0	

Note: Let  $X_{p,q}$  denote the cell at row  $p$  and column  $q$ . The lower and upper bounds for all suppressed cells except  $X_{2,c}$  and  $X_{3,c}$  are  $-\infty$  and  $+\infty$ . The lower and upper bounds for  $X_{2,c}$  and  $X_{3,c}$  are 0 and 9.5.

FIG. 1.2. *A Published Table.*

row column index	a	b	c	row sum
1	<span style="border: 1px solid black;">0</span>	<span style="border: 1px solid black;">9</span>	1	10
2	<span style="border: 1px solid black;">9</span>	<span style="border: 1px solid black;">9</span>	<span style="border: 1px solid black;">0</span>	18
3	6	<span style="border: 1px solid black;">0</span>	<span style="border: 1px solid black;">5</span>	11
column sum	15	18	6	



In the above  $3 \times 3$  table, the number in each cell is the value of that cell. A cell with a box is a suppressed cell. The lower and upper bounds of the suppressed cells are 0 and 9. The graph below the table is the suppressed graph of the table. Vertex  $R_p$  corresponds to row  $p$ , and vertex  $C_q$  to column  $q$ .

FIG. 2.1. A Table and Its Suppressed Graph.