Quality & Quantity manuscript No.
(will be inserted by the editor)

Reasoning under Uncertainty and Multi-Criteria Decision Making in Data Privacy

Bice Cavallo · Gerardo Canfora · Livia D'Apuzzo · Massimo Squillante

Received: date / Accepted: date

Abstract By means of an integration of decision theory and probabilistic models, we explore and develop methods for improving data privacy. Our work encompasses disclosure control tools in statistical databases and privacy requirements prioritization; in particular we propose a Bayesian approach for the on-line auditing in Statistical Databases and Pairwise Comparison Matrices for privacy requirements prioritization. The first approach is illustrated by means of examples in the context of statistical analysis on the census and medical data, where no salary (resp. no medical information), that could be related to a specific employee (resp. patient), must be released; the second approach is illustrated by means of examples, such as an e-voting system and an e-banking service that have to satisfy privacy requirements in addition to functional and security ones.
Several fields in the social sciences, economics and engineering will benefit from the advances in this research area: e-voting, e-government, e-commerce, e-banking, e-health, cloud computing and risk management are a few examples of applications for the findings of this research.

Keywords Statistical Databases · Bayesian Networks · Multi-criteria Decision Support Methods · Pairwise Comparison Matrices

Bice Cavallo, Livia D'Apuzzo
Department of Architecture, University of Naples "Federico II", Via Toledo 402, 80134 Naples, Italy
E-mail: bice.cavallo@unina.it, liviadap@unina.it

Gerardo Canfora
Department of Engineering, University of Sannio, Viale Traiano 1, 82100 Benevento, Italy
E-mail: gerardo.canfora@unisannio.it

Massimo Squillante
Department of Economic, Legal and Social Studies, University of Sannio, Via delle Puglie 82, 82100 Benevento, Italy
E-mail: squillan@unisannio.it

1 Introduction

We are in an era in which a huge rate of information of physical, biological, environmental, social and economic systems is produced. Recording, accessing and disseminating this information affect in a crucial way the progress of knowledge and the productivity of economy.
Public opinion has shown a growing awareness of privacy issues over the last few years. High-profile losses of personal information and growing concerns about the nature and extent of personal information collected by organizations has led to a growing debate about the impact of ICT pervasiveness on privacy.
The goal of data privacy is to provide efficient and effective solutions for releasing data, while providing guarantees that the identities and other sensitive information of the individuals, who are the subjects of the data, are protected. While data security ensures that the user has the authority to receive a given piece of information, data privacy addresses disclosures based on inferences that can be drawn from released data.

The main task of the National Statistical Offices (NSO) is the collection of data and dissemination of this information to government, researchers, media, businesses, and no-profit organizations. 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, then it is needed to balance the collection and dissemination of data with the public expectation of privacy and the legal obligations. The lack of privacy undermines most of all other fundamental rights (e.g. freedom of speech and democracy); from a societal and legal point of view, an effort is needed to define and standardize technical concepts and methods regarding this issue.
Beyond law and ethics, there are also practical reasons for data collectors to care about confidentiality: unless users are convinced that their privacy is being adequately protected, they are unlikely to co-operate and supply their data for statistical analysis and for the research advancing.
Below, is a real-life example, taken from the literature, of how a privacy breach occurs.

Example 1 (Sweeney, 2002) "*The National Association of Health Data Organizations (NAHDO) reported that 37 states in the USA have legislative mandates to collect hospital level data and that 17 states have started collecting ambulatory care data from hospitals, physicians offices, clinics, and so forth (NAHDO, 1996). The leftmost circle in Figure 1 contains a subset of the fields of information, or attributes, that NAHDO recommends these states collect; these attributes include the patients ZIP code, birth date, gender, and ethnicity.*
In Massachusetts, the Group Insurance Commission (GIC) is responsible for purchasing health insurance for state employees. GIC collected patient specific data with nearly one hundred attributes per encounter along the lines of the those shown in the leftmost circle of Figure 1 for approximately 135,000 state employees and their families."

After removing fields containing name, address, social security number, and other

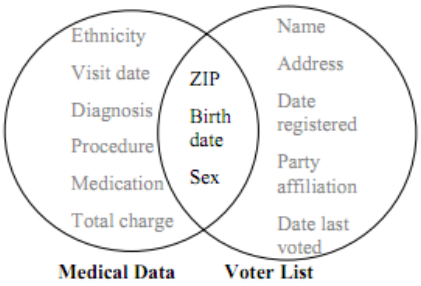

Fig. 1 Governor's health records are disclosed (Figure from (Sweeney, 2002))

explicit identifiers, GIC assumed it protected patient privacy, and gave a copy of the data to researchers and sold a copy to industry (GIC, 1997).

"*For twenty dollars, I purchased the voter registration list for Cambridge Massachusetts and received the information on two diskettes (Cambridge, 1997). The rightmost circle in Figure 1 shows that these data included the name, address, ZIP code, birth date, and gender of each voter. This information can be linked using ZIP code, birth date and gender to the medical information, thereby linking diagnosis, procedures, and medications to particularly named individuals. For example, William Weld was governor of Massachusetts at that time and his medical records were in the GIC data. Governor Weld lived in Cambridge Massachusetts. According to the Cambridge Voter list, six people had his particular birth date; only three of them were men; and, he was the only one in his 5-digit ZIP code.*"

Thus, Sweeney (2002) disclosed the Governor's health records (which included diagnosis and prescriptions).

In a world where information can uncover new possibilities, research in data privacy is essential. IBM Privacy Research Institute and UNESCO Chair in Data Privacy are two example of institutions that promote research in privacy:

- IBM Privacy Research Institute (IBM, 2001) is an organization within IBM Research to promote and advance research in privacy and data protection technology. His goal is to develop technologies for enterprises to conduct e-business in privacy-enabling ways. The institute's research focuses on technologies for commercial applications, particularly for e-business;
- UNESCO Chair in Data Privacy (CRISES, 2007) does research, training and dissemination in a field considered relevant by UNESCO for the welfare of humankind (data privacy). Chair general objectives are: train public administrations of transition countries in privacy-preserving techniques and legal aspects of privacy; promote the adoption of privacy-preserving technologies by the private sectors (IT manufacturers, banks, telecom operators, etc.); raise the public awareness of the need, the right and the preserving the individual privacy in the information society; promote cross-national research on privacy technologies; exchange master and doctoral students on privacy between transition and developed countries.

Our research work includes new paradigms for:

- Disclosure control tools in Statistical Databases (SDBs). A SDB system enables its users to retrieve only aggregate statistics (e.g., min, max, sum, mean and count) for a subset of the entities represented in the database. SDBs are useful for statistical analysis and building models. We use Bayesian models (Pearl, 1998) for addressing disclosures based on probabilistic inferences that can be drawn from released data (Canfora and Cavallo, 2008a,b,c, 2009a,b, 2010; Cavallo and Canfora, 2012);
- Privacy requirements prioritization. Data privacy plays a central role in many modern applications (e.g. social network, e-voting, cloud computing); thus it is needed to plan and design a system with privacy in mind. We use a generalization (Cavallo and D'Apuzzo, 2009a, 2010, 2012a,b; Cavallo et al, 2009, 2010, 2012) of the Analytic Hierarchy Process (Saaty, 1977, 1980, 1986, 1988) for privacy requirement prioritization.

The research idea, proposed by Cavallo (2011), received several acknowledgements and awards as TR35-GI (2011), organized by Technology Review, journal of Massachusetts Institute of Technology (MIT), and 'Forum della Ricerca Innovazione Imprenditorialità'(RIEForum).

The paper is organized as follows: Section 2 proposes a Bayesian model for reasoning under uncertainty in SDBs; Section 3 describes a multi-criteria method for privacy requirements prioritization; Section 4 provides concluding remarks and directions for future work.

2 Disclosure Control tools in Statistical Databases

Privacy is an important issue in the statistical analysis of human-related data.

Example 2 For checking whether in a certain geographic area, there is a gender-based discrimination, we can use the census data to check, e.g., whether for all people from this area who have the same level of education, there is a correlation between salary and gender. One can think of numerous possible questions of this type related to different sociological, political, medical, economic, and other questions. From this viewpoint, it is desirable to give researchers the tools to perform statistical analysis for their specific research. On the other hand, we do not want to give them direct access to the raw census data, because a large part of the census data is sensitive. For example, for most people, salary information is sensitive.

A SDB system enables its users to retrieve only aggregate statistics (e.g., min, max, sum, mean and count) for a subset of the entities represented in the database.

Example 3 Let us consider a dataset with attributes $(name, age, salary)$ supporting statistical queries of the form "give me the maximum of salaries of all individuals whose age x satisfies condition $C(x)$", where C is an arbitrary predicate on the domain of age, such as $30 \leq x \leq 40$. Assume further that the pair $(name, age)$ is publicly available, but the attribute salary is confidential. What measures suffice to protect the confidentiality of the salary information?

This is the classical statistical database security problem, studied extensively since the 1970's; Adam and Wortmann (1989) provide a comprehensive survey. Several methods for protecting privacy in SDBs have been presented in the literature.

These methods are classified under three general approaches: data perturbation, output perturbation and query restriction.

Data perturbation approach introduces noise in the data. The original database is perturbed and this modified database is then made available to researchers (Matloff, 1986; Traub et al, 1984; Warner, 1965).

Output perturbation approach evaluates the query on the original data, but returns a perturbed version of the answer (Beck, 1980; Blum et al, 2005; Dalenius, 1981; Denning, 1980; Dinur and Nissim, 2003).

The techniques in the *query restriction* category specify the set of queries that should not be answered to ensure that privacy is not breached. None of the answers to legal queries are perturbed (Fellegi, 1972; Schlorer, 1975; Denning et al, 1979; Dobkin et al, 1979).

An example of query restriction is the *auditing* (Chin, 1986; Kenthapadi et al, 2005; Kleinberg et al, 2003; Malvestuto et al, 2006; Nabar et al, 2006). Two kinds of auditing have been studied (Chin, 1986):

- *On-line auditing*. The queries are answered one by one in sequence and the auditor has to determine whether the SDB is compromised when answering a new query;
- *Off-line auditing*. Given a set of queries $\{q_1, ..., q_t\}$, the auditor has to identify a maximum subset of queries that can be answered simultaneously without compromising the SDB.

We focus on the on-line max-min and sum auditing, and assume that:

- T is a dataset with n records;
- y_i, for $i \in \{1, \ldots, n\}$, is a sensitive data of the i-th individual in T.

With a max, min or sum query, a user asks the max or min sensitive value, or the sum of sensitive values, respectively, in a subset of records in the dataset. In the on-line auditing problem, given a set of statistical queries $\{q_1, q_2, ..., q_{t-1}\}$, the corresponding answers $\{m_1, m_2, ..., m_{t-1}\}$ and the current query q_t, the auditor provides the answer to q_t if and only if there is not a privacy breach.

Example 4 Let us assume that an user submits the max query $q_1 = max\{y_1, y_2, y_3\}$ (e.g. the maximum of salaries of three individuals) and the auditor provides the answer $m_1 = 8$. Then, if the user submits the max query $q_2 = max\{y_1, y_2\}$, with answer $m_2 = 6$, then the auditor has to deny the answer m_2.

We propose a probabilistic definition of privacy breach, so that a privacy breach occurs if and only if a sensitive value is disclosed with probability greater than or equal to a given tolerance probability tol. Thus, given the answers $m_1, \ldots, m_t$ to the queries $q_1, \ldots, q_t$, and a prior-knowledge α about the sensitive data, the privacy is breached if and only if the user knows a sensitive value y_i with probability greater than or equal to tol, that is:

$$Pr(y_i = s | m_1, \ldots, m_t, \alpha) \geq tol. \quad (1)$$

We address disclosures based on probabilistic inferences that can be drawn from released data and manage user prior-knowledge, in addition to released data, by means of a Bayesian Network (BN).

2.1 Bayesian networks

A BN is a probabilistic graphical model that represents a set of variables and their probabilistic dependencies (Pearl, 1998). 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:

$$size(CPT) = s \cdot \prod_{j=1}^{n} s_j. \tag{2}$$

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.

A key application of a BN is the computation of posterior probabilities of the form $P(x|\epsilon)$, where, in general, ϵ is evidence (i.e., information) received from external sources about the (possible) states/values of a subset of the variables of the network. For a set E of discrete evidence variables, the evidence appears in the form of a likelihood distribution over the states of E; also often called an evidence function for E.

In the next section, we will describe a suitable BN able to deal with the on-line auditing; by inserting evidence on the nodes encoding the answers to the queries, we will compute the probability of a privacy breach in equation (1).

Independence of causal influence (ICI) (Srinivas, 1993) among local parent-child or cause-effect relationship allows for further factoring. ICI has been used to reduce the complexity of knowledge acquisition. The size of conditional distribution that encodes the max (or min) operator can be reduced when the n-ary max (resp. min) operator is decomposed into a set of binary max (resp. min) operators. Two well known approaches to the decomposition are: parent divorcing (Olesen et al, 1989) and temporal transformation (Heckerman, 1993). Parent divorcing constructs a binary tree in which each node encodes a binary operator. Temporal transformation constructs a linear decomposition tree in which each node encodes a binary operator.

2.2 Using Bayesian networks for on-line max-min and sum auditing

In the literature, on-line max-min auditing (combination of max and min queries) has been addressed with some restrictive assumptions, primarily that sensitive values must be all distinct and the sensitive field has an uniform distribution. Canfora and Cavallo (2009b) propose a Bayesian approach able to:

- remove these limitations;

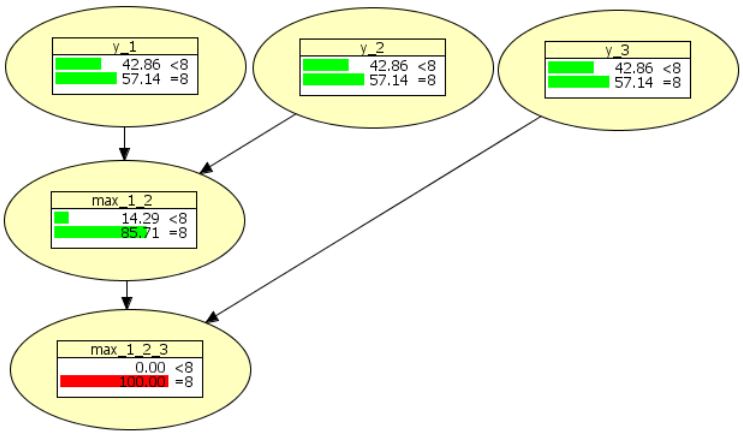

Fig. 2 Temporal transformation for the max-query $max\{y_1, y_2, y_3\}$. For each $i = 1, 2, 3$, $Pr(y_i = 8|max\{y_1, y_2, y_3\} = 8) = \frac{4}{7} = 0.5714$.

Table 1 Table with a duplicated sensitive value.

EMPLOYEE	Alice	Bob	Carl	David	Evelyn
Z (*1000 €)	9	8	8	5	4

- deal with max-min auditing;
- manage the implicit delivery of information that derives from denying the answer to a query;
- provide a graphical representation of user knowledge;
- capture user background knowledge.

In the approach, a BN is able to deal with the on-line auditing: indeed, by means of inference on the BN, the auditor decides whether or not to answer the query. Both the structure and the CPTs of the BN are updated at run-time, that is after each user query; for reducing the CPT size of the BN, each query is represented by means of a temporal transformation.

Example 5 The temporal transformation for the max query $max\{y_1, y_2, y_3\}$, with answer equal to 8, is represented in Figure 2. Thus, if the auditor provides the answer to the query, then each sensitive data is disclosed with probability equal to 0.5714.

A unique aspect of our work is that the BN is built dynamically as a new query is submitted; thus, after adding the nodes for the current query, we insert evidence and provide the answer if and only if the privacy is not breached.

Example 6 Let Table 1 be a dataset containing salaries (Z) of 5 employees. Then, the sequence of the max-min queries $max\{z_2, z_4\}$, $max\{z_3, z_5\}$, $min\{z_2, z_5\}$ is modelled by means of the BN in Figure 3; we stress, for instance, that Carl's salary is disclosed with probability equal to 0.8571.

Canfora and Cavallo (2009a) deal with the on-line max auditing in dynamic databases. A static database is one that never changes after it has been created.

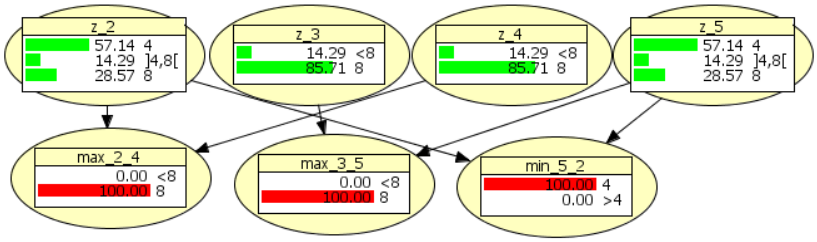

Fig. 3 BN for the sequence: $q_1 = max\{z_2, z_4\}$, $q_2 = max\{z_3, z_5\}$, $q_3 = min\{z_2, z_5\}$. In absence of user background knowledge, $Pr(z_3 = 8|m_1 = 8, m_2 = 8, m_3 = 4) = Pr(z_4 = 8|m_1 = 8, m_2 = 8, m_3 = 4) = 0.8571$; thus, if *tol* is chosen smaller or equal to this value then the privacy is breached.

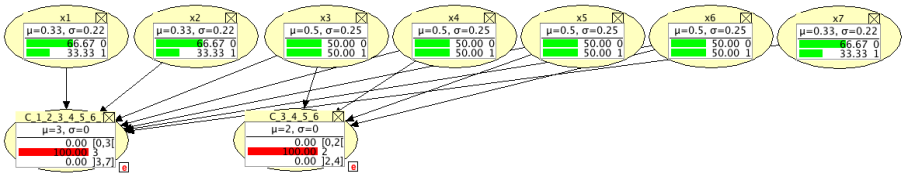

Fig. 4 $P(x_i|\sum_{i=1}^{7} x_i = 3,\ \sum_{i=3}^{6} x_i = 2)$

Most census are static: whenever a new version of the database is created, the new version is considered to be another static database. A dynamic databases can change over time. This feature can complicate the privacy problem considerably, because frequent releases of new versions may enable users to make use of the differences among the versions in ways that are difficult to foresee.

Canfora and Cavallo (2010) consider the problem of auditing databases which support statistical sum-queries to protect the security of sensitive information and focus on the case in which the domain of the sensitive information is a discrete set, in particular a boolean domain. Principles and techniques developed for the privacy of SDBs in the case of continuous attributes do not apply here. A probabilistic framework is proposed for the on-line sum-auditing and the sum-queries are audited by means of a BN; Cavallo and Canfora (2012) propose an optimization of this BN, based on parent divorcing or temporal transformations, in such way that the CPT size of the BN for a sum-query of length l decreases from the complexity $O(2^l)$ to $O(l^3)$.

Example 7 Let us consider a hospital that collects data from patients and provides patients' information to an external medical center; it would be desirable to help advance medical research by allowing statistical analysis of collected data, while no medical information, that could be related to a specific patient, must be released. Let us assume that $x_i = 1$ encodes "HIV=YES"for the i-th patient (resp. $x_i = 0$ encodes "HIV=NO"); of course x_i is a sensitive data. Since the domain of the sensitive field is the boolean one, a sum query is equivalent to a count query. If the medical center submits the sum/count queries $\sum_{i=1}^{7} x_i$ and $\sum_{i=3}^{6} x_i$, with answers 3 (i.e. 3 patients with HIV) and 2 (i.e. 2 patients with HIV) respectively, the BN, encoding medical center's knowledge, is shown in Figure 4. Thus, if the tolerance value *tol* is chosen greater than 0.6666 then the privacy is not breached.

The BN, proposed for dealing with the on-line auditing, is able to manage with several kinds of user background knowledge (e.g. additional knowledge about one or more sensitive values, sensitive field with known probability distribution); thus we will be able to deal with privacy problems in several social contexts.

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, in (Canfora and Cavallo, 2009b, 2010), a preliminary set of experimental trials is performed; it is aimed at assessing the usefulness of auditing system in terms of probability to deny with respect to a fixed tolerance value, to a fixed number of queries and with respect to the queries length. However an exact analysis of usefulness of our auditing system is an open problem and we forecast the use of a suitable metric and the utility theory for evaluating the trade-off between usefulness and risk of the Bayesian approach.

3 Privacy requirements prioritization

Data privacy plays a central role in many modern applications; thus it is needed to plan and design a system with privacy in mind. In the same way in which software functional requirements are analysed up-front to make sure that the planned system will meet user needs in terms of services being delivered, privacy requirements must be defined up-front to satisfy the needs of the customers and to comply with laws, standards and service policies.

Capturing and modelling privacy requirements in the early stages of system development is essential to provide high assurance of privacy protection to both stakeholders and consumers. A mechanism is needed to model privacy requirements and user privacy preferences in a systematic way so that privacy policies can be enforced in the software system.

What happens if engineers fail to develop adequate privacy requirements? Requirements influence an overall implementation. Therefore, lacking adequate privacy requirements, the system may not prevent violations of human rights, such as leaking personal information. Legislative and economic penalties for such violations are more serious these days.

Example 8 Let us consider an e-voting system. A fundamental objective for democratic elections is secrecy of the vote. It requires that only the voter knows his voting decision and nobody else is able to gain information about it. Thus, the system has to satisfy privacy requirements in addition to functional and security ones. The privacy requirements for this system include:

– the privacy of the vote has to be guaranteed during the casting, transfer, reception, collection, and tabulation of vote. For instance, in the TCP/IP protocol suite, an Internet packet carries the IP addresses of the source and the destination machines. In e-voting systems the awareness by the destination machine of the IP address of the source machine, may compromise the anonymity requirement. Even if a client sends a vote without any identifying information

the identity of the voter can be extracted from the IP address. Voting protocols seek to overcome this problem by implementing an anonymous channel whereby a server can reliably and securely receive messages but cannot determine the identity of the sending machine;
– none of the participants involved in the voting process (organizers, election officials, trusted third parties, voters, etc) should be able to link a vote to an identifiable voter.

3.1 Pairwise Comparison Matrices

Related to the phase of privacy requirements prioritization, different structured techniques can be used. As proposed by Bijwe and Mead (2010), we will use Pairwise Comparison Matrices (PCMs); this method allows us to:

– apply a pairwise comparison to assess the preference intensity of a requirement over an other;
– determine the priority of privacy requirements;
– measure the consistency of the preferences.

More exactly, let $X = \{x_1, x_2, ..., x_n\}$ be a set of alternatives or requirements, a Decision Maker(DM) may state his/her preferences, for the set X, by means of a preference relation

$$\mathcal{A} : (x_i, x_j) \in X \times X \mapsto \mathcal{A}(x_i, x_j) = a_{ij} \in \mathbb{R},$$

where a_{ij} represents the preference intensity of x_i over x_j. The preference relation is represented by the PCM:

$$A = \begin{pmatrix} a_{11} & a_{12} & ... & a_{1n} \\ a_{21} & a_{22} & ... & a_{2n} \\ ... & ... & ... & ... \\ a_{n1} & a_{n2} & ... & a_{nn} \end{pmatrix}. \tag{3}$$

A condition of reciprocity is assumed for a PCM in such way that the intensity of preference of x_i over x_j, expressed by the entry a_{ij}, can be exactly read by means of the element a_{ji}.
In literature, several kinds of PCMs are proposed, thus the entry a_{ij} assumes different meanings:

– in multiplicative PCMs, a_{ij} represents a preference ratio in $]0, +\infty[$ and the reciprocity is expressed by $a_{ji} = \frac{1}{a_{ij}}$, for each $i, j \in \{1, \ldots, n\}$;
– in additive PCMs, a_{ij} is a preference difference in $]-\infty, +\infty[$ and the reciprocity is expressed by $a_{ji} = -a_{ij}$;
– in fuzzy PCMs, a_{ij} is a preference degree in [0,1] and the reciprocity is expressed by $a_{ji} = 1 - a_{ij}$.

In an ideal situation, in which the DM is strongly coherent when stating his/her preferences, the PCM satisfies the consistency property, that, in the multiplicative case, is expressed as follows:

$$a_{ik} = a_{ij} \cdot a_{jk} \quad \forall\, i, j, k = 1, \ldots, n. \tag{4}$$

Under condition of consistency for $A = (a_{ij})$, the preference value a_{ij} can be expressed by means of the components of a suitable vector, called consistent vector for $A = (a_{ij})$; for a multiplicative PCM, it is a positive vector $\underline{w} = (w_1, w_2, ..., w_n)$ verifying the condition

$$\frac{w_i}{w_j} = a_{ij} \quad \forall\, i, j = 1, \ldots, n.$$

Thus, if $A = (a_{ij})$ is a consistent PCM, then it is reasonable to choose a priority vector in the set of consistent vectors, while if $A = (a_{ij})$ is an inconsistent PCM then we look for a vector that is close to be a consistent vector, for the multiplicative case:

$$\frac{w_i}{w_j} \approx a_{ij} \quad \forall\, i, j = 1, \ldots, n.$$

In literature, the notions of consistency and consistent vector are provided for additive and fuzzy cases too.
Multiplicative PCMs play a basic role in the Analytic Hierarchy Process (AHP) (Saaty, 1980), it is a structured technique for dealing with complex decisions. Rather than prescribing an ideally correct decision, the AHP helps the DMs to find the one that best suits their needs and their understanding of the problem. The AHP provides a comprehensive and rational framework for structuring a decision problem, for representing and quantifying its elements, for relating those elements to overall goals, and for evaluating alternative solutions.

To make a decision in a systematic way, we need to decompose the decision into the following steps (Saaty, 2008):

1. *"Define the problem and determine the kind of knowledge sought.*
2. *Structure the decision hierarchy from the top with the goal of the decision, then the objectives from a broad perspective, through the intermediate levels (criteria on which subsequent elements depend) to the lowest level (which usually is a set of the alternatives).*
3. *Construct a set of pairwise comparison matrices. Each element in an upper level is used to compare the elements in the level immediately below with respect to it.*
4. *Use the priorities obtained from the comparisons to weigh the priorities in the level immediately below. Do this for every element. Then for each element in the level below add its weighed values and obtain its overall or global priority. Continue this process of weighing and adding until the final priorities of the alternatives in the bottom most level are obtained."*

AHP is used around the world in a wide variety of decision situations, in fields such as government, business, industry, healthcare and education; recently, Saaty and Zoffer (2011) propose AHP for negotiating the Israeli-Palestinian Controversy.

In the last twenty years many features of Saaty's AHP have been criticised (e.g. Barzilai, 1998):

- the assumption of the Saaty scale $S = \{\frac{1}{9}, \frac{1}{8}, \frac{1}{7}, \frac{1}{6}, \frac{1}{5}, \frac{1}{4}, \frac{1}{3}, \frac{1}{2}, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$ restricts the DM's possibility to be coherent (consistent): for instance, if $a_{12} = 4$, that is the privacy requirement x_1 is preferred 4 times to privacy requirement x_2, and $a_{23} = 3$, that is the privacy requirement x_2 is preferred 3 times to privacy requirement x_3, then the DM is coherent if and only if

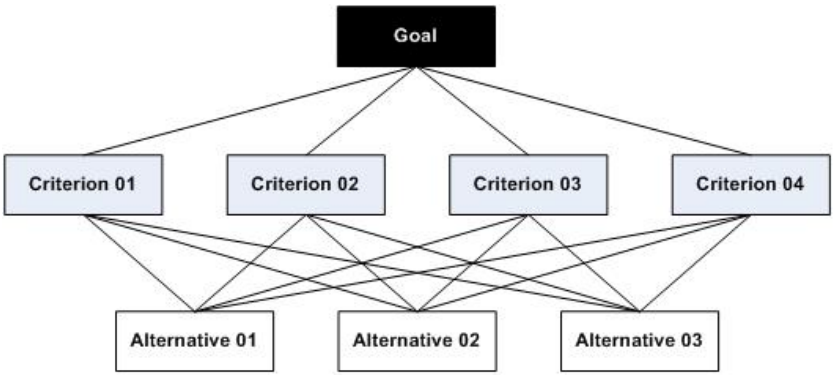

Fig. 5 Example of hierarchy (Figure from (Vargas, 2010)).

$a_{13} = 12$, that is the privacy requirement x_1 is preferred 12 times to privacy requirement x_3, but as 12 is not in the Saaty scale, the DM will not be coherent. The assumption of any limited and closed set of values presents the same shortcoming (e.g. the interval $[0, 1]$ for fuzzy PCMs);

- a measure of closeness to the consistency for a PCM has been provided by Saaty in terms of the principal eigenvalue λ_{max}. This measure has been questioned because it is not easy to compute, has not a simple and geometric meaning and, in some cases, seems to be unfair;
- the right eigenvector associated to λ_{max} has been considered as a prioritization vector, unfortunately it does not satisfy the independence of scale-inversion condition. So it may be that if the DM expresses the preference values respect to the privacy requirements over a scale or over its inverse scale, then he/she may obtain two different prioritizations;
- the conventional AHP leads to the possibility of occurrence of the rank reversal phenomenon (adding or deleting an alternative or criteria may cause a reversal in the ranking of the old ones). Since the set of the privacy requirements is often not fixed ex-ante ante but is variable and is constructed in accordance with reasons of relevance, simplicity and privacy laws, the application of AHP is potentially flawed.

Thus, Cavallo and D'Apuzzo (2009a, 2010, 2012a) and Cavallo et al (2012) propose PCMs over an abelian linearly ordered group $\mathcal{G} = (G, \odot, \leq)$; this approach generalizes multiplicative, additive and fuzzy approach and is able to remove the first three shortcomings of the classical AHP; currently, we are investigating the rank reversal phenomenon. In detail, the notion of $\odot$-consistency expressed by means of the group operation $\odot$:

$$a_{ik} = a_{ij} \odot a_{jk} \qquad \forall i, j, k = 1, \ldots, n, \tag{5}$$

allows us to remove the first drawback; we stress that equation (5) is equal to equation (4) if $\odot$ is the usual multiplication. Under the assumption of divisibility of $\mathcal{G}$, for each $A = (a_{ij})$, a consistency measure $I_{\mathcal{G}}(A)$, expressed in terms of $\odot$-mean of $\mathcal{G}$-distances, is provided and a $\odot$-mean vector $\underline{w}_m(A) = (w_1, \ldots w_n)$, satisfying the independence of scale-inversion condition, is associated to A.
A link between $\underline{w}_m(A)$ and $I_{\mathcal{G}}(A)$ gives more validity to $I_{\mathcal{G}}(A)$ and more meaning

to $\underline{w}_m(A)$; in fact, it ensures that if $I_{\mathcal{G}}(A)$ is close to the identity element then, from a side A is close to be a consistent PCM and from the other side $\underline{w}_m(A)$ is close to be a consistent vector; thus, it can be chosen as a priority vector for the alternatives. Others properties of $\underline{w}_m(A)$ and $I_{\mathcal{G}}(A)$ are described by Cavallo and D'Apuzzo (2012a) and Cavallo et al (2012), respectively.

Example 9 Let us consider a bank that has to develop an e-banking service for allowing his customers to conduct financial transactions on a secure website. The bank will be directly responsible for the safety and privacy of the e-banking service. The system will have to satisfy privacy requirements in addition to functional and security ones, such as:

x_1 clearly articulate the level of customer privacy and at what extent his/her information (e.g. salary, credit cards number, electronic funds transfer) will be exposed internally within the bank (e.g., operator, database administrator, manager);
x_2 a customer must be given the option of not giving their personally identifiable information if the information collected is not related to the primary purpose for which the information was collected;
x_3 the customers choice about personally identifiable information being disclosed to third parties must be honoured. The customer must also have the means to change their choice;
x_4 provide information about how personally identifiable information collected by the site.

Each website should satisfy x_2, x_3 and x_4 (PrivacyTrust, 2002).

Let us suppose that a bank's DM expresses his/her preference intensities by means of the following multiplicative PCM:

$$A = \begin{pmatrix} 1 & 2 & 4 & 5 \\ \frac{1}{2} & 1 & 2 & \frac{5}{2} \\ \frac{1}{4} & \frac{1}{2} & 1 & \frac{5}{4} \\ \frac{1}{5} & \frac{2}{5} & \frac{4}{5} & 1 \end{pmatrix};$$

we stress, for instance, that $a_{14} = 5$ means that the DM prefers the privacy requirement x_1 5 times with respect to x_4.

Since A is a consistent PCM (i.e. $a_{ik} = a_{ij} \cdot a_{jk}\ \forall i, j, k = 1, \ldots, 4$), the DM is perfectly coherent when expressing his/her preferences.
The $\odot$-mean vector (i.e. its i-th component w_i is the geometric mean of the elements of $i - th$ row of A)

$$\underline{w}_m(A) = (\sqrt[4]{40}, \sqrt[4]{\frac{5}{2}}, \sqrt[4]{\frac{5}{32}}, \sqrt[4]{\frac{8}{125}})$$

provides a privacy requirements prioritization: indeed, by $w_1 > w_2 > w_3 > w_4$, we have $x_1 \succ x_2 \succ x_3 \succ x_4$, where $x_i \succ x_j$ means that x_i is strictly preferred to x_j; moreover w_i represents the weight of the privacy requirement x_i.
We stress that $\underline{w}_m(A)$ is a consistent vector (i.e. $\frac{w_i}{w_j} = a_{ij}, \quad \forall i, j = 1, \ldots, 4$).

Example 10 Let us consider the e-voting system in Example 8 and let $\{x_1, x_2, x_3\}$ be a set of privacy requirements (e.g. privacy of the vote during the transfer, collection and tabulation of the vote) and

$$B = \begin{pmatrix} 1 & 2 & 7 \\ \frac{1}{2} & 1 & 3 \\ \frac{1}{7} & \frac{1}{3} & 1 \end{pmatrix}$$

a multiplicative PCM encoding the preference intensities of the privacy requirements. Then, B is an inconsistent PCM (e.g. $b_{13} \neq b_{12} \cdot b_{23}$) and

$$\underline{w}_m(B) = (\sqrt[3]{14}, \sqrt[3]{\frac{3}{2}}, \sqrt[3]{\frac{1}{21}}).$$

Since B is inconsistent, the ratio $\frac{w_i}{w_j}$ gives only an approximation of b_{ij} (e.g. $\frac{w_1}{w_2} = 2.105 \approx b_{12} = 2$ and $\frac{w_2}{w_3} = 3.16 \approx b_{23} = 3$).

To assess the effectiveness of our approach for privacy requirements prioritization, we propose the application of the proposed methodology to different contexts, such as e-banking, e-government, e-voting and cloud computing services.

4 Conclusions and future work

We propose new paradigms for disclosure control in statistical databases and privacy requirements prioritization by means of techniques such as Bayesian networks and pairwise comparison matrices.
Several fields in the social sciences, economics and engineering will benefit from the advances in this research area: e-voting, e-government, e-commerce, e-banking, e-health, cloud computing and risk management are a few examples of applications for the findings of this research.
With regard to disclosure control tools in statistical databases, there are several directions for further investigation. A first goal is to quantify the usefulness of the auditing scheme. For usefulness, we mean the probability to deny with respect to a fixed tolerance value, to a fixed number of queries and with respect to the queries length. As discussed in Section 2.2, the exact analysis of usefulness for the auditing system is an open problem. Finally, we are exploring combinations of different statistical queries (sum, mean, count, etc.) and we are modelling the collision of multiple users in the system; currently, the system builds a BN for each user, and this could entail scalability problem for large applications. Our long-term research goal is to use a BN as a unifying framework including the interactions among the various domains of uncertainty that affect query auditing in statistical databases.

Our future work will be directed also to investigate, in the general framework of pairwise comparison matrices over a divisible alo-group, the following topics:

- to look for conditions weaker than consistency that allow us to identify the actual qualitative ranking on a set of criteria, alternatives or requirements. Then, the problem will be to find a vector agreeing with this ranking (Basile and D'Apuzzo, 2002; Cavallo and D'Apuzzo, 2009b);

– to define the decision procedure of the AHP;
– to deal with the rank reversal phenomenon.

To assess the effectiveness of this approach for privacy requirements prioritization, we propose the application of the proposed methodology to different contexts, such as e-banking, e-government, e-voting and cloud computing services. However, others multi-criteria techniques will be experimented.
Moreover our future research tasks include, but are not limited to new paradigms for:

– Privacy-Preserving Record Linkage. The privacy problem becomes more complex when different archives, managed by different organizations, must be crossed through techniques of record linkage. We will address the limitations of the current approaches by proposing novel methods, for instance a combination of cryptographic techniques and anonymization methods will be able to evaluate the trade-off of the solutions along three dimensions: privacy, cost and accuracy;
– Privacy metrics. These metrics measure the achieved privacy protection. The task includes the study of inter-temporal metrics, measuring the utility and risk as a function of the time horizon, and the extension of privacy metrics to unstructured data (e.g. police reports, court decisions).

While the above topics have been presented separately, there are obvious connections among them. For example, graphical models, such as the Bayesian networks, are useful in Privacy-Preserving Record Linkage. Another example involves the use of a suitable metric to evaluate the trade-off between utility and risk of the Bayesian approach.

Finally, a business plan has been drawn up for planning research and development activities of an enterprise acting in the field of data privacy.

References

Adam NR, Wortmann JC (1989) Security-control methods for statistical databases: A comparative study. ACM Comput Surv 21(4):515–556

Barzilai J (1998) Consistency measures for pairwise comparison matrices. J Multi-Crit Decis Anal 7:123–132

Basile L, D'Apuzzo L (2002) Weak consistency and quasi-linear means imply the actual ranking. Int J of Uncert, Fuzziness and Knowledge-Based Systems 10(3):227–239

Beck LL (1980) A security mechanism for statistical databases. ACM Trans Database Syst 5(3):316–338

Bijwe A, Mead NR (2010) Adapting the square process for privacy requirements engineering. Tech. rep., Software Engineering Institute, Carnegie Mellon University

Blum A, Dwork C, Mcsherry F, Nissim K (2005) Practical privacy: The sulq framework. In: Proceedings of the International Conference on Principles of Data Systems (PODS)

Cambridge (1997) Cambridge voters list database City of Cambridge, Massachusetts.

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

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

Canfora G, Cavallo B (2008c) Reasoning under uncertainty in on-line auditing. In: Privacy in Statistical Databases, Lecture Notes in Computer Science, Springer-Verlag Berlin Heidelberg, vol 5262, pp 257–269

Canfora G, Cavallo B (2009a) A bayesian approach for on-line max auditing of dynamic statistical databases. In: EDBT/ICDT Workshops, pp 107–116

Canfora G, Cavallo B (2009b) A bayesian model for disclosure control in statistical databases. Data Knowl Eng 68(11):1187–1205

Canfora G, Cavallo B (2010) A probabilistic approach for on-line sum-auditing. In: ARES, pp 303–308

Cavallo B (2011) Metodi, modelli e tecnologie per la data privacy. ZeroUno (353):90–93

Cavallo B, Canfora G (2012) A bayesian approach for on-line sum/count/max/min auditing on boolean data. In: Privacy in Statistical Databases, Lecture Notes in Computer Science, Springer-Verlag Berlin Heidelberg, pp 295–307

Cavallo B, D'Apuzzo L (2009a) A general unified framework for pairwise comparison matrices in multicriterial methods. Int J Intell Syst 24(4):377–398

Cavallo B, D'Apuzzo L (2009b) Recent advances in Applied Mathematics, WSEAS, chap Transitive pairwise comparison matrices over abelian linearly ordered groups, pp 207–212

Cavallo B, D'Apuzzo L (2010) Characterizations of consistent pairwise comparison matrices over abelian linearly ordered groups. Int J Intell Syst 25(10):1035–1059

Cavallo B, D'Apuzzo L (2012a) Deriving weights from a pairwise comparison matrix over an alo-group. Soft Computing 16(2):353–366

Cavallo B, D'Apuzzo L (2012b) Investigating properties of the $\odot$-consistency index. In: IPMU (4), pp 315–327

Cavallo B, D'Apuzzo L, Squillante M (2009) Building consistent pairwise comparison matrices over abelian linearly ordered groups. In: ADT, pp 237–248

Cavallo B, DApuzzo L, Marcarelli G (2010) Pairwise comparison matrices: Some issue on consistency and a new consistency index. In: Greco S, Marques Pereira R, Squillante M, Yager R, Kacprzyk J (eds) Preferences and Decisions, Studies in Fuzziness and Soft Computing, vol 257, Springer Berlin / Heidelberg, pp 111–122

Cavallo B, D'Apuzzo L, Squillante M (2012) About a consistency index for pairwise comparison matrices over a divisible alo-group. Int J Intell Syst 27(2):153–175

Chin FY (1986) Security problems on inference control for sum, max, and min queries. J ACM 33(3):451–464

CRISES (2007) Unesco chair in data privacy. `http://unescoprivacychair.urv.cat/presentacio.php`, Accessed 20 December 2012

Dalenius T (1981) A simple procedure for controlled rounding. Statistik Tidskrift 3:202–208

Denning DE (1980) Secure statistical databases with random sample queries. ACM Trans Datab Syst 5(3):291–315

Denning DE, Denning PJ, Schwartz MD (1979) The tracker: A threat to statistical database security. ACM Trans Datab Syst 4(1):76–96

Dinur I, Nissim K (2003) Revealing information while preserving privacy. In: Proceedings of the International Conference on Principles of Data Systems (PODS), pp 202–210

Dobkin DP, Jones AK, Lipton RJ (1979) Secure databases: Protection against user influence. ACM Trans Datab Syst 4(1):76–96

Fellegi IP (1972) On the question of statistical confidentiality. J Amer Statis Asso 67(337):7–18

GIC (1997) Group insurance commission testimony before the massachusetts health care committee See Session of the Joint Committee on Health Care, Massachusetts State Legislature

Heckerman D (1993) Causal independence for knowledge acquisition and inference. In: Proceedings of Ninth Conference on Uncertainty in Artificial Intelligence, pp 122–127

IBM (2001) Ibm privacy research institute. `http://www.research.ibm.com/privacy/`, Accessed 20 December 2012

Kenthapadi K, Mishra N, Nissim K (2005) Simulatable auditing. In: PODS, pp 118–127

Kleinberg J, Papadimitriou C, Raghavan P (2003) Auditing boolean attributes. J of Computer and Syst Sciences 66(1):244–253

Malvestuto FM, Mezzini M, Moscarini M (2006) Auditing sum-queries to make a statistical database secure. ACM Trans on Inf and Syst Security 9(1):31–60

Matloff NE (1986) Another look at the use of noise addition for database security. In: Proceedings of IEEE Symposium on Security and Privacy, pp 173–180

Nabar SU, Marthi B, Kenthapadi K, Mishra N, Motwani R (2006) Towards robustness in query auditing. In: International Conference on Very Large Data Bases, pp 151–162

NAHDO (1996) A guide to state-level ambulatory care data collection activities Falls Church: National Association of Health Data Organizations (NAHDO)

Olesen K, Kjaerulff U, Jensen F, Jensen F, Falck B, Andreassen S, Andersen S (1989) A munin network for the median nerve–a case study in loops. Appl Artificial Intelligence 3(2–3):385–403

Pearl J (1998) Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann, San Francisco, CA, USA

PrivacyTrust (2002) Privacy policy requirements. `http://www.privacytrust.org/certification/privacy/privacy_requirements.html`, Accessed 20 December 2012

Saaty TL (1977) A scaling method for priorities in hierarchical structures. J Math-Psychology 15:234–281

Saaty TL (1980) The Analytic Hierarchy Process. McGraw-Hill, New York

Saaty TL (1986) Axiomatic foundation of the analytic hierarchy process. Management Science 32(7):841–855

Saaty TL (1988) Decision Making for Leaders. University of Pittsburgh

Saaty TL (2008) Decision making with the analytic hierarchy process. Int J Services Sciences 1(1):83–98

Saaty TL, Zoffer HJ (2011) Negotiating the israeli-palestinian controversy from a new perspective. Int J of Inf Technology & Decision Making 10(1):5–64

Schlorer J (1975) Identification and retrieval of personal records from a statistical bank. Methods Inform Medicine

Srinivas S (1993) A generalization of the noise-or-model. In: Ninth Annual Conference of Uncertainty on AI, pp 208–218

Sweeney L (2002) k-anonymity: A model for protecting privacy. Int J of Uncert, Fuzziness and Knowledge-Based Systems 10(5):557–570

TR35-GI (2011) Protezione della privacy dei dati mediante reti bayesiane. Technology Review, edizione italiana (2):10

Traub JF, Yemini Y, Wozniakowski H (1984) The statistical security of a statistical database. ACM Trans Database Syst 9(4):672–679

Vargas RV (2010) Using the analytic hierarchy process (ahp) to select and prioritize projects in a portfolio. PMI Global Congress 2010 North America `http://www.ricardo-vargas.com/wp-content/uploads/downloads/articles/ricardo_vargas_ahp_project_selection_en.pdf`, Accessed 20 December 2012

Warner SL (1965) Randomized response: A survey technique for eliminating evasive answer bias. J Am Stat Assoc 60(309):63–69