Granulation as a Privacy Protection Mechanism[⋆]

Da-Wei Wang, Churn-Jung Liau, and Tsan-sheng Hsu

Institute of Information Science
Academia Sinica, Taipei 115, Taiwan
{wdw,liaucj,tshsu}@iis.sinica.edu.tw

Abstract. How to achieve a balance between data publication and privacy protection has been an important issue in information security for several years. When microdata is released to users, attributes that clearly identify individuals are usually removed. Nevertheless, it is still possible to link released data with some public or easy-to-access databases to obtain confidential information. To safeguard privacy, numerous techniques, such as generalization, suppression, and microaggregation, have been proposed to modify the to-be-released data. In this paper, we propose attribute-oriented granulation as a data protection mechanism that can integrate both generalization and microaggregation into a uniform framework. We address the computational issue of searching for the most specific granulation that satisfies confidentiality requirements. A breadth-first search algorithm with basic pruning strategies is presented and its properties are investigated. The properties can be used to improve the efficiency of our algorithm. We also define some quantitative measures of data quality and security, and apply evolutionary computation techniques to find the optimal granulation for privacy protection.

1 Introduction

Privacy protection is one of the main concerns in the field of data security. In recent years, statistical disclosure control [2] has become increasingly important due to the requirements of data security. One of the major issues in disclosure control is the database linking problem. Generally speaking, the problem is how to prevent users[1] obtaining confidential information about an individual[2] by linking to some public or easy-to-access database with data they can obtain legally from a data center.

Though the protection of privacy is very important, over-restriction of access to a database may render the data useless. Therefore, the main challenge is how to achieve a balance between privacy protection and data availability. One

[⋆] A preliminary version of this paper was published in [1]. This work was partially supported by the Taiwan Information Security Center (TWISC) and NSC (Taiwan). NSC Grants: 95-2221-E-001-019 (D.W. Wang), 95-2221-E-001-029-MY3 (C.J. Liau), and 95-2221-E-001-004 (T-s. Hsu).

[1] In this paper, a user refers to anyone receiving data and having the potential to breach the privacy of individuals.

[2] An individual refers to a person whose privacy should be protected.

J.F. Peters et al. (Eds.): Transactions on Rough Sets VII, LNCS 4400, pp. 256–273, 2007.

possibility is to modify the data before it is released by generalizing the values of some data cells to a coarser level of precision. To do this, we can partition the domain of attributes according to a certain level of precision, and generalize the data from the finest to the coarsest level until the privacy requirement is met. This kind of operation is called *attribute-oriented granulation* (AOG). In this paper, we investigate the application of AOG to privacy protection. It is shown that AOG can integrate generalization[3,4,5,6,7] and microaggregation[8] into a uniform framework. To address the computational issue of searching for the most specific granulation that satisfies confidentiality requirements, a breadth-first search algorithm with basic pruning strategies is presented and its properties are investigated. The properties can be used to improve the efficiency of the basic algorithm. We also define the quantitative measures of data quality and security and apply evolutionary computation (EC) techniques to find the optimal granulation for privacy protection.

The remainder of the paper is organized as follows: In Section 2, we use an example to illustrate the concept of AOG, and formally introduce the AOG operation. The logical security of AOG and its computational aspects are explored in Section 3. We also present several properties of AOG that are used to improve the search algorithm. In Section 4, we discuss the security and quality of AOG and apply an EC approach to the search for an optimal AOG. We then present our conclusions in Section 6.

2 Attribute-Oriented Granulation

2.1 A Running Example

In this paper, we investigate the privacy protection problem that may arise when a data table [9] is released. The data in many application domains, such as medical records, financial transaction records, employee information, and so on, can be organized as data tables. A data table consists of a set of records, each of which corresponds to an individual and has some attributes.

The attributes of a data table can be divided into three sets [10,11]. The first consists of *identifiers* that can be used to identify to whom a data record belongs. Therefore, these attributes are always masked off in response to a query. Let us equate a set of identifiers with a set of individuals. Throughout this paper, a set of individuals (or identifiers) is denoted by U. Second, we have a set of *quasi-identifiers*, the values of which are known to the public. For example, in [12], it is pointed out that some attributes like birth-date, gender, ethnicity, etc. are included in some public databases, such as those that contain census data or voter registration lists. These attributes, if not appropriately processed, may be used to re-identify an individual's record in a data table, thus causing a privacy violation. The last kind of attribute is the *confidential attribute*, the values of which we have to protect. It is often the case that an asymmetry exists between the values of a confidential attribute. For example, if the attribute is a HIV test result, then the revelation of a '+' value may cause a serious invasion of privacy, whereas it does not matter to know that an individual has a '−' status. In this

ID	D.O.B.	ZIP	Height	Income	Health
u_1	24/09/56	24126	161	400K	1
u_2	06/09/56	24129	167	300K	1
u_3	30/09/56	24133	163	300K	1
u_4	23/03/56	10427	160	300K	0
u_5	18/03/56	10431	165	100K	2
u_6	05/03/56	10466	168	100K	2
u_7	20/04/55	26015	175	400K	2
u_8	18/04/55	26032	170	300K	1
u_9	09/04/55	26617	173	100K	0
u_{10}	01/04/55	26628	171	400K	0
u_{11}	23/04/55	26328	176	400K	0

Fig. 1. A data table

1	09/56	24***	[160,170)	400K	1
2	09/56	24***	[160,170)	300K	1
3	09/56	24***	[160,170)	300K	1
4	03/56	10***	[160,170)	300K	0
5	03/56	10***	[160,170)	100K	2
6	03/56	10***	[160,170)	100K	2
7	04/55	26***	[170,180)	400K	2
8	04/55	26***	[170,180)	300K	1
9	04/55	26***	[170,180)	100K	0
10	04/55	26***	[170,180)	400K	0
11	04/55	26***	[170,180)	400K	0

Fig. 2. A generalized data table

paper, let T denote a data table for a set of individuals U, and t_{ij} denote the value of an attribute j of an individual u_i.

We use the data table in Figure 1 as our running example[7]. In the table, $U = \{u_1, \cdots, u_{11}\}$ is a set of individuals (or identifiers); the quasi-identifiers are date of birth, zip code, and height; and the confidential attributes are income and health status. The values of "Health" are denoted by "normal"(0), "slightly ill"(1), and "seriously ill"(2) respectively.

In [11,5,12], the notion of *bin size* is proposed to resolve the database linkage problem. A *bin* is defined as an equivalence class based on the quasi-identifiers, and the bin's size is its cardinality. To be deemed secure, a table must satisfy the condition that the size of any bin is sufficiently large. The security criterion is called k-anonymity if each bin is required to contain at least k individuals. Though, in general, the chance of a user obtaining confidential information is smaller if the bin size is larger, it is well-known that controlling the bin size alone is not sufficient to stop inference attacks [13]. To fully protect privacy, we must consider some alternative criteria to complement the bin size.

One technique of protecting privacy is to release the data in a coarser granularity. For example, the date of birth may be given as only the year and month, or only the first two digits of the ZIP code may be given. In addition, "Height" can be expressed as a range, instead of a precise value. A concrete generalization of the data table in Figure 1 is given in Figure 2. The first column denotes the serial numbers of the released data records.

From the generalized data table, we observe that the bin containing u_1, u_2, and u_3 is size 3. However, since the health status attribute of the rows in this bin has the value 1, the recipient of the table can infer that u_1, u_2, and u_3 are all slightly ill, though he does not know which of them has an income of 400K.

2.2 AOG Operations

In this section, we formally define the modification operation that can be applied to a data table to enhance privacy protection. As the operation is based on partitioning the domain of attributes according to different granular scales, it is called attribute-oriented granulation (AOG). We first recall the basic definition of a partition. Let V be a domain of values for some attribute; then, a *partition* π of V is a set $\{s_1, s_2, \ldots, s_k\}$ of mutually disjoint subsets of V such that $\cup_{i=1}^{k} s_i = V$. Each s_i is called an equivalence class of the partition, and we use $\pi(v)$ to denote the equivalence class containing v. Let π_1 and π_2 be two partitions of V. Then, π_1 is a *refinement* of π_2, written as $\pi_1 \preceq \pi_2$ if, for $s \in \pi_1$ and $t \in \pi_2$, either $s \subseteq t$ or $s \cap t = \emptyset$. Let $\pi_1 \prec \pi_2$ denote $\pi_1 \preceq \pi_2$ and $\pi_1 \neq \pi_2$. For a given set V, we use $\perp$ and $\top$ (possibly with indices) to denote the finest partition $\{\{v\} \mid v \in V\}$ and the coarsest partition $\{V\}$ respectively.

Let us assume that the set of quasi-identifiers is $\{1, 2, \ldots, m\}$ and denote the domain of attribute i by V_i for $1 \leq i \leq m$. Then, an AOG operation is specified by a tuple $(\pi_1, \pi_2, \ldots, \pi_m)$, where for $1 \leq i \leq m$, π_i is a partition of V_i. Let $\tau_1 = (\pi_1, \pi_2, \ldots, \pi_m)$ and $\tau_2 = (\pi'_1, \pi'_2, \ldots, \pi'_m)$ be two AOG operations. Then, τ_1 is *at least as specific as* τ_2, denoted by $\tau_1 \preceq \tau_2$, if for $1 \leq i \leq m$, $\pi_i \preceq \pi'_i$; and τ_1 is *more specific than* τ_2, denoted by $\tau_1 \prec \tau_2$, if $\tau_1 \preceq \tau_2$ and $\tau_1 \neq \tau_2$.

Since the number of possible partitions of a domain may be prohibitively large, we sometimes focus on a subset of *admissible partitions*. Let us define Π_i as the set of admissible partitions of V_i such that $\perp_i$ and $\top_i \in \Pi_i$ for $1 \leq i \leq m$; then, the set of *admissible AOGs* is $\Pi = \Pi_1 \times \Pi_2 \times \cdots \times \Pi_m$. τ_2 is called a direct successor of τ_1 in Π if $\tau_1 \prec \tau_2$ and there does not exist any $\tau \in \Pi$ such that $\tau_1 \prec \tau \prec \tau_2$.

2.3 The Running Example

Figure 3 shows a set of admissible partitions for our running example, where the partitions for the dates of birth and zip codes are obvious, and the partitions for height are defined as

$$\begin{aligned}
I_1 &= \{\cdots, \{160\}, \{161\}, \cdots, \{174\}, \{175\}, \cdots\}, \\
I_5 &= \{\cdots, [160, 165), [165, 170), [170, 175), \cdots\}, \\
I_{10} &= \{\cdots, [160, 170), [170, 180), \cdots\}, \\
I_{20} &= \{\cdots, [160, 180), [180, 200), \cdots\}.
\end{aligned}$$

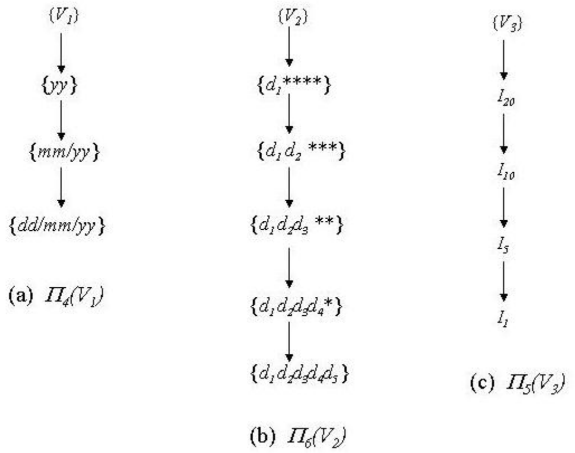

Fig. 3. Admissible partitions for the quasi-identifiers in our running example

3 Logical Security

3.1 Security of AOG

To decide whether an AOG is secure, we use Pawlak's decision logic (DL, [9]) to describe confidential information. The set of *atomic sentences* for DL is $\mathcal{P} = \{(j, v) \mid j \in J, v \in V_j\}$, where J is the set of confidential attributes. The intuitive meaning of the atomic sentence (j, v) is that an individual's attribute j has value v. The set of sentences is the smallest set containing $\mathcal{P}$ that is closed on the Boolean connectives $\neg, \wedge$, and $\vee$. If $\alpha \subseteq V_j$, we abbreviate $\vee_{v\in\alpha}(j, v)$ as (j, α). We assume that the information an individual u wants to keep confidential is represented by a set of DL sentences, $CON(u)$. As usual, the sentences are evaluated inductively with respect to the data table T and each individual in U as follows:

1. $u_i \models (j, v)$ iff $t_{ij} = v$.
2. $u \models \neg\varphi$ iff $u \not\models \varphi$.
3. $u \models \varphi \wedge \psi$ iff $u \models \varphi$ and $u \models \psi$.
4. $u \models \varphi \vee \psi$ iff $u \models \varphi$ or $u \models \psi$.

The meaning set of a sentence φ, $\|\varphi\|_T = \{u \in U \mid u \models \varphi\}$, is the set of individuals that satisfies φ in the data table T. The subscript T is usually omitted when it is clear from the context.

Let π be a partition of the domain of an attribute k; then, the π-*indiscernibility relation* with respect to the data table T, denoted by $ind_T(\pi)$, is an equivalence relation on U defined by $(u_i, u_j) \in ind_T(\pi) \Leftrightarrow \pi(t_{ik}) = \pi(t_{jk})$. Again, the subscript T is usually omitted for convenience. Let $\tau = (\pi_1, \pi_2, \ldots, \pi_m)$ be an AOG operation; then, the τ-*indiscernibility relation* with respect to the data table T is defined as

$$ind(\tau) = \cap_{1\leq k\leq m} ind(\pi_k).$$

An AOG operation, $\tau = (\pi_1, \pi_2, \ldots, \pi_m)$, determines how the data is modified before it is released. The requirement is that, for any $u_i, u_j \in U$ and attribute k, $\pi_k(t_{ik}) = \pi_k(t_{jk})$ iff t_{ik} and t_{jk} are replaced by the same value in the modified data table. For example, the generalization method in [4,11] replaces a table entry t_{ik} with $\pi_k(t_{ik})$, whereas the microaggregation method in [8] replaces it with some statistics, such as the mean, median, or mode of the multiset[3] $\{t_{jk} \mid (u_i, u_j) \in ind(\pi_k)\}$. Thus, the AOG method subsumes both generalization and microaggregation. In this paper, we do not specify any particular modification method for the AOG operation. We simply use $\tau(T)$ to denote the table derived by modifying the data table T with τ.

Given the τ-indiscernibility relation, the standard definition of the lower approximation in rough set theory is used to define the logical security of an AOG. The lower approximation for any set $X \subseteq U$ is defined as

$$\underline{ind(\tau)}X = \{u \mid \forall (u, u') \in ind(\tau), u' \in X\}.$$

The AOG operation τ is *logically secure* (or simply secure) for u if $u \notin \underline{ind(\tau)}[\![\varphi]\!]$ for any $\varphi \in CON(u)$, and secure for U if it is secure for all $u \in U$. Once the data table to be released has been modified by an AOG τ, the user can not distinguish the records of two individuals who are indiscernible in the relation $ind(\tau)$. Therefore, even though the user knows the values of all the quasi-identifiers of an individual, as well as how the values are modified, he can not deduce that the individual satisfies a confidential property φ, provided that τ is secure.

3.2 The Running Example

Let us consider an AOG $\tau = (\pi_1, \pi_2, \pi_3) = (mm/yy, d_1, d_2 * **, I_{10})$. Then

$$ind(\tau) = ind(\pi_1) = ind(\pi_2) = \{\{u_1, u_2, u_3\}, \{u_4, u_5, u_6\}, \{u_7, u_8, u_9, u_{10}, u_{11}\}\},$$

$$ind(\pi_3) = \{\{u_1, u_2, u_3, u_4, u_5, u_6\}, \{u_7, u_8, u_9, u_{10}, u_{11}\}\}.$$

By using the generalization method to modify the data table in Figure 1, we obtain the generalized table in Figure 2. On the other hand, if the microaggregation method is used to modify the data table, and the arithmetical mean and median are taken as the statistical operators of the continuous and ordinal attributes respectively, then we can obtain the modified data table in Figure 4. To understand how the table is derived, let us consider the individual u_1. First, for the continuous attribute height, $[u_1]_{ind(\pi_3)} = \{u_1, u_2, u_3, u_4, u_5, u_6\}$; thus, t_{13} is replaced by the arithmetical mean of the multiset $\{161, 167, 163, 160, 165, 168\}$, which is equal to 164. Second, for the ordinal attributes, date of birth and zip code, $[u_1]_{ind(\pi_1)} = [u_1]_{ind(\pi_2)} = \{u_1, u_2, u_3\}$; thus, t_{11} and t_{12} are replaced by the median of $\{24/09/56, 06/09/56, 30/09/56\}$ and $\{24126, 24129, 24133\}$, which are 24/09/56 and 24129, respectively.

Note that the tables produced by generalization and microaggregation are structurally isomorphic[14]. It is shown in [15] that isomorphic tables have the

[3] A multiset is a set that allows the multiple occurrence of its elements.

1	24/09/56	24129	164	400K	1
2	24/09/56	24129	164	300K	1
3	24/09/56	24129	164	300K	1
4	18/03/56	10431	164	300K	0
5	18/03/56	10431	164	100K	2
6	18/03/56	10431	164	100K	2
7	18/04/55	26617	173	400K	2
8	18/04/55	26617	173	300K	1
9	18/04/55	26617	173	100K	0
10	18/04/55	26617	173	400K	0
11	18/04/55	26617	173	400K	0

Fig. 4. Our running example modified by the microaggregation technique

same granular data model, defined as (U, Q), such that Q is a set of equivalence relations induced by the attributes.

Now, if $\varphi = (\text{Health}, 2)$ is a confidential sentence, then $\underline{ind(\tau)}\|\varphi\| = \emptyset$, since $\|\varphi\| = \{u_5, u_6, u_7\}$. Thus, τ is secure for U if $CON(u) = \{\overline{(\text{Health}, 2)}\}$. On the other hand, if $\varphi' = (\text{Health}, 1) \in CON(u_1)$, then τ is insecure for u_1, since $\underline{ind(\tau)}\|\varphi'\| = \{u_1, u_2, u_3\}$. The result is matched by our intuition.

3.3 The Basic Search Algorithm

Since the goal of privacy protection is to find a secure and maximally informative AOG operation, we can achieve it by a bottom-up search of all possible AOGs. The algorithm proposed in this section is based on a breadth-first search through the set of admissible AOGs using basic pruning strategies. For simplicity, we present the basic algorithm in this section, and discuss improvements that make it more efficient in the next section. Our previous experiments show that the performance of the basic algorithm is acceptable in non-realtime environments [6].

Although it is sufficient to find *a* secure and maximally informative AOG for a data table, for the sake of flexibility, our search algorithm is designed to find *all* secure and maximally informative AOGs for a given data table. We start from the most specific AOG $(\bot_1, \cdots, \bot_m)$ and test its security according to our definition. If this operation is secure, we stop searching. Otherwise, we have to climb the search tree according to the partial order $\preceq$ between AOG operations. Each new AOG must be tested to evaluate its security. If it is secure, then all AOG operations above it can be pruned, since our purpose is to find the maximally informative (i.e., $\preceq$-minimal) AOGs. Thus, the pruning operation substantially reduces the number of AOGs that must be visited. In the search algorithm presented in Figure 5, the function GET-FROM-QUEUE returns the first element of a queue, whereas the procedure PUT-INTO-QUEUE adds an AOG to the end of a queue. These operations are standard and can be found in algorithm textbooks. Also, we record the status of each τ in a Boolean array F, where $F(\tau) = 1$ means that it is not necessary to check τ any further.

```
Procedure SEARCH($\Pi, T, CON$)

1. Initialize a Boolean array $F[\tau] := 0$ for all $\tau \in \Pi$;
2. Initialize a queue of AOG operations $Q \leftarrow \{(\bot_1, \bot_2, \cdots, \bot_m)\}$;
3. while $Q \neq \emptyset$ do
     begin
     repeat $\tau \leftarrow$GET-FROM-QUEUE($Q$) until $F[\tau] = 0 \vee Q = \emptyset$;
     if $F[\tau] = 1 \wedge Q = \emptyset$ then exit;
     $F[\tau] \leftarrow 1$;
     if SECURITY($\tau, T, CON$)
       then  begin
         Output($\tau$);
         $F[\tau'] \leftarrow 1$ for all $\tau'$ such that $\tau \preceq \tau'$
         end
       else for each direct successor $\tau'$ of $\tau$ do
         if $F[\tau'] = 0$ then PUT-INTO-QUEUE($Q, \tau'$)
     end
```

Fig. 5. The search algorithm for AOG

```
Function SECURITY($\tau, T, CON$)

1. Find $ind(\tau)$ by sorting $\tau(T)$;
2. Initialize Boolean $SF \leftarrow 1$;
3. for each $u \in U$ do
     begin
     (a) $US[u] \leftarrow 0$;
     (b) for each $\varphi \in CON[u]$ do
         begin
         $KN(u, \varphi) \leftarrow 1$;
         for each $u' \in [u]_{ind(\tau)}$ do $KN(u, \varphi) \leftarrow KN(u, \varphi) \wedge (u' \models \varphi)$;
         $US(u) \leftarrow US(u) \vee KN(u, \varphi)$
         end;
     (c) $SF \leftarrow SF \wedge \neg US[u]$
     end
```

Fig. 6. The security test function for AOG

The SECURITY function takes an AOG τ, a data table T, and the confidential data function CON as its arguments and returns 1 if τ is secure with respect to T according to the confidential requirement specified by CON; otherwise, it returns 0. The SECURITY function is presented in Figure 6. By sorting $\tau(T)$ according to its quasi-identifiers, we can partition U into $ind(\tau)$-equivalence classes. Then, we use a Boolean variable, SF, and two Boolean arrays, US and KN, indexed by U and $U \times \mathcal{L}_0$ respectively, to compute the output, where $\mathcal{L}_0$ denotes the set of confidential sentences. Here SF, which is initialized to 1, denotes the security of τ, whereas $US[u] = 1$ means that τ is not secure for u; hence, the final security level is computed by repeat conjunction of SF with $\neg US[u]$ for all $u \in U$. The array KN denotes the user's knowledge about individuals, so $KN(u, \varphi) = 1$ means the user knows that u satisfies φ, i.e., $u \in \underline{ind(\tau)}\|\varphi\|$. $KN(u, \varphi)$ is computed by repeat conjunction of its initial value 1 with $u' \models \varphi$ for all $u' \in [u]_{ind(\tau)}$, where $u' \models \varphi$ means that φ is satisfied by u'. Furthermore, τ is not secure for u if for some $\varphi \in CON[u]$, $KN(u, \varphi) = 1$; consequently, $US(u)$ is computed by repeat disjunction of its initial value 0 with $KN(u, \varphi)$ for all $\varphi \in CON[u]$.

The complexity of the SECURITY function can be analyzed as follows. First, Step 1, the sorting step, needs $O(n \log n)$ time using standard algorithms, where n is the cardinality of U. Let us assume the evaluation $u' \models \varphi$ can be performed in constant-bounded time; then, the total execution time of Step 3 is

$$\sum_{u \in U} |CON[u]| \cdot |[u]_{ind(\tau)}|.$$

Assuming the size of each $CON[u]$ is bounded above by a constant C, the total execution time of Step 3 is at most

$$C \cdot \sum_{u \in U} |[u]_{ind(\tau)}|,$$

which is in $O(n^2)$ time, since $|[u]_{ind(\tau)}| \leq n$ for all $u \in U$. The $O(n^2)$ bound is quite loose, because $|[u]_{ind(\tau)}|$ may be much smaller than n. Furthermore, in the special case where all individuals have the same set of confidential data (or at least in the case where, for all $u_1, u_2 \in U$, $(u_1, u_2) \in ind(\tau)$ implies $CON[u_1] = CON[u_2]$), Step 3(b) is only executed once for each individual corresponding to a different $ind(\tau)$-equivalence class, which reduces its computation time to $O(n)$. Therefore, the total time complexity of the security test procedure is $O(n^2)$ in general, and $O(n \log n)$ in special cases.

3.4 Computational Improvement

As noted earlier, our algorithm for finding maximally informative AOGs is based on a breadth-first search with basic pruning strategies. Recently, a more efficient algorithm for full-domain k-anonymity, called Incognito, has been proposed [16]. It employs more advanced pruning strategies based on the generalization, rollup, and subset properties.

The generalization and subset properties still hold if k-anonymity is replaced by our security criterion, whereas the rollup property can be easily extended to our framework if the frequency set used in [16] is replaced by the characteristic functions of confidential sentences. Therefore, Incognito can be easily adapted to find all maximally informative AOGs for a data table. In the following, we show that the generalization, rollup, and subset properties hold in our framework.

First, the generalization property is an obvious fact that is used in our basic search algorithm.

Property 1 (Generalization property). *If $\tau_1 \preceq \tau_2$ and τ_1 is secure, then τ_2 is also secure.*

Second, to demonstrate the subset property, we have to define an AOG for an arbitrary subset of quasi-identifiers. So far, we have only defined an AOG for the set of *all* quasi-identifiers. Let J be a subset of $\{1, 2, \cdots, m\}$, then an AOG for J is specified by an m-tuple $(\pi_1, \pi_2, \ldots, \pi_m)$ such that $\pi_i = \top_i$ iff $i \notin J$. If $J_1 \subseteq J_2$, and $\tau_1 = (\pi_1, \pi_2, \ldots, \pi_m)$ and $\tau_2 = (\pi'_1, \pi'_2, \ldots, \pi'_m)$ are the respective AOGs for J_1 and J_2, we say that τ_1 is a restriction of τ_2, denoted by $\tau_1 = \tau_2 | J_1$, if $\pi_i = \pi'_i$ for $i \in J_1$. It is obvious that $\tau_2 \preceq \tau_1$ if τ_1 is a restriction of τ_2. The subset property is therefore a corollary of the generalization property.

Property 2 (Subset property). *Any restriction of a secure AOG is also secure.*

Third, the rollup property must be adapted to our framework. We note that it is necessary to count the records with each unique combination of values of quasi-identifiers in order to check k-anonymity. The rollup property is used to execute the count efficiently. To check our security criterion, we do not have to count the number of records. Instead, we only need to check whether a confidential sentence is falsified for individuals with a combination of quasi-identifiers values. Thus, we define a characteristic function as a mapping from each equivalence class of τ to the subset of confidential sentences falsified by some individuals in the class. Then, the rollup property can be reformulated as follows.

Property 3 (Rollup property). *If $\tau_1 \succeq \tau_2$, then we can generate each set of falsified sentences for the characteristic function of τ_2 by a set union from the characteristic function of τ_1.*

4 Security and Data Quality

4.1 Security Measure

The security criterion defined in the preceding section is purely qualitative. Thus, even though the security condition is satisfied, there is still a sufficiently high probability that the user could infer an individual's confidential information. To assess the security of a protection mechanism more precisely, a number of quantitative criteria have been proposed [6,17]. One criterion that measures how much confidential information is leaked is called the *average benefit criterion*, because it was originally used to assess the benefit a user derives when he receives released data. It is especially appropriate for AOG operations and can also be used to measure risk, since the lower the average benefit, the less an individual's privacy can be breached.

To define such a risk measure, we examine the difference between a user's a priori and a posteriori knowledge. Consider a data table containing an $ind(\tau)$-equivalence class, where 99 percent of the individuals in that class have the same confidential value for one specific attribute. It is tempting to conclude that personal privacy could be violated easily. However, if this distribution is close to the prior distribution of the attribute value of the entire population, release of the above-mentioned data would not be a threat to personal privacy, since a user could not learn much about the distribution by database linking. It is therefore important to consider the original distribution of attribute values in a database.

We now propose an information-theoretic approach that measures information gain after receiving $\tau(T)$. The user's a priori knowledge about φ can be modeled by the prior probability $Pr(\varphi)$, which is the statistical probability of φ for the whole population. If the set U is sufficiently representative of the whole population, then

$$Pr(\varphi) = \frac{|\{x \mid x \models \varphi\}|}{|U|}.$$

On the other hand, the user's a posteriori knowledge about whether u satisfies φ is the percentage of individuals satisfying φ in the $ind(\tau)$-equivalence class $[u]_{ind(\tau)}$, written as

$$Pr_\tau(\varphi|u) = \frac{|\{x \mid x \in [u]_{ind(\tau)} \wedge x \models \varphi\}|}{|[u]_{ind(\tau)}|}.$$

Note that $Pr_\tau(\varphi|u)$ is the rough membership [18] of u in $[\![\varphi]\!]$. Let $dm(u, \varphi)$ be a positive real number denoting the potential damage to an individual u if his/her confidential information φ is breached. We assume the damage values of the individuals are normalized so that $\sum_{\varphi \in CON(u)} dm(u, \varphi) = 1$ for each $u \in U$. Thus, the risk to u due to the release of $\tau(T)$, denoted by $ri(\tau, u)$, is

$$\sum_{\varphi \in CON(u)} dm(\varphi) \cdot \max(\frac{\log Pr(\varphi) - \log Pr_\tau(\varphi|u)}{\log Pr(\varphi)}, 0),$$

and the security measure of τ is defined as

$$sf(\tau) = 1 - \frac{\sum_{u \in U} ri(\tau, u)}{|U|}.$$

4.2 Quality Measure

Privacy protection mechanisms inevitably reduce the quality of released data. We should therefore assess how data quality is affected by AOG operations. Since such operations are based on the partition of quasi-identifier domains, we can use Shannon's entropy to measure data quality. First, the entropy of a partition π of a domain V is defined as

$$h(\pi) = \sum_{s \in \pi} -\frac{|s|}{|V|} \cdot \log \frac{|s|}{|V|}.$$

Second, we consider the significance of the quasi-identifiers. Let $w_i \in [0, 1]$ denote the importance of the quasi-identifiers in data utilization. We also assume that

$\sum_{1 \le i \le m} w_i = 1$. In Section 3, we only considered the case where all quasi-identifiers are equally important, i.e., $w_i = 1/m$ for $1 \le i \le m$. Thus, the quality measure defined in this section is more flexible. Finally, the quality of an AOG, $\tau = (\pi_1, \pi_2, \cdots, \pi_m)$, is defined as

$$ql(\tau) = \sum_{1 \le i \le m} w_i \cdot \frac{h(\pi_i)}{\log(|V_i|)},$$

where V_i is the domain of the quasi-identifier i.

4.3 The Search for Optimal AOGs

Once we can quantitatively measure the security and quality of released data, the search for the optimal AOG for privacy protection becomes an optimization problem. In other words, we have to find τ in the set of admissible AOGs that maximizes the objective function $sf(\tau) \cdot ql(\tau)$. There are numerous techniques for solving such problems. Here, we use the EC approach to find the optimal AOG.

The EC approach is a class of nature-inspired methodologies that can solve hard problems. By this approach, a population of possible solutions is initially given. Then, three basic mechanisms of evolution, i.e., *reproduction*, *mutation*, and *selection*, are applied to the population of solutions to produce the next generation of the population. The process is repeated until satisfactory solutions are found, or a pre-determined number of iterations is reached. A basic scheme of the EC algorithm is presented in Figure 7. The algorithm is adapted from the approach introduced in [19].

The initial population of the algorithm is a randomly selected subset of admissible AOGs. At every step t, also called a generation, each AOG in the population $P(t)$ is evaluated according to some predefined fitness function. Then, a subset of AOGs is selected from $P(t)$ according to the result of the evaluation. The selected subset, known as the *mating pool*, is denoted by $P'(t)$. Next, reproduction and mutation operations are applied to AOGs in $P'(t)$ to produce a

```
Procedure EC

  1. Initialize a population $P(0) \subseteq \Pi$
  2. while not done do
       begin
       Evaluate $P(t)$;
       $P'(t) \leftarrow$ Select$[P(t)]$;
       $P''(t) \leftarrow$ Genetic-Op$[P'(t)]$;
       $P(t+1) \leftarrow$ Intro$[P''(t), P(t)]$;
       $t \leftarrow t + 1$
       end
```

Fig. 7. An EC algorithm for AOG

01001|11001	010|0111|001
10001|00111	100|0100|111
$\downarrow$	$\downarrow$
10001|11001	100|0100|001
01001|00111	010|0111|111
1-point crossover	2-points crossover

Fig. 8. Typical reproduction operations in GA

new population $P''(t)$. The AOGs in $P''(t)$ are offspring of those in $P'(t)$. Finally, $P''(t)$, together with $P(t)$, is introduced into the next-generation of the population $P(t+1)$; usually $P(t)$ is simply replaced by $P''(t)$ to form $P(t+1)$.

A concrete implementation of the skeleton in Figure 7 can be achieved by the standard genetic algorithm (GA). In the GA implementation, we assume that each admissible set of partitions, Π_i, is identified by a set of integers $\{0, 1, \cdots, |\Pi_i| - 1\}$; therefore, each partition in Π_i can be encoded as a binary string of length $\lceil \log |\Pi_i| \rceil$, and each AOG can be encoded as a binary string of length $\sum_{1 \leq i \leq m} \lceil \log |\Pi_i| \rceil$. The fitness function of GA is simply the objective function $sf(\cdot)\overline{ql}(\cdot)$. There are a number of ways to perform the selection. The most popular is the roulette wheel method, where each AOG is selected with a probability proportional to its fitness. The typical reproduction operation for GA is *crossover*, which is performed with a fixed probability, called the *crossover rate*, between two selected AOGs. Figure 8 shows two kinds of crossover operation. The mutation operation is performed by flipping bits at random with some small probability, i.e., the mutation rate. Note that the crossover and mutation operations may produce illegal codes that do not correspond to any AOG, so post-processing is necessary to adjust the codes to legal AOGs.

As an example, we use the admissible partitions in Figure 3. We need an 8-bit string to encode an AOG (2 bits for the date of birth, 3 bits for the zip code, and 3 bits for the height). Thus, for example, (01011010) denotes the AOG $(\{mm/yy\}, d_1 d_2 * **, I_{10})$. If a crossover operation

010|11010
001|00101
$\downarrow$
010|00101
001|11010

is carried out, the resultant codes correspond to (1,0,5) and (0,7,2). However, these are not legal encodings of any AOG, since 5 is not a legal code for height and 7 does not correspond to any partition of zip codes. To transform them into legal encodings, we can change 5 to 5mod5=0 and 7 to 7mod6=1; therefore, the offspring of the crossover operation should be 01000000 and 00001010.

5 Related Works

As mentioned in Section 3.1, the granulation approach subsumes two important data protection techniques, generalization[7,11,5,12] and microaggregation[8]. Moreover, rough set theory has been applied to privacy protection previously [20,6,21]. In this section, we further discuss several works related to our approach.

The main concept of logical security models a user's knowledge based on indiscernibility. Traditionally, epistemic logic has been used to represent such knowledge. The relationship between epistemic operators and rough set approximation has been studied extensively[22,23]. Epistemic logic has also been applied to the analysis of security [24,25,26]. The security logic (SL) developed in [24] is a permission-based approach that specifies the knowledge a user is allowed to have, which contrasts with our prohibition-based approach based on the set of confidential sentences. The logic of security (LS) proposed in [25,26] is applied to the analysis of dynamic systems with multiple subjects, where each subject is permitted to know different levels of confidential information according to his role. SL and LS can be applied to the analysis of general security problems; however, our framework is specifically tailored for the database linking problem.

While we are concerned with the issue of *attribute disclosure*, many previous works have addressed the issue of *identity disclosure*. Attribute disclosure occurs when some characteristic of an individual can be inferred more accurately because of the released data, whereas identity disclosure means that an individual can be uniquely identified. The issue of identity disclosure in the database linking context has been addressed in [27,17,11,5,12]. In those works, the main goal of privacy protection is to maintain the anonymity of data records, i.e., to prevent the user from knowing which data record belongs to a specific individual. The k-anonymity criterion mentioned earlier is designed to prevent identity disclosure. However, it has been observed that k-anonymity is not sufficient for attribute disclosure control, so a logical criterion has been formulated to remedy the problem[4]. A similar problem, called *homogeneity attack*, is also observed in [28]. In this case, the l-diversity criterion is proposed to prevent such attacks.

The protection of confidential information has been widely studied in the contexts of disclosure control [29], inference control [30,13,31], access control [32], and data mining [33,34,35,36]. The works most closely related to our approach are those on disclosure control, which modifies data to prevent users from recognizing individual identities in the data or discovering private information about the individuals. Various techniques have been applied in disclosure control. In addition to the granulation approach, whereby released data is made less precise than the original data, other techniques, such as data perturbation [37] or lying [38,39], distort the data to be released. Data perturbation adds noise to the released data, while ensuring that some statistical properties of the whole data set are preserved; whereas lying distorts the truth, i.e., the negation of the correct answer to the user's query to prevent the user from inferring confidential information.

Another important aspect of disclosure control is the assessment of disclosure risk and data quality. A variety of measures for assessing disclosure risk and information loss have been proposed in [40,41,42,17,43,44,28]. Some information measures associated with data tables may also be useful in such assessments. Measures of interest include: Shannon's entropy [45], Kolmogorov's complexity [46], and uncertainty-based information measures [47]. Based on the assessment of disclosure risk and data quality, we can achieve a balance between data availability and privacy protection.

In contrast to our framework for the database linking context, some models have been proposed for dealing with the confidentiality problem in more general contexts [48,49,50,24,51]. Also, complementary to the approach proposed in this paper, some probabilistic or decision-theoretic approaches to data security have been proposed in [20,42,52,53,6,44].

6 Conclusion

Granular computing (GrC) is an emerging computing paradigm developed from Pawlak's rough set theory. In recent years, it has had a strong impact on many application domains. In this paper, we apply GrC techniques to privacy protection in the context of data release. Granulation of the domains of quasi-identifiers makes it possible to release microdata without invading individuals' privacy. An attribute-oriented technique is employed to modify the to-be-released data. To achieve a balance between the quality of the released data and privacy protection, we present a basic search algorithm to find the maximally specific AOGs that satisfy the security requirements. We also discuss the properties that can be utilized to improve the efficiency of the algorithm. Then, we define quantitative measures to assess the security and quality of an AOG, and show that EC techniques can be employed to find the optimal granulation for privacy protection.

To demonstrate the performance of the proposed approach, further theoretical analysis and experimental verification of the proposed optimization algorithm are needed. Moreover, to improve the optimization algorithm, more criteria of data quality and security measures could be considered. In the longer term, we will explore the possibility of applying other GrC techniques, such as reduct computation and dependency analysis, to resolve practical data security problems.

References

1. Wang, D., Liau, C., Hsu, T.: Attribute-oriented granulation for privacy protection. In: Proceedings of the 2nd IEEE International Conference on Granular Computing. (2006) 726–731
2. Willenborg, L., de Waal, T.: Statistical Disclosure Control in Practice. Springer (1996)
3. Chiang, Y., Hsu, T.s., Kuo, S., Liau, C., Wang, D.: Preserving confidentiality when sharing medical database with the Cellsecu system. International Journal of Medical Informatics **71** (2003) 17–23

4. Hsu, T., Liau, C., Wang, D.: A logical model for privacy protection. In: Proceedings of the 4th International Conference on Information Security. LNCS 2200, Springer-Verlag (2001) 110–124
5. Sweeney, L.: Achieving k-anonymity privacy protection using generalization and suppression. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems **10** (2002) 571–588
6. Wang, D., Liau, C., Hsu, T.: Medical privacy protection based on granular computing. Artificial Intelligence in Medicine **32** (2004) 137–149
7. Wang, D., Liau, C., Hsu, T.: An epistemic framework for privacy protection in database linking. Data and Knowledge Engineering (2006)
8. Domingo-Ferrer, J., Torra, V.: Ordinal, continuous and heterogeneous k-anonymity through microaggregation. Data Mining and Knowledge Discovery **11** (2005) 195–212
9. Pawlak, Z.: Rough Sets–Theoretical Aspects of Reasoning about Data. Kluwer Academic Publishers (1991)
10. Dalenius, T.: Finding a needle in a haystack - or identifying anonymous census records. Journal of Official Statistics **2** (1986) 329–336
11. Samarati, P.: Protecting respondents' identities in microdata release. IEEE Transactions on Knowledge and Data Engineering **13** (2001) 1010–1027
12. Sweeney, L.: k-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems **10** (2002) 557–570
13. Denning, D.: Cryptography and Data Security. Addison-Wesley Publishing Company (1982)
14. Grzymala-Busse, J.: Algebraic properties of knowledge representation systems. In: Proceedings of the ACM SIGART International Symposium on Methodologies for Intelligent Systems, ACM Press (1986) 432–440
15. Lin, T.: Mining associations by linear inequalities. In: Proceedings of the 4th International Conference on Data Mining, IEEE Press (2004) 154–161
16. LeFevre, K., DeWitt, D., Ramakrishnan, R.: Incognito: efficient full-domain k-anonymity. In: Proceedings of the 24th ACM SIGMOD International Conference on Management of Data. (2005) 49–60
17. Iyengar, V.: Transforming data to satisfy privacy constraints. In: Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mmining. (2002) 279–288
18. Pawlak, Z.: Rough sets and fuzzy sets. Fuzzy Sets and Systems **17** (1985) 119–123
19. Pena-Reyes, C., Sipper, M.: Evolutionary computation in medicine: An overview. Artificial Intelligence in Medicine **19** (2000) 1–23
20. Chiang, Y., Chiang, Y., Hsu, T., Liau, C., Wang, D.: 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 2475, Springer-Verlag (2002) 226–233
21. Wang, D., Liau, C., Hsu, T., Chen, J.P.: Value versus damage of information release: A data privacy perspective. International Journal of Approximate Reasoning **43** (2006) 179–201
22. Orłowska, E.: Logic for reasoning about knowledge. Zeitschrift f. Math. Logik und Grundlagen der Math **35** (1989) 559–572
23. Orłowska, E.: Kripke semantics for knowledge representation logics. Studia Logica **XLIX** (1990) 255–272
24. Glasgow, J., MacEwen, G., Panangaden, P.: A logic for reasoning about security. ACM Transactions on Computer Systems **10** (1992) 226–264

25. Bieber, P., Cuppens, F.: A definition of secure dependencies using the logic of security. In: Proc. of the 4th IEEE Computer Security Foundations Workshop. (1991) 2–11
26. Cuppens, F.: A logical formalization of secrecy. In: Proc. of the 6th IEEE Computer Security Foundations Workshop. (1993) 53–62
27. Dawson, S., di Vimercati, S.D.C., Lincoln, P., Samarati, P.: Maximizing sharing of protected information. Journal of Computer and System Sciences **64** (2002) 496–541
28. Machanavajjhala, A., Gehrke, J., Kifer, D., Venkitasubramaniam, M.: l-diversity: privacy beyond k-anonymity. In: Proceedings of The 22nd International Conference on Data Engineering. (2006)
29. Bethlehem, J., Keller, W., Pannekoek, J.: Disclosure control of microdata. Journal of the American Statistical Association **85** (1990) 38–45
30. Brodsky, A., Farkas, C., Jajodia, S.: Secure databases: Constraints, inference channels, and monitoring disclosures. IEEE Transactions on Knowledge and Data Engineering **12** (2000) 900–919
31. Morgenstern, M.: Controlling logical inference in multilevel database systems. In: Proc. of the IEEE Symposium on Security and Privacy. (1988) 245–255
32. Bonatti, P., Damiani, E., di Vimercati, S.D.C., Samarati, P.: An access control model for data archives. In: Proceedings of the 16th International Conference on Information Security: Trusted Information: The New Decade Challenge. (2001)
33. Agrawal, D., Aggarwal, C.: On the design and quantification of privacy preserving data mining algorithms. In: Proceedings of the 12th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems. (2001) 247–255
34. Clifton, C., Kantarcıoğlu, M., Vaidya, J.: Privacy-preserving data mining. In Chu, W., Lin, T., eds.: Foundations and Advances in Data Mining. Springer-Verlag (2005) 313–344
35. Saygin, Y., Verykios, V., Clifton, C.: Using unknowns to prevent the discovery of association rules. SIGMOD Record **30** (2001) 45–54
36. Srikant, R.: Privacy preserving data mining: challenges and opportunities. In: Proceedings of the 6th Pacific-Asia Conference on Knowledge Discovery and Data Mining. LNCS 2336, Springer-Verlag (2002) 13
37. Muralidhar, K., Sarathy, R.: Security of random data perturbation methods. ACM Transactions on Database Systems **24** (1999) 487–493
38. Biskup, J., Bonatti, P.: Confidentiality policies and their enforcement for controlled query evaluation. In: Proceedings of the 2nd European Symposium on Research in Computer Security. LNCS 2502, Springer-Verlag (2002) 39–55
39. Bonatti, P., Kraus, S., Subrahmanian, V.: Foundations of secure deductive databases. IEEE Transactions on Knowledge and Data Engineering **7** (1995) 406–422
40. Damiani, E., di Vimercati, S.D.C., Jajodia, S., Paraboschi, S., Samarati, P.: Balancing confidentiality and efficiency in untrusted relational dbmss. In: Proceedings of the 10th ACM Conference on Computer and Communication Security. (2003) 93–102
41. Domingo-Ferrer, J.: Advances in inference control in statistical databases: An overview. In: Inference Control in Statistical Databases: From Theory to Practice. LNCS 2316, Springer-Verlag (2002) 1–7
42. Hsu, T., Liau, C., Wang, D., Chen, J.: Quantifying privacy leakage through answering database queries. In: Proceedings of the 5th International Conference on Information Security. LNCS 2433, Springer-Verlag (2002) 162–175

43. Truta, T., Fotouhi, F., Barth-Jones, D.: Privacy and confidentiality management for the microaggregation disclosure control method: disclosure risk and information loss measures. In: Proceeding of the ACM Workshop on Privacy in the Electronic Society. (2003) 21–30
44. Wang, D., Liau, C., Hsu, T.s., Chen, J.: On the damage and compensation of privacy leakage. In: Proceedings of the 18th Annual IFIP WG 11.3 Working Conference on Data and Applications Security, Kluwer Academic Publisher (2004) 311–324
45. Shannon, C.: The mathematical theory of communication. The Bell System Technical Journal **27** (1948) 379–423,623–656
46. Li, M., Vitanyi, P.: An introduction to Kolmogorov Complexity and its Applications. Springer-Verlag (1993)
47. Klir, G., Wierman, M.: Uncertainty-Based Information : Elements of Generalized Information Theory. Physica-Verlag (1998)
48. Cholvy, L., Cuppens, F.: Analysing consistency of security policies. In: Proc. of the IEEE Symposium on Security and Privacy. (1997) 103–112
49. Cuppens, F., Demolombe, R.: A deontic logic for reasoning about confidentiality. In Brown, M., Carmo, J., eds.: Deontic logic, agency, and normative systems: ΔEON'96, Third International Workshop on Deontic Logic in Computer Science. (1996) 66–79
50. Cuppens, F., Demolombe, R.: A modal logical framework for security policies. In Ras, Z., Skowron, A., eds.: Proc. of the 10th International Symposium on Methodologies for Intelligent Systems. LNAI 1325, Springer-Verlag (1997) 579–589
51. Syverson, P., Stubblebine, S.: Group principals and the formalization of anonymity. In: Proc. of the 1999 World Congress on Formal Methods. LNCS 1708 (1999) 814–833
52. III, J.G., Syverson, P.: A logical approach to multilevel security of probabilistic systems. Distributed Computing **11** (1998) 73–90
53. Syverson, P., III, J.G.: The epistemic representation of information flow security in probabilistic systems. In: Proc. of the 8th IEEE Computer Security Foundations Workshop. (1995) 152–166