Quantifying Privacy Leakage Through Answering Database Queries

Tsan-sheng Hsu, Churn-Jung Liau, Da-Wei Wang, and Jeremy K.-P. Chen

Institute of Information Science
Academia Sinica, Taipei, Taiwan
{tshsu,liaucj,wdw}@iis.sinica.edu.tw

Abstract. We assume a database consists of records of individuals with private or sensitive fields. Queries on the distribution of a sensitive field within a selected population in the database can be submitted to the data center. The answers to the queries leak private information of individuals though no identification information is provided. Inspired by decision theory, we present a quantitative model for the privacy protection problem in such a database query or linkage environment in this paper. In the model, the value of information is estimated from the viewpoint of the querier.

To estimate the value, we define the information state of the data user by a class of probability distributions on the set of possible confidential values. We further define the usefulness of information based on how easy the data user can locate individuals that fit the description given in the queries. These states and the usefulness of information can be modified and refined by the user's knowledge acquisition actions. The value of information is then defined as the expected gain of the privacy receiver and the privacy is protected by imposing costs on the answers of the queries for balancing the gain.

Key words: Privacy, Data table, Decision logic, Quantitative model, Value of information.

1 Introduction

There are many technical problems to be addressed for privacy protection. The most basic one is to avoid unauthorized users access to the confidential information. Some significant works on controlling disclosure of private information from databases have been done[4–6, 8–10, 15, 22, 23]. Recently, an epistemic logic has been proposed to model the privacy protection problem in a database linking context[11]. In [2], a prototype system is designed to implement this logical model in the context of querying a medical database. The safety criteria of the data is defined rigorously in the logic and data to be disclosed must be generalized to meet this requirement. The safety criteria defined in the requirement are purely qualitative so we can only identify the situation in which the exact confidential information came to the users' knowledge. However, in many cases, even if the private information is only known with some imprecision, there is still

a risk of privacy leakage. Therefore it is very important to have the capability of risk assessment in the model of privacy protection.

Someone may benefit from the privacy leakage, but it may also be harmful for others. For example, the health information of a customer would be valuable in the decision-making of an insurance company. However, the dissemination of an individual's health information without his consent in advance is definitely an invasion of his privacy. Thus the value of confidential information would be an incentive towards invasion of privacy. The information brokers may try to collect and sell personal information for their own interest. On the other hand, it is usually difficult to estimate the damage caused by privacy leakage. However, to discourage the invasion of privacy, the damage of the victim must be appropriately compensated by the one disseminating the information. Therefore, the evaluation of gain and loss of privacy leakage is a crucial problem in privacy protection.

In this paper, we try to tackle the problem from the aspects of information value. We focus on the following database query environment. In a data center, private information about individuals are collected. There are private or sensitive fields as well as identification fields in each record. Queries on the distribution of a sensitive field within a selected population in the database can be submitted to the data center. The answers to the queries leak private information of individuals though no identification information is provided.

We study a quantitative model for the privacy protection problem in such a database query environment. It is for modeling the value of information from the viewpoint of the querier. We will model the value of information as the expected gain of knowledge of the information. In the model, we need to represent the knowledge states of an user receiving some kind of information. We further define the usefulness of information based on how easy the data user can locate individuals that fit the description given in the queries. The knowledge states and the usefulness of information can be changed or refined by receiving some answer to the user's query. Thus we also need a formalism to represent the data to be protected and a language to describe which kinds of queries are allowed. The data table and decision logic proposed in [17] will be employed as the data representation formalism and the query language respectively.

In the rest of the paper, we first review the data table formalism and the decision logic in our application context. The basic components of our models—the information states and knowledge acquisition actions—is defined in section 3. In section 4, the model for information value and its use in privacy protection are presented. Finally, the results are summarized in the concluding section.

2 Data Representation and Query Language

To state the privacy protection problem, we must first fix the data representation. The most popular data representation is by data table([17]). The data in many application domains, for example, medical records, financial transactions, employee data, etc., can be represented as data tables. A data table can be seen

as a simplification of a relational database, since the latter in general consists of a number of data tables. A formal definition of data table is given in [17].

Definition 1 *A data table*[1] *is a pair* $T = (U, A)$ *such that*

- U *is a nonempty finite set of individuals, called the population or the universe,*
- A *is a nonempty finite set of primitive attributes, and*
- *every primitive attribute* $a \in A$ *is a total function* $a : U \to V_a$*, where* V_a *is the set of values of* a*, called the domain of* a*.*

The attributes of a data table can be divided into three sets. The first contains the *key attributes*, which can be used to identify to whom a data record belongs, therefore these attributes are always masked off in response to a query. Since the key attributes uniquely determine the individuals, we can assume that they are associated with elements in the universe U and omit them from this point. Second, we have a set of *easy-to-know attributes*, the values of which are easily discovered by the public. For example, in [21], it is pointed out that some attributes like birth-date, gender, ethnicity, etc., are included in some public databases such as census data or voter registration lists. The last kind of attributes is the *confidential type*, the values of which are mainly the goals we have to protect. Sometimes, there is an asymmetry between the values of a confidential attribute. For example, if the attribute is a HIV test result, the revelation of a '+' value may cause a serious privacy invasion, whereas it does not matter to know that an individual has a '−' value. For simplification, we assume there is exactly one confidential attribute in a data table. Thus a data table is usually written as $T = (U, A \cup \{c\})$ where A is the set of easy-to-know attributes and c is the confidential one.

Let $V_c = \{s_0, s_1, \ldots, s_{t-1}\}$ be the set of possible values for the confidential attribute c. It is assumed that the *a prior* information of the user is the probability distribution of the population on V_c. In other words, we assume that the user knows the value $\frac{|\{u \in U | c(u) = s_i\}|}{|U|}$ for all $0 \leq i \leq t-1$. Then the user can improve his knowledge by investigating some sampled individuals of the population or querying the data center that stores the data table. By investigation, the user can discover the exact value of the confidential attribute of the chosen individuals. However, much effort is necessary to do the investigation. On the other hand, a query may ask for the probability distribution of sensitive fields in a specific subset of the population. Once the query is correctly answered, the user not only knows the probability distribution of the specific sub-population, but also that of its complement on V_c. Thus we need a language to specify a subset of individuals. To achieve this purpose, we suggest to use the decision logic(DL) proposed in [17]. The DL is originally designed for the representation of rules induced from a data table by data mining techniques. However, it is also perfectly suitable for the query of a data table since each formula of the logic is satisfied by some individuals in the data table.

[1] Also called knowledge representation system, information system, or attribute-value system

Syntactically, an atomic formula for the data table $T = (U, A \cup \{c\})$ is of the form (a, v), where $a \in A$ is an easy-to-know attribute and $v \in V_a$ is a possible value of the attribute a. The well-formed formulas (wff) of the logic is closed under the Boolean connectives negation($\neg$), conjunction($\wedge$), disjunction($\vee$), and implication($\rightarrow$). For the semantics, an individual $u \in U$ satisfies an atomic formula (a, v), written as $u \models_T (a, v)$ iff $a(u) = v$. Intuitively, any individual satisfying (a, v) has v as the value of his attribute a. The satisfaction of other wffs can then be defined recursively as usual.

From the semantics of decision logic, we define the truth set of a wff φ with respect to the data table T, denoted by $|\varphi|_T$, as $\{u \in U \mid u \models_T \varphi\}$. Thus each wff φ specifies a subset of individuals $|\varphi|_T$ in the data table. When a query φ is submitted by an user to the data center, this means this user wants to know the distribution of the sub-population $|\varphi|_T$ on V_c. If the query is correctly answered, the user would also simultaneously know the distribution of the sub-population $U - |\varphi|_T$ by the axioms of probability. In other words, a correctly answered query would partition the population into two sub-populations and the distributions thereof on the confidential attribute values are known respectively. In this way, the user can subsequently query the data center to refine his knowledge regarding the distributions of the different sub-populations on the confidential attribute values. To model the evolution of the user's information after different queries, we need a formal representation of user's information states. The next section will be devoted to the definitions of such representation.

3 The Information States

From here on, let us fix a data table $T = (U, A \cup \{c\})$. Let $V_c = \{s_0, s_1, \ldots, s_{t-1}\}$ be the set of possible values for the confidential attribute and let $U = \{u_1, \ldots, u_n\}$ be the set of individuals. A *logical partition* of U is a subset of DL wffs $\Pi = \{\varphi_1, \varphi_2, \ldots, \varphi_m\}$ such that $|\varphi_1|_T \cup \cdots \cup |\varphi_m|_T = U$ and $|\varphi_i|_T \cap |\varphi_j|_T = \emptyset$ if $i \neq j$. Each $|\varphi_i|_T$ is called an equivalence class of Π. A piece of information (or knowledge) known to the user is given by a logical partition of U, a set of probability distributions indexed by the wffs of the partition, and the number of investigated individuals. In the following, we use $|\varphi|$ to denote the cardinality of $|\varphi|_T$.

Definition 2 *An information state (or a knowledge state) $\mathcal{I}$ for the set of possible private attribute values V_c and the set of individuals U is a triplet*

$$(\Pi, (\mu_i)_{0 \leq i \leq t-1}, (\kappa_i)_{0 \leq i \leq t-1})$$

where Π is a logical partition on U and for all $0 \leq i \leq t-1$, $\mu_i : \Pi \rightarrow [0, 1]$ and $\kappa_i : \Pi \rightarrow \mathcal{N}$ ($\mathcal{N}$ denotes the set of natural number) are functions satisfying the following constraints for any $\varphi \in \Pi$,

(i) $\sum_{i=0}^{t-1} \mu_i(\varphi) = 1$,
(ii) $|\varphi| \cdot \mu_i(\varphi)$ *is a natural number, and*

(iii) $\kappa_i(\varphi) \leq |\varphi| \cdot \mu_i(\varphi)$

For ease of description, we use the vector notations in denoting μ_i's and κ_i's. Thus $\boldsymbol{\mu} = (\mu_0, \ldots, \mu_{t-1})$ and $\boldsymbol{\kappa} = (\kappa_0, \ldots, \kappa_{t-1})$ denotes vector mappings which can be applied to elements of Π and the result of such application is a vector consisting of the results of applying its component functions to the element. The dimension of each vector will be self-evident from the context and not explicitly specified. By the vector notation, an information state defined above can be denoted by $(\Pi, \boldsymbol{\mu}, \boldsymbol{\kappa})$. Let $\mathcal{I}$ be such an information state, then $(\Pi, \boldsymbol{\mu})$ is called a *partial knowledge state* compatible with $\mathcal{I}$. Note that a partial knowledge state may be compatible with various information states.

Within an information state, the user partitions the population into a number of subpopulations. He knows the probability distribution of each subpopulation on the confidential attribute values. Intuitively, $\mu_i(\varphi)$ is the proportion of the individuals in sub-population $|\varphi|_T$ which have confidential attribute value s_i, whereas $\kappa_i(\varphi)$ is the number of investigated individuals in sub-population $|\varphi|_T$ which have confidential attribute value s_i. Since each DL wff φ is composed from atomic formulas with easy-to-know attributes only, it can be assumed that it takes little effort for the user to verify whether a given individual satisfies φ. Furthermore, it can also be assumed that the cardinality of the truth set of each φ is known to the public. However, note that it may sometimes be very difficult for the user to locate an individual satisfying a specific φ from the whole population U.

The information states of an user can be subsequently changed by his investigation of some individuals in a specific sup-population and by his queries posed to and the answers obtained from the data center. This is a process of knowledge refinement and can be modeled by the knowledge acquisition actions as follows.

A logical partition Π_2 is a refinement of another logical partition Π_1, denoted by $\Pi_2 \sqsubseteq \Pi_1$, if for all $\varphi_2 \in \Pi_2$, there exists $\varphi_1 \in \Pi_1$ such that $|\varphi_2|_T \subseteq |\varphi_1|_T$. It is clear that if $\Pi_2 \sqsubseteq \Pi_1$, then each $|\varphi_1|_T$ such that $\varphi_1 \in \Pi_1$ can be written as a union of the truth sets of some wffs in Π_2.

Definition 3 *Let $\mathcal{I}_1 = (\Pi_1, \boldsymbol{\mu_1}, \boldsymbol{\kappa_1})$ and $\mathcal{I}_2 = (\Pi_2, \boldsymbol{\mu_2}, \boldsymbol{\kappa_2})$ be two information states. $\mathcal{I}_2$ is a refinement of $\mathcal{I}_1$, also denoted by $\mathcal{I}_2 \sqsubseteq \mathcal{I}_1$, if both of the following conditions are satisfied:*

1. $\Pi_2 \sqsubseteq \Pi_1$.
2. *For each $\varphi \in \Pi_1$, if $|\varphi|_T = \bigcup_{1 \leq i \leq l} |\varphi_i|_T$ for some set $\{\varphi_1, \ldots, \varphi_l\} \subseteq \Pi_2$, then*

$$|\varphi| \cdot \boldsymbol{\mu_1}(\varphi) = \sum_{i=1}^{l} |\varphi_i| \cdot \boldsymbol{\mu_2}(\varphi_i),$$

and

$$\boldsymbol{\kappa_1}(\varphi) \leq \sum_{i=1}^{l} \boldsymbol{\kappa_2}(\varphi_i).$$

Note that the arithmetics (addition and multiplication) and comparison between vectors (and scalars) are defined as usual. For example, the addition of two vectors is carried out point-wise and results in a vector of the same dimension.

In our framework, there are two kinds of knowledge acquisition actions which can refine the user's information states. The first one is the query action. Each query action is represented by a wff φ in DL. The intended answer of the query is the distribution of the confidential values within the selected population $|\varphi|_T$ in the database. The other is the investigation action, which is specified by a wff φ and a positive integer number k. This means that the user have investigated k individuals from the set $|\varphi|_T$ in this action. For the uniformity of the representation, each knowledge acquisition action is written as $\alpha = (\varphi, k)$ for some DL wff φ and $k \geq 0$. When $k > 0$, it is an investigation action, whereas it is a query one if $k = 0$.

Definition 4 *1. A knowledge acquisition action $(\varphi, 0)$ is applicable under the information state $\mathcal{I}_1 = (\Pi_1, \boldsymbol{\mu_1}, \boldsymbol{\kappa_1})$ and results in a state $\mathcal{I}_2 = (\Pi_2, \boldsymbol{\mu_2}, \boldsymbol{\kappa_2})$ if*

(a) there exists $\varphi' \in \Pi_1$ such that $|\varphi|_T \subseteq |\varphi'|_T$,
(b) $\Pi_2 = \Pi_1 - \{\varphi'\} \cup \{\varphi, \varphi' \wedge \neg\varphi\}$,
(c) $\mathcal{I}_2$ is a refinement of $\mathcal{I}_1$,
(d) $\boldsymbol{\kappa_2}(\psi) = \boldsymbol{\kappa_1}(\psi)$ for any $\psi \in \Pi_1 - \{\varphi'\}$, and
(e) $\boldsymbol{\kappa_2}(\varphi) + \boldsymbol{\kappa_2}(\varphi' \wedge \neg\varphi) = \boldsymbol{\kappa_1}(\varphi')$.

2. A knowledge acquisition action (φ, k) where $k > 0$ is applicable under the information state $\mathcal{I}_1 = (\Pi_1, \boldsymbol{\mu_1}, \boldsymbol{\kappa_1})$, and $\mathcal{I}_2 = (\Pi_2, \boldsymbol{\mu_2}, \boldsymbol{\kappa_2})$ is a resultant state of the application if

(a) $\varphi \in \Pi_1$ and $k \leq |\varphi| - \sum_{i=0}^{t-1} \kappa_i(\varphi)$
(b) $\Pi_1 = \Pi_2$,
(c) $\boldsymbol{\mu_1} = \boldsymbol{\mu_2}$,
(d) $\boldsymbol{\kappa_2}(\psi) = \boldsymbol{\kappa_1}(\psi)$ for any $\psi \neq \varphi$, and
(e) $\sum_{i=0}^{t-1} \kappa_{2i}(\varphi) = \sum_{i=0}^{t-1} \kappa_{1i}(\varphi) + k$.

Since the goal of the user is to refine his knowledge by the queries, a rational user would pose his queries so that his knowledge would be improved by the answers of the queries. Thus if the user's information state is $(\Pi_1, \boldsymbol{\mu_1}, \boldsymbol{\kappa_1})$, then he poses a query about a subset of an equivalence class in Π_1. This is the requirement of Condition 1a in Definition 4. Then, after the query is answered, the corresponding equivalence class is partitioned into two parts — one satisfying φ and the other not, so we have the Condition 1b in Definition 4. Condition 1c in Definition 4 further requires that the answer is correct so that the resultant information state is a refinement of the original one. Furthermore, since the query action does not cause any new individuals being investigated, the $\boldsymbol{\kappa_2}$ function agrees with $\boldsymbol{\kappa_1}$ in the part of the population which is not split by the query, while for the split part, the number of investigated individuals is not changed in total. This is reflected respectively in Conditions 1d and 1e of the definition.

In the case of investigation action, we assume the user will only investigate the individuals in a sub-population represented by a wff in Π_1. The assumption

is inessential, since, if the investigated individuals are across some different subpopulations, the corresponding investigation action can be decomposed into a sequence of actions satisfying the applicability condition. Since it is assumed that the user knows the total number of individuals in $|\varphi|_T$ and those which have been investigated by him so far is equal to $\sum_{i=0}^{t-1} \kappa_i(\varphi)$, he would not try to investigate more individuals than all remaining ones. This is exactly required by the applicability condition of Definition 4.2a. Conditions 2b to 2d are obvious since these values are not affected by the investigation. What the investigation can affect is the total number of the investigated individuals in $|\varphi|_T$ and this is reflected in Condition 2e.

4 The Value of Information

To quantitatively determine the value of information, we must have a user model. Let us consider the case where the user is an agent trying to use the private information to aid his decision in a game. The game is played between the agent and individuals in the population U. The agent can decide the rate he want to charge an individual for playing the game (i.e., the admission fee). The rate is decided on a personalized basis so that each individual may be charged with different rates. However, once an individual agrees to play the game with the agent and pay the fee asked by the agent, he will have a chance to get back some reward which will be the loss of the agent. The reward of an individual is determined by his confidential attribute value. Let r_i denote the reward of an individual with the confidential attribute value s_i for $0 \leq i \leq t-1$, then $\boldsymbol{\rho} = (r_0, r_1, \ldots, r_{t-1}) \in \Re^t$ is called the loss vector of the agent.

Let $\mathcal{I}_0 = (\{\top\}, \boldsymbol{\mu_0}, \boldsymbol{\kappa_0})$ be the initial information state of the user, where $\top$ denotes any tautology in the DL and $\boldsymbol{\kappa_0}(\varphi) = (0, \ldots, 0)$. Let $\boldsymbol{\rho}$ be a given loss vector. The agent first decides the *base rate* of the game on the expected loss according to his initial information state, i.e., $R_0 = \boldsymbol{\rho} \cdot \boldsymbol{\mu_0}(\top)$. Thus, in the initial state, the expected payoff of the agent for playing the game is zero. However, once he acquires pieces of information and reaches a new information state, he can utilize the acquired information for making some profit.

We further assume that each individual will go into the game if he is charged with the base rate. However, he can refuse to do so if the agent charges him with a rate higher than the base one. The higher the rate, the more likely the individual refuses to play the game. If the information state is $\mathcal{I} = (\Pi, \boldsymbol{\mu}, \boldsymbol{\kappa})$, where $\Pi = \{\varphi_1, \ldots, \varphi_m\}$, a reasonable decision of the agent for the rate of an individual u satisfying φ is as follows:

1. if u has been investigated and it is known that the confidential attribute value of u is s_i, then the most profitable decision of the agent would be to charge the individual with $\max(R_o, r_i)$ so that the agent's payoff is $\max(R_o - r_i, 0)$;
2. if the individual has not been investigated, the agent knows the probability of the confidential attribute value of u being s_i to be

$$p_i(\varphi) = \frac{|\varphi| \cdot \mu_i(\varphi) - \kappa_i(\varphi)}{|\varphi| - \sum_{i=0}^{t-1} \kappa_i(\varphi)}. \tag{1}$$

In this case, the most reasonable decision of the agent would be to charge the individual with $\max(R_o, \sum_{i=0}^{t-1} p_i(\varphi) \cdot r_i)$ so that the agent's expected payoff would be $\max(R_o - \sum_{i=0}^{t-1} p_i(\varphi) \cdot r_i, 0)$

Thus, in average, the agent can have the following expected payoff B_φ in playing the game with an individual satisfying φ:

$$\begin{aligned} B_\varphi = {} & \max(R_o - \textstyle\sum_{i=0}^{t-1}(p_i(\varphi) \cdot r_i), 0) \cdot \frac{|\varphi| - \sum_{i=0}^{t-1} \kappa_i(\varphi)}{|\varphi|} \\ & + \textstyle\sum_{i=0}^{t-1} \max(R_o - r_i, 0) \cdot \frac{\kappa_i(\varphi)}{|\varphi|} \end{aligned} \tag{2}$$

Thus, by using the knowledge about the individuals' confidential attributes, the agent can raise the rates of those who may incur a greater loss to him in order to avoid the possible loss. The value of the information is then dependent on how much he can benefit from obtaining the information. The expected gain of the agent with regard to each individual is computed by

$$B_{\mathcal{I}} = \sum_{\varphi \in \Pi} B_\varphi \cdot \frac{|\varphi|}{|U|},$$

if he decides the rates according to the two principles above.

Example 1 The scenario described above usually occurs between an insurance company and its customers. The base rate is applied to a typical customer if the company does not have any further information about his health condition. However, for the customers of high risk, the company would raise their rates. Thus the health information of the customers would be valuable to the insurance company. To avoid the leakage of privacy, the data center may correspondingly raise the cost of answering a query so that the information value for the company is counter-balanced. The company would not have the incentive to obtain the private information. ■

The notions of the value of information have been extensively studied in decision theory[7, 14]. In our model above, if investigation actions are not allowed, all information states are of the form $(\Pi, \boldsymbol{\mu}, \boldsymbol{\kappa_0})$, so $\kappa_{0i}(\varphi) = 0$ and $p_i(\varphi) = \mu_i(\varphi)$ for all $0 \leq i \leq t-1$ and $\varphi \in \Pi$. Consequently, $B_{\mathcal{I}}$ would be simplified into

$$\sum_{\varphi \in \Pi} \max(R_o - \boldsymbol{\mu}(\varphi) \cdot \boldsymbol{\rho}, 0) \cdot \frac{|\varphi|}{|U|}$$

which is the value of partial information defined in [14] if our user model is appropriately formulated as a decision problem of the agent. While in our case the partial information is obtained by querying the data center, another approach for obtaining partial information by sampling is suggested in [14]. Though sampling is similar to investigation, the information obtained from these two kinds of actions are quite different. For the sampling actions, even though the chosen individuals may be thoroughly investigated, only the statistical information of

these investigated individuals would be kept. In fact, it is the statistical information which would be used in the prediction of the status of the whole population. However, for the investigative actions, the user would indeed keep the personal information of each investigated individual and not do the statistical inference from the investigated individuals to the whole population.

On the other hand, if no query actions are possible, the information states are always of the form $(\{\top\}, \boldsymbol{\mu_0}, \boldsymbol{\kappa})$. Once all individuals have been fully investigated (though this is hardly possible in any practical case) the information state becomes a perfect state $\mathcal{I} = (\{\top\}, \boldsymbol{\mu_0}, \boldsymbol{\kappa})$, where $\kappa_i(\top) = \mu_{0i}(\top) \cdot |U|$, so $p_i(\top) = 0$ for all $0 \leq i \leq t-1$. Consequently, $B_{\mathcal{I}}$ would be simplified into

$$\sum_{i=0}^{t-1} \max(R_o - r_i, 0) \cdot \mu_{0i}(\top)$$

which is precisely the value of perfect information defined in [14]. Thus we have modeled the value of hybrid information in the above-defined framework.

5 Privacy Protection by Pricing Mechanism

5.1 Basic Scheme

According to the user model above, the user can improve his payoff from 0 to $B_{\mathcal{I}}$ when his information state is evolved from the initial state to $\mathcal{I}$. If the information is free of charge, the user would gladly receive it and consequently, the privacy of the individuals may be invaded. Thus, one approach to privacy protection is to impose costs on the answers of the queries so that the user cannot make a profit from obtaining the private information. This can be achieved by including a pricing mechanism in the data center. However, since the answer to a query may have different effects under different information states, the pricing mechanism must be adaptive according to the query history of the user. In general, it is very difficult to design an adaptive pricing mechanism since the users may have to pay different prices for the same queries under different situations. Therefore, instead of charging each query separately, we shall consider a more restricted setting. Assume that each user is allowed to ask a batch of queries only once. Afterward, he can do any investigative actions he wants. However, the data center would not answer his queries afterwards. Thus the pricing mechanism of the data center is to decide the cost of each batch of queries so that the user cannot benefit from receiving the answers of the queries.

Let $(\Pi, \boldsymbol{\mu}, \boldsymbol{\kappa})$ be the information state of the user after a sequence of queries and follow-up investigative actions, where $\Pi = \{\varphi_1, \varphi_2, \ldots, \varphi_m\}$. Since the data center has no control on how the user will carry out his investigation after receiving the answers, it can only guarantee that the cost is high enough so that the user cannot make a profit from the answers of the queries, no matter what investigation be done. Thus, based only on the partial knowledge state $\mathcal{P} = (\Pi, \boldsymbol{\mu})$, the data center must estimate the maximum payoff the agent can

have under different information states compatible with $\mathcal{P}$. Let $\mathbf{k} = (k_1, \ldots, k_m)$ be an m-tuple of non-negative integers and define

$$F_{\mathbf{k}} = \{\boldsymbol{\kappa} \mid \sum_{i=0}^{t-1} \kappa_i(\varphi_j) = k_j, \forall 1 \leq j \leq m\}$$

as the set of $\boldsymbol{\kappa}$ functions which denote the possible investigation results when a specific number of individuals has been investigated. The set of information states compatible with $\mathcal{P}$ and $\mathbf{k}$ is defined as

$$\mathcal{IS}(\mathcal{P}, \mathbf{k}) = \{(\mathcal{P}, \boldsymbol{\kappa}) \mid \boldsymbol{\kappa} \in F_{\mathbf{k}}\}$$

and the maximal value of information of the agent under $\mathcal{P}$ and $\mathbf{k}$ is defined as

$$B(\mathcal{P}, \mathbf{k}) = \max\{B_{\mathcal{I}} \mid \mathcal{I} \in \mathcal{IS}(\mathcal{P}, \mathbf{k})\}.$$

We now further assume that a cost function $\gamma_{inv} : \Phi \times \mathcal{Z}^+ \rightarrow \Re^+$ is available to both the user and the data center, where Φ is the set of DL wffs and $\mathcal{Z}^+$ and $\Re^+$ are respectively the set of positive integer and real numbers. The intended meaning of $\gamma_{inv}(\varphi, k)$ is the cost of the investigation of k individuals satisfying φ. It can be assumed that γ_{inv} is a super-linear function in its second argument. Thus, when the user poses a batch of queries Q, the data center can know what the resultant partial knowledge state $\mathcal{P}$ would be once the answer is released. Therefore, the price of Q must be decided before releasing the information. The price $price(Q)$ of the answers to the batch of queries should be decided such that

$$|U| \cdot B(\mathcal{P}, \mathbf{k}) - \sum_{i=1}^{m} \gamma_{inv}(\varphi_i, k_i) \leq price(Q) \tag{3}$$

holds for any $\mathbf{k}$. The lowest solution of $price(Q)$ for (3) is

$$\max_{\mathbf{k}} |U| \cdot \max\{B_{\mathcal{I}} \mid \mathcal{I} \in \mathcal{IS}(\mathcal{P}, \mathbf{k})\} - \sum_{i=1}^{m} \gamma_{inv}(\varphi_i, k_i) \tag{4}$$

where the domain of $\mathbf{k}$ is finite since $0 \leq k_i \leq |\varphi_i|$.

5.2 Usefulness of Information

In our pricing mechanism, the data center assumes that the user can play the above-mentioned game with all individuals in U and charge them based on the total gain he can achieve. However, this may be an over-estimation since the user cannot play the game with all individuals when the population is large. To circumvent the problem, we may assume that the user must spend some resources for playing the game with the individuals. Let $\gamma_{ply} : \Phi \times \mathcal{Z}^+ \rightarrow \Re^+$ be another cost function such that $\gamma_{ply}(\varphi, l)$ denotes the cost of the user playing the

game with l individuals satisfying φ. Given an m-tuple of non-negative integers $\mathbf{l} = (l_1, \ldots, l_m)$ and an information state $\mathcal{I}$, define

$$B_{\mathcal{I}}^{\mathbf{l}} = \sum_{i=1}^{m} B_{\varphi_i} \cdot l_i.$$

The price in (4) can be replaced by

$$\max_{\mathbf{k},\mathbf{l}}(\max\{B_{\mathcal{I}}^{\mathbf{l}} \mid \mathcal{I} \in \mathcal{IS}(\mathcal{P}, \mathbf{k})\} - \sum_{i=1}^{m} \gamma_{inv}(\varphi_i, k_i) - \sum_{i=1}^{m} \gamma_{ply}(\varphi_i, l_i)) \tag{5}$$

where both the domains of $\mathbf{k}$ and $\mathbf{l}$ are restricted to $0 \leq k_i, l_i \leq |\varphi_i|$.

Intuitively, each l_i and k_j represent the *usefulness* of information. Given two equivalent classes in a logical partition, it may be easier to find potential members in one equivalence class than in the other depending on the conditions each equivalence class satisfied. It may also be true that it is easier, and thus cost-effective, to investigate members in one equivalence class than in the other. These two may be closely related, but not necessarily the same.

Example 2 Assume we again use the insurance company model mentioned in Example 1. Assume the world population is represented by all adults in the country. An equivalence class may be characterized as being the people living in the same county while another equivalence class is described as the people with weight between 60 to 65 kilograms. It is easy for the first group of people to be investigated and then to be added as customers, while it is relatively difficult for the second group of people. ■

Thus the data center can decide the price of the answers to the batch of queries Q by a two-level maximization procedure in (4) or (5). The outer level maximization would depend on the form of the cost functions γ_{inv} and/or γ_{ply}, so it is unlikely to find an analytic solution for it. However, the inner maximization can be reduced to a set of m maximization of B_φ for each $\varphi \in \Pi$. More specifically, given φ and $0 \leq k \leq |\varphi|$, it is to find $\boldsymbol{\kappa}(\varphi)$ which maximizes B_φ among all $\boldsymbol{\kappa}$ satisfying $\sum_{i=0}^{t-1} \kappa_i(\varphi) = k$ and $\kappa_i(\varphi) \leq |\varphi| \cdot \mu_i(\varphi)$ for all $0 \leq i \leq t-1$. This is in turn equivalent to the following constraint optimization problem in the integer domain:

$$\begin{array}{c} \text{Maximize } \max(R_0 - \sum_{i=0}^{t-1} \frac{n_i - x_i}{N-k} \cdot r_i, 0) \cdot \frac{N-k}{N} + \sum_{i=0}^{t-1} \max(R_0 - r_i, 0) \cdot \frac{x_i}{N} \\ \text{s.t.} \\ x_0 + x_1 + \cdots + x_{t-1} = k \\ 0 \leq x_i \leq n_i \ \ (0 \leq i \leq t-1) \end{array} \tag{6}$$

where N and n_i's correspond to $|\varphi|$ and $|\varphi| \cdot \mu_i(\varphi)$'s respectively. The solution of Equation (6) can be given by the following proposition for $k \leq N$. Without loss of generality, we assume $r_0 \geq r_1 \geq \cdots \geq r_{t-1}$ for the loss vector in the proposition.

Proposition 1 *Assume* $N = \sum_{i=0}^{t-1} n_i$

1. *if* $k = N$*, then the solution of Equation (6) is* $x_i = n_i$ *for* $0 \leq i \leq t-1$ *and its maximum value is*

$$\sum_{i=0}^{t-1} \max(R_0 - r_i, 0) \cdot \frac{n_i}{N};$$

2. *if if* $k < N$ *and* l *is the smallest natural number such that* $\sum_{i=0}^{l} n_i > k$*, then the solution of Equation (6) is*

$$x_i = \begin{cases} n_i & \text{if } \ i < l, \\ k - \sum_{i=0}^{l-1} n_i & \text{if } \ i = l, \\ 0 & \text{if } \ i > l, \end{cases}$$

and its maximum value is

$$\max(R_0 - \sum_{i=l+1}^{t-1} \frac{n_i}{N-k} \cdot r_i + \frac{\sum_{i=0}^{l} n_i - k}{N-k} \cdot r_l, 0) \cdot \frac{N-k}{N}$$
$$+ \sum_{i=0}^{l-1} \max(R_0 - r_i, 0) \cdot \frac{n_i}{N} + \max(R_0 - r_l, 0) \cdot \frac{k - \sum_{i=0}^{l-1} n_i}{N}.$$

The individuals who will incur more loss to the agent are high risk ones. For the low risk individuals, the investigation cannot improve the payoff for the agent. However, for the high risk ones, the investigation can indeed decrease the loss for the agent by raising their admission fees appropriately. The more high risk individuals have been investigated, the more loss the agent can avoid, so the maximum payoff occurs when the investigation is carried out from the most risky individuals to the least risky ones. This intuition is verified by the preceding proposition.

6 Related Works

To quantify the value of information is by no means a novel problem. However, the quantitative models for privacy protection provides a new angle to look at the problem. As shown in section 4, our model for the value of information has generalized a standard notion in decision theory[14, 7]. While the decision-theoretic analysis [14] emphasizes the value of information from the decision maker's viewpoint, our model is mainly concerned with privacy protection by the information provider. For the former, a decision maker can decide if he will purchase a piece of information according to the value of the information. For the latter, the information provider can charge the user of the information with appropriate rates.

An alternative model for the value of information in the privacy protection context is proposed in [1, 2]. In their model, the value of information is estimated

by the expected cost the user must pay for achieving the perfect knowledge state from the given information. The estimation is based on the rationale that the more investigation efforts a piece of information can reduce, the more valuable it is. However, without regarding the user model, the value of information defined there may not reflect the real situation.

Besides the decision theoretic analysis, the value of information can also be estimated by some information theoretic measures. The central notion of such measures is the entropy introduced by Shannon[20]. In the machine learning literatures, it is used to define the information gain of an attribute for a classification problem[16]. Though the information gain is an useful index in selecting the most informative features for the classification problem, it still suffers the same problem as the value of information defined in [1, 2] since it does not take into account the fact that some confidential attribute values are more sensitive than others.

The sensitivity of different attribute values are taken into account in the average benefit and average cost models proposed in [3]. However, while only query actions are allowed there, we also consider the investigative actions in modelling the value of information.

In contrast with the quantitative approach of this paper, some qualitative criteria for privacy protection have been proposed in [11, 12, 18, 19, 21]. These criteria are designed to protect personal sensitive information in the release of a microdata set, i.e. a set of records containing information on individuals. The main objective is to avoid the re-identification of individuals or in other words, to prevent the possibility of deducing which record corresponds to a particular individual even though the explicit identifier of the individual is not contained in the released information. On the other hand, our models are concerned with the release of statistical information which is less specific than microdata in general. However, microdata release can also be handled in our framework when the queries are specific enough. Let us define a complete specification formula (CSF) as a DL wff of the form $\wedge_{a \in A}(a, v_a)$, where A is the set of all easy-to-know attributes and v_a is a value in the domain of A. The answer to the batch of queries Q consisting of all CSF's is equivalent to the microdata release of the whole data table T. Therefore, our models are applicable in a more general context.

7 Conclusion

In this paper, we present a quantitative model for privacy protection. In the model, a formal representation of the user's information states is given, and we estimate the value of information for the user by considering a specific user model. Under the user model, the privacy protection task is to ensure that the user cannot profit from obtaining the private information.

It must be emphasized that the value of information is defined with respect to the particular user model. When other user models are considered, the value of information may be different. Some examples can be seen in [13]. A problem

for the pricing mechanism arises naturally since different users may put different values on the same information. This means that we may have to set different prices for different kinds of users on the same information. However, this is not so odd as it seems at first glance. In fact, differentiating prices have been employed in the software market. Differences usually occur in educational and commercial uses.

There are further complicated problems in privacy protection which can not be resolved from a purely technical aspect. For example, our schemes cannot prevent a group of users from collectively investigating private information by individually querying the data center. This must be considered from a legal aspect. Upon releasing data to a user, the user must sign a contract prohibiting him from revealing the data to others. The legal possibility for the collusion of a group of users is thus blocked. In future works, we would like to investigate how technology and law can be fully combined in the protection of privacy.

References

1. Y.-C. Chiang. Protecting privacy in public database (in Chinese). Master's thesis, Graduate Institute of Information Management, National Taiwan University, 2000.
2. Y.-C. Chiang, T.-s. Hsu, S. Kuo, and D.-W. Wang. Preserving confidentially when sharing medical data. In *Proceedings of Asia Pacific Medical Informatics Conference*, 2000.
3. Y.T. Chiang, Y.C. Chiang, T.-s. Hsu, C.-J. Liau, and D.-W. Wang. How much privacy? - a system to safe guard personal privacy while releasing database. In *Proceedings of the 3rd International Conference on Rough Sets and Current Trends in Computing*, LNCS. Springer-Verlag, 2002.
4. F. Y. Chin and G. Özsoyoğlu. Auditing and inference control in statistical databases. *IEEE Transactions Software Engineering*, 8:574–582, 1982.
5. L. H. Cox. Suppression methodology and statistical disclosure control. *Journal of the American Statistical Association*, 75:377–385, 1980.
6. D. E. R. Denning. *Cryptography and Data Security*. Addison-Wesley, 1982.
7. G.D. Eppen and F.J. Gould. *Quantitative Concepts for Management*. Prentice Hall, 1985.
8. F. Duarte de Carvalho, N. P. Dellaert, and M. de Sanches Osório. Statistical disclosure in two-dimensional tables: General tables. *Journal of the American Statistical Association*, 428:1547–1557, 1994.
9. D. Gusfield. A graph theoretic approach to statistical data security. *SIAM Journal on Computing*, 17:552–571, 1988.
10. T.-s. Hsu and M. Y. Kao. Security problems for statistical databases with general cell suppressions. In *Proceedings of the 9th International Conference on Scientific and Statistical Database Management*, pages 155–164, 1997.
11. T.-s. Hsu, C.-J. Liau, and D.-W. Wang. A logical model for privacy protection. In *Proceedings of the 4th International Conference on Information Security*, LNCS 2200, pages 110–124. Springer-Verlag, 2001.
12. A.J. Hundepool and L.C.R.J. Willenborg. "μ- and τ-ARGUS: Software for statistical disclosure control". In *Proceedings of the 3rd International Seminar on Statistical Confidentiality*, 1996.

13. J. Kleinberg, C.H. Papadimitriou, and P. Raghavan. "On the value of private information". In *Proc. 8th Conf. on Theoretical Aspects of Rationality and Knowledge*, 2001.
14. D.V. Lindley. *Making Decisions*. John Wiley & Sons, 1985.
15. T.S. Mayer. Privacy and confidentiality research and the u.s. census bureau recommendations based on a review of the literature. Technical Report RSM2002/01, U.S. Bureau of the Census, 2002.
16. T. Mitchell. *Machine Learning*. McGraw-Hill, 1997.
17. Z. Pawlak. *Rough Sets–Theoretical Aspects of Reasoning about Data*. Kluwer Academic Publishers, 1991.
18. P. Samarati. "Protecting respondents' identities in microdata release". *IEEE Transactions on Knowledge and Data Engineering*, 13(6):1010–1027, 2001.
19. P. Samarati and L. Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Technical report SRI-CSL-98-04, Computer Science Laboratory, SRI International, 1998.
20. C.E. Shannon. "The mathematical theory of communication". *The Bell System Technical Journal*, 27(3&4):379–423,623–656, 1948.
21. L. Sweeney. "Guaranteeing anonymity when sharing medical data, the Datafly system". In *Proceedings of American Medical Informatics Association*, 1997.
22. W. E. Winkler. The state of record linkage and current research problems. Technical Report RR99/04, U.S. Bureau of the Census, 1999.
23. W. E. Winkler. Record linkage software and methods for merging administrative lists. Technical Report RR01/03, U.S. Bureau of the Census, 2001.