Journal of Zhejiang University SCIENCE A
ISSN 1673-565X (Print); ISSN 1862-1775 (Online)
www.zju.edu.cn/jzus; www.springerlink.com
E-mail: jzus@zju.edu.cn

ε-inclusion: privacy preserving re-publication of dynamic datasets

Qiong WEI[†], Yan-sheng LU, Lei ZOU

(*School of Computer Science and Technology, Huazhong University of Science and Techndogy, Wuhan 430074, China*)

[†]E-mail: weijoan@gmail.com

Received Nov. 17, 2007; revision accepted Mar. 28, 2008

Abstract: This paper presents a novel privacy principle, *ε*-inclusion, for re-publishing sensitive dynamic datasets. *ε*-inclusion releases all the quasi-identifier values directly and uses permutation-based method and substitution to anonymize the microdata. Combined with generalization-based methods, *ε*-inclusion protects privacy and captures a large amount of correlation in the microdata. We develop an effective algorithm for computing anonymized tables that obey the *ε*-inclusion privacy requirement. Extensive experiments confirm that our solution allows significantly more effective data analysis than generalization-based methods.

Key words: Privacy preservation, Re-publication, *ε*-inclusion, Privacy principle
doi:10.1631/jzus.A071595 **Document code:** A **CLC number:** TP391; TP311

INTRODUCTION

The need to preserve private information while publishing data for statistical processing is a widespread problem (Sweeney, 2000; Byun *et al*., 2006; Machanavajjhala *et al*., 2006). Conventional methods for privacy-preserving data publishing are to remove the attributes that clearly identify individuals. However, recent research has shown that a large fraction of the US population can be identified using non-key attributes such as date of birth, gender, and zip code (Sweeney, 2000). The attributes that can be utilized in a linking attack to recover individuals' identities are called 'quasi-identifier' (QI) attributes.

Generalization (Samarati and Sweeney, 1998a; Sweeney, 2002a) is a popular methodology to thwart linking attacks. It divides the microdata into 'QI-groups', and then transforms the QI-values in each group to a uniform format. *k*-anonymity (Iyengar, 2002; Sweeney, 2002b; Bayardo and Agrawal, 2005; Fung *et al*., 2005; LeFevre *et al*., 2005; 2006b) is the first anonymization principle based on generalization in the literature. *k*-anonymity requires each QI-group to contain at least *k* records. Due to its pioneering nature, however, *k*-anonymity places no constraint on the sensitive attribute values in each QI-group. Absence of such constraints may result in a 'homogenous' QI-group, where all records possess exactly the same sensitive attribute value. So once an adversary realizes that a victim is in a homogeneous QI-group, he/she can correctly infer the sensitive value of the victim with a probability 100%.

Prevention of homogeneity is equivalent to ensuring adequate diversity in the sensitive attribute values of a QI-group. This is the motivation of *l*-diversity (Machanavajjhala *et al*., 2006), which requires each QI-group to contain at least *l* 'well-represented' sensitive values. Thus, *l*-diversity can ensure that an adversary correctly infers the sensitive information of any individual involved in the microdata with a probability of at most 1/*l*. The analysis of (Xiao and Tao, 2006a) has proved that *l*-diversity always guarantees stronger privacy preservation than *k*-anonymity.

The existing principles for generalization include variance control (LeFevre *et al*., 2006a), *t*-closeness (Li *et al*., 2007), (*k*, *e*)-anonymity (Zhang *et al*., 2007), (*c*, *k*)-safety (Martin *et al*., 2007), privacy skyline (Chen *et al*., 2007), *δ*-presence (Nergiz *et al*., 2007), perturbed generalization (Tao *et al*., 2008) and

(ε, m)-anonymity (Li *et al.*, 2008). They achieve different types of privacy protection; therefore, the choice of a principle depends on the needs of the underlying application. Anonymized publication can also be achieved by other methodologies. Kifer and Gehrke (2006) developed marginal publicalion, which releases the anonymized versions of the projections of the microdata on different subsets of attributes. Xiao and Tao (2006b) advocated anatomy that publishes the QI and sensitive attribute values directly in two different tables.

However, these methods focus on the problems in static data publication. They cannot prevent adversaries from threatening the privacy of dynamic data re-publication. Byun *et al.*(2006) first proposed a solution to effectively prevent adversaries from inferring individuals' sensitive information by combining several released datasets. However, this solution only supports insertions, and has some drawbacks (Xiao and Tao, 2007).

A new generalization-based principle, m-invariance, has been proposed by Xiao and Tao (2007) to limit the risk of privacy disclosure in re-publication of fully dynamic datasets, which can be modified by any sequence of insertions and deletions. m-invariance uses counterfeited generalization (CG) to ensure that a record's signatures at different release time are the same in the presence of critical absence. However, previous work (Meyerson and Williams, 2004; Aggarwal, 2005) has proved that the essential drawback of generalization-based anonymization approaches is considerable information loss from the microdata.

To maximize the utility of the released data, m-invariance needs to minimize the number of counterfeit records and the amount of generalization on the QI attributes. However, we observe that once given the microdata table $T(j)$ ($1\leq j\leq n$), the deleted records and the inserted records are fixed, namely the number of critical absences is fixed. In order to ensure the m-invariance in all QI-groups of the same record at different snapshots, the number of counterfeit records must be equal to the number of critical absences. Consequently, the number of counterfeit records cannot be minimized. Furthermore, CG will insert some counterfeit records into the released table, which in turn increases the space cost. CG needs to find an appropriate general function for each record in the microdata, which also increases the computational cost of the released tables.

In this paper, we propose a novel privacy principle, ε-inclusion, which directly releases the exact QI values and uses permutation-based anonymizaton approach to anonymize the microdata. To prevent an adversary from inferring the sensitive value of any individual by combining several released tables, we use substitution to ensure ε-inclusion in all the QI-groups of the same record at different release time.

To illustrate the idea, consider the microdata Tables 1a and 2a at different release time. The hospital has published Table 1b with respect to the microdata Table 1a and tries to release an anonymized version of Table 2a. Our method leads to publication of Table 2b, and an auxiliary Table 2c, which indicates that an individual's sensitive value is replaced in QI-group 1.

Without any other information, the adversary cannot confirm which record's sensitive value is replaced in QI-group 1. Assume that the adversary who has the precise QI details of Bob attempts to infer the disease of Bob from Tables 1b, 2b and 2c. He/She knows that the record of Bob must be in QI-group 1 of Tables 1b and 2b. S_1={dyspepsia, bronchitis} and S_2={dyspepsia, gastritis, flu, bronchitis}, where S_1 and S_2 are the sets of Bob's probable sensitive values according to Tables 1b and 2b, respectively. $S_1\subseteq S_2$, so the adversary cannot eliminate any disease that Bob cannot have contracted. Note that, although the adversary learns that an individual's sensitive value is replaced in QI-group 1 of Table 2b, he/she still cannot know which disease is replaced.

Table 1 Microdata T(1) and released data T^*(1)

(a) Microdata T(1)

Name	Age	Zip.	Disease
Bob	21	12k	dyspepsia
Alice	22	14k	bronchitis
Andy	24	18k	flu
David	23	25k	gastritis
Gary	41	20k	flu
Helen	36	27k	gastritis
Jane	37	33k	dyspepsia
Ken	40	35k	flu

(b) Released data T^*(1)

Group-ID	Age	Zip.	Disease
1	21	12k	bronchitis
1	22	14k	dyspepsia
2	24	18k	gastritis
2	23	25k	flu
3	41	20k	gastritis
3	36	27k	flu
4	37	33k	flu
4	40	35k	dyspepsia

Table 2 Microdata *T*(2) and released data *T(2)**

(a) Microdata *T*(2)

Name	Age	Zip.	Disease
Bob	21	12k	dyspepsia
David	23	25k	gastritis
Mary	24	30k	pneumonia
Emily	25	21k	flu
Jane	37	33k	dyspepsia
Gary	41	20k	flu
Tom	60	44k	gastritis
Vince	65	36k	flu
Ray	22	27k	colitis
Paul	52	34k	pneumonia
Steve	56	40k	gastritis

(b) Released data *T**(2)

Group-ID	Age	Zip.	Disease
1	21	12k	flu
1	24	30k	dyspepsia
1	25	12k	gastritis
1	56	40k	bronchitis
2	23	25k	flu
2	41	20k	gastritis
3	22	27k	gastritis
3	37	33k	flu
3	52	34k	dyspepsia
3	60	44k	pneumonia
3	65	36k	colitis

(c) Substitution statistics

Group-ID	Count
1	1

In summary, we make the following contributions in this paper. First, we release the exact QI values to avoid generalization on QI attributes, which in turn avoids considerable information loss from the microdata, and use substitution and permutation-based anonymization approach to provide privacy preservation. Second, we propose a privacy principle, ε-inclusion, to deal with privacy disclosure in data re-publication, and use substitution to facilitate the enforcement of ε-inclusion in the presence of critical absence. Third, based on ε-inclusion, we propose an effective algorithm to compute the released tables. Finally, extensive experiments show that our method significantly outperforms *m*-invariance in both effectiveness of data analysis and computational cost.

PRELIMINARIES

Let *T* be a microdata table. *T* contains *d* QI attributes denoted as A_i^{QI} $(1\le i\le d)$. Assume that *T* contains one sensitive attribute denoted as A^{S}. Each QI attribute can be either numerical or categorical. The sensitive attribute is categorical. For any record $r\in T$, we denote $r.A_i^{\mathrm{QI}}$ $(1\le i\le d)$ as the QI value of *r*, and $r.A^{\mathrm{S}}$ as its sensitive value.

As time evolves, *T* is updated with insertions and deletions which can arrive in any order. We use an integer *j* to denote the timestamp of the *j*th publication. Let *T*(*j*) be the snapshot of *T* at time *j*. The publisher releases a pair of tables $\{T^*(j), S(j)\}$, where $T^*(j)$ anonymizes *T*(*j*), and *S*(*j*) is an auxiliary table providing some substitution information about $T^*(j)$. In particular, the ε-inclusion of all the released tables is achieved through replacing some sensitive values in the presence of critical absence. Before formalizing this new concept, we clarify several basic notions as follows.

Definition 1 (QI-group/Partition) For a microdata table *T*(*j*), a QI-group is a subset of the records in *T*(*j*). A partition of *T*(*j*) consists of all the QI-groups $QI_1(j)$, $QI_2(j)$, …, $QI_{N_j}(j)$, specifically, $\forall QI_i(j)$, $QI_k(j)$, $1\le i\ne k\le N_j$, $QI_i(j)\cap QI_k(j)=\varnothing$ and $\bigcup_{i=1}^{N_j} QI_i(j)=T(j)$, where N_j is the number of QI-groups. Each QI-group is assigned a unique ID in the partition.

Definition 2 (Permutation) Let $\{QI_1(j), QI_2(j), \ldots, QI_{N_j}(j)\}$ be a partition of *T*(*j*). A random permutation over $B=\{1, 2, \ldots, |QI_k(j)|\}$ $(1\le k\le N_j)$, where $|QI_k(j)|$ is the number of the records contained in $QI_k(j)$, can be defined as a mapping f: $B\rightarrow B$. It satisfies the following requirements:

(1) It maps each value $b_i\in B$ to another value $b_j\in B$, which is represented as $f(b_i)=b_j$;

(2) $\forall b_i$, $b_j\in B$ and $b_i\ne b_j$, $f(b_i)\ne f(b_j)$;

(3) $f(B)=B$.

It is obvious that a QI-group $QI_i(j)$ in *T*(*j*) is transformed to a QI-group $QI_i^*(j)$ $(1\le i\le N_j)$ in $T^*(j)$ after performing permutation. $\forall r\in T(j)$, $r.QI^*(j)$ denotes the QI-group in $T^*(j)$ that contains the released record of *r*. We refer to $r.QI^*(j)$ as the released group of *r*.

Definition 3 (Signature) (Xiao and Tao, 2007) Let $QI_i^*(j)$ $(1\le i\le N_j)$ be a QI-group in $T^*(j)$ for any $j\in[1, n]$. The signature of $QI_i^*(j)$ is the set of distinct sensitive values in $QI_i^*(j)$, denoted as $QI_i^*(j).S(A^{\mathrm{S}})$.

Definition 4 (Substitutionary publication) The anonymized version $T^*(j)$ of *T*(*j*) is computed based on the partition of *T*(*j*), and has the following properties:

(1) $T^*(j)$ contains a column A^{g} named 'Group-ID', and all QI attributes and the sensitive attribute in *T*(*j*);

(2) Each record $r\in T(j)$ has a released record $r^*\in T^*(j)$ such that $r^*.A_i^{\mathrm{QI}}=r.A_i^{\mathrm{QI}}$ $(1\le i\le d)$, $r^*.A^{\mathrm{g}}$ is the

ID of $QI(j)$ in $T(j)$ that contains r, and $r^*.A^{\mathrm{S}}=v$, where $v \in r.QI(j).S(A^{\mathrm{S}})$;

(3) For each QI-group $QI^*(j)$ in $T^*(j)$, if $r \in QI(j)$ and $S_{\mathrm{ca}}=r.QI^*(j-1).S(A^{\mathrm{S}})-r.QI(j).S(A^{\mathrm{S}}) \neq \varnothing$, there must exist a record $r_{\mathrm{s}} \in QI(j)$ such that $QI^*(j)$ contains r_{s}' substitutionary records r_{s}^* and $r_{\mathrm{s}}^*.A^{\mathrm{S}}=v'$, where $v' \in S_{\mathrm{ca}}$, $r_{\mathrm{s}}^*.A^{\mathrm{g}}$ is the ID of $QI(j)$, and $r_{\mathrm{s}}^*.A_i^{\mathrm{QI}} = r_{\mathrm{s}}.A_i^{\mathrm{QI}}$ for $1 \le i \le d$.

Definition 5 (Auxiliary table) The auxiliary table $S(j)$ accompanying $T^*(j)$ has two columns 'Group-ID' and 'Count'. In order to ensure ε-inclusion of all the released tables, for any record $r \in T(j) \cap T(j-1)$ $(2 \le j \le n)$, if $r.QI^*(j-1).S(A^{\mathrm{S}}) \not\subseteq r.QI(j).S(A^{\mathrm{S}})$, $QI^*(j)$ contains some substitutionary records.

Example 1 We set $T(1)$ and $T(2)$ to Tables 1a and 2a, $T^*(1)$ and $T^*(2)$ to Tables 1b and 2b, respectively, and $S(2)$ to Table 2c. Consider r the record <Bob, 21, 12k, dyspepsia> in $T(2)$. $r.QI^*(1).S(A^{\mathrm{S}})=$ {bronchitis, dyspepsia} and $r.QI(2).S(A^{\mathrm{S}})=${dyspepsia, pneumonia, flu, gastritis}. $r.QI^*(2)$ must contain one substitutionary record, since $r.QI^*(1).S(A^{\mathrm{S}}) \not\subseteq r.QI(2).S(A^{\mathrm{S}})$. In particular, the record <Mary, 24, 30k, bronchitis> is a substitutionary record, and the actual record in $T(2)$ is <Mary, 24, 30k, pneumonia>. $S(2)$ summarizes the number of substitutionary records in each QI-group.

An adversary may also possess 'background knowledge' that does not exist in those released tables. In this paper, we assume that the adversary has the background knowledge described as follows:

Definition 6 (Prior knowledge) At time n, an adversary's prior knowledge includes:

(1) The deployed privacy principle;

(2) The identities and QI-values of all the records in $U(n)=\bigcup_{j=1}^{n} T(j)$ and the lifespan of each record.

ANALYSIS

When the publisher is preparing $\{T^*(n), S(n)\}$, where $n \ge 1$, he/she must take into account the information that the adversary can combine with $T^*(n)$ to intrude privacy. Apparently, such information includes all the data in $T^*(1)$, $T^*(2)$, …, $T^*(n-1)$ released previously and the prior knowledge. Before releasing $T^*(n)$, the publisher must guarantee that the privacy of every record in $U(n)$ has been adequately protected. According to the analysis of (Xiao and Tao, 2007), we know that the reason of privacy disclosure in dynamic data publishing is critical absence, and that the cardinality of the intersection of the signatures of a record's released groups in different release time decreases. Our method uses substitution to ensure ε-inclusion of the sequence of released tables $T^*(1)$, $T^*(2)$, …, $T^*(n)$.

If an adversary correctly infers the sensitive value of any record $r \in U(n)$ by combining $T^*(1)$, $T^*(2)$, …, $T^*(n)$ and the prior knowledge, we think that the privacy of r is breached. So we should prevent the adversary from combining $T^*(1)$, $T^*(2)$, …, $T^*(n)$ and the prior knowledge. Considering this case: $\forall r \in U(n)$ and $[x, y]$ is the lifespan of record r, we assume that $QI^*(x)$, $QI^*(x+1)$, …, $QI^*(y)$ are the released groups of record r at different release time. Let $|QI^*(x).S(A^{\mathrm{S}}) \cap QI^*(x+1).S(A^{\mathrm{S}}) \cap \ldots \cap QI^*(y).S(A^{\mathrm{S}})|=\alpha$. Since $|QI^*(x).S(A^{\mathrm{S}})|$ is ascertained first, if $\alpha<|QI^*(x).S(A^{\mathrm{S}})|$, we think the privacy of record r is breached. Specifically, he/she can correctly infer the individual's sensitive information when $\alpha=1$. So $|QI^*(x).S(A^{\mathrm{S}}) \cap QI^*(x+1).S(A^{\mathrm{S}}) \cap \ldots \cap QI^*(y).S(A^{\mathrm{S}})|$ determines the degree of privacy disclosure, and should be maximized. It is obvious that α is the largest when $QI^*(x).S(A^{\mathrm{S}}) \subseteq QI^*(x+1).S(A^{\mathrm{S}}) \subseteq \ldots \subseteq QI^*(y).S(A^{\mathrm{S}})$.

On the other hand, even though the sensitive values in each QI-group are distinct, the released table cannot withstand similarity attacks if the sensitive values in each QI-group are semantically similar. According to these observations, we should make the distinct sensitive values in each QI-group be different enough. We give the definition of ε-distinctness as follows:

Definition 7 (ε-distinctness) A released table $T^*(j)$ $(1 \le j \le n)$ is ε-distinct, if each QI-group in $T^*(j)$ contains at least λ categories of the sensitive values and each category contains at least τ distinct sensitive values. Specifically, each QI-group in $T^*(j)$ contains at least $\varepsilon=\lambda*\tau$ records and all the records have different sensitive values.

Clearly, Definition 7 describes the privacy principle for each released table, but the adversary may combine the released tables $T^*(1)$, $T^*(2)$, …, $T^*(n)$ to infer the individuals' sensitive values. So we should require that all the released tables satisfy some restrictions.

Definition 8 (ε-inclusion) A sequence of released tables $T^*(1)$, $T^*(2)$, …, $T^*(n)$ ($n\geq1$) is ε-inclusive if the following conditions hold:

(1) $T^*(j)$ is ε-distinct $\forall j\in[1, n]$;

(2) For any record $r\in U(n)$ with lifespan $[x, y]$, $QI^*(x).S(A^S)\subseteq QI^*(x+1).S(A^S)\subseteq\ldots\subseteq QI^*(y).S(A^S)$, where $QI^*(j)$ is the released group of r at time $j\in[x, y]$.

The rational of ε-inclusion is that, if a record r is published several times, the signature of its released group in the later released table must include the signature of its released group in the earlier released table. In fact, ε-inclusion implies m-invariance, but not the vice versa.

A good publication method should preserve both privacy and correlation between QI and sensitive attributes. Obviously, for any record $r\in T(j)$ ($1\leq j\leq n$), every publication method will lose certain information of r. On the other hand, the method should permit development of an approximate modeling of r. Hence, the quality of correlation preservation depends on how accurate the re-constructed modeling is. ε-inclusion can not only provide enough privacy protection for the individuals but also ensure that the correlation between QI attributes and sensitive attributes does not deviate significantly from the actual correlation.

Lemma 1 Given a microdata table T, let T_g^* denote the released table obtained by generalization and T_p^* denote the released table obtained by ε-inclusion, then $E_g\geq E_p$, where E_g denotes the correlation deviation between T and T_g^*, and E_p denotes the correlation deviation between T and T_p^*.

Proof The proofs of Lemma 1 and the following lemmas 2 and 3 can be found in the appendix.

According to the definition of ε-inclusion, we can calculate the risk of privacy disclosure quantificationally.

Lemma 2 If $\{T^*(1), T^*(2), \ldots, T^*(n)\}$ is ε-inclusive, then an adversary can correctly infer any individual's sensitive value with a probability of at most $1/\varepsilon$ and any individual's category of the sensitive value with a probability of at most $1/\lambda$.

The algorithm is an incremental approach to performing re-publication. Specially, the publisher only needs the microdata tables $T(n-1)$, $T(n)$ and the last released version $T^*(n-1)$ to prepare $T^*(n)$. The microdata tables $T(1)$, $T(2)$, …, $T(n-2)$ and their released versions do not need to be retained.

Lemma 3 If $\{T^*(1), T^*(2), \ldots, T^*(n-1)\}$ is ε-inclusive, then $\{T^*(1), T^*(2), \ldots, T^*(n-1), T^*(n)\}$ is also ε-inclusive if and only if

(1) $T^*(n)$ is ε-distinct;

(2) For any record $r\in T(n-1)\cap T(n)$, $r.QI^*(n-1).S(A^S)\subseteq r.QI^*(n).S(A^S)$.

ALGORITHM

This section elaborates the computation of the tables $\{T^*(n), S(n)\}$ released at the nth publication. The objective of the algorithm is to compute the released tables which can provide enough privacy protection for any record contained in $U(n)$. Namely, $\forall r\in U(n)$, assume that $[x, y]$ is the lifespan of record r, our algorithm make the condition $r.QI^*(x).S(A^S)\subseteq\ldots\subseteq r.QI^*(y).S(A^S)$ hold.

The released dataset $T^*(n)$ can withstand similarity attacks and privacy disclosure, only if at most $1/\lambda$ of the records in $T(n)-T(n-1)$ and $T(j)$ ($1\leq j\leq n$) have the most frequent sensitive category and at most $1/(\lambda*\tau)=1/\varepsilon$ of the records in $T(n)-T(n-1)$ and $T(j)$ ($1\leq j\leq n$) have the most frequent sensitive value. According to Lemma 3, calculation of $T^*(n)$ requires only microdata tables $T(n-1)$, $T(n)$ and the last released table $T^*(n-1)$. We divide the records in $T(n)$ into two disjoint sets $S_{\text{old}}=T(n)\cap T(n-1)$ and $S_{\text{new}}=T(n)-T(n-1)$. Our algorithm ensures two properties as follows:

(1) For any record $r\in S_{\text{old}}$, $r.QI^*(n-1).S(A^S)\subseteq r.QI^*(n).S(A^S)$;

(2) For any record $r\in S_{\text{new}}$, its released record r^* in $T^*(n)$ is contained in an ε-distinct QI-group.

We produce $T^*(n)$ in three phases: division, assignment and permutation. The rest of this section elaborates each phase in turn. Algorithm 1 presents the ε-inclusion algorithm.

1. Division

For each $r\in S_{\text{old}}$, we define its signature as the signature of its released group in $T^*(n-1)$. This phase simply partitions S_{old} into several clusters. The records belonging to S_{old}, with the same signatures in $T^*(n-1)$ and possessing different sensitive values, are assigned to the same cluster and the cluster possesses the same signature as the records in $T^*(n-1)$. We say that a cluster C is balanced, if and only if every sensitive value in its signature is owned by only one record in C.

Algorithm 1 The ε-inclusion algorithm

```
Input: $T(n-1)$, $T(n)$, $T^*(n-1)$
Output: $T^*(n)$
$\forall r \in S_{\text{old}}$, the signatures are $r.QI^*(n-1).S(A^S)$;
The records belonging to $S_{\text{old}}$, with the same signatures in
  $T^*(n-1)$ and possessing different sensitive values, are
  assigned to the same cluster;
$S_1$=the set of clusters which are not balanced;
for (each cluster $C$ in $S_1$) do
  if $\exists r \in S_{\text{new}}$ and $r.A^S=v \in C.S(A^S)$, and $v$ is not owned by
    any record contained in $C$
    $C=C \cup \{r\}$;
  else
    select an appropriate record $r$ in $S_{\text{new}}$ as a substitu-
      tionary record, then $C=C \cup \{r\}$;
  endif
endfor
$S_2$=the remaining records in $S_{\text{new}}$;
assign the records in $S_2$ to some new clusters which are
  $\varepsilon$-distinct;
$S_3$=the remaining records in $S_2$;
for (each record $r \in S_3$) do
  assign $r$ to a cluster $C_i$ which satisfies $r.A^S \notin C_i.S(A^S)$ and
    the cluster is $\varepsilon$-distinct after inserting the record $r$;
endfor
for ($k$=0 to $N_n$) do
  define $pt$ as a random permutation over $\{1, 2, \ldots, |C_k|\}$;
  for ($i$=0 to $|C_k|$)
    $r'_i.A_j^{\text{QI}} = r_i.A_j^{\text{QI}}$ $(1 \le j \le d)$; $r'_i.A^{\text{S}} = r_{pt(i)}.A^{\text{S}}$;
    insert $(r'_i.A_1^{\text{QI}}, r'_i.A_2^{\text{QI}}, \ldots, r'_i.A_d^{\text{QI}}, r'_i.A^{\text{S}}, k)$ into $T^*(n)$;
  endfor
endfor
```

2. Assignment

In this phase, we assign the records in S_{new} to the clusters, subject to four rules. Firstly, each record $r \in S_{\text{new}}$ can be placed only in one cluster whose signature includes $r.A^{\text{S}}$. Secondly, for any cluster which is not balanced and there is no record $r \in S_{\text{new}}$ such that $r.A^{\text{S}}$ belongs to the signature of the cluster, we should select some appropriate records as substitutionary records from S_{new} and assign them to the cluster. Thirdly, for the remaining records in S_{new}, we group them into some new clusters which are ε-distinct and for any record $r \in S_{\text{new}}$ that cannot be assigned to the new clusters, we can find a cluster whose signature does not include $r.A^{\text{S}}$ and has at most $1/\varepsilon$ records possessing the most frequently used sensitive value, and then assign the record r to it. Fourthly, at the end of the phase, all clusters contain at least ε records and all records have different sensitive values. Each cluster is regarded as a QI-group in $T(n)$.

3. Permutation

After assignment, for each cluster, we define pt as a random permutation over $\{1, 2, \ldots, |C_k|\}$, and for each record $r_i \in C_k$, we obtain a record $r_{pt(i)} \in C_k$ according to pt, and make $r_i.A^{\text{S}}=r_{pt(i)}.A^{\text{S}}$. After permutation, each cluster is regarded as a QI-group in $T^*(n)$.

EXPERIMENTAL RESULTS

In this section, we focus on investigating the utility of the anonymized datasets computed according to ε-inclusion principle. The main experiments we report in this paper focus on query answering accuracy and the computational cost of the released datasets.

The dataset used in our experiments is the real dataset CENSUS (http://www.ipums.org) containing personal information of 600k American adults. In order to compare ε-inclusion with m-invariance, we also create the same two datasets OCC and SAL from CENSUS as m-invariance. OCC includes four QI attributes—age, gender, education, and birthplace, and a sensitive attribute—occupation. SAL contains the same four QI attributes, but a different sensitive attribute—income.

A dynamic microdata table T_{OCC} (T_{SAL}) is created from OCC (SAL). The first version $T_{\text{OCC}}(1)$ contains 200k records randomly sampled from OCC. The other 400k records in OCC are contained in a pool. At the jth ($j \ge 2$) timestamp, $T_{\text{OCC}}(j)$ is obtained by arbitrarily deleting δ records from $T_{\text{OCC}}(j-1)$, and then inserting the same number of records randomly removed from the pool. Here, the so-called 'update volume' δ controls the update rate. We repeat this process up to timestamp H=1+400k/δ. $T_{\text{SAL}}(j)$ is obtained like $T_{\text{OCC}}(j)$.

The experiments in (Xiao and Tao, 2007) have shown that l-diversity fails to support re-publication because it results in a large number of vulnerable records. The number of vulnerable records decreases as δ grows in l-diversity. So in this section, we ignore the comparison with l-diversity. It is obvious that the level of privacy protection of ε-inclusion and that of m-invariance are the same when $\varepsilon=m$. So we focus on comparing our method with m-invariance about the utility of the released data and the computational cost of the released data.

Utility of the released data

Any anonymization method needs to trade off between privacy and utility. In the previous section, we have proved that the correlation structure of the permuted datasets is not significantly affected by ε-inclusion, and that ε-inclusion can provide enough privacy protection for the individuals presented in the microdata. In addition, aggregate queries are important during microdata analysis in a variety of domains and Zhang *et al.*(2007) have verified that a permuted table permits more accurate analysis than a generalized table. So in this paper, we focus on the accuracy of aggregate query results.

In the following set of experiments, we use $T_x^*(j)$, where $1\leq j\leq H$ and x=OCC or SAL, to answer aggregate queries about the original microdata. We regard the accuracy of aggregate query results as the measure of utility of the released tables. Specifically, each query has the form

Select *count*(*) from $T_x(j)$,
where $pred(A_i^{\mathrm{QI}})$ $(i=1,\ 2,\ 3,\ 4)$ and $pred(A^{\mathrm{S}})$.

For each attribute A, $pred(A)$ has a length of $|A|\cdot s^{1/(d+1)}$, where $|A|$ is the domain size of A and s is a query parameter called 'expected query selectivity'. In our experiment, d=4. A workload consists of 10000 queries with the same j and s.

Given a query, we obtain its actual result ACT from the microdata table $T_x(j)$, and compute an estimated answer EST from $T_x^*(j)$. The relative error of a query equals $|ACT-EST|/ACT$. We measure the workload error as the average relative error of all the queries.

Adopting $m=\varepsilon=10$, Fig.1 plots the workload error as a function of time for the released tables of T_x with δ=5k and 40k, respectively. From Fig.1, we know that the average relative error of ε-inclusion is smaller than that of m-invariance. Furthermore, the error does not vary significantly with time, and the error of m-invariance is not sensitive to the update volume δ, but the error of ε-inclusion is smaller when δ=40k. This is expected because critical absence is less likely when a larger number of records are inserted at each timestamp.

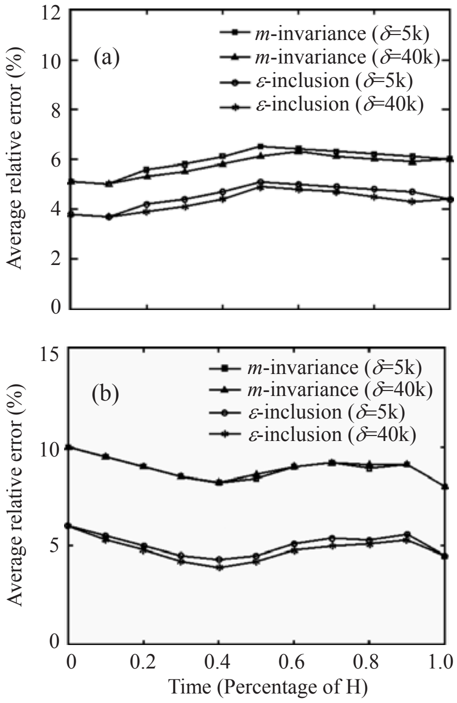
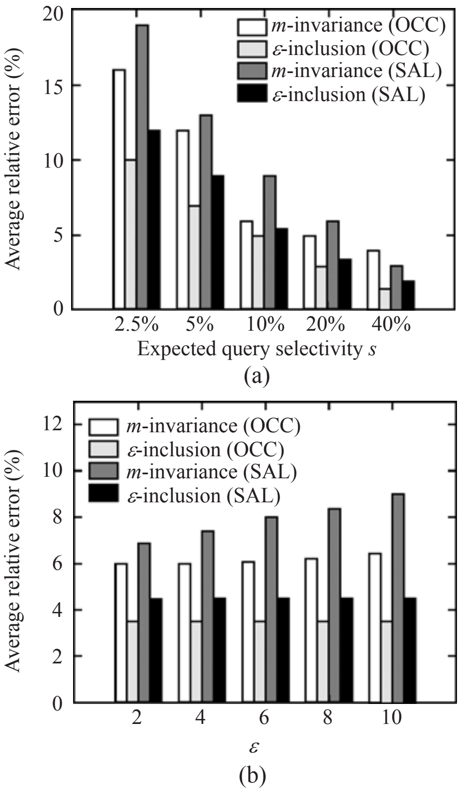


Fig.1 Average query error vs. time. (a) OCC; (b) SAL

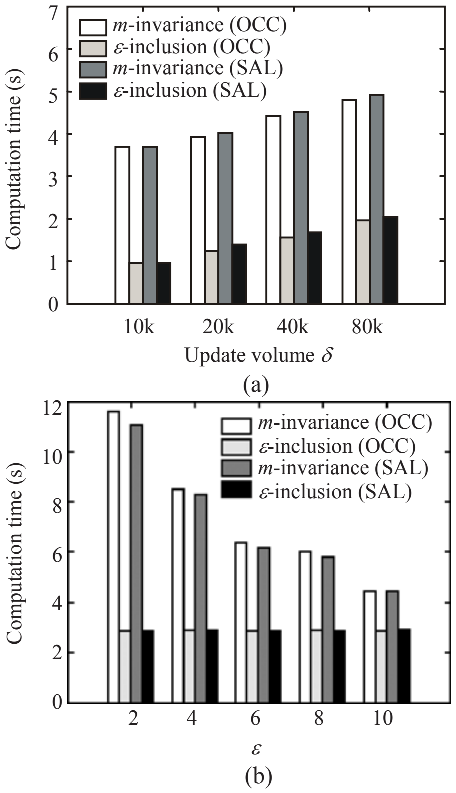
In the next experiment, we focus on the released tables with δ=5k, and the results are given in Fig.2. We measure the average relative error of all workloads performed at each timestamp in the history of the employed released tables. Fig.2a plots the average relative error as a function of the expected query selectivity s for the released tables with $m=\varepsilon=10$. In m-invariance, the accuracy improves as s increases, because a higher s leads to larger query results, whereas in general, aggregate analysis is effective for sizable queries. In ε-inclusion, the accuracy also improves as s increases, because the increase of s results in the increasing of estimated answer EST. Fixing s at 10%, Fig.2b illustrates the average relative error with respect to ε and m. In m-invariance, a smaller m requires less generalization, and hence permits even more accurate analysis. However, in ε-inclusion, the average relative error is not sensitive to ε because the QIs are directly released in ε-inclusion.



(a)

(b)

Fig.2 (a) Average query error vs. expected query selectivity *s*; (b) Average query error vs. *ε*



(a)

(b)

Fig.3 (a) Computation time vs. update volume *δ*; (b) Computation time vs. *ε*

Computational cost

This experiment compares the efficiency of our algorithm and CG algorithm (Xiao and Tao, 2007). First, we set $m=\varepsilon=10$ and measure the average time of computing a released table of different update volume δ. Fig.3a demonstrates the time as a function of δ. The computation time increases as δ increases, this is because that both algorithms need to process more newly inserted records at each timestamp. However, the cost of our algorithm is smaller than that of CG algorithm because our algorithm does not need to balance and split the clusters. Then, we fix δ to 5k and plot the cost as a function of ε and m in Fig.3b. The computation time of CG algorithm decreases as m increases, because a larger m necessitates fewer clusters, and requires a smaller number of cluster splits. The cost of our algorithm does not vary significantly with ε, because fixing δ means that the number of the records our algorithm needs to process is also fixed.

CONCLUSION

This paper proposes a new privacy principle, ε-inclusion, to reduce the risk of privacy disclosure in dynamic data re-publication. In order to enhance the accuracy of data analysis, our method releases the exact QI values without generalization. To protect individual privacy, we use a permutation-based anonymization approach to anonymize the microdata. Experimental results proved that the released tables computed according to our method allow more effective data analysis than m-invariance.

This work also initiates several directions for future work. First, extending our technique to multiple sensitive attributes is an interesting topic. Secondly, it would be exciting to extend the proposed technique to tackle alternative forms of background knowledge.

References

Aggarwal, C.C., 2005. On *k*-anonymity and the Curse of Dimensionality. Proc. Very Large Data Bases, Trondheim, Norway, p.901-909.

Bayardo, R.J., Agrawal, R., 2005. Data Privacy through Optimal *k*-anonymization. Proc. Int. Conf. on Data Engineering, Tokyo, Japan, p.217-228.

Byun, J.W., Sohn, Y., Bertino, E., Li, N., 2006. Secure Anonymization for Incremental Dataset. Secure Data Management, Seoul, Korea, p.48-63. [doi:10.1007/11844662_4]

Chen, B.C., Ramakrishnan, R., LeFevre, K., 2007. Privacy Skyline: Privacy with Multidimensional Adversarial Knowledge. Proc. Very Large Data Bases, Vienna, Austria, p.770-781.

Fung, B.C.M., Wang, K., Yu, P.S., 2005. Top-down Specialization for Information and Privacy Preservation. Proc. Int. Conf. on Data Engineering, Tokyo, Japan, p.205-216.

Iyengar, V.S., 2002. Transforming Data to Satisfy Privacy Constraints. Proc. ACM Knowledge Discovery and Data Mining, Edmonton, Alberta, Canada, p.279-288.

Kifer, D., Gehrke, J., 2006. Injecting Utility into Anonymized Datasets. Proc. ACM Management of Data, Chicago, Illinois, USA, p.217-228.

LeFevre, K., DeWitt, D.J., Ramakrishnan, R., 2005. Incognito: Efficient Full-domain *k*-anonymity. Proc. ACM Management of Data, Baltimore, Maryland, USA, p.49-60.

LeFevre, K., DeWitt, D., Ramakrishnan, R., 2006a. Mondrian Multidimensional *k*-anonymity. Proc. Int. Conf. on Data Engineering, Atlanta, Georgia, USA, p.25.

LeFevre, K., DeWitt, D., Ramakrishnan, R., 2006b. Workload-aware Anonymization. Proc. ACM Knowledge Discovery and Data Mining, Philadelphia, PA, USA, p.277-286.

Li, J., Tao, Y., Xiao, X., 2008. Preservation of Proximity Privacy in Publishing Numerical Sensitive Data. Proc. ACM Management of Data, Vancouver, BC, Canada.

Li, N., Li, T., Venkatasubramanian, S., 2007. *t*-closeness, Privacy Beyond *k*-anonymity and *l*-diversity. Proc. Int. Conf. on Data Engineering, Istanbul, Turkey, p.106-115.

Machanavajjhala, A., Gehrke, J., Kifer, D., Venkitasubramaniam, M., 2006. *l*-diversity: Privacy beyond *k*-anonymity. Proc. Int. Conf. on Data Engineering, Atlanta, Georgia, USA, p.24.

Martin, D., Kifer, D., Machanavajjhala, A., Gehrke, J., Halpern, J., 2007. Worst-case Background Knowledge in Privacy. Proc. Int. Conf. on Data Engineering, Istanbul, Turkey.

Meyerson, A., Williams, R., 2004. On the Complexity of Optimal *k*-anonymity. Proc. ACM Symp. on Principles of Database Systems, Paris, France, p.223-228.

Nergiz, M.E., Atzori, M., Clifton, C., 2007. Hiding the Presence of Individuals from Shared Databases. Proc. ACM Management of Data, Beijing, China, p.665-676.

Samarati, P., Sweeney, L., 1998a. Protecting Privacy When Disclosing Information: *k*-anonymity and Its Enforcement through Generalization and Suppression. Technical Report. SRI International, Carnegie Mellon University.

Samarati, P., Sweeney, L., 1998b. Generalizing Data to Provide Anonymity When Disclosing Information. Proc. ACM Symp. on Principles of Database Systems, Seattle, Washington, USA, p.188.

Sweeney, L., 2000. Uniqueness of Simple Demographics in the U.S. Population. Technical Report, LIDAP-WP4. Laboratory for International Data Privacy, Carnegie Mellon University, PA.

Sweeney, L., 2002a. *k*-anonymity: a model for protecting privacy. *Int. J. Uncert. Fuzz. Knowl.-Based Syst.*, **10**(5): 557-570. [doi:10.1142/S0218488502001648]

Sweeney, L., 2002b. Achieving *k*-anonymity privacy protection using generalization and suppression. *Int. J. Uncert. Fuzz. Knowl.-Based Syst.*, **10**(5):571-588. [doi:10.1142/S021848850200165X]

Tao, Y., Xiao, X., Li, J., Zhang, D., 2008. On Anti-corruption Privacy Preserving Publication. Proc. Int. Conf. on Data Engineering, Cancun, Mexico.

Xiao, X., Tao, Y., 2006a. Personalized Privacy Preservation. Proc. ACM Management of Data, Chicago, Illinois, USA, p.229-249.

Xiao, X., Tao, Y., 2006b. Anatomy: Simple and Effective Privacy Preservation. Proc. Very Large Data Bases, Seoul, Korea, p.139-150.

Xiao, X., Tao, Y., 2007. *m*-Invariance: Towards Privacy Preserving Re-publication of Dynamic Datasets. Proc. ACM Management of Data, Beijing, China, p.689-700.

Zhang, Q., Koudas, N., Srivastava, D., Yu, T., 2007. Aggregate Query Answering on Anonymized Tables. Proc. Int. Conf. on Data Engineering, Istanbul, Turkey, p.116-125.

APPENDIX: PROOFS OF LEMMAS 1~3

Proof of Lemma 1

Each record *r* in the microdata *T* can be regarded as a point in a (*d*+1)-dimensional space *QS*. We model *r* as a probability density function (PDF) $\varsigma_r(x): QS \rightarrow [0,1]$:

$$\varsigma_r(x) = \begin{cases} 1, & x = r, \\ 0, & \text{otherwise}, \end{cases}$$

where *x* is a random variable in *QS*, and the condition *x*=*r* implies $x.A_i^{\text{QI}} = r.A_i^{\text{QI}}$ $\forall i \in [1,\ d]$ and $x.A^{\text{S}}=r.A^{\text{S}}$. In a generalized table, assume *r* belongs to a QI-group QI_k. According to the definition of generalization, the generalized form of *r* is $(\vartheta_k^1, \vartheta_k^2, \cdots, \vartheta_k^d, r.A^{\text{S}})$, where ϑ_k^i $(1 \le i \le d)$ is an interval enclosing $r.A_i^{\text{QI}}$. Denote the length of ϑ_k^i as $|\vartheta_k^i|$. So the reconstructed PDF $\tilde{\varsigma}_r^{\text{g}}(x)$ of *r* is

$$\tilde{\varsigma}_r^{\mathrm{g}}(x)=\begin{cases}\left(\prod_{i=1}^{d}|\vartheta_k^i|\right)^{-1}, \forall i\in[1,d], r.A_i^{\mathrm{QI}}\in\vartheta_k^i,\\ 0, \qquad\qquad \text{otherwise.}\end{cases}$$

Next we discuss the permuted table obtained by ε-inclusion. Also assume QI_k as the released QI-group containing r. According to the definition of ε-distinctness, QI_k contains ε distinct sensitive values $v_1, v_2, \dots, v_\varepsilon$ and the count value in the permuted table corresponding to v_j ($1\le j\le\varepsilon$) equals one. The reconstructed PDF $\tilde{\varsigma}_r^{\mathrm{p}}(x)$ of r is

$$\tilde{\varsigma}_r^{\mathrm{p}}(x)=\begin{cases}\dfrac{c(v_j)}{|QI_k|}=\dfrac{1}{|QI_k|}, & x=(r.A_1^{\mathrm{QI}},r.A_2^{\mathrm{QI}},\cdots,r.A_d^{\mathrm{QI}},v_\alpha)\\ & \wedge\alpha\in[1,\varepsilon-c_1],\\ 0, & \text{otherwise,}\end{cases}$$

where $|QI_k|$ denotes the number of records in QI_k and c_1 denotes the number of substitutionary records, and the QI-values $r.A_1^{\mathrm{QI}},r.A_2^{\mathrm{QI}},\cdots,r.A_d^{\mathrm{QI}}$ of r are directly released in the permuted table. Given an approximate PDF $\tilde{\varsigma}_r$, we quantify its error from the actual ς_r as $E_r=\int_{x\in QS}[\tilde{\varsigma}_r(x)-\varsigma_r(x)]^2\mathrm{d}x$. So we can quantify the correlation deviation of generalized datasets as $E^{\mathrm{g}}=\sum_{\forall r\in T}\int_{x\in QS}[\tilde{\varsigma}_r^{\mathrm{g}}(x)-\varsigma_r(x)]^2\mathrm{d}x$ and the correlation deviation of permuted datasets as $E^{\mathrm{p}}=\sum_{\forall r\in T}\int_{x\in QS}[\tilde{\varsigma}_r^{\mathrm{p}}(x)-\varsigma_r(x)]^2\mathrm{d}x$. Notice that $\tilde{\varsigma}_r^{\mathrm{p}}(x)$ is greater than 0, only when x lies at one of the $\varepsilon-c_1$ points in QS. On the other hand, in practice, $\tilde{\varsigma}_r^{\mathrm{g}}(x)$ typically takes a small value when x distributes across a large region. Obviously, $[\tilde{\varsigma}_r^{\mathrm{g}}(x)-\varsigma_r(x)]^2\ge[\tilde{\varsigma}_r^{\mathrm{p}}(x)-\varsigma_r(x)]^2$, thus $E^{\mathrm{g}}\ge E^{\mathrm{p}}$. Namely, the quality of correlation preservation of ε-inclusion is better than that of generalization-based approaches.

Proof of Lemma 2

It is obvious that there are two types of records in $U(n)$: one is the records whose lifespan is $[x, x]$ ($x\in[1, n]$), and the other is the records whose lifespan is $[x, y]$ ($x<y$, x, $y\in[1, n]$).. Any record r whose lifespan is $[x, x]$ must belong to the table $T^*(x)$. According to the QI-values of r, the adversary can find out that the record r belongs to the QI-group $QI^*(x)$. Without any other information, he/she assumes that every record in $QI^*(x)$ has an equal chance to carry any A^{S} value relevant to $QI^*(x)$. So the adversary may infer the record r's sensitive value with a probability $P_{\mathrm{r}}\{r.A^{\mathrm{S}}=v\}=c(v)/|r.QI^*(x).S(A^{\mathrm{S}})|$ and infer the record r's sensitive category with a probability $P_{\mathrm{r}}\{r.A^{\mathrm{S}}\in C_{\mathrm{s}}\}=c(v)/|QI^*(x).C(A^{\mathrm{S}})|$, where $c(v)$ denotes the number of the sensitive values v involved in $QI^*(x)$, C_{s} denotes a certain category of the sensitive values and $QI^*(x).C(A^{\mathrm{S}})$ denotes the set of distinct categories of the sensitive values in $QI^*(x)$. ε-inclusion ensures that $|r.QI^*(x).S(A^{\mathrm{S}})|\ge\varepsilon$ and $|QI^*(x).C(A^{\mathrm{S}})|\ge\lambda$, and $c(v)=1$. Thus, $P_{\mathrm{r}}\{r.A^{\mathrm{S}}=v\}\le 1/\varepsilon$ and $P_{\mathrm{r}}\{r.A^{\mathrm{S}}\in C_{\mathrm{s}}\}\le 1/\lambda$. Any record r whose lifespan is $[x, y]$ must belong to the tables $T^*(x)$, $T^*(x+1)$, …, $T^*(y)$. According to the QI-values of r, the adversary can find out that the record r belongs to the QI-groups $QI^*(x)$, $QI^*(x+1)$, …, $QI^*(y)$. Without any other information, he/she assumes that the record r has an equal chance to carry any A^{S} value relevant to $r.QI^*(x).S(A^{\mathrm{S}})\cap\dots\cap r.QI^*(y).S(A^{\mathrm{S}})$. So from the adversary's perspective, $P_{\mathrm{r}}\{r.A^{\mathrm{S}}=v\}=c(v)/|S|$, where $S=r.QI^*(x).S(A^{\mathrm{S}})\cap\dots\cap r.QI^*(y).S(A^{\mathrm{S}})$, and $P_{\mathrm{r}}\{r.A^{\mathrm{S}}\in C_{\mathrm{s}}\}=c(v)/|S.C(A^{\mathrm{S}})|$, where $c(v)$ denotes the number of the sensitive values v involved in S and $S.C(A^{\mathrm{S}})$ denotes the set of distinct categories of the sensitive values in S. ε-inclusion ensures that $|S|=|r.QI^*(x).S(A^{\mathrm{S}})|\ge\varepsilon$, $c(v)=1$ and $|r.QI^*(x).C(A^{\mathrm{S}})|\ge\lambda$. Thus, $P_{\mathrm{r}}\{r.A^{\mathrm{S}}=v\}\le 1/\varepsilon$ and $P_{\mathrm{r}}\{r.A^{\mathrm{S}}\in C_{\mathrm{s}}\}\le 1/\lambda$.

Proof of Lemma 3

By Definition 8, if $\{T^*(1), T^*(2), \dots, T^*(n-1), T^*(n)\}$ is ε-inclusive, then the two conditions in Lemma 3 hold. Conversely, since all $T^*(j)$ ($1\le j\le n$) are m-distinct, they satisfy the first requirement in Definition 8. Next we show that the second requirement holds for any record $r\in U(n)$. If $r\notin T(n)$ or $r\in T(n)$ but $r\notin T(n-1)$, it is obvious that the second requirement in Definition 8 trivially holds for r. Consider the case where $r\in T(n-1)\cap T(n)$, since $\{T^*(1), T^*(2), \dots, T^*(n-1)\}$ is ε-inclusive, the released groups of r satisfy $r.QI^*(1).S(A^{\mathrm{S}})\subseteq\dots\subseteq r.QI^*(n-1)$. $S(A^{\mathrm{S}})$. Based on the second condition, we know that $r.QI^*(1).S(A^{\mathrm{S}})\subseteq\dots\subseteq r.QI^*(n-1).S(A^{\mathrm{S}})\subseteq r.QI^*(n).S(A^{\mathrm{S}})$. Therefore, the second requirement in Definition 8 is also satisfied for r.