Working Paper No. 2 (Summary)

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

Dipartimento di Informatica, Università "La Sapienza", Via Salaria 113, 00198 Roma, Italy

Statistical databases raise concerns on the compromise of individual privacy, a statistical database [1] being an ordinary database which contains information about individuals (persons, companies, organisations etc.) but its users are only allowed to access sums of individual data provided that they do not lead to the disclosure of confidential data. Consider a file F with scheme R = {NAME, SSN, AGE, DEPARTMENT, SALARY} and a statistical-query system which answers only queries such as: "What is the sum of salaries of the individuals qualified by the condition P" where P is a "category predicate", that is, a condition on the domain of the pair {AGE, DEPARTMENT} of "category" attributes such as DEPARTMENT ? Direction & AGE = 40. Such a query is called a (*categorical*) *sum query* on SALARY. Assume further that SALARY is a confidential attribute. If the set T of records from F selected by P is a singleton, say $T = \{t\}$, then the response to the sum query above (if released) would allow the salary of the individual corresponding to the tuple t to be disclosed and, therefore, it should be denied. The response to such a sum query is called a *sensitive sum* [13, 14]. Indeed, more sophisticated sensitivity criteria exist which broaden the class of sum queries that should be left unanswered. A (memoryless) security measure that leaves unanswered only sum queries whose responses are sensitive sums is not adequate for it does not exclude the possibility of computing some sensitive sum by combining nonsensitive responses. What measures suffice to avoid the disclosure of sensitive sums? This is the classical version of the *security control problem* for sum queries. Indeed, a stronger version of this problem is of practical interest since the producers of statistical databases (e.g., Census Bureaux) demand that an "accurate" estimate of no sensitive sum should be possible in the sense that, if s is a sensitive sum, then the possible values of s consistent with the information released by the statistical-query system should span an interval that is not entirely contained in the interval $[(1-p)\ s, (1+p)\ s]$ where p is fixed percentage, called the *protection level*. Note that for $p = 0$, one has the classical version of the security problem: the exact value of no sensitive sum should be disclosed. The program of the statistical query system that should ensure the protection of sensitive sums is called the "auditor" and it should work as follows. Let R be a relation scheme containing a confidential attribute S with domain D, and let K be the set of category attributes in R. Suppose that sum queries $Q(1), \ldots, Q(n-1)$ on S have been already answered when a new sum query $Q(n)$ on S arrives. Let $P(v)$ and $q(v)$ be the category predicate and the response to $Q(v)$, $1 = v = n$. Without loss of generality, we assume that each $P(v)$ is of the form "K in $C(v)$" where $C(v)$ is a nonempty subset of the domain of K. The amount of information that would be released to the users if $Q(n)$ were answered will represented by a model which consists of a semantic part and an analytic part. The semantic part of the model contains the overlap relationships among the sum queries $Q(v)$ and consists of a hypergraph $G = (V, E)$ where $V = \{1, \ldots, n\}$ and E contains the nonempty subsets e of V such that the subset of the domain of K

$$c(e) = «_{v Œ e}\, C(v)\ -\ »_{v œ e}\, C(v)$$

is not empty. The analytic part of the model consists of the system of linear constraints

$$\begin{cases} \sum_{e \in E(v)} x(e) = q(v) & (v \in V) \\ \qquad x(e) \in D & (e \in E) \end{cases}$$

where $E(v) = \{e\ Œ\ E\text{: } v\ Œ\ e\}$. Here, variable $x(e)$ stands for the unknown (to the users) sum of S over the set of tuples selected by the predicate "K in $c(e)$". Thus, a "snooper" can compute the two quantities $\min_X x(e)$ and $\max_X x(e)$ for each hyperedge e of G, where X is the solution set of constraint system (1). In order to decide if the response to $Q(n)$ leads to the disclosure of some sensitive sum, the auditor will compute the actual value $s(e)$ of each $x(e)$ and apply the sensitivity criterion in use to decide if $s(e)$ is sensitive; next, for each e having $s(e)$ sensitive, it will apply the safety test which checks that

$$\min_X x(e) < (1–p)\, s(e) \quad \text{and} \quad \max_X x(e) > (1+p)\, s(e)$$

where p is the protection level in use. If each sensitive sum $s(e)$ is protected at level p, then (and only then) the auditor will decide that $Q(n)$ can be answered safely.
Suppose that the weighted hypergraph (G, s), where s is the function on E that assigns to each hyperedge e of G the corresponding sum $s(e)$, has been constructed and suppose that a number of sensitive weights have been found. What remains to do is testing each sensitive sum $s(e)$ for safety, which requires computing $\min_X x(e)$ and $\max_X x(e)$. In what follows, the weighted hypergraph (G, s) will be referred to as the *map* of $\{Q(1), \ldots, Q(n)\}$ and the two quantities $\min_X x(e)$ and $\max_X x(e)$ are called the *tightest lower bound* and the *tightest upper bound* on the weight of the hyperedge e of the map (G, s), respectively.
What makes the work of the auditor hard is that the number of hyperedges of the map of $\{Q(1), \ldots, Q(n)\}$ may be exponential in n. To overcome this difficulty, a query-overlap restriction can be introduced which, for a fixed positive integer r, requires that the response to the (current) sum query $Q(n)$ is soon denied if there are r previously answered sum queries $Q(v_1), \ldots, Q(v_r)$ such that

$$(«_{i=1,\ldots,r}\, C(v_i))\; «\; C(n)\; ?\; Ø.$$

The simplest nontrivial case is $r = 2$, which implies that the map of $\{Q(1), \ldots, Q(n)\}$ is a graph (where loops are allowed). Now, the problem of computing the tightest bounds on the weight of an edge of a graph under the assumption that edge weights are nonnegative reals can be solved using linear programming methods. However, a more efficient solution algorithm can be obtained by transforming the graph into a flow network; then, the tightest bounds on the weight of an edge can be found with two or four maximum-flow computations depending on whether the edge is or is not a loop.