A Probabilistic Approach for on-Line Sum-Auditing

Gerardo Canfora, Bice Cavallo
University of Sannio, Italy
{gerardo.canfora, bice.cavallo}@unisannio.it

Abstract

In this paper we consider the problem of auditing databases which support statistical sum-queries to protect the security of sensitive information. We study the special case in which the domain of the sensitive information is a discrete set; in particular, we focus on a boolean domain. Principles and techniques developed for the security of statistical databases in the case of continuous attributes do not apply here. We provide a probabilistic framework for the on-line sum-auditing and we show that sum-queries can be audited by means of a Bayesian network. Finally, we provide a preliminary analysis of the usefulness of the probabilistic approach.

1. Introduction

The rapid growth of the Internet makes it easier than ever to collect data on a large scale. For example, the main task of the National Statistical Offices (NSO) is to collect information from individuals and organizations and disseminate this information to researchers, media, government agencies, businesses, and nonprofit organizations. In general, the information can include medical, voter registration, census, and customer data. For instance, references [8] and [14] contain public resources including population, economics, industry, and geography. Applications include studying the effects of treatments on disease, tracking disease outbreaks, allocating public funds, building economic models (from census data), and performing trend analysis.

Since such data sets also contain sensitive information, such as the disease of an individual, or the salary of an employee, it is needed to balance the collection and dissemination of data with the public expectation of privacy and the legal obligations. As an example, hospitals collect a large amount of data concerning the medical history of some population. Thus, it would be desirable to help advancing medical research by allowing statistical analysis of collected data. However, no medical information, that could be related to a specific patient, must be released. In order to preserve the privacy, while distributing statistical information, data collectors use Statistical Databases (SDB). These are database systems that enable users to retrieve only aggregate statistics (e.g., mean, sum, max, min, and count) for a subset of the entities represented in the database. Consider, for example, a company database containing the salaries of employees. A user may want to determine the sum of the salaries in a subset of records in the database. He/she cannot, however, be allowed to glean the salary of any one employee in particular.

A number of disclosure control methods to protect a SDB have been proposed in the literature (see [1] for a survey). We focus on auditing [6], [7], [9], [10], [11], [13], and particularly on the on-line auditing of sum-queries over discrete-valued data. On-line auditing entails that queries are answered one by one in sequence and an auditor has to determine whether the SDB is compromised when answering a new query.

Reference [6] considers the on-line sum, max, and mixed sum/max auditing problems. Both the on-line sum and the on-line max problem have efficient auditing algorithms. However, the mixed sum/max problem is NP-hard. References [2], [4], [5] and [9] deal with on-line max/min auditing.

Most of the work in this area assumes that the confidential data are real-valued and essentially unbounded (see [11]). In certain important applications, however, data may have discrete values, or have maximum or minimum values that are fixed a priori and frequently attainable. In these cases, traditional methods for maintaining security are inadequate.

Consider a dataset with m records; let us assume that there is a sensitive field X. For each $i \in \{1, \dots m\}$, the value x_i must not be disclosed. For instance, the system

$$\begin{cases} x_1 + x_2 + x_4 = 1, \\ x_2 + x_3 = 1, \\ x_1 + x_3 = 1, \end{cases}$$

is secure if the variables are real, but it is not secure if they are boolean, because in this case the values of all variables are determined. In reference [10], the authors study the

sum-auditing problem over boolean attributes and propose an algorithm that approximates the auditing problem.

Following references [2], [3], [4] and [5], the original contribution of this paper is fourfold:

1. to extend the classical notion of privacy to a probabilistic notion in the sum-auditing and to provide a formal base for probabilistic sum-auditing on a discrete domain;
2. to show how to deal with the on-line sum-auditing by means of a Bayesian network;
3. to manage user prior-knowledge in addition to the past history of the user sum-queries;
4. to provide, for the boolean case, the results of a preliminary set of experimental trials aimed at assessing the usefulness of the approach.

The paper is organized as follows: Section 2 introduces the notation and definitions useful in the next sections; Section 3 provides a formal base for probabilistic sum-auditing on a discrete domain; Section 4 proposes a Bayesian network for the on-line sum-auditing; Section 5 discusses a set of experiments aimed at assessing the usefulness of the probabilistic approach; finally, Section 6 provides concluding remarks and directions for future work.

2. Notation and preliminaries

We assume that:

- T is a dataset with m records;
- X is the sensitive field;
- D is the domain of the sensitive field X;
- a query $q = \{q_1, \ldots, q_m\}$ is represented by a sequence of size equal to m, with elements $q_j \in \{0, 1\}$ for each $j \in \{1, \ldots, m\}$. For instance, let assume $m = 5$, then $q = \{0, 1, 1, 0, 1\}$ encodes $x_2 + x_3 + x_5$;
- $J_q = \{j \in \{1, ..., m\} | q_j = 1\}$;
- $l = |J_q|$ is the length of the query q, that is the number of elements in q equal to 1. For instance, let assume $m = 5$ and $q = \{0, 1, 1, 0, 1\}$ then $J_q = \{2, 3, 5\}$ and $l = 3$;
- s is the answer to a sum-query q, that is $\sum_{j \in J_q} x_j = s$;
- $\Omega = \{(\ldots, x_j \ldots) | x_j \in D \; and \; j \in J_q\}$ the sample space;
- $S = \{(\ldots, x_j, \ldots) \in \Omega | \sum_{j \in J_q} x_j = s\}$.

We consider the following definition of probabilistic compromise:

Definition 1. *A privacy breach occurs if and only if a sensitive data is disclosed with probability greater or equal to a given tolerance probability tol. If a sensitive data is disclosed with $tol = 1$, then the SDB is fully compromised (see [5]).*

3. A probabilistic approach for on-line sum-auditing

In this section, we present a probabilistic approach to deal with sum-queries.

In Section 3.1, we assume that the domain of the sensitive field is the set of the non-negative integers, while in Section 3.2 the domain is the set of the first n non-negative integers; we focus on the boolean domain $D = \{0, 1\}$.

3.1 The sensitive values are non-negative integers

In this section, we assume that the user has no knowledge about the probability distribution of the sensitive field; thus each state of a sensitive data has the same probability. In the next sections, we consider the general and more realistic case in which the user has background knowledge about the probability distribution of the sensitive data.

Let $D = \{0, 1, \ldots, ...\}$ be the domain of the sensitive field X, $q = \{q_1, \ldots, q_m\}$ a sum-query, $|J_q| = l > 1$, then the following propositions hold:

Proposition 1. *The cardinality of S is:*

$$|S| = \binom{l-2}{0} + \binom{(l-2)+1}{1} + \ldots + \binom{(l-2)+s}{s}. \quad (1)$$

Proposition 2. *Let $k \in \{0, 1, \ldots, s\}$, then $\forall j \in J_q$, the following posterior probabilities hold:*

$$P(x_j = s - k | \sum_{j \in J_q} x_j = s) = \frac{\binom{(l-2)+k}{k}}{|S|}. \quad (2)$$

Example 1. *Let us assume $q = \{1, 1, 1, 0, \ldots 0\}$, that is $x_1 + x_2 + x_3 = 2$, then $l = 3$ and $s = 2$. We have:*

$$S = \{(2, 0, 0), (0, 2, 0), (0, 0, 2), (1, 1, 0), (1, 0, 1), (0, 1, 1)\}$$

and $|S| = 6$. Moreover, for $j \in J_q = \{1, 2, 3\}$:

$$P(x_j = 2 | \sum_{j \in J_q} x_j = 2) = \frac{\binom{1}{0}}{6} = \frac{1}{6},$$

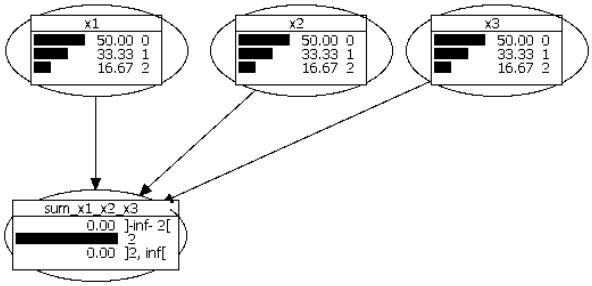

Figure 1. Probabilities computed in Example 1. $P(x_j = 2|\sum_{j\in J_q} x_j = 2) = \frac{1}{6}$, $P(x_j = 1|\sum_{j\in J_q} x_j = 2) = \frac{1}{3}$, $P(x_j = 0|\sum_{j\in J_q} x_j = 2) = \frac{1}{2}$

$$P(x_j = 1|\sum_{j\in J_q} x_j = 2) = \frac{\left(\begin{array}{c} 2 \\ 1 \end{array}\right)}{6} = \frac{1}{3},$$

$$P(x_j = 0|\sum_{j\in J_q} x_j = 2) = \frac{\left(\begin{array}{c} 3 \\ 2 \end{array}\right)}{6} = \frac{1}{2}.$$

3.2 Boolean sensitive values

Let $D = \{0, 1, \ldots, ...n-1\}$ be the domain of the sensitive field X. In this paper, we focus on the case $n = 2$.

As in reference [5], we propose a general approach suitable in the case the user has, or has not, background knowledge about the probability distribution of the sensitive data.

Let $D = \{0, 1\}$, $q = \{q_1, \ldots, q_m\}$ a sum-query, $|J_q| = l > 0$, then $\Omega = \{(\ldots, x_j \ldots)|j \in J_q \; and \; x_j = 0,\ 1\}$.

The following propositions hold:

Proposition 3. *Given $p \in [0, 1]$, the prior probabilities $P(x_j = 0) = 1 - p$ and $P(x_j = 1) = p$, then:*

$$P(\sum_{j\in J_q} x_j = s) = \left(\begin{array}{c} l \\ s \end{array}\right) \cdot p^s \cdot (1-p)^{l-s}.$$

In particular, if $p = \frac{1}{2}$ we have:

$$P(\sum_{j\in J_q} x_j = s) = \frac{\left(\begin{array}{c} l \\ s \end{array}\right)}{2^l}$$

Proof. It is enough to consider that we deal with the binomial distribution. □

Proposition 4. *For each x_j with $j \in J_q$, the posterior probabilities hold:*

$$P(x_j = 0|\sum_{j\in J_q} x_j = s) = 1 - \frac{s}{l} \tag{3}$$

$$P(x_j = 1|\sum_{j\in J_q} x_j = s) = \frac{s}{l} \tag{4}$$

Proof. For a fixed $\bar{j} \in J_q$, applying Bayes theorem and Proposition 3, we have that:

$$\begin{aligned} P(x_{\bar{j}} = 1|\sum_{j\in J_q} x_j = s) = \\ = &\frac{P(\sum_{j\in J_q} x_j = s|x_{\bar{j}} = 1) \cdot P(x_{\bar{j}} = 1)}{P(\sum_{j\in J_q} x_j = s)} = \\ = &\frac{P(\sum_{j\in J_q\setminus\{\bar{j}\}} x_j = s-1) \cdot P(x_{\bar{j}} = 1)}{P(\sum_{j\in J_q} x_j = s)} = \\ = &\frac{\left(\begin{array}{c} l-1 \\ s-1 \end{array}\right) \cdot p^{s-1} \cdot (1-p)^{(l-1)-(s-1)} \cdot p}{\left(\begin{array}{c} l \\ s \end{array}\right) \cdot p^s \cdot (1-p)^{l-s}} = \\ = &\frac{s}{l}. \end{aligned}$$

□

Example 2. *Let us assume $q = \{1, 1, 1, 1, 1, 1, 1, 0, \ldots 0\}$, that is*

$$x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 = 3,$$

then $l = 7$ and $s = 3$. If each state has the same probability $p = \frac{1}{2}$, we have:

$$P(\sum_{j=1}^{7} x_j = 3) = \frac{\left(\begin{array}{c} 7 \\ 3 \end{array}\right)}{2^7} = \frac{35}{128} = 0.2734, \tag{5}$$

else if, for instance, $p = \frac{1}{3}$ then:

$$P(\sum_{j=1}^{7} x_j = 3) = \left(\begin{array}{c} 7 \\ 3 \end{array}\right) \cdot (\frac{1}{3})^3 \cdot (\frac{2}{3})^4 = \frac{560}{2187} = 0.2561. \tag{6}$$

Moreover:

$$P(x_j = 0|\sum_{j=1}^{7} x_j = 3) = \frac{4}{7} = 0.5714, \tag{7}$$

$$P(x_j = 1|\sum_{j=1}^{7} x_j = 3) = \frac{3}{7} = 0.4286. \tag{8}$$

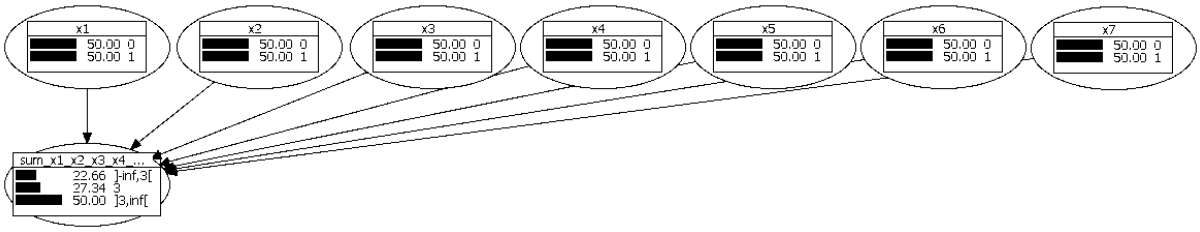

Figure 2. Probabilities computed in Example 2 before of the evidence on the sum-node, with $p = \frac{1}{2}$. $P(\sum_{j=1}^{7} x_j = 3) = 0.2734$

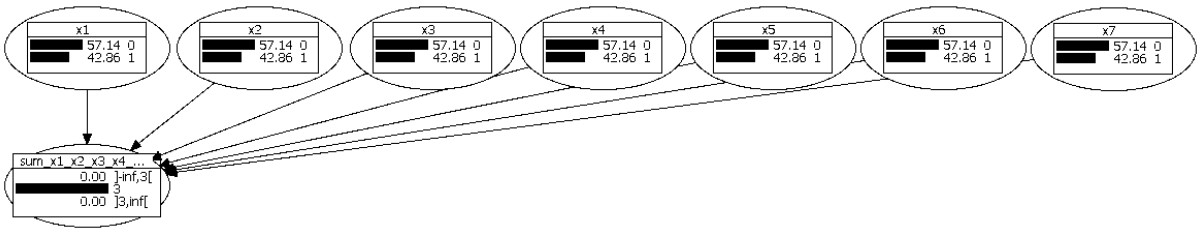

Figure 3. Probabilities computed in Example 2 after the evidence on the sum-node. $P(x_j = 0 | \sum_{j=1}^{7} x_j = 3) = 0.5714$**,** $P(x_j = 1 | \sum_{j=1}^{7} x_j = 3) = 0.4286$.

4. A Bayesian approach

In this section, we present a Bayesian network (BN) able to represent an user's uncertain knowledge after a sequence of sum-queries: in Section 4.1 we introduce the BNs; in Section 4.2 we show how a BN is able to compute all the probabilities and dependencies among variables described above for a one sum-query; in Section 4.3 we deal with the on-line sum-auditing.

4.1 Bayesian networks

A BN is a probabilistic graphical model that represents a set of variables and their probabilistic dependencies [12]. A BN, also called a belief net, is a directed acyclic graph (DAG), which consists of nodes to represent variables and arcs to represent dependencies between variables. Arcs, or links, also represent causal influences among the variables. The strength of an influence between variables is represented by the conditional probabilities which are summarized in a conditional probability table (CPT). If there is an arc from node *A* to another node *B*, *A* is called a parent of *B*, and *B* is a child of *A*. The set of parent nodes of a node X is denoted $parents(X)$. The size of the CPT of a node X depends on the number *s* of its states, the number *n* of $parents(X)$, and the number s_j of parent states, in the following way:

$$size(CPT) = s \cdot \prod_{j=1}^{n} s_j. \tag{9}$$

For every possible combination of parent states, there is an entry listed in the CPT. Thus, for a large number of parents the CPT will expand drastically. If the node X has no parents, its local probability distribution is said to be *unconditional*, otherwise it is *conditional*. If the value of a node is observed, then the node is said to be an *evidence* node.

In order to add prior knowledge on a BN we can add likelihood; adding likelihood is what we do when the user learns something about the state of the BN, which can be entered into a node. The simplest form is the evidence, that is, the probability that a state is 1 while the probability of each other states is 0. In general, likelihood has value in $[0, 1]$ and represents the probability of a state. Obviously, the sum of all probabilities is necessarily 1.

4.2 A Bayesian network for a sum query

Given D the domain of the sensitive field X, $q = \{q_1, \ldots, q_m\}$ a sum-query with $|J_q| = l > 1$, and s such that $\sum_{j \in J_q} x_j = s$, then we build the BN in the following way:

- in a first level there are nodes encoding the sensitive variables. If D is the domain in Section 3.1 then,

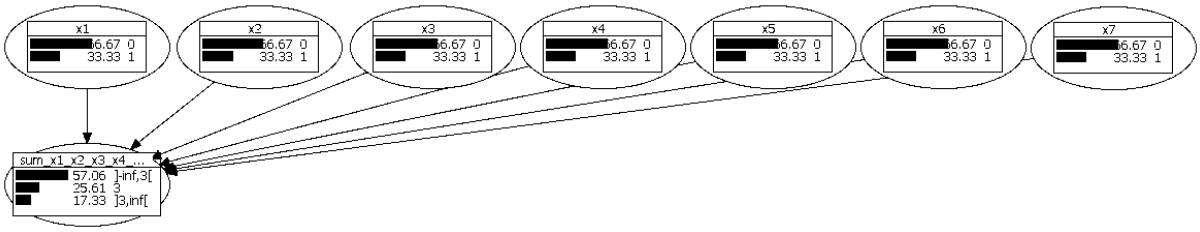

Figure 4. Probabilities computed in Example 2 before of the evidence on the sum-node, with $p = \frac{1}{3}$. $P(\sum_{j=1}^{7} x_j = 3) = 0.2561$

fixed $s_q = s$, these nodes have $s_q + 1$ states, that are $\{0, 1, 2, \ldots, s\}$, else if D is the domain in Section 3.2 then the nodes has $s_q + 1$ states with $s_q = min\{s, n-1\}$, so the states are $\{0, \ldots, s\}$ if $s_q = s$ and $\{0, \ldots, n-1\}$ if $s_q = n - 1$.

- in a second level there is a node encoding the sum $\sum_{j \in J_q} x_j$. This node is has three states: $]-\infty, s[$, s, $]s, +\infty[$.

Inserting evidence on the second state of the sum-node, the BN computes $P(x_j = k | \sum_{j \in J_q} x_j = s)$, for each $k \in \{0, \ldots, s_q\}$.
For instance, the probabilities in Example 1 are computed by means of the BN in Figure 1, the probabilities in (5) are computed by means of the BN in Figure 2 and the probabilities in (7) and in (8) by means of the BN in Figure 3.

4.2.1 Managing user background knowledge

In this section, we assume that the user knows the probability distribution of the sensitive field. For instance, we assume that the sensitive field X is such that $P(x_i = 0) = \frac{2}{3}$ and $P(x_i = 1) = \frac{1}{3}$, for each $i \in \{1, \ldots, m\}$. Thus, adding likelihood on the nodes x_i, the BN is able to capture this additional knowledge (see Figure 4); for the sum-node we obtain the same probability in (6). Inserting evidence on the sum-node, we obtain again the BN in Figure 3.
Finally, our model is able to capture also user prior knowledge about one or more sensitive values. We suppose that the user submits $x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7$ and he/she knows the prior probability $P(x_1 = 1) = 0.9$. Thus, it is enough to add likelihood only on the node encoding x_1.

4.3 A Bayesian network for the sum auditing

We build the BN for the on-line sum-auditing problem at run-time, that is we update the BN after each user query and decide whether or not to answer the query.

Example 3. *We continue the scenario described in Example 2 and we suppose that the user asks a second query* $q = \{0, 0, 1, 1, 1, 1, 0, 0, \ldots 0\}$*, that is* $x_3 + x_4 + x_5 + x_6$*. If the answer is* $s = 3$ *then the privacy is breached with probability equal to 1 (see Figure 5 a)) and the auditor denies the answer.*
If the answer is $s = 2$*, and the tolerance* tol *is greater than* 0.666*, then the privacy is not breached and the auditor provides the answer (see Figure 5 b)).*

5. Experimentation

An interesting question is to define a notion of usefulness for an auditing scheme; obviously if an auditor always denies, the privacy is never breached, but the user has not usefulness from the SDB. Intuitively, it seems that the more an auditing scheme denies, the less useful it is. On the other hand, the more an auditing scheme answers, the less secure it might be. Thus, we provide the results of a preliminary set of experimental trials aimed at assessing the usefulness of auditing system in terms of probability to deny respect to a fixed tolerance value (see Definition 1).

We use a boolean dataset with 300 records, and, for each tolerance value, we generate in a random way 150 different queries of length less than 6 and we assume that the prior probabilities are $P(x_i = 0) = P(x_i = 1) = \frac{1}{2}$. The results of the trials are shown in the following table:

Tolerance	Probability to deny
1	0.37
0.9	0.45
0.8	0.56
0.7	0.62
0.6	0.76
0.5	1

Of course, we have the maximum usefulness when the tolerance value is equal to 1, in fact the auditor denies the answer to a query if and only if the SDB is fully compromised (see

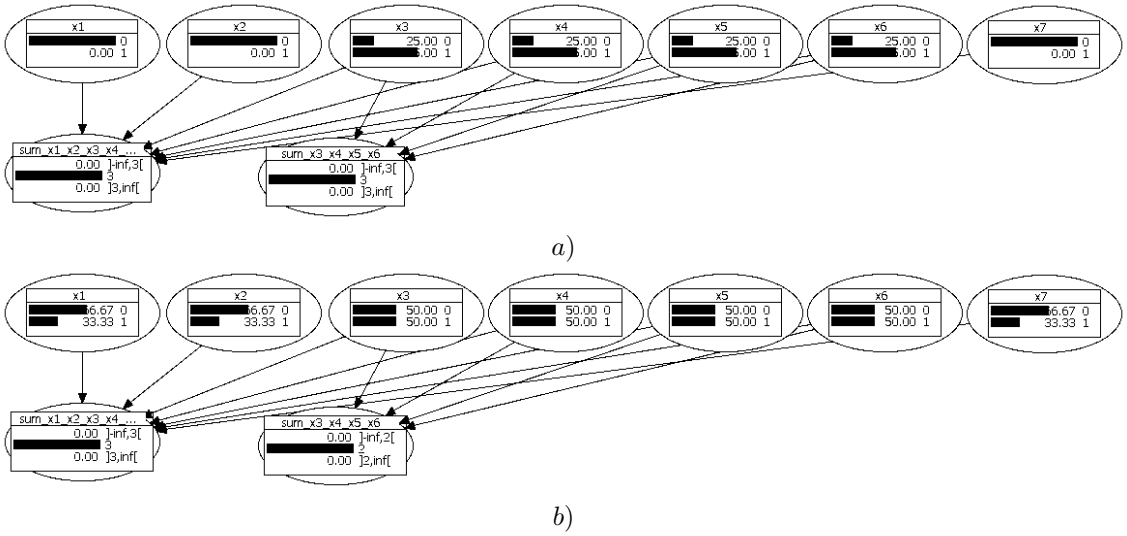

Figure 5. a) The privacy is breached with probability equal to 1. b) If $tol > 0.666$ then the privacy is not breached.

Definition 1). There is not usefulness if the tolerance value is equal to 0.5 because the auditor always denies.

6. Conclusions and future work

We have proposed a novel approach to reasoning under uncertainty in the on-line sum auditing of Statistical Databases. We have provided a formal base for probabilistic on-line sum auditing and we have proposed a Bayesian network as a disclosure tool based on probabilistic inferences that can be drawn from released data, focusing on boolean sensitive values.

Moreover, we have provided the results of a preliminary set of experimental trials aimed at assessing the usefulness of auditing system in terms of probability to deny.

Our future work will be to provide an exact analysis of usefulness for sum-auditing and to assess the scalability of the approach, in terms of complexity of time and memory requirements, and eventually to optimize the Bayesian model.

References

[1] N. R. Adam and J. C. Worthmann. Security-control methods for statistical databases: a comparative study. *ACM Computing Surveys (CSUR)*, Volume 21(4), December 1989.

[2] G. Canfora and B. Cavallo. A bayesian approach for on-line max and min auditing. In *Proocedings of International workshop on Privacy and Anonymity in Information Society (PAIS)*, pages 12–20. ACM DL, 2008.

[3] G. Canfora and B. Cavallo. A bayesian approach for on-line max auditing. In *Proocedings of The Third International Conference on Availability, Reliability and Security (ARES)*, pages 1020–1027. IEEE Computer Society Press, 2008.

[4] G. Canfora and B. Cavallo. Reasoning under uncertainty in on-line auditing. In *Privacy in Statistical Databases, Lecture Notes in Computer Science*, volume 5262, pages 257–269. Springer-Verlag Berlin Heidelberg, 2008.

[5] G. Canfora and B. Cavallo. A bayesian model for disclosure control in statistical databases. *Data & Knowledge Engineering*, 68(11):1187–1205, November 2009.

[6] F. Y. Chin. Security problems on inference control for sum, max, and min queries. *Journal of the ACM*, 33(3):451–464, July 1986.

[7] F. Y. Chin and G. Ozsoyoglu. Auditing and inference control in statistical databases. *IEEE Transaction on Software Engineering*, SE-8(6):574–582, November 1982.

[8] Istat. $http://www.istat.it/$. Last access: 26/09/2009.

[9] K. Kenthapadi, N. Mishra, and K. Nissim. Simulatable auditing. *In PODS*, pages 118–127, June 2005.

[10] J. Kleinberg, C. Papadimitriou, and P. Raghavan. Auditing boolean attributes. *Journal of Computer and System Sciences*, 66(1):244–253, February 2003.

[11] F. M. Malvestuto, M. Mezzini, and M. Moscarini. Auditing sum-queries to make a statistical database secure. *ACM Transactions on Information and System Security (TISSEC)*, 9(1):31–60, February 2006.

[12] J. Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann, San Francisco, CA, USA, 1998.

[13] S. P. Reiss. Security in databases: A combinatorial study. *Journal of the ACM*, 26(1):45–57, January 1979.

[14] U.S. $http://www.census.gov$. Census Bureau. Last access: 26/09/2009.