Working Paper No. 2
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

Joint ECE/Eurostat work session on statistical data confidentiality
(Luxembourg, 7-9 April 2003)

Topic (i): New theories and emerging methods

STATISTICAL DATABASE SECURITY UNDER A QUERY-OVERLAP RESTRICTION

Invited Paper

Submitted by the University of Rome, Italy[1]

[1] Prepared by Franco Malvestuto.

Statistical Database Security

under a Query-Overlap Restriction

Francesco M. Malvestuto

Università "La Sapienza", Roma, Italy

Abstract. In a statistical database, the query-answering system should leave unanswered sum-queries that could lead to the disclosure of confidential data. To this end, each sum-query and previously answered sum-queries should be audited. We give a general framework for controlling the amount of information released when sum-queries are answered, both from the viewpoint of the user and from the viewpoint of the query-answering system. Moreover, we show that, under a suitable query-overlap restriction, an auditing procedure can be efficiently worked out using flow-network computation.

1 Introduction

A statistical database [1] is an ordinary database which containing information on individuals (persons, companies, organisations etc.) but its users are allowed to only access summary statistics over "categories" of individuals. For example, consider a statisticaldatabase containing a file R with scheme {NAME, SSN, SEX, AGE, DEPARTMENT, SALARY}. The users can ask for totals of statistics on SALARY for groups of individuals but these cannot be selected using the attributes NAME and SSN which are private. In this paper, we focus on *sum-queries* such as

q: «What is the sum of salaries of employees with AGE ≥ 40 and DEPARTMENT ≠ Direction or Sex = Male ?».

Here, SALARY is the *summary attribute* and the three attributes AGE, DEPARTMENT and SEX are *category attributes*. Answering sum-queries (and, more in general, statistical queries) raises concerns on the compromise of individual privacy and protection of confidential data should be afforded. Sum-queries whose answers could lead to the disclosure of confidential data are called *sensitive* [3, 7, 8]. In our example, if SALARY is a confidential attribute and q is sensitive, then the response to q should be denied. However, a (memoryless) security measure that only leaves unanswered sensitive sum-queries is not adequate for a confidential data could be disclosed by combining the responses to nonsensitive sum-queries. In order to make the database

'secure', its query-answering system (QAS) should be endowed with a security procedure which is called into play every time a sum-query is asked to decide if its response could lead to the disclosure of some sensitive statistic. The situation can be depicted as a competitive game played by the QAS, which has as its opponent a hypothetical user, henceforth referred to as the *snooper*, who at all times is well-informed of all answered sum-queries and attempts to pry sensitive statistics out of their responses. In this paper we address the following two problems.

(*The snooper problem*) Given a set of answered sum-queries and a statistic of interest, find an optimal estimate of its total.

(*The QAS problem*) Given the set of answered sum-queries and the set of sensitive statistics, decide if a new sum-query can be safely answered, that is, if it can be answered without running the risk that some sensitive statistic can be disclosed.

A non-perturbative solution is given by the "auditing" technique [1], which consists in releasing the exact values of sum-queries if the amount of information that is (implicitly and explicitly) released by the QAS never allows the totals of sensitive statistics to be disclosed. Most implementations [1, 6] of the auditing technique are neither realistic nor efficient, since they are based on the snooper's knowledge of the sets of records selected by sum-queries so that their computational complexity increases with the size of the underlying database. In this paper, we present an implementation of the auditing technique which works with categories and, therefore, its computational complexity is independent of the size of the underlying database.

2 Basic Definitions

Let R be a file of the statistical database. Let β be a summary attribute (of additive type) in the scheme of R and let $\grave{A} = \{\alpha_1, \ldots, \alpha_k\}$ be the set of category attributes in the scheme of R that are used to ask for summary statistics on β. The category attributes may be either independent or dependent; they are independent if every tuple $a = (a_1, \ldots, a_k)$ on $\grave{A}$ is meaningful. An example of dependent category attributes is given by `SEX` and `DIVISION` in a hospital database: since there cannot be any male patient in the gynaecological division, the couple (`Male`, `Gynaecology`) is meaningless. By A we denote the set of meaningful tuples on $\grave{A}$. The tuples in A and the subsets of A will be referred to as *elementary categories* and *categories*, respectively; moreover, elementary categories are assumed to be mutually exclusive and globally exhaustive. If K is an arbitrary category, by the statistic $\beta(K)$ we mean the collection of the value of β over the set of records in R that fall into the category K. Typically, *a sum-query* q on β asks for the total of some statistic $\beta(K)$, $K \subseteq A$, written $q = \beta(K)$. In order to speed up the processing of sum-queries on β, the QAS will make use of a table,

referred to as the *summary table* on β, which is created by the QAS once and for all and reports, for each elementary category a, the total $b(a)$ of $\beta(\{a\})$. Thus, the value of the sum-query $q = \beta(K)$ is computed as $\sum_{a \in K} b(a)$ without accessing the file R. When a sum-query q on β is answered by the QAS, we shall see that the snooper is always able to infer from the value of q and the values of previously answered sum-queries on β the tightest lower bound l and the tightest upper bound u on the value of every statistic $\beta(S)$. The pair $[l, u]$ will be referred to as the *interval-estimate* of the value of $\beta(S)$. In what follows, we only consider the case that the summary attribute β is of nonnegative real type, so that $l \geq 0$ and $u \leq +\infty$. So, if $\beta(S)$ is a sensitive statistic and the interval $[l, u]$ is narrow, then $\beta(S)$ is not protected. We assume that the security policy adopted by our QAS to avoid the disclosure of sensitive statistics requires that, when a sum-query is answered, for each sensitive statistic $\beta(S)$, the width of the interval $[l, u]$ is greater than a threshold value Δ, we call its *protection level* of $\beta(S)$; that is, if $u - l > \Delta$. It is understood that all sensivite statistics are initially identified and the protection level of each of them (if any) is fixed.

Example 1. Consider a file with scheme {NAME, DEPARTMENT, SALARY}, where SALARY is the summary attribute and DEPARTMENT is the category attribute. Hencefeorth, we assume that the domain of SALARY is $_+$, the domain of DEPARTMENT is {a, b, c, d, e, f, g}, the summary table on SALARY contains the following data

DEPARTMENT	SALARY
a	15.0
b	9.0
c	7.5
d	6.5
e	5.5
f	1.5
g	1.0

and that the three categories {a}, {a, f} and {a, g} are only the only sensitive catgeories for SALARY with protection levels 3.0, 3.3 and 3.2, respectively.

3 The snooper at work

Assume that sum-queries $q_1 = \beta(K_1)$, …, $q_n = \beta(K_n)$, $n \geq 1$, have been answered. Let b_i be the value of q_i, $1 \leq i \leq n$. Without loss of generality, we assume that each K_i is not empty; however, it may happen that $K_i = K_{i'}$ even if $i \neq i'$. Let $A_n = \bigcup_{i=1,\ldots,n} K_i$ and $\boldsymbol{K} = \{K_1, \ldots, K_n\}$. Then, it is uniquely determined the coarsest of the partitions of A_n such

that each K_i can be recovered by taking the union of one or more classes of the partition. This partition will be referred to as the *categorization scheme* generated by $\boldsymbol{K}$.

Example 2. Let $K_1 = \{a, b, f, g\}$, $K_2 = \{b, c, d, g\}$ and $K_3 = \{d, e, f, g\}$. The categorization scheme generated by $\boldsymbol{K} = \{K_1, K_2, K_3\}$ is formed by seven categories, each of which is a singleton: {a}, {b}, {c}, {d}, {e}, {f} and {g}.

Let $\boldsymbol{C} = \{C_1, \ldots, C_m\}$ be the categorization scheme generated by $\boldsymbol{K}$, and let $M = \{1, \ldots, m\}$ and $N = \{1, \ldots, n\}$. For each $i \in N$, let $M(i) = \{j \in M: C_j \subseteq K_i\}$. The amount of information conveyed by the responses to $q_1, \ldots, q_n$ can be described by the constrained system of linear equations

$$\begin{cases} \sum_{j \in M(i)} x_j = b_i & (i \in N) \\ x_j \geq 0 & (j \in M) \end{cases} \qquad (1)$$

where variable x_j stands for the (unknown) total of the statistic $\beta(C_j)$. Let $\mathbf{X}$ be the solution set of constraint system (1). Of course, $\mathbf{X}$ is not empty since constraint system (1) is consistent. For each subset J of M, let

$$l(J) = \min \{\textstyle\sum_{j \in J} x_j: \mathbf{x} \in \mathbf{X}\} \qquad u(J) = \max \{\textstyle\sum_{j \in J} x_j: \mathbf{x} \in \mathbf{X}\}.$$

Suppose now that the snooper wants to get information on statistic $\beta(S)$ as detailed as possible, where S is an arbitrary category. The following three cases will be examined separately:

Case 1. S is *covered* by $\boldsymbol{C}$, that is, either $S = \emptyset$ or there is a nonempty subset J of M such that $S = \bigcup_{j \in J} C_j$. Then, the interval-estimate of the total of $\beta(S)$ is given by $[l(J), u(J)]$.
Case 2. S is a subset of A_n but is not covered by $\boldsymbol{C}$. Then the interval-estimate is $[l(J), u(J')]$, where $J = \{j \in M: C_j \subseteq S\}$ and $J' = \{j \in M: C_j \cap S \neq \emptyset\}$.
Case 3. S is not a subset of A_n. Then the interval-estimate is $[l(J), +\infty]$ where $J = \{j \in M: C_j \subseteq S\}$.

So, if S is sensitive, $\beta(S)$ runs the risk of being disclosed only if S is a subset of A_n and is definitely disclosed (no matter what protection level has been fixed by the QAS) if $l(J) = u(J)$, that is, if the total of $\beta(S)$ can be exactly evaluated.

Example 1 (continued). Consider four sum-queries q_1 = SALARY(K_1), …, q_4 = SALARY(K_4), where K_1 = {a, b}, K_2 = {a, c, d}, K_3 = {b, c, e} and K_4 = {d, f}. The categorization scheme $\boldsymbol{C}$ generated by $\boldsymbol{K}$ = {K_1, K_2, K_3, K_4} is formed by the following six categories: C_1 = {a}, C_2 = {b}, C_3 = {c}, C_4 = {d}, C_5 = {e} and C_6 = {f}. Let b_1 = 24, b_2 = 29, b_3 = 18 and b_4 = 12 be the values of q_1, …, q_4, respectively. Constraint system (1) reads

$$\begin{cases} x_1 + x_2 = 24 \\ x_1 + x_3 + x_4 = 29 \\ x_2 + x_3 + x_5 = 18 \\ x_4 + x_6 = 12 \\ x_1, x_2, x_3, x_4, x_5, x_6 \geq 0 \end{cases}$$

Suppose that the snooper attempts to pry the total of the sensitive statistic SALARY(S) where S = {a, e}. Since $S = C_1 \quad C_5$, S is covered by $\boldsymbol{C}$ and he can get the interval-estimate [11.5, 42.5] using standard linear-programming methods.

From a computational point of view, it should be noted that the number (m) of variables in constraint system (1) may be exponential in the number (n) of its equations, that is, in the number of answered sum-queries (see Example 2). Therefore, computing interval-estimates can be very expensive. However, sometimes it is possible to reduce the amount of computation as follows. Suppose that there is $i \quad N$ such that, for each $i' \neq i$, either $K_i \cap K_{i'} = \emptyset$ or $K_i \quad K_{i'}$; then, K_i itself belongs to the categorization scheme, say $K_i = C_j$, and then the i.th equation in constraint system (1) is $x_j = b_i$. Note that, if there is another i' such that $K_{i'} = K_i$, then the corresponding equation $x_j = b_{i'}$ is redundant and can be deleted. At this point, each occurence of x_j in the remaining equations can be deleted since it can be assigned the value b_i. If this procedure is repeated, ultimately one obtain an equivalent constraint system formed by a certain number of equations of the form

$$x_j = \text{const} \ (= c_j) \qquad (j \quad M^*)$$

where M^* is a subset of M, and by a constraint system of the form

$$\begin{cases} \sum_{j \quad M(i) - M^*} x_j = w_i & (i \quad N^*) \\ x_j \geq 0 & (j \quad M - M^*) \end{cases}$$

where N^* is a subset of N and w_i is the 'revised' valued of sum-query q_i. Note that if $M^* = M$ then the total of every statistic $\beta(S)$ can be exactly and directly evaluated whenever S is covered by $\boldsymbol{C}$.

4 How to repel the attacks of the snooper

When a new sum-query $q = \beta(K)$ arrives, the QAS must decide if it can be safely answered, that is, if each sensitive statistic is protected when the value of the sum-query will be released. Needless to say, if q is sensitve then the response to q will be denied. In what follows, we assume that q is not sensitive. Assume that constraint system (1) — possibly in its reduced form — represents the amount of information released by the responses to previously answered sum-queries, and that $\boldsymbol{C} = \{C_1, \ldots, C_m\}$ is the underlying categorization scheme. Of course, if q can be exactly evaluated then it will be answered. Otherwise, the QAS will compute the categorization scheme $\boldsymbol{C}' = \{C'_1 \quad \ldots \quad C'_{m'}\}$ generated by $\boldsymbol{C} \quad \{K\}$ to take the response to q into account, and can *refine* constraint system (1). After doing that, the QAS will test each sensitive statistic to see if it is protected, and only if this is the case q will be answered. Indeed, it is sufficient to test sensitive categories thare are subsets of $C'_1 \quad \ldots \quad C'_{m'}$. We call such categories the *sensitive targets* for the refinement of constraint system (1). So, if each sensitive target is protected (i.e., $u - l > \Delta$), then the refinement of constraint system (1) is secure and q can be safely answered. We shall also make use of a weaker security criterion which requires that, for each $C'_{i'}$, $1 \leq ' \leq m'$, if $C'_{i'}$ is sensitive then it is protected. What remains to clarify is how $\boldsymbol{C}'$ is computed. It is easy to see that $\boldsymbol{C}'$ can be obtained from $\boldsymbol{C}$ by replacing each C_j having $K \cap C_j \neq \emptyset$ by the two categories $C_j - K$ and $K \cap C_j$, and by adding the category $K - (\ _{j=1,\ldots,m}\ C_j)$. More precisely, $\boldsymbol{C}'$ is obtained from the set family

$$(\ _{j=1,\ldots,m}\ \{C_j - K, K \cap C_j\}) \quad \{K - (\ _{j=1,\ldots,m}\ C_j)\}$$

As noticed above, the number of variables in constraint system (1) may be exponential in the number of its equations so that the auditing procedure may be time-consuming. To overcome this difficulty, we introduce a query-overlap restriction which, for a fixed positive integer r, requires that at all times the number of variables per equation in constraint systems such as (1) is not greater than r. It is not difficult to see that, if constraint system (1) is submitted to the query-overlap restriction of order 2, then its refinement is not if and only if there is a category C_j in $\boldsymbol{C}$ such that (*i*) either $K - C_j \neq \emptyset$ or K is a proper subset of C_j, and (*ii*) the variable x_j occurs in r equations of constraint system (1). The simplest nontrivial case is $r = 2$. Despite its simplicity, the query-overlap restriction of order 2 is powerful enough to deal with the security problem for two-dimensional tables as shown in [5]. In the next section, we address the problem of

computing interval-estimates for any constraint system submitted to the query-overlap restriction of order 2. To this end, we now introduce a graphical representation of such a constraint system, we call a *sum map*. For example, the sum map associated with constraint system (1) is the graph, say G, having as its (vertex-edge) incidence matrix the coefficient matrix of the equation system. Note that loops are allowed. Moreover, the vertices of G will be weighted by the constant terms of the equations featured in constraint system (1).

Example 1 (continued). The sum maps $G_1, \ldots, G_5$ given the responses to sum-queries $q_1, \ldots, q_i$ for $i = 1, \ldots, 5$, are shown in Fig. 1.

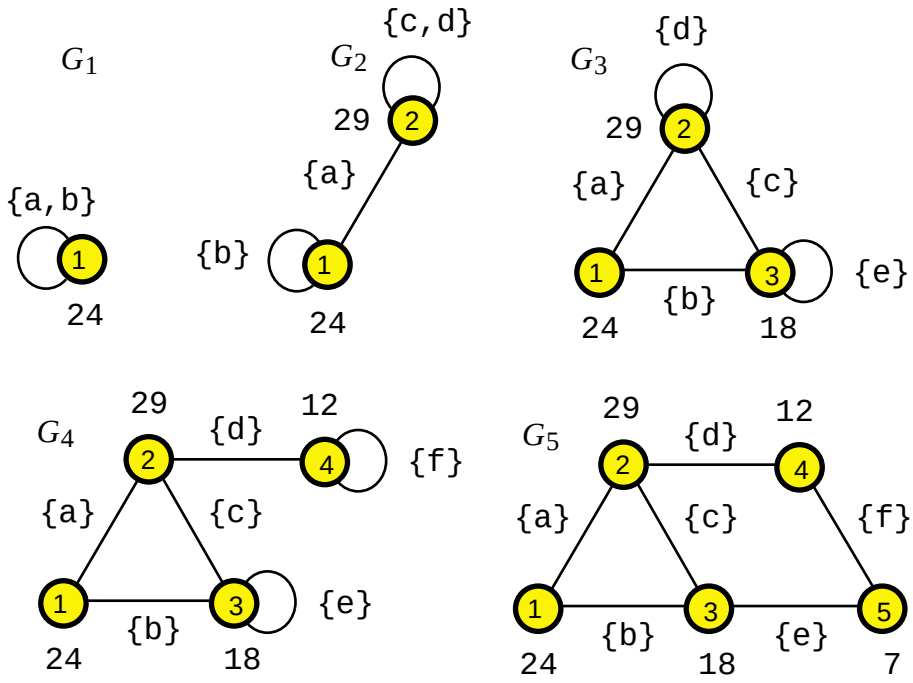

Figure 1

5 Computing interval-estimates

Let $G = (N, E)$ be a sum map associated with constraint system (1). Let $N = \{1, \ldots, n\}$, $E = \{e_1, \ldots, e_m\}$ and $M = \{1, \ldots, m\}$. Without loss of generality, we assume that G is connected. The problem we deal with is how to find the minimum and the maximum of the function $\sum_{j \in J} x_j$ for a given (nonempty) subset J of M. We shall show that the

problem of minimizing or maximizing the function $\sum_{j \in J} x_j$ can be converted into a *bipartite transportation problem* [2], that is, in the form

minimize $\sum_{j \in J} u_j x_j$

subject to $\mathbf{Ax} = \mathbf{d}, \mathbf{x} \in \mathbb{R}^m_+$

where $\mathbf{A}$ is the incidence matrix of a bipartite digraph D. If (U, V) is the bipartition of D, each vertex $i \in U$ of D is as a *source* with supply $-d_i$ and each vertex $i \in V$ of D is as a *sink* with demand d_i. The vector $\mathbf{d}$ is called the *demand vector*. Finally, it is well-known that every transportation problem can be efficiently solved with the *network simplex method* [2].

We now separately discuss two cases depending on whether G is or is not bipartite.

Case 1. G is bipartite. Let (U, V) be the bipartition of G. Direct each edge of G from U to V; thus, if $e = (i, k)$ is an edge of G, $i \in U$ and $k \in V$, i is the tail and k is the head of the directed edge, we denote by $\vec{e} = i \rightarrow k$. Let D be the resulting bipartite digraph and let $\mathbf{A}$ be the incidence matrix of D; thus, each directed edge $\vec{e}$ of D corresponds to a column $\mathbf{a}$ of $\mathbf{A}$ defined by

$$a_i = \begin{cases} -1 & \text{if } i \text{ is the tail of } \vec{e} \\ +1 & \text{if } i \text{ is the head of } \vec{e} \\ 0 & \text{otherwise} \end{cases}$$

Let $\mathbf{d}$ be defined by

$$d_i = \begin{cases} -b_i & (i \in U) \\ b_i & (i \in V) \end{cases}$$

Then, constraint system (1) is like the constraint system of a bipartite transportation problem. By taking $\mathbf{u}$ as the incidence vector of the edge set $\{e_j : j \in J\}$ (or as its opposite), we have converted the problem of minimizing (or maximizing, respectively) into a bipartite transportation problem. So, the problem of deciding if constraint system (1) is secure can be solved in an efficient way. Finally, Gusfield [4] proved that the problem of deciding if constraint system (1) is weakly secure can be solved with maximum-flow computation.

Case 2. G is not bipartite. The edges of G that are not loops will be referred to as links. If G contains p links, it is convenient to order the edges of G as $e_1, \ldots, e_p, e_{p+1}, \ldots, e_m$ where $e_1, \ldots, e_p$ are all links. Moreover, we always write an edge of G as (i, j) where $i \le j$. We now transform G into a bipartite graph $H = (P, F)$ with $2n$ vertices and $m+p$ edges. The graph H is constructed as follows [5]. The vertex set of H is taken to be $P = \{1, \ldots, 2n\}$. Vertex $n + i$ is meant to be a 'copy' of i. The edge set F of H is defined as follows. Arbitrarily choose a spanning tree T of G, and let G' be the bipartite graph obtained from T by adding all non-tree edges of G that do not create odd cycles. The edges $f_1, \ldots, f_{m+p}$ of H are defined as follows:

— if $e_j = (i, k)$ is an edge of G' then $f_j = e_j$ and $f_{m+j} = (n+i, n+k)$, $1 \le j \le p$;

— if $e_j = (i, k)$ is a link of G but not an edge of G' then $f_j = (i, n+k)$ and $f_{m+j} = (k, n+i)$, $1 \le j \le p$;

— if $e_j = (i, i)$ is a loop of G then $f_j = (i, n+i)$, $p < j \le m$.

Let (U, V) be the bipartition of G' and let $U' = \{n+i: i \quad U\}$ and $V' = \{n+i: i \quad V\}$. Note that, since G is a nonbipartite, connected graph, H is a bipartite, connected graph with bipartition (R, S) where $R = U \quad V'$ and $S = U' \quad V$. Finally, the weights w_p of vertices of H are taken to be

$$w_p = \begin{cases} b_p & (p = 1, \quad, n) \\ b_{p-n} & (p = n+1, \quad, 2n) \end{cases}$$

Example 2. Consider the map G_3 of Fig. 1. Fig. 2 shows one of the possible bipartite map associated with G_3.

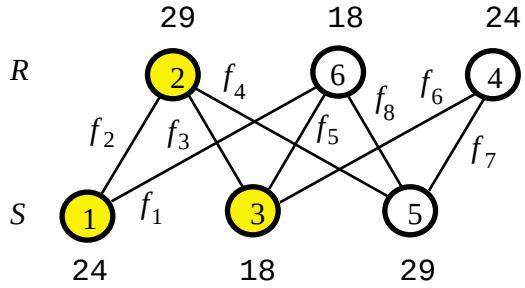

Figure 11

Let **H** be the incidence matrix of H and let **w** be the vector of weights of vertices of H. Consider the constraint system

$$\mathbf{Hy} = \mathbf{w}, \mathbf{y} \quad {}_{+}^{m+p} \qquad (2)$$

The following obvious fact show that the solution set **X** of constraint system (1) and the solution set **Y** of constraint system (2) are closely related to each other.

Fact. For every solution **x** of constraint system (1), the vector **y** with

$$y_j = \begin{cases} x_j & \text{if } j \le m \\ x_{j-m} & \text{if } m < j \le m+p \end{cases}$$

is a solution of constraint system (2), and for every solution **y** of constraint system (2), the vector **x** with

$$x_j = \begin{cases} \frac{1}{2}(y_j + y_{m+j}) & \text{if } j \le p \\ y_j & \text{if } p < j \le m \end{cases}$$

is a solution of constraint system (1).

Consider now an arbitrary subset J of M. Let $J' = \{j \quad J: j \le p\}$ and $J'' = J - J'$. By Fact, one has that the function $\sum_j \ {}_J x_j$ over the solution set of constraint system (1) has the same range as the function

$$(1/2)\,[\textstyle\sum_j \ {}_{J'}(y_j + y_{m+j})] + \sum_j \ {}_{J''} y_j \qquad (3)$$

over the solution set of constraint system (2). So, minimizing (or maximizing) the function $\sum_j \ {}_J x_j$ subject to constraint system (1) can be obtained by minimizing (maximizing, respectively) function (3) subject to constraint system (2). Finally, as we saw above, each of these two problems can be converted into a bipartite transportation problem and, hence, solved in an efficient way. Finally, Malvestuto and Mezzini [5] proved that the problem of deciding if constraint system (1) is weakly secure can be solved with maximum-flow computation.

6 Future research

We discussed the case of a summary attribute of nonnegative real type. However, by virtue of the total unimodularity of the incidence matrix of a bipartite graph, the results proved in Section 5 for a bipartite map also apply to summary attributes of nonnegative integer type. The case of a nonbipartite map is open and left to future research.

References

1. Adam, N.R., Wortmann, J.C.: Security control methods for statistical databases: a comparative study. ACM Computing Surveys **21** (1989) 515-556.
2. Ahuja, R.K., Magnanti, T.L., Orlin, J.B.: Network flows. Prentice Hall, Englewood Cliffs, 1993.
3. Cox, L.H.: Suppression methodology and statistical disclosure control. J. American Statistical Association **75** (1980) 377-385.
4. Gusfield, D.: A graph-theoretic approach to statistical data security. SIAM J. Computing **17** (1988) 552-571.
5. Malvestuto, F.M., Mezzini, M.: A linear algorithm for finding the invariant edges of an edge-weighted graph. SIAM J. on Computing **31** (2002) 1438-1455.
6. Malvestuto, F.M., Moscarini, M.: An audit expert for large statistical databases. In Statistical Data Protection. EUROSTAT (1999) 29-43.
7. Willenborg, L., de Waal, T.: Statistical Disclosure Control in Practice. Lecture Notes in Statistics, Vol. 111. Springer-Verlag, New York (1996).
8. Willenborg, L., de Waal, T.: Elements of Statistical Disclosure. Lecture Notes in Statistics, Vol. 155. Springer-Verlag, New York (2000).