# Privacy Preserving Serial Data Publishing By Role Composition

Yingyi Bu[1] Ada Wai-Chee Fu[1] Raymond Chi-Wing Wong[2] Lei Chen[2] Jiuyong Li[3]

[1] The Chinese University of Hong Kong
yybu,adafu@cse.cuhk.edu.hk

[2] Hong Kong University of Science and Technology
raywong,leichen@cse.ust.hk

[3] University of South Australia
Jiuyong.Li@unisa.edu.au

## ABSTRACT

Previous works about privacy preserving serial data publishing on dynamic databases have relied on unrealistic assumptions of the nature of dynamic databases. In many applications, some sensitive values changes freely while others never change. For example, in medical applications, the disease attribute changes with time when patients recover from one disease and develop another disease. However, patients do not recover from some diseases such as HIV. We call such diseases permanent sensitive values. To the best of our knowledge, none of the existing solutions handle these realistic issues. We propose a novel anonymization approach called HD-composition to solve the above problems. Extensive experiments with real data confirm our theoretical results.

## 1. INTRODUCTION

Data mining on databases is quite useful. However, publishing data related to individuals to public may compromise individual privacy. For example, a hospital may release patient diagnosis records so that data analysts and researchers can study the characteristics of various diseases. The raw data, also called *microdata*, contain the identities of individuals such as names or keys which should not be released in order to protect individual privacy. However, if an adversary has access to the publicly available voter registration list[1], s/he can discover a large portion of patients' identities by joining the published table and the voter registration list on some attributes such as Age, Sex and Zipcode, which are called *quasi-identifier attributes* (*QID*). In recent years, studies [15, 8, 18, 20, 19, 5, 3, 7] have been made to ensure that the sensitive information of individuals cannot be easily identified in a static table by generalizing *QIDs* to form *anonymized groups* and only publishing those cloaked anonymized groups.

### 1.1 Motivation

Serial publishing for dynamic databases is often necessary when there are insertions, deletions and updates in the microdata. To our best knowledge, there are only three works targeting *parts* of these scenarios. The first one, proposed by Byun et, al. [1], uses delayed publishing to avoid problems caused by insertions, but does not consider deletions and updates. The second work [2] also considers only insertions. The third one, $m$-invariance [21], considers both insertion and deletion and requires that each individual is linked to a fixed set of at least $m$ distinct sensitive values. Counterfeit records sometimes are added into the published table in order to protect privacy in data with deletions. However, data updates have not been considered in $m$-invariance. Our work is motivated by four main challenges.

Firstly, both the QID value and sensitive value of an individual can change, while some special sensitive values should remain unchanged. For example, after a move, the postal code of an individual changes. That is, the external table such as a voter registration list can have multiple releases and changes from time to time. Also, a patient may recover from one disease but develop another disease. The motivating example for [1] and [21] is that the adversary may notice a neighbor being sent to hospital, from which s/he knows that a record for the neighbor must exist in two or more consecutive releases. They further assume that the disease attribute of the neighbor must remain the same in these releases. However, the presence of the neighbor in multiple data releases does not imply that the records for the neighbor will remain the same in terms of the sensitive value. At the same time, some sensitive values that once linked to an individual can never be unlinked. For instance, in medical records, sensitive diseases such as HIV, diabetes and cancers are to this date incurable, and therefore they are expected to persist. We call these values *permanent sensitive values*. Permanent sensitive values can be found in many domains of interest. Some examples are "having a pilot's qualification" and "having a criminal record".

We take $m$-invariance as a representative to illustrate the inadequacy of traditional approaches for the scenarios above, by the following example. In Tables 1, $\mathcal{RL}_1$, $\mathcal{RL}_2$ and $\mathcal{RL}_3$ are snapshots of a voter registration list at times 1, 2 and 3, respectively. The microdata table $T_1$, $T_2$ and $T_3$ are to be anonymized at times 1, 2 and 3, respectively. In Table 2, three tables $T_1^*$, $T_2^*$ and $T_3^*$ are published serially at times 1, 2, and 3, respectively. It is easy to see that $T_1^*$, $T_2^*$ and $T_3^*$ satisfy 3-invariance. This is because in

[1] There are many sources of such an external table. *Most municipalities sell population registers that include the identifiers of individuals along with basic demographics; examples include local census data, voter lists, city directories, and information from motor vehicle agencies, tax assessors, and real estate agencies* [12]. In the voter list, 87% of the voters were identifiable with just the full postal code, gender and birth date [13]. From [15], it is reported that a city's voter list in two diskettes was purchased for twenty dollars, and was used to re-identify medical records.

*VLDB '08,* August 24-30, 2008, Auckland, New Zealand

| PID | Age | Zip. |
|---|---|---|
| $p_1$ | 23 | 16355 |
| $p_2$ | 22 | 15500 |
| $p_3$ | 21 | 12900 |
| $p_4$ | 26 | 18310 |
| $p_5$ | 25 | 25000 |
| $p_6$ | 20 | 29000 |
| $p_7$ | 24 | 33000 |
| ... | ... | ... |
| $p_{\|RL\|}$ | 31 | 31000 |

(a) $\mathcal{RL}_1$

| PID | Age | Zip. |
|---|---|---|
| $p_1$ | 23 | 16355 |
| $p_2$ | 22 | 15500 |
| $p_3$ | 21 | 12900 |
| $p_4$ | 26 | 18310 |
| $p_5$ | 25 | 25000 |
| $p_6$ | 20 | 29000 |
| $p_7$ | 24 | 33000 |
| ... | ... | ... |
| $p_{\|RL\|}$ | 31 | 31000 |

(b)$\mathcal{RL}_2$

| PID | Age | Zip. |
|---|---|---|
| $p_1$ | 23 | 16355 |
| $p_2$ | 22 | 15500 |
| $p_3$ | 21 | 12900 |
| $p_4$ | 26 | 18310 |
| $p_5$ | 25 | *15000* |
| $p_6$ | 20 | 29000 |
| $p_7$ | 24 | 33000 |
| ... | ... | ... |
| $p_{\|RL\|}$ | 31 | 31000 |

(c)$\mathcal{RL}_3$

| PID | Disease |
|---|---|
| $p_1$ | Flu |
| $p_2$ | HIV |
| $p_3$ | Fever |
| $p_4$ | HIV |
| $p_5$ | Flu |
| $p_6$ | Fever |

(d) $T_1$

| PID | Disease |
|---|---|
| $p_1$ | Flu |
| $p_2$ | HIV |
| $p_3$ | *Flu* |
| $p_4$ | HIV |
| $p_5$ | *Fever* |
| $p_6$ | Fever |

(e) $T_2$

| PID | Disease |
|---|---|
| $p_1$ | Flu |
| $p_2$ | HIV |
| $p_3$ | Flu |
| $p_4$ | HIV |
| $p_5$ | Fever |
| $p_6$ | Fever |

(f) $T_3$

**Table 1: Voter Registration List($RL$) and Microdata($T$)**

| PID | G.ID | Age | Zip. | Disease |
|---|---|---|---|---|
| $p_1$ | 1 | [21, 23] | [12k, 17k] | Flu |
| $p_2$ | 1 | [21, 23] | [12k, 17k] | HIV |
| $p_3$ | 1 | [21, 23] | [12k, 17k] | Fever |
| $p_4$ | 2 | [20, 26] | [18k, 29k] | HIV |
| $p_5$ | 2 | [20, 26] | [18k, 29k] | Flu |
| $p_6$ | 2 | [20, 26] | [18k, 29k] | Fever |

(a)First Publication $T_1^*$

| PID | G.ID | Age | Zip. | Disease |
|---|---|---|---|---|
| $p_2$ | 1 | [20, 22] | [12k, 29k] | HIV |
| $p_3$ | 1 | [20, 22] | [12k, 29k] | Flu |
| $p_6$ | 1 | [20, 22] | [12k, 29k] | Fever |
| $p_1$ | 2 | [23, 26] | [16k, 25k] | Flu |
| $p_4$ | 2 | [23, 26] | [16k, 25k] | HIV |
| $p_5$ | 2 | [23, 26] | [16k, 25k] | Fever |

(b)Second Publication $T_2^*$

| PID | G.ID | Age | Zip. | Disease |
|---|---|---|---|---|
| $p_2$ | 1 | [21, 25] | [12k, 16k] | HIV |
| $p_3$ | 1 | [21, 25] | [12k, 16k] | Flu |
| $p_5$ | 1 | [21, 25] | [12k, 16k] | Fever |
| $p_1$ | 2 | [20, 26] | [16k, 29k] | Flu |
| $p_4$ | 2 | [20, 26] | [16k, 29k] | HIV |
| $p_6$ | 2 | [20, 26] | [16k, 29k] | Fever |

(c)Third Publication $T_3^*$

**Table 2: Published Tables $T^*$ satisfying 3-invariance**

any release, for each individual, the set of 3 distinct sensitive values that the individual is linked to in the corresponding anonymized group remains unchanged. Note that HIV is a *permanent* disease but Flu and Fever are *transient* diseases. Furthermore, assume that from the registration lists, one can determine that $p_1, p_2, ..., p_6$ are the only individuals who satisfy the QID conditions for the groups with G.ID = 1 and G.ID = 2 in all the three tables of $T_1^*, T_2^*$ and $T_3^*$. Then surprisingly, the adversary can determine that $p_4$ has HIV with 100% probability. The reason is based on *possible world exclusion* from all published releases. First, we show that $p_1$ and $p_6$ cannot be linked to HIV. Suppose that $p_1$ suffers from HIV. In $T_1^*$, since $p_1, p_2$ and $p_3$ form an anonymized group containing one HIV value, we deduce that both $p_2$ and $p_3$ are not linked to HIV. Similarly, in $T_2^*$, since $p_1, p_4$ and $p_5$ form an anonymized group containing one HIV value, $p_4$ and $p_5$ are non-HIV carriers. Similarly, from $T_3^*$, we deduce that $p_4$ and $p_6$ are not linked to HIV. Then, we conclude that $p_2$, $p_3$, $p_4$, $p_5$ and $p_6$ do not contract HIV. However, in each of the releases $T_1^*$, $T_2^*$ and $T_3^*$, we know that there are two HIV values. This leads to a contradiction. Thus, $p_1$ cannot be linked to HIV. Similarly, by the same inductions, $p_6$ cannot be an HIV carrier. Finally, from the anonymized group with G.ID = 2 in $T_3^*$, we figure out that $p_4$ must be an HIV carrier! No matter how large $m$ is, this kind of possible world exclusion can appear after several publishing rounds. Note that even if the registration list remains unchanged, the same problem can occur since the six individuals can be grouped in the same way as in $T_1^*$, $T_2^*$ and $T_3^*$ at 3 different time, according to the algorithm in [21].

Secondly, the anonymization mechanism for serial publishing should provide *individual-based protection*. Yet previous works [1] and [21] focus on *record-based protection*. In $m$-invariance [21], each record is associated with a lifespan of contiguous releases and a signature which is an *invariant* set of sensitive values linking to $r_j$ in the published table. If a record $r_j$ for individual $p_i$ appears at time $j$, disappears at time $j+1$ (e.g. $p_i$ may discontinue treatment or may switch to another hospital), and reappears at time $j+2$, the appearance at $j+2$ is treated as a new record $r_{j+2}$ in the anonymization process of [21]. There is no memory of the previous signature for $r_j$, and a new signature is created for $r_{j+2}$. Let us take a look at $T_1^*$ in Table 2. From $T_1^*$, we can find that by 3-invariance, the signature of the records for $p_1$ and $p_3$ in $T_1^*$ is {Flu, HIV, Fever}. If $p_1$ and $p_3$ recover from Flu and Fever at time 2 (not in $T_2$), and reappears due to other disease at time 3(in $T_3$), the reappearance of $p_1$ and $p_3$ in $T_3$ is treated as new records $r'_1$, $r'_3$ and by $m$-invariance, there is no constraint for their signature. Thus, at time 3, if the signatures for $r'_1$ and $r'_3$ do not contain HIV, $p_1$ and $p_3$ will be excluded from HIV. Consequently, $p_2$ will be found to have HIV! However, it is not easy to extend $m$-invariance to individual-based protection. For example, binding invariant signatures to individuals is not feasible, since an individual may develop new diseases that are not in the signature.

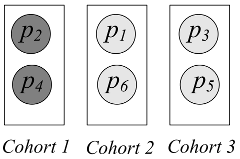
Thirdly, the knowledge model for the adversary should be realistic. Literature [1] and [21] assume that it is trivial to obtain the background knowledge of each individual's presence or absence in every snapshot of the microdata. However, gaining this kind of *participation knowledge* can be as hard as knowing the individual's sensitive values, because one's participation in a microdata snapshot is also confidential. For example, [10] deals with protecting the information about the presence of individuals in a data release. A more plausible scenario is that an adversary knows the participation knowledge of a few close friends.

Fourthly, suppression (removing a record that should exist) is better than counterfeit (a deleted record is not deleted). If a rare but deadly disease such as SARS is suppressed in the published data, we can still broadcast the existence of such a disease without breaching privacy. On the other hand, if a record of SARS is counterfeited but is known to be wiped out, the counterfeit will be discovered and lose its protection power for the previous releases.

## 1.2 Contribution

This paper presents a first study of the problem of privacy preserving serial data publishing with permanent sensitive values and dynamic registration lists. We analyze the difficulty and show that without permanent sensitive values, traditional models such as $\ell$-diversity [8] are enough to protect privacy for serial publishing. However, with permanent sensitive values, the problem becomes very difficult. No generalization is possible if we assume the adversary possesses full participation knowledge. Fortunately, such an assumption is not realistic and we can assume limited participation knowledge instead.

We propose an anonymization method called *HD-composition* which involves two major roles, namely *holder* and *decoy*. The objective is to bound the probability of linkage between any individual and any sensitive value by a given threshold, e.g., 1/$\ell$. Suppose an individual $p_i$ has a sensitive value $s$ in the microdata. One major technique used for anonymizing static data is to form



(a) Cohorts

| PID | G.ID | Age | Zip. | Disease |
|---|---|---|---|---|
| $p_1$ | 1 | [22, 25] | [15k, 17k] | Flu |
| $p_2$ | 1 | [22, 25] | [15k, 17k] | HIV |
| $p_5$ | 1 | [22, 25] | [15k, 17k] | Fever |
| $p_3$ | 2 | [20, 26] | [12k, 29k] | Flu |
| $p_4$ | 2 | [20, 26] | [12k, 29k] | HIV |
| $p_6$ | 2 | [20, 26] | [12k, 29k] | Fever |

(b) $T_3^*$'

**Table 3: Generalization $T_3^*{}'$ by HD-composition**

an anonymized group mixing $p_i$ and other individuals whose sensitive values are not $s$. Merely having the published anonymized groups, the adversary cannot establish strong linkage from $p_i$ to $s$. We also follow this basic principle, where we call the individual to be protected a holder and some other individuals for protection decoys.

We propose two major principles for partitioning: *role-based partition* and *cohort-based partition*. By role-based partition, in every anonymized group of the published data, for each holder of a permanent sensitive value $s$, $\ell - 1$ decoys which are not linked to $s$ can be found. Thus, each holder is masked by $\ell - 1$ decoys. By cohort-based partition, for each permanent sensitive value $s$, we construct $\ell$ cohorts, one for holders and the other $\ell - 1$ for decoys; restrict that decoys from the same cohort cannot be placed in the same partition, this is to imitate the properties of true holders.

Consider the example in Table 1 and Table 2, where $\ell = 3$. Since $p_2$ and $p_4$ are HIV-holders in $T_1$. In Table 3(a), they are both in cohort 1 where all HIV-holders are stored. In $T_1^*$, $p_1$ and $p_3$ form an anonymized group with an HIV-holder (i.e. $p_2$), and they are HIV-decoys. $p_1$ and $p_3$ are inserted into cohort 2 and 3, respectively. Similarly, $p_6$ and $p_5$ are decoys for $p_4$ in $T_1^*$, and are inserted into cohorts 2 and 3, respectively. With the constraints of cohort-based partition, we get anonymized table $T_3^*{}'$(Table 3(b)) rather than the problematic $T_3^*$( Table 2(c)). This is because in $T_3^*$, decoys $p_1$ and $p_6$ are grouped with holder $p_4$, but $p_1$ and $p_6$ are from cohort 2, which violates the constraint of cohort-based partition.

It should be noted that our solution involves a novel idea of utilizing the assumption of a trusted data publisher and the benefit of making the anonymization process known to the public. It is obvious that the data publisher must be trusted. Otherwise, the published data cannot be used for data mining purpose. Hence, the publisher is assumed to follow the anonymization algorithm exactly. Interestingly, the knowledge of the algorithm is then helpful to protect the privacy by ensuring a proper analysis of the probability of linking individuals to sensitive values. More details will be given in Section 3.3.2. Therefore it turns out that providing more knowledge to the adversary actually helps in privacy protection, which is a surprising finding. Another interesting idea is to utilize the external information such as the voter registration lists to facilitate the anonymization. This is first suggested in [10]. The main observation is that there will typically be a lot of individuals who are in the registration list but are "absent" in a data release. These individuals can then help to anonymize the ones who are present.

The rest of this paper is organized as follows. Section 2 formalizes the related concepts and states the problem of serial data publishing. Section 3 presents the solution HD-composition and the corresponding analytical study. Section 4 is an empirical study, demostrating the inadequacy of conventional privacy models and the cost of the proposed techniques. Section 5 surveys the previous work related to ours. Section 6 concludes the paper.

## 2. PROBLEM DEFINITION

Assume that microdata tables $T_1$, $T_2$, ..., $T_n$ are generated at times 1, 2, ..., $n$. Let $\mathcal{RL}_j$ be the snapshot of a publicly available registration list (e.g., a voter registration list) at time $j$. The attributes of $\mathcal{RL}_j$ include: (1) an individual identity attribute $\mathcal{RL}_j.pid$, which is the primary key of $\mathcal{RL}_j$, and (2) $d$ quasi-identifier attributes (QID): $\mathcal{RL}_j.qid_1$, $\mathcal{RL}_j.qid_2$, ..., and $\mathcal{RL}_j.qid_d$. Assume that $\mathcal{I}$ is the set of all individuals, thus at time $j$, $\mathcal{RL}_j$ corresponds to a subset of $\mathcal{I}$. Tables 1(a) to (c) are examples of registration list snapshots. $T_j$ can be seen as a snapshot of a dynamic dataset $T$ at time $j$. The attributes of $T_j$ include: (1) an individual identity attribute $T_j.pid$, which is a *foreign key* referencing $\mathcal{RL}_j$, and (2) a sensitive attribute $T_j.S$. Each instance of $T_j.S$ is a sensitive value. Following the literature's convention [8], the value in $T_j.S$ should be categorical, while the other attributes can be either numerical or categorical. For each individual $p_i \in \mathcal{RL}_j$, $T_j.S(p_i)$ denotes $\pi_S(\sigma_{pid=p_i.pid}(T_j))$, the set of sensitive values associated with individual $p_i$ in $T_j$. Assume that at each time $j$, the publisher releases $T_j$'s anonymized version $T_j^*$ to the public. For example, Tables 1(d) to (f) are 3 snapshots of $T$, while Tables 2(a) to (c) are published anonymized data at times 1, 2, and 3.

Let $\mathcal{S}$ be the set of all sensitive values that can appear in the sensitive attributes. In the set $\mathcal{S}$, some values are *permanent sensitive values*, which form the set $PS$. A permanent sensitive value is a value that will stay with an individual permanently. Hence if $s$ is a permanent sensitive value, and for an individual $p_i$, $s \in T_j.S(p_i)$, then for all times $j' > j$, either $p_i$ has no record in $T_{j'}$, or $s \in T_{j'}.S(p_i)$. The remaining values in $\mathcal{S}$ are called *transient sensitive values* and they form the set $TS$. For example, diseases like Alzheimer, HIV and cancer are incurable and are permanent sensitive values, while diseases like Flu, Diarrhea and SARS are examples of transient sensitive values.

DEFINITION 1 (ANONYMIZED GROUP). *At time j, let $P(\mathcal{RL}_j)$ be a* **partitioning** *of the individuals in $\mathcal{RL}_j$, each partition $\mathcal{P}$ in $P(\mathcal{RL}_j)$ is related to an* **anonymized group** *$\mathcal{AG}_j(\mathcal{P})$. Each $\mathcal{AG}_j(\mathcal{P})$ is associated with (1) a unique group ID, and (2) a predicate Q on the QID attributes covering the $QID$ values of all individuals in $\mathcal{P}$. We also say that $\mathcal{AG}_j(\mathcal{P})$ is the* **hosting group** *of $p_i$ for each $p_i \in \mathcal{P}$. $\mathcal{AG}_j(p_i)$ is also used to refer to $p_i$'s hosting group.*

An example of the predicate $Q$ is a bounding box for numerical QID values. In Table 2(a), there are two anonymized groups, group 1 and group 2, and their intervals of age and zipcode correspond to the predicate $Q$.

DEFINITION 2 (GENERALIZATION). *A* **generalization** *$T_j^*$ of $T_j$ is generated upon a partitioning of $\mathcal{RL}_j$ denoted by $P(\mathcal{RL}_j)$. The attributes of $T_j^*$ include: (1) an anonymized group ID gid, (2) attribute predicate set Q for predicates over the QID attributes, and (3) a sensitive value attribute S. For every individual $p_i$ with a tuple in $T_j$, let $\mathcal{P}$ denote the partition in $P(\mathcal{RL}_j)$ that $p_i$ belongs to, a tuple $t_i^*$ in $T_j^*$ is generated by: (1) $t_i^*.gid$ = $\mathcal{AG}_j(\mathcal{P}).id$, (2) $t_i^*.Q = \mathcal{AG}_j(\mathcal{P}).Q$, and (3) $t_i^*.S$=$T_j.S(p_i)$. Let $T_j^*.S(\mathcal{AG}_j(\mathcal{P}))$ denote $\bigcup_{p_i \in \mathcal{AG}_j(\mathcal{P})} T_j.S(p_i)$*

For example, Table 2 shows three generalizations: $T_1^*$ for $T_1$, $T_2^*$ for $T_2$, and $T_3^*$ for $T_3$.

In many applications (such as the medical databases) most of the individuals in the registration list will not have any record in a particular data release (since most people are healthy). In fact, such absent individuals will play an important role in privacy preservation by being possible culprits. Hence, we introduce different states of existence for individuals. For $p_i \in \mathcal{RL}_j$, when there exists a tuple in $T_j$ with $T_j.pid = p_i$, we say that $p_i$ is **present** in $T_j$ and

| PID | Disease. |
|---|---|
| $p_1$ | Flu |
| $p_2$ | HIV |
| $p_3$ | Flu |
| $p_5$ | Fever |
| $p_6$ | Fever |
| $p_7$ | Ulcer |

(a) Microdata $T_4$

| PID | G.ID | Age | Zip. | Disease |
|---|---|---|---|---|
| $p_1$ | 1 | [20, 26] | [12k, 33k] | Flu |
| $p_2$ | 1 | [20, 26] | [12k, 33k] | HIV |
| $p_3$ | 1 | [20, 26] | [12k, 33k] | Flu |
| $p_5$ | 1 | [20, 26] | [12k, 33k] | Fever |
| $p_6$ | 1 | [20, 26] | [12k, 33k] | Fever |
| $p_7$ | 1 | [20, 26] | [12k, 33k] | Ulcer |

(b) Topmost Generalization $T_4^*$

**Table 4: Example with no $\ell$-scarce generalization**

| PID | Disease. |
|---|---|
| $p_1$ | Flu |
| $p_2$ | Fever |
| $p_3$ | HIV |
| $p_4$ | HIV |
| $p_5$ | Flu |
| $p_6$ | Fever |

(a) $T_1^p$

| PID | Disease. |
|---|---|
| $p_1$ | Flu |
| $p_2$ | Fever |
| $p_3$ | HIV |
| $p_4$ | HIV |
| $p_5$ | Fever |
| $p_6$ | Flu |

(b)$T_2^p$

| PID | Disease. |
|---|---|
| $p_1$ | Fever |
| $p_2$ | Flu |
| $p_3$ | HIV |
| $p_4$ | HIV |
| $p_5$ | Fever |
| $p_6$ | Flu |

$T_3^p$

**Table 5: Possible Table Series**

also is **present** in any *generalization* $T_j^*$ of $T_j$; otherwise, $p_i$ is **absent**. A special case of absence is when an individual cease to exist forever. $p_i$ is said to be **extinct**. We assume that the extinction of individuals is public knowledge.

Next, we need to clarify the knowledge that an adversary may possess for privacy attack. Typically, there are 2 kinds of knowledge considered: *prior knowledge* and *participation knowledge*.

Prior knowledge is known to the public. At time $j$, **prior knowledge** includes: (1) all publicly available registration lists (the one at time $j$ is $\mathcal{RL}_j$), (2) the permanent sensitive value set $PS$, (3) the published table series $\{T_1^*, T_2^*, ..., T_j^*\}$, and (4) the list of extinct individuals.

An adversary can be familiar with their close friends and have the knowledge of their participation information (presence or absence in microdata snapshots) at any time. This is called the **participation knowledge**. Different adversary may have participation knowledge about a different set of individuals, but fortunately the adversary cannot be familiar with every individual thus her/his participation knowledge is always bounded. We assume that the size of any adversary's participation knowledge does not exceed $\mathcal{K}$. That is, s/he can know at most $\mathcal{K}$ individuals' presence or absence information at any time.

At any time $j$, an adversary $\mathcal{A}$ can use her/his prior knowledge and participation knowledge to infer the sensitive values associated with each individual. In order to quantify the attack on serial data publishing, we define the term *privacy disclosure risk*.

DEFINITION 3 (PRIVACY DISCLOSURE RISK). *Let $B_n$ be the prior knowledge at time $n$, and $\mathcal{A}$ be an adversary with the specific participation knowledge $P_{\mathcal{A}}$. The* ***privacy disclosure risk*** *at time $n$ with respect to individual $p_i$ and sensitive value $s$ is given by*

$$risk(p_i, s, n) = \max_{\mathcal{A}} \max_{1 \leq j \leq n} Prob(p_i, s, j | B_n, P_{\mathcal{A}})$$

*where $Prob(p_i, s, j|B_n, P_{\mathcal{A}})$ is the probability that $p_i$ is linked to $s$ at time $j$ given the knowledge of $B_n$ and $P_{\mathcal{A}}$.*

DEFINITION 4 (PROBLEM). *At any time $n$, generate and publish an anonymization $T_n^*$ for $T_n$, the anonymization must ensure that $\forall p_i \in \mathcal{I}$, $\forall s \in S$, $risk(p_i, s, n) \leq \frac{1}{\ell}$. Such a series of anonymization $T_1^*, T_2^*, ..., T_n^*$ is called an $\ell$-scarce anonymization series. We also say that the anonymization series satisfies $\ell$-***scarcity***.

Generalization is one way of anonymization. In Table 2, generalizations $T_1^*$, $T_2^*$ and $T_3^*$ all satisfy $\ell$-diversity[2] ($\ell$=3) and $m$-invariance ($m$=3) but violate $\ell$-scarcity($\ell$=3), as discussed in the introduction part. Yet generalizations $T_1^*$, $T_2^*$ and $T_3^{*\prime}$ satisfy not only $\ell$-diversity ($\ell$=3) but also $\ell$-scarcity ($\ell$=3).

This problem is very difficult when the adversary can have participation knowledge of every individual. Generalization is not possible in some cases. A generalization $T^*$, is called a *topmost generalization* if $T^*$ has only one *anonymized group* $\mathcal{AG}$ where all individuals in $T$ are covered by $\mathcal{AG}$'s *QID* predicate. Obviously, if the topmost generalization could not guarantee privacy, then no generalization can. In Table 4, $T_4$ is the microdata snapshot at time 4, consequent to $T_1, T_2$ and $T_3$ in Table 2, while $T_4^*$ is the topmost generalization for $T_4$. Even if we adopt a topmost generalization for each of $T_1, T_2, T_3$ and $T_4$, we deduce that $T_4^*$ still cannot guard privacy if the adversary knows everyone's participation information (i.e., $\mathcal{K}$=$|\bigcup_j \mathcal{RL}_j|$). This is because $p_4$ has been found to be absent while $p_1, p_2, p_3, p_5, p_6$ are still present at time 4, and one of the two HIV occurrences has disappeared. Since HIV is a permanent sensitive value, $p_4$ must be linked to HIV.

We can adopt the conventional random world assumption on the probabilistic analysis. In serial data publishing, a random world can be defined by a possible series of microdata tables, where by possible, we mean that the series could have been the original microdata tables which have generated the observed published anonymized tables and which do not violate the known knowledge. Hence, we introduce the definitions of *possible table series*.

DEFINITION 5 (POSSIBLE TABLE SERIES). *At time $n$, table series $TS_{possible}^{\mathcal{A}}$=$\{T_1^p, T_2^p, ..., T_n^p\}$ is called a* **possible table series** *for adversary $\mathcal{A}$ if the following requirement is satisfied:*

1. *$\forall T_j^p \in TS_{possible}$, (1) $T_j^*$ is a generalization of $T_j^p$ and (2) $T_j^p$ does not violate the prior knowledge nor participation knowledge of $\mathcal{A}$,*
2. *$\forall p_i \in \mathcal{I}$, $\forall ps \in PS$, if $\exists j$, $ps \in T_j^p.S(p_i)$, then $\forall j' > j$, either $p_i$ is absent in $T_{j'}^p$, or $ps \in T_{j'}^p.S(p_i)$.*

For the example, consider the table series in Table 2. Assume an adversary $\mathcal{A}$ have participation knowledge of individuals $p_1, ..., p_6$. At time 3, besides $\{T_1, T_2, T_3\}$ in Table 1, $\{T_1^p, T_2^p, T_3^p\}$ in Table 5 is also a possible table series.

At time $n$, the **privacy disclosure risk** of linking individual $p_i$ to sensitive value $s$ is:

$$risk(p_i, s, n) = \max_{\mathcal{A}} \max_{1 \leq j \leq n} n_{link}(p_i, s, j)/n_{total}(TS_{possible}^{\mathcal{A}}) \quad (1)$$

where $n_{total}(TS_{possible}^{\mathcal{A}})$ is the number of possible table series for $\mathcal{A}$ at time $n$, and $n_{link}(p_i, s, j)$ is the number of possible table series $\{T_1^p, T_2^p, ..., T_n^p\}$ for $\mathcal{A}$ where $s \in T_j^p.S(p_i)$.

In the above we adopt the random world assumption. A possible world is given by a possible table series. The risk is related to the number of random worlds with assignment of the value $s$ to individuals $p_i$.

.

[2] In this paper, $\ell$-diversity refers to the requirement that in any anonymized group, at most $1/\ell$ of the records contain any sensitive value $v$.

# 3. HD-COMPOSITION

We propose a generalization technique called *HD-composition* to solve the problem in Definition 4. In HD-composition, each permanent sensitive value holder is protected by a number of decoys, which all act like holders to confuse the attacker. This basic idea is similar to that used in $k$-anonymity and $\ell$-diversity, yet it is a dynamic setting here. It is possible that an individual is linked to more than one permanent sensitive values. Thus we shall use a permanent sensitive value as the prefix for roles, because one may serve as roles for different permanent sensitive values concurrently.

Let us formally clarify 3 basic roles in HD-composition. At any time $j > 1$, we specify the time immediately before the anonymization of $T_j$ as time $j^-$, and the time immediately after anonymization with possibly some new role assignments as time $j^+$.

For $p_i \in \mathcal{I}$ and $s \in PS$: if $\exists 1 \leq j' < j$, $s \in T_{j'}.S(p_i)$, then $p_i$ is called an **$s$-holder** from time $j^-$. If $\forall s' \in PS$, $s' \notin T_j.S(p_i)$ and $p_i$ is not a $s$-decoy at time $j^-$, then $p_i$ is called an **$s$-clean individual**. In Sections 3.1 and 3.2, we will describe when and how to select individuals to serve as **$s$-decoys**.

For example, in Table 2, initially no one is assigned any role, at time $1^-$, $p_2$ and $p_4$ are and HIV-holders, while $p_1$, $p_3$, $p_5$ and $p_6$ are HIV-clean individuals. At time $1^+$, after the anonymization process, the roles of HIV-decoys may be assigned to $p_1, p_3, p_5, p_6$.

## 3.1 Role-based Partition

For the first microdata snapshot $T_1$, we generate $T_1^*$ that satisfies $\ell$-diversity [8]. Then, at any time $j > 1$, after $T_{j-1}^*$ is published, before generalizing $T_j$ to $T_j^*$, the roles are updated as follows. For each $p_i$ that becomes $s$-holder at time $j$, $\ell - 1$ individuals are selected to serve as $s$-decoys. After roles are updated, we can anonymize $T_j$ to its generalization $T_j^*$, the constraints of which are specified by a basic principle called *role-based partition* as follows.

PRINCIPLE 1 (ROLE-BASED PARTITION). *At time $j > 1$, each anonymized group $\mathcal{AG}$ in $T_j^*$ satisfies the following conditions:*

1. *For each permanent sensitive value $s$ in $PS$, let $N_d$ be the number of $s$-decoy in $\mathcal{AG}$, and $N_s$ be the number of $s$-holders in $\mathcal{AG}$. Then $N_d = (\ell - 1) \cdot N_s$.*
2. *For each transient sensitive value $s' \in TS$, let $N_{s'}$ be the count of $s'$ in $\mathcal{AG}$. Then $N_{s'} \leq \frac{1}{\ell}|\mathcal{AG}|$,*

   *where $|\mathcal{AG}|$ is the number of individuals that are present in $\mathcal{AG}$.*

By condition 1 in the above, there should always be $\ell - 1$ $s$-decoys for each $s$-holder. Note that if there is no $s$-holder in $\mathcal{AG}$, there should be no $s$-decoys; otherwise, those $s$-decoys will lose their protection functionality since they can be excluded from the possibility of being an $s$-holders, and consequently disclose the $s$-holders they have protected in earlier table publications. Condition 2 requires that each transient sensitive value satisfies a simplified form of $\ell$-diversity. Note that if the adversary has no participation knowledge the value of $|\mathcal{AG}|$ can be defined as the number individuals in $\mathcal{AG}$, both present or absent. However, if in the worst case the adversary has participation knowledge of the individuals then only such individuals can help to provide for the uncertainty, and $|\mathcal{AG}|$ should be the number of individuals that are present in $\mathcal{AG}$.

The following lemma shows a necessary condition for $\ell$-scarcity, which is enforced by the role-based partition principle.

LEMMA 1. *Let $\mathcal{AG}$ be an anonymized group at time $j$. For any $s \in S$, let $count(s, \mathcal{AG})$ be the count of $s$ in $\mathcal{AG}$, if $\frac{count(s,\mathcal{AG})}{|\mathcal{AG}|} > 1/\ell$, then $\exists p_i$, $risk(p_i, s, n) > 1/\ell$.*

| PID | G.ID | Age | Zip. | Disease |
|---|---|---|---|---|
| $p_5$ | 1 | [20, 25] | [25k, 33k] | Fever |
| $p_6$ | 1 | [20, 25] | [25k, 33k] | Flu |
| $p_7$ | 1 | [20, 25] | [25k, 33k] | Ulcer |

**Table 6:** $T_5^*$

**Proof**: At time $n$, in each possible table series $\{T_1^p, T_2^p, ...T_n^p\}$, some individual in $\mathcal{AG}$ will be assigned to each of the $s$ occurrences in $AG$, that is, $\forall s \in S$, $s \in T_j^p(p_i)$, for some $p_i \in \mathcal{AG}$. If there are $k$ possible series, then the total number of such assignment is $count(ps, \mathcal{AG}) \times k$. Hence at least one individual is assigned more than $count(ps, \mathcal{AG}) \times k/|\mathcal{AG}|$ times. Therefore one or more of the individuals will be assigned to $s$ in more than $\frac{1}{\ell}$ of all the series. Note that $s$ can be either permanent or non-permanent.■

For transient sensitive values, their association with individuals in one data release has no impact on their associations with individuals in another release. Since role-based partition ensures $\ell$-diversity for transient sensitive values, we can derive the following lemma.

LEMMA 2. *If the anonymization mechanism follows role-based partition principle, then at time $n$, $\forall$ $s' \in TS$, $\forall p_i \in \mathcal{I}$, $risk(p_i$, $s'$, $n) \leq \frac{1}{\ell}$.*

In other words, if there are no permanent sensitive values, $\ell$-diversity at each release is sufficient for the required protection. From Lemma 2, in order to satisfy the privacy requirement, on top of role-based partition, we only need to make sure that there is no privacy breach on the permanent sensitive values, meaning that $\forall p_i \in \mathcal{I}$ and $\forall s \in PS$, $risk(p_i, s, n) \leq 1/\ell$.

Let us return to the example in Table 2. After $T_1^*$ is published, before $T_2^*$ is generated, roles should be updated. In order to achieve 3-scarcity, two HIV-decoys should be selected from Group 1 and Group 2 respectively. Therefore, HIV-decoy candidates $p_1$, $p_3$, $p_5$, and $p_6$ are all selected to be HIV-decoys. From then on, $p_5$ and $p_6$ must not disclose the fact that they are not HIV infected. Role-based partition will not allow publications like $T_5^*$ in Table 6, since in Group 1 of $T_5^*$, there are 2 HIV-decoys but no HIV-holder. If so happens that no other individuals satisfy the predicate of Group 1, then $p_5$, $p_6$, $p_7$ will be identified to be the members of this group. Hence, $p_5$ and $p_6$ will be found to be non-HIV individuals. By excluding the linkage of $p_5$ and $p_6$ to HIV, and by cross referencing table $T_1^*$, $p_4$ will be discovered to be an HIV-carrier with 100% certainty. Thus role-based partition prevents such a privacy breach.

## 3.2 Cohort-based Partition

Unfortunately, role-based partition cannot prevent the linkage exclusion problem as shown in our first motivation example in Section 1.1.

Therefore, besides role-based partition, we propose another partition principle: *cohort-based partition.* Our goal is that an adversary could never reach a contradiction by assuming that an $s$-decoy is an $s$-holder. Without such contradiction, no linkage exclusion will be possible. The solution is to distribute the $s$-decoys into $\ell - 1$ *cohorts*, and ensure that $s$-decoys from the same group could never share the linkage to one $s$ appearance, which is a basic property if they were indeed $s$-holders. Naturally, all the $s$-holders can form one cohort too because they can never share linkage to only one $s$. Including a cohort for $s$-holders, we have in total $\ell$ cohorts. The structure of *cohorts* is utilized for this purpose.

To better present the ideas, we first assume no participation knowledge on the adversary's side. Also we assume no change in the registration list over time. These restrictions will be relaxed in Section 3.4.

### 3.2.1 Container and Cohort

From the above discussion individuals would be assigned to cohorts. However, an individual that is initially assigned the role of a decoy may become a $s$-holder and in which case it must pass its role onto some other individual who then acts as if s/he has been the decoy all the time. To find such eligible replacement, we do not enter individuals directly to the cohorts. Instead, we enter a structure related to a region, where the region can contain multiple individuals so that replacements can be found within the region. Such a region is called a container.

Definition 6 (*s*-container). *An $s$-container $C$ at time $j$ is defined by a predicate $\mathcal{Q}_\mathcal{C}$ over the attributes of the QID. We say that $C$ contains an individual $p_i$ in a registration list $\mathcal{RL}_j$ if the QID value of $p_i$ in $\mathcal{RL}_j$ satisfies the predicate $\mathcal{Q}_\mathcal{C}$.*

Definition 7 (*s*-container instance/owner/*s*-buddy). *A set $\mathcal{CI} = \{C, p_o\}$ is called an **$s$-container instance**, where $C$ is an $s$-container containing a special individual $p_o$ which is either an $s$-holder or an $s$-decoy, who is the **owner** of $\mathcal{CI}$. We also say that $\mathcal{CI}$ is owned by $p_o$. $C$ may also contain $s$-clean individuals. An $s$-clean individual in an $s$-container instance $\mathcal{CI}$ that has existed since the creation of $\mathcal{CI}$ is called an **$s$-buddy**.*

An $s$-buddy is a potential eligible replacement for taking up the role of an $s$-decoy. Example 1 below helps to illustrate the above definitions. Note that at any time, there is a one-to-one mapping between the set of $s$-holders plus $s$-decoys to the set of $s$-container instances.

At time $j$, a set $\mathcal{C}$ of individuals is called an **$s$-clique** if $\mathcal{C}$ only contains one $s$-holder and $\ell - 1$ $s$-decoys.

Definition 8 (*s*-cohort/*s*-cohort formation). *A set consisting of a number of $s$-container instances is called an **$s$-cohort**. An **$s$-cohort formation** is a set of $\ell$ disjoint $s$-cohorts, in which one of the $s$-cohorts consists of container instances owned by $s$-holders, and the remaining $s$-cohorts consist of container instances owned by $s$-decoys. Moreover, in each $s$-cohort, there are exactly $N_s$(the number of $s$-holders) container instances.*

Let us see how the cohorts are initialized. At time 2, from every partition $\mathcal{P} \in P(T_1)$, for each $s \in PS$ that is contained in $T_1^*.S(\mathcal{AG}(\mathcal{P}))$, we can form exactly $count(s, \mathcal{AG}(\mathcal{P}))$ $s$-cliques. $\forall s \in PS$, $\ell$ empty $s$-cohorts are initialized, and $CH(s)$ is a set which stores those $\ell$ $s$-cohorts. For each $s$-clique $\mathcal{C}$, we put the $s$-holder's $s$-container instance into the first $s$-cohort $CH(s)[1]$, and place the $\ell - 1$ *$s$-decoys*' $s$-containers in the same $s$-clique into the $\ell - 1$ $s$-cohorts $CH(s)[2], ..., CH(s)[\ell]$, (one for each) respectively. For all $s \in PS$, we construct $CH(s)$ by repeating the above procedure. The predicates in the corresponding containers in the construction must follow the principle of *cohort-based partition*.

Principle 2 (Cohort-based Partition). *At time $j$, each anonymized group $\mathcal{AG}$ in $T_j^*$ satisfies the following conditions:*

1. *In $\mathcal{AG}$, if there are $m$ $s$-holders, then $\forall i$ where $2 \leq i \leq \ell$, there are exactly $m$ $s$-decoys whose $s$-container instances are in $CH(s)[i]$,*

2. *Let $Q$ be the predicate on the QID values of $\mathcal{AG}$ and let $\mathcal{Q}$ be the predicate of the container for the container instance owned by any $s$-decoy or $s$-holder $p_i$ in $\mathcal{AG}$, then $\mathcal{Q}$ implies $Q$.*

| PID | Age | Zip. | Role |
|---|---|---|---|
| $p_1$ | 23 | 16355 | HIV-decoy, Alzheimer-clean |
| $p_2$ | 22 | 15500 | HIV-holder |
| $p_3$ | 21 | 12900 | HIV-decoy, Alzheimer-clean |
| $p_4$ | 26 | 18310 | HIV-holder |
| $p_5$ | 25 | 25000 | HIV-decoy, Alzheimer-buddy |
| $p_6$ | 20 | 29000 | HIV-decoy, Alzheimer-clean |
| $p_7$ | 23 | 16910 | HIV-buddy, Alzheimer-clean |
| $p_8$ | 24 | 15505 | HIV-buddy, Alzheimer-clean |
| $p_9$ | 22 | 13055 | HIV-buddy, Alzheimer-clean |
| $p_{10}$ | 25 | 18870 | HIV-buddy, Alzheimer-clean |
| $p_{11}$ | 26 | 25500 | HIV-buddy, Alzheimer-buddy |
| $p_{12}$ | 20 | 28500 | HIV-buddy, Alzheimer-buddy |
| $p_{13}$ | 23 | 26950 | HIV-clean, Alzheimer-decoy |
| $p_{14}$ | 24 | 25855 | Alzheimer-holder |
| $p_{15}$ | 21 | 29355 | HIV-clean, Alzheimer-decoy |
| ... | ... | ... | ... |
| $p_{30}$ | 24 | 28000 | HIV-clean, Alzheimer-clean |
| ... | ... | ... | ... |
| $p_{\lvert RL \rvert}$ | 31 | 31000 | HIV-clean, Alzheimer-clean |

(a) $\mathcal{RL}_6$

| PID | Disease |
|---|---|
| $p_1$ | Flu |
| $p_2$ | HIV |
| $p_3$ | Fever |
| $p_4$ | HIV |
| $p_5$ | Flu |
| $p_6$ | Fever |
| $p_{13}$ | Ulcer |
| $p_{14}$ | Alzheimer |
| $p_{15}$ | Diarrhea |
| $p_{30}$ | Flu |

(b)$T_6$

**Table 7: Registration list $\mathcal{RL}_6$ and microdata snapshot $T_6$**

| Container(s) | PIDs | Instances | Age | Zip. |
|---|---|---|---|---|
| $C_1$(HIV) | $p_1, p_7, p_2, p_8$ | $CI_h(p_1), CI_h(p_2)$ | [22, 24] | [15k, 17k] |
| $C_2$(HIV) | $p_3, p_9$ | $CI_h(p_3)$ | [21, 22] | [12k, 14k] |
| $C_3$(HIV) | $p_4, p_{10}$ | $CI_h(p_4)$ | [25, 26] | [18k, 19k] |
| $C_4$(HIV) | $p_5, p_{11}$ | $CI_h(p_5)$ | [25, 26] | [25k, 26k] |
| $C_5$(HIV) | $p_6, p_{12}$ | $CI_h(p_6)$ | 20 | [28k, 29k] |
| $C_6$(Alz.) | $p_5, p_6, p_{13}, p_{14}$ | $CI_a(p_{13}), CI_a(p_{14})$ | [23, 26] | [25k, 27k] |
| $C_7$(Alz.) | $p_{12}, p_{15}$ | $CI_a(p_{15})$ | [20, 21] | [28k, 30k] |

**Table 8: Containers and container instances**

At any time $j > 1$, after $T_{j-1}^*$ is published, before generalizing $T_j$ to $T_j^*$, the roles are updated. Then, fresh $s$-decoys' container instances are distributed into cohorts $CH(s)[o](2 \leq o \leq \ell)$ while fresh $s$-holders' container instances are entered into cohort $CH(s)[1]$.

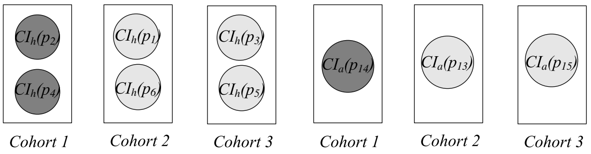
Example 1. *Suppose at time 6, the registration table $\mathcal{RL}_6$ and microdata $T_6$ are given in Table 7. There are two permanent sensitive values: HIV and Alzheimer. By HD-composition, each individual is assigned some roles, as shown in Table 7(a). By the container construction procedure ( in Section 3.2.2), the containers at time 6 are maintained as in Table 8, five of which are HIV-containers ($C_1$ to $C_5$), and two of which are Alzheimer-containers ($C_6$ to $C_7$). Note that these containers are defined by the predicates on Age and Zip Code as shown in the last two columns of the table. Then, we have HIV-cohorts and Alzheimer-cohorts as in Figure 1(a) and (b) respectively. Finally, the publication $T_6^*$ in Table 9 is obtained, which satisfies both the role-based partition principle and cohort-based partition principle.*

### 3.2.2 Container Maintenance

An $s$-decoy may become an $s$-holder and there will be replacement for such a decoy. The concept of containers enables us to easily identify a possible replacement. In the next part, we discuss details of how to maintain container instances after such updates.

To simplify our discussion of container maintenance, we make the following assumption.This assumption will be relaxed in Section 3.2.5.

Assumption 1 (Buddy Abundance). *At time $j$, each $s$-container instance $CI$ from time $j-1$ contains at least one unique $s$-buddy $B_p(CI)$ which is present and one $s$-buddy $B_a(CI)$ which is absent. By uniqueness, we mean that for two instances $CI_1$ and $CI_2$, $B_p(CI_1) \neq B_p(CI_2)$ and $B_a(CI_1) \neq B_a(CI_2)$.*



(a) HIV-cohorts (b) Alzheimer-cohorts

**Figure 1: Cohorts for HIV and Alzheimer**

| PID | G.ID | Age | Zip. | Disease |
|---|---|---|---|---|
| $p_1$ | 1 | [21, 24] | [12k, 17k] | Flu |
| $p_2$ | 1 | [21, 24] | [12k, 17k] | HIV |
| $p_3$ | 1 | [21, 24] | [12k, 17k] | Fever |
| $p_4$ | 2 | [20, 26] | [18k, 29k] | HIV |
| $p_5$ | 2 | [20, 26] | [18k, 29k] | Flu |
| $p_6$ | 2 | [20, 26] | [18k, 29k] | Fever |
| $p_{13}$ | 3 | [20, 26] | [25k, 30k] | Ulcer |
| $p_{14}$ | 3 | [20, 26] | [25k, 30k] | Alzheimer |
| $p_{15}$ | 3 | [20, 26] | [25k, 30k] | Diarrhea |
| $p_{30}$ | 3 | [20, 26] | [25k, 30k] | Flu |

**Table 9: Published Table $T_6^*$**

**Introducing new $s$-holder**: At any time $j$ if there is a new $s$-holder $p_i$, then a new container instance for $p_i$ is added to the cohort $CH(s)[1]$, also one new decoy container instance need to be made available from each of the remaining $\ell - 1$ cohorts.

In the following, we show how to initialize or maintain the $s$-container instances according to how $p_i$'s role changes:

[CASE M1]: $p_i$ is not an $s$-decoy at time $j - 1$ and becomes a new $s$-holder at time $j$:
$\mathcal{C}$ = *allocateContainer($p_i$)*; $\mathcal{CI}(p_i) \leftarrow \{\mathcal{C}, p_i\}$; enter $\mathcal{CI}(p_i)$ into $CH(s)[1]$. (Note that the decoys for $p_i$ will be introduced by CASE M2 or CASE M4.)

[CASE M2]: $p_i$ becomes an $s$-decoy that owns a new container instance at time $j$ in cohort $\mathcal{H}$:
$\mathcal{C}$ = *allocateContainer($p_i$)*; $\mathcal{CI}(p_i) \leftarrow \{\mathcal{C}, p_i\}$; enter $\mathcal{CI}(p_i)$ into cohort $\mathcal{H}$.

[CASE M3]: an $s$-decoy $p_i$ that owns a container instance $CI$ at time $j-1$ becomes an $s$-holder at time $j$: $\mathcal{C}$ = *allocateContainer($p_i$)*; $\mathcal{CI}(p_i) \leftarrow \{\mathcal{C}, p_i\}$; enter $\mathcal{CI}(p_i)$ into $CH(s)[1]$. (Note that the decoys for $p_i$ will be introduced by CASE M2 or CASE M4.)
One $s$-buddy $p_{i'}$ is chosen from $CI$ to become an $s$-decoy and the new owner of $CI$ (CASE M4).

[CASE M4]: an $s$-buddy $p_i$ for a container instance $CI$ becomes an $s$-decoy which owns $CI$.

In the above, the routine *allocateContainer($p_i$)* is to construct a new container $C$ which contains $p_i$.

It is important to note that with no participation knowledge, the adversary cannot determine which individual is present or absent. Hence with an $s$-decoy $p_i$ in a container instance $CI$, there is no contradiction from the published data that $p_i$ has been present or absent together with the $s$-holder since the creation of $CI$. It is because when $p_i$ is absent(present), there must be some other individual $p_j$ that is present(absent) and the adversary cannot distinguish between $p_i$ and $p_j$.

**Absence of individuals.** An individual that is present at one data release may become absent in another. It becomes an issue when such an individual has been an $s$-holder or an $s$-decoy. Suppose $p_i$ is present in $T_{j-1}$ but is absent in $T_j$.

[CASE A1]: If $p_i$ is an $s$-holder at time $j - 1$, then when it is absent we must also make $\ell - 1$ $s$-decoys absent. In order to make a decoy absent, an absent $s$-buddy in an $s$-decoy $p_d$'s container instance takes over the role of this $s$-decoy. $p_d$ becomes $s$-clean. This corresponds to CASE M4. We say that the container instance becomes **dormant**. In this case, there is no explicit changes of containers and cohorts.

[CASE A2]: If $p_i$ is an $s$-decoy at time $j - 1$, then it is replaced by an $s$-buddy $p_b$ who is present in the same container. This corresponds to CASE M4. $p_i$ becomes an $s$-clean individual again.

When an $s$-holder that has been absent become present, absent $s$-decoys can be made present by switching the $s$-decoy role to a present $s$-buddy. If an $s$-decoy that has been absent becomes present, and the container instance should remain dormant, then the role of $s$-decoy is passed to another $s$-buddy who is currently absent.

**Clique Extinction: exit of container instances.** If an individual $p_i$ becomes extinct, it is a special case of absence for $p_i$ since $p_i$ will never be present in the future. Whenever there is any $s$-holder $p_i$ that is extinct and there exist $\ell - 1$ $s$-decoy container instances with the $s$-decoy or any $s$-buddy that is extinct, we say that a *clique extinction* has occurred, the $\ell$ corresponding container instances will be removed from their respective cohorts together. The secret of the $s$-holder will be kept forever.

### 3.2.3 Anonymization step

HD-composition integrates role-based partition and cohort-based partition: a partition should be a role-based partition and a cohort-based partition simultaneously. With such a partition, anonymization (generalization) is applied, following Definition 2.

### 3.2.4 Discussion

Note that an individual $p_i$ may be linked to *multiple sensitive values* in $T_j$. Suppose $\{s_1, s_2\} \in T_j.S(p_i)$. $p_i$ will be both an $s_1$-holder and an $s_2$-holder. In such a case, $p_i$ is in both the $s_1$-cohort and the $s_2$-cohort. Similarly an individual can be a decoy for multiple sensitive values.

One basic assumption for the data releases is that the number of permanent sensitive value occurrences is not large compared to the dataset size. In the worst case, $\ell - 1$ decoys are required to cover each such occurrence, which is possible only if the occurrences do not exceed $1/\ell$ of the data size. This is justified in the medical data scenario, where sensitive and persistent diseases are relatively rare, with evidence from a real dataset used in our empirical studies.

### 3.2.5 Relaxing the Buddy Assumption

At time $j$ if there exists an $s$-decoy $p_i$ in the cohort formation at $j - 1$ that has turned into an $s$-holder (CASE M3), if Assumption 1 does not hold for the container instance of $p_i$, there will be no $s$-buddy found, it is not possible to find a replacement for $p_i$, in this case $p_i$ is **pinned** as an $s$-decoy. The $s$ value of $p_i$ is suppressed and replaced by a transient or non-sensitive value which appears in the microdata $T_j$. When an $s$-decoy $p_d$ is pinned, it will act as an $s$-decoy until it is extinct. When its QID is changed, if necessary, the old container instance owned by $p_d$ is replaced by a new container instance $C_{new}$ which is owned by $p_d$, and in $C_{new}$, $p_d$ remains pinned.

Pinning an individual as an $s$-decoy may involve a suppression of an occurrence of the $s$ value in the case where the $s$-decoy is actually linked to $s$. However, even if an adversary knows that some $s$ occurrence is suppressed, it is difficult to pinpoint an individual that has been suppressed. Hence no extra privacy breach can be gained from such knowledge.

| PID | Dis. |
|---|---|
| $p_1$ | * |
| $p_2$ | HIV |
| $p_3$ | * |
| $p_4$ | HIV |
| $p_5$ | * |
| $p_6$ | * |

(a) $P_1$

| PID | Dis. |
|---|---|
| $p_1$ | * |
| $p_2$ | * |
| $p_3$ | HIV |
| $p_4$ | HIV |
| $p_5$ | * |
| $p_6$ | * |

(b) $P_2$

| PID | Dis. |
|---|---|
| $p_1$ | * |
| $p_2$ | HIV |
| $p_3$ | * |
| $p_4$ | * |
| $p_5$ | HIV |
| $p_6$ | * |

(c) $P'_1$

| PID | Dis. |
|---|---|
| $p_1$ | HIV |
| $p_2$ | * |
| $p_3$ | * |
| $p_4$ | * |
| $p_5$ | * |
| $p_6$ | HIV |

(c) $P'_2$

| PID | Dis. |
|---|---|
| $p_1$ | * |
| $p_2$ | * |
| $p_3$ | HIV |
| $p_4$ | * |
| $p_5$ | HIV |
| $p_6$ | * |

(c) $P'_3$

**Table 10: Possible Assignments**

## 3.3 Privacy Guarantee

Next let us examine the privacy guarantee of HD-composition. First, we need the definition of *possible assignments*.

At time $n$, $T'_n$ is possible table for $T_n$ if $T'_n$ is the $n$-th table in a possible table series. For a possible table $T'_n$, the set of individuals that are linked to a sensitive value $s$ is called a *possible s-assignment* at time $n$.

### *3.3.1 Preliminary Analysis*

Let us return to the example in Table 2. HIV is a permanent sensitive value but Flu and Fever are not. The objective is to publish a generalization series that satisfies persistent $\ell$-scarcity, where $\ell = 3$. We first build three HIV-cohorts. In this example, each container instance contains only one individual which is also the owner of the container instance. Let the three cohorts be $(p_1, p_6)$, $(p_2, p_4)$, and $(p_3, p_5)$, respectively. By HD-composition, the cohorts remain unchanged for times 1, 2. Let $T^*_1, T^*_2$ be the anonymizations based on these cohorts. Given these published tables, the adversary may analyze as follows. From the registration lists, the set of individuals belonging to each group in the two tables can be determined. From the group with G.ID=1 in $T^*_1$, if $p_1$ is linked to HIV, then neither $p_2$ nor $p_3$ could be associated with HIV, and therefore from the group with G.ID=1 in $T^*_2$, $p_6$ must link to HIV. Therefore, one possible assignment is that $p_1$ and $p_6$ are HIV carriers. Following this inference procedure, there are 5 *possible assignments* at time 2: $P_1$, $P_2$, $P'_1$, $P'_2$ and $P'_3$ in Table 10.

Note that there are fewer *possible assignments* linking either $p_1$ or $p_6$ to HIV (only $P'_3$) than that linking each of $p_2, p_3, p_4, p_5$ to HIV. For example for $p_2$, there are two such assignments $P_1$ and $P'_1$. Under each possible $s$ assignment, to construct the *possible tables* of $T_1$ or $T_2$ is straightforward: occurrences of the other 2 sensitive values could be arbitrary assigned to the remaining individuals since they are transient. Therefore, totally there are 20 possible tables for $T_1$, in which 4 tables link $p_1$ to HIV($P'_2$), 8 link $p_2$ to HIV($P_1$ and $P'_1$), and 8 link $p_3$ to HIV($P'_2$ and $P'_3$). Combining the two tables, there are $5(4^2)$ possible table series in total, we have $4^2$ possible table series for $p_1$ or $p_6$ being linked to HIV, $2(4^2)$ possible table series for each of $p_2, p_3, p_4, p_5$ being linked to HIV. Thus, the linkage probability of $p_2, p_3, p_4$ or $p_5$ to HIV is 2/5, which exceeds the required threshold of $\frac{1}{3}$($\ell$=3).

However, this result is counter-intuitive, given that generalization is based on three cohorts and each cohort has equal chance of being the HIV-holder cohort. It is not logical that some individual, such as $p_1$ or $p_6$, gets a smaller chance of being linked to HIV.

### *3.3.2 Trusted Publisher based Analysis*

What is missing in the above analysis is that it does not take into account the anonymization mechanism. In our approach, the adversary or the public will be provided with the additional knowledge of the HD-composition mechanism. Then, the above random possible world assumption is not accurate nor complete. The possible world model should be enriched with possibilities or random variables made relevant by the mechanism. In particular, the $s$-cohort formations for all $s$ becomes one of the random variables. The same sensitive value assignment to the individuals under different cohort formations will be considered as two different possible worlds. The random world assumption is then applied on top of this model.

With the mechanism knowledge, but not knowing exactly the cohorts, first an adversary would need to consider the possible formations of the cohorts. From $T^*_1$ and $T^*_2$, there are only two possible cohort formations, namely, $F_1 = \{(p_1, p_6), (p_2, p_4), (p_3, p_5)\}$ and $F_2 = \{p_1, p_6), (p_3, p_4), (p_2, p_5)\}$. For formation $F_1$, there are 3 possible assignments, namely, $\{p_1, p_6\}$, $\{p_2, p_4\}$, and $\{p_3, p_5\}$. Each individual is linked to HIV in one out of the 3 assignments. For $F_2$, the possible assignments are $\{p_1, p_6\}$, $\{p_3, p_4\}$, and $\{p_2, p_5\}$. Again, each individual is linked to HIV in one out of the 3 assignments. Under each of these six scenarios, there are the same number of possible table series by considering all possible assignments of the transient values to $T^p_j.S(p_i)$ for the remaining $p_i$'s. Therefore, by the random world assumption, each of the two formation has equal probability, and the number of possible table series where each individual is linked to HIV is one-third of all the possible table series.

The above example shows that we should consider possible worlds as defined by a possible cohort formation and a possible assignment of sensitive values under the formation. We first define the term possible $s$-cohort formation series which is a series of consistent $s$-cohort formation from which $T^*_1, ...T^*_n$ can be generated.

DEFINITION 9 ( POSSIBLE S-COHORT FORMATION SERIES). *At time $n$, a series of s-cohort formation $\{L_1, L_2, L_3, ..., L_n\}$, where $L_j$ is an s-cohort at time $j$, is a* ***possible $s$-cohort formation series*** *if and only if the followings hold:*
*Let $L_i = \{C^i_1, C^i_2, ...C^i_\ell\}$ and $L_j = \{C^j_1, C^j_2, ...C^j_\ell\}$,*

1. *If 2 different s-container instances $CI_1$ and $CI_2$ occur in $L_i$ and $L_j$, then $CI_1$ and $CI_2$ are in the same s-cohort in $L_i$ iff they are in the same s-cohort in $L_j$.*
2. *If an s-container instance occurs in both $C^i_1$ and $C^j_1$ (both are cohorts for s-holders), then if $p_o$ is the owner of the instance in $C^i_k$, it is also the owner of the instance in $C^j_k$.*

The first condition in the above governs that a container instance will never switch from one cohort to another cohort. The second condition says that for $s$-holder cohorts, the owner of a container instance is never changed. With HD-composition, the container instance may become dormant but is never removed from a cohort until it joins a clique extinction, once a $s$-holder is assigned to be the owner of a container instance, it will not give up its ownership to another individual until the container instance goes extinct. Therefore we have the following lemma.

LEMMA 3. *HD-composition generates only possible s-cohort formation series for each s.*

We call the combination of possible cohort formation series for all permanent sensitive values a global possible cohort formation series. A random possible world is defined by a global possible cohort formation series $G$ and a possible table series generated by $G$. With the HD-composition mechanism, the definition of $risk(p_i, s, n)$ is defined to be the fraction of random possible worlds where $p_i$ is assigned to $s$.

THEOREM 1. *With HD-composition, the risk of any individual $p_i$ being linked to $s$ at time $n$ is given by*

$$risk(p_i, s, n) \leq 1/\ell$$

**Proof Sketch**: Let $\mathcal{CF}_{possible}$ be the set of all possible $s$-cohort formation series at time $n$. $\mathcal{CF}_{possible}$ can be partitioned into $\ell - 1$ subsets: $\mathcal{CF}^1, ..., \mathcal{CF}^{\ell-1}$, where $\mathcal{CF}^k$ corresponds to the set of $s$-cohort formation series with exactly $k$ cohorts which have never experienced any change of ownership for any container instance that appears at two different times. That is, there are $\ell - k$ cohorts which have experienced some changes in such ownerships at or before time $n$.

The risk of individual $p_i$ being linked to $s$ is related to the number of random worlds with assignment of the value $s$ for $p_i$. For $\mathcal{CF}^k$, it is easy to see that there are $k$ cohorts that could be the $s$-holder cohort, and there are $\ell - k$ cohorts that cannot be the $s$-holder cohort. In $\mathcal{CF}^k$, let $\mathcal{CF}^k_{p_i}$ be the subset of cohort formation series where $p_i$ is the owner of some container instance $\mathcal{CI}(p_i)$ in the published data.

Since each container instance conforms to the Cohort-based partition principle, and all possible instances are constructed from the anonymized groups in the published tables, each anonymized group has the same effect on each cohort. Hence each container instance has an equal chance of being included in each cohort.

Within $\mathcal{CF}^k_{p_i}$, there is a probability of $k/\ell$ where $CI_i$ is within one of the possible $s$-holder cohorts, and for each such cohort, there is a chance of $1/k$ where the cohort is indeed the $s$-holder cohort. Hence the probability that $p_i$ is linked to $s$ is given by $k/\ell \times 1/k = 1/\ell$. This probability is conditioned on $\mathcal{CF}^k_{p_i}$, in other words,

$$Prob(p_i, s, n|\mathcal{CF}^k_{p_i}) = 1/\ell \qquad (2)$$

Let $Prob(\mathcal{CF}^k) = \frac{|\mathcal{CF}^k|}{|\mathcal{CF}_{possible}|}$. The number of possible worlds generated by each of the possible $s$-cohort formation series is the same because there are the same number of remaining linkages of values to individuals in each anonymization group in the published tables, and the individuals are not distinguishable from the adversary viewpoint. Therefore from Equation 2 the probability of $p_i$ being linked to $s$ at time $n$ is

$$Prob(p_i, s, n|\mathcal{CF}^k) = Prob(p_i, s, n|\mathcal{CF}^k_{p_i}) \times \frac{|\mathcal{CF}^k_{p_i}|}{|\mathcal{CF}^k|} \leq 1/\ell$$

Since $\mathcal{CF}^1, ..., \mathcal{CF}^{\ell-1}$ constitute all the possibilities, we have

$$\begin{aligned} Prob(p_i, s, n) &= \sum_{k=1}^{\ell-1} Prob(p_i, s, n|\mathcal{CF}^k) \times Prob(\mathcal{CF}^k) \\ &\leq 1/\ell \times \sum_{k=1}^{\ell-1} Prob(\mathcal{CF}^k) = 1/\ell \end{aligned}$$

■

## 3.4 Refinement of HD-composition

In the previous discussion, for clarity, there is no consideration of QID value changes and participation knowledge of the adversary. In this subsection we refine HD-composition to address these issues.

### *3.4.1 QID value updates*

From the multiple releases of $\mathcal{RL}_l$, one can determine the changes of individuals' QID values from time to time, the changes lead to changes in the $s$-cohorts. Suppose $p_i$ has changed the QID value at release $T_j$ so that it is different from that in $T_{j-1}$. If $p_i$ is an $s$-holder and, after the QID change, it is not contained in the container instance $CN$ at time $j - 1$, then a new container instance owned by $p_i$ is created, which replaces the old container. If $p_i$ is an $s$-decoy and, after the QID value change, $p_i$ is not in the original container $CN$, then $CN$ should still contain an $s$-decoy. Hence, the role of $p_i$ as an $s$-decoy is replaced by an $s$-buddy in the container instance. The role of $p_i$ is changed and it becomes an $s$-clean individual.

### *3.4.2 Participation Knowledge*

Here we consider that the adversary has the knowledge of the presence and/or absence of up to $\mathcal{K}$ individuals. We replace the buddy abundance assumption in Section 3.2.2 with the following assumption ($\mathcal{K}$ buddy abundance assumption): given $T_j$, and the cohort formation at time $j - 1$, each container in any cohort should contain at least $\mathcal{K}+1$ $s$-buddies which are present and at least $\mathcal{K}+1$ $s$-buddies which are absent.

With at least $\mathcal{K} + 1$ $s$-buddies present, an adversary will not have knowledge of at least one $s$-buddy $p_x$ that is present, with the QID anonymization, from the adversary's viewpoint, $p_x$ can be any individual present or absent in $T_j$. Therefore, $p_x$ can become an $s$-decoy if needed, and the $s$-decoy role of $p_x$ can also be swapped in the future with the $s$-buddy role of other $s$-buddies.

With at least $\mathcal{K} + 1$ $s$-buddies absent, an adversary will not have knowledge of at least one absent $s$-buddy $p_y$. Hence, $p_y$ can be the $s$-decoy who is absent in a dormant container instance, and the role of decoy of $p_y$ can be swapped with another $s$-buddy in the future.

In case the condition in the $\mathcal{K}$ buddy abundance assumption above is not satisfied, it is possible for the adversary to know everyone that is present, then an adversary may limit the scope of possible $s$-decoy to the current set $X$ of present individuals. Therefore, the available pool of $s$-buddies must be within $X$, and can be further restricted if only a subset of $X$ is present at any time. The final available $s$-buddy will be pinned as described in Section 3.2.5.

Let us call the scheme of HD-composition with the refinements in this section the refined HD-composition, the proof of the following lemma is similar to that for HD-composition.

THEOREM 2. *Anonymization by refined HD-composition satisfies $\ell$-scarcity.*

## 3.5 Anonymization Algorithm

Existing anonymization algorithms [3, 5] can be adopted for the task of HD-composition. The existing anonymization algorithms generate a number of anonymized groups, each of which satisfies some privacy requirement such as $\ell$-diversity. In this paper, we can borrow one of these algorithms by making each group in the output of these algorithms satisfy the properties of both role-based partition and cohort-based partition instead. This flexibility is a good feature since it opens the opportunity of leveraging the known anonymization algorithms to optimize HD-composition. In this paper, we adopt a top-down strategy [3] to perform the anonymization algorithm because it is found that a top-down approach often introduce less information loss in data publishing.

Specifically, each time we generate a published table $T^*_j$ from $T_j$, we perform the following steps. Firstly, for each $s \in PS$, update the $\ell$ $s$-cohorts. Secondly, for each $s \in PS$, $s$-cliques are created from all the $s$-holders and $s$-decoys. If there are $N$ container instances in every s-cohort, then we will create $N$ $s$-cliques. At the beginning, all the $s$-holders and $s$-decoys are set as available (not involved in any clique). We perform the following process $m$ times. We want to form an $s$-clique $\mathcal{C}$. Initially, we set $\mathcal{C} = \emptyset$. Randomly pick an $s$-cohort among the $\ell$ $s$-cohorts. In this chosen cohort, randomly select an *available* container instance in which the owner $p_j$ is picked and inserted into $\mathcal{C}$. For each of all other $s$-cohorts, we find an *available* container instance and a corresponding owner $p_{j'}$ which is the *closest*[3] to $p_j$, insert $p_{j'}$ into $\mathcal{C}$, and mark $p_{j'}$ as unavailable. The final $\mathcal{C}$ forms a final $s$-clique. Finally, with every formed clique as an atomic unit, an existing anonymization algorithm is applied to form a number of groups each of which satisfy the condition 2 of role-based partition. Note that when an anonymized group contains a clique, the QID predicates of the container instances(whose owner in the clique) will be implied by the QID predicate of the group.

# 4. EXPERIMENTS

All the experiments are performed on a Linux workstation with a 3.2Ghz CPU and 2 Giga-byte memory. We deploy one public available real hospital database CADRMP[4]. In the database, there are 8 tables: *Reports*, *Reactions*, *Drugs*, *ReportDrug*, *Ingredients*, *Outcome*, and *Druginvolve*. *Reports* consists of some patients' basic personal information. Therefore, we take it as the voter registration list. *Reactions* has a foreign key *PID* referring to the attribute *ID* in *Reports* and another attribute to indicate the person's disease. After removing tuples with missing values, *Reactions* has 105,420 tuples while *Reports* contains 40,478 different individuals. In *Reactions*, an individual can have multiple diseases, which is quite coherent with our problem setting. We take *Breast Cancer*(15 occurrences in *Reactions*), *Cavity Cancer*(1 occurrences in *Reactions*) and *Diabetes*(60 occurrences in *Reactions*) as permanent sensitive values and other diseases as transient sensitive values.

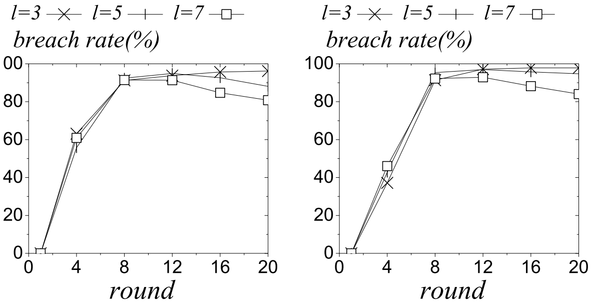
Dynamic microdata table series $TS_{exp}$=$\{T_1, T_2, ..., T_{20}\}$ is created from *Reactions*. $T_1$ is randomly selected from *Reactions* with a rate of 10%. In later rounds, $T_i$ are generated in this way: 10% of the tuples in $T_{i-1}$ randomly change their sensitive values, 10% of the tuples in $T_{i-1}$ are removed, and 10% of tuples in *Reactions* are inserted into $T_i$. Besides, dynamic voter registration list series $RLS_{exp}$=$\{\mathcal{RL}_1, \mathcal{RL}_2, ..., \mathcal{RL}_{20}\}$ is made from *Reports*. $\mathcal{RL}_1$ is firstly set as a duplicate of *Reports*. In later rounds, $\mathcal{RL}_i$ are generated in this way: from $\mathcal{RL}_{i-1}$, $vol$% people change their values on QIDs($vol$ is a parameter), 0.2% people become extinct, and 0.35% people are freshly generated(randomly) and inserted[5]. In all our experiments, there are 20 publishing rounds in total.

## 4.1 Failures of Conventional Generalizations

In this section, we confirms the significance of the privacy breach problems in traditional generalization principles. We use $\ell$-diversity [8] and $m$-invariance [21] as the representative principles, since $\ell$-diversity is widely adopted while $m$-invariance is the latest work offering protection on serial data publishing. Following $m$-invariance, when someone changes one of her/his sensitive value, we treat it as an articulate deletion and re-insertion.

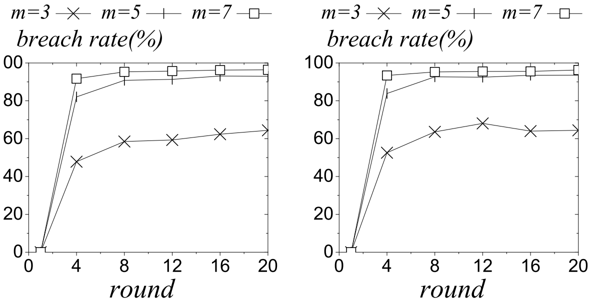
Given $TS_{exp}$, we adopt the $\ell$-diversity and $m$-invariance generalizations respectively. Then, we capture the individuals whose privacy is breached: those with a higher privacy disclosure risk than the threshold 1/$\ell$(or 1/$m$). Those individuals are called vulnerable individuals. It is noted that the vulnerable individuals must once appear in an anonymized group linkable to a permanent sensitive value. As in Table 2, at time 3, $p_1$, ..., $p_6$ are all vulnerable individuals: $p_1$, $p_5$ and $p_6$ are excluded from HIV, $p_2$ and $p_3$ are found to have 50% probability to contract HIV, $p_4$ is disclosed. At time $j$, given the number of vulnerable individuals: $N_{vul}(j)$, and the number of individuals once sharing the linkage probability to a permanent sensitive value: $N_{link}(j)$, the rate of privacy breach is calculated as follows:

$$\text{rate of privacy breach at time } j = \frac{N_{vul}(j)}{N_{link}(j)}$$



(a) $vol\%$ = 2% (b) $vol\%$ = 5%

**Figure 2: Rate of privacy breach vs. round: $\ell$-diversity**



(a) $vol\%$ = 2% (b) $vol\%$ = 5%

**Figure 3: Rate of privacy breach vs. round: $m$-invariance**

In Figure 2 and Figure 3, we plot the rate of privacy breaches at each publishing round in $\ell$-diversity and $m$-invariance generalizations, with $vol$%=2% and $vol$%=5% respectively. Each curve corresponds to the result obtained with a different $\ell$ or $m$. From the figures, it could be found that both $\ell$-diversity and $m$-invariance fail to support serial publication on fully dynamic databases, because they result in a high privacy breach rate. For example, when $\ell$=5(or $m$=5), after about 5 publishing rounds, there are more than 90% individuals once linkable to a permanent sensitive value suffering a privacy breach risk higher than 1/5.

## 4.2 Evaluations of HD-Composition

In this part, we test HD-composition in terms of effectiveness and efficiency. Given $TS_{exp}$ and $RLS_{exp}$, let $vol$%=2%, we apply the anonymization algorithm in Section 3.5 to sequentially compute the generalized table $T_j^*$ for microdata table $T_j$. In order to well evaluate HD-composition, we examine 4 aspects: the portion of perturbed tuples, the utility of published tables, the computation overheads, and the additional costs for dealing with participation knowledge, as follows. In the following container size $C$ refers to the number of individuals that are present in the container. Also the default values of $C, \ell, \mathcal{K}$ are 5, 4, 0, respectively[6].

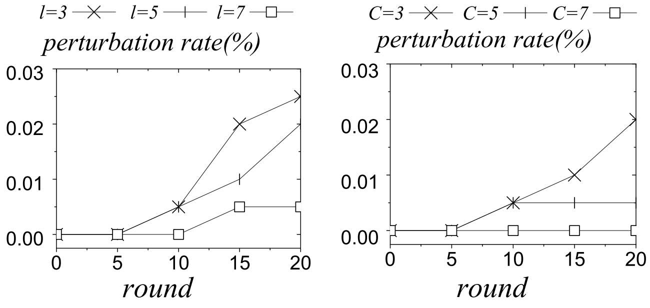
**Perturbation Rate.** As described in Section 3.2.5, perturbing sensitive values sometimes is necessary. Thus, we want to investigate

[3] The closeness is measured by Euclidean distance on the QIDs. For categorical attributes, one can adopt the distance definitions in [6].

[4] http://www.hc-sc.gc.ca/dhp-mps/medeff/databasdon/index_e.html

[5] The insertion and deletion rate of registration list is according to the death and birth rate per quarter of US. in year 2007, http://www.indexmundi.com/united_states/

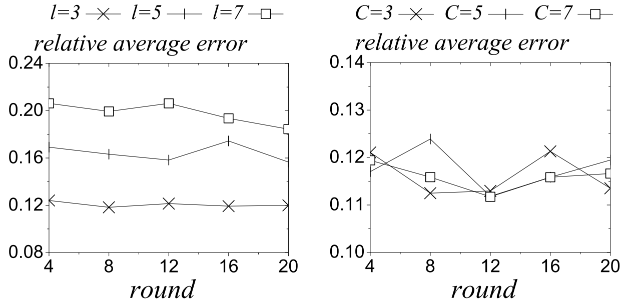
[6] In another suit of experiments, we set default $\mathcal{K}$ to be 3, and get very similar results.



(a) Effect of $\ell$

(b) Effect of container size $C$

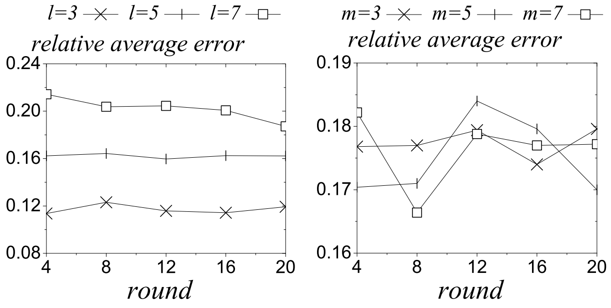
**Figure 4: Perturbation ratio vs. round: HD-composition**



(a) Effect of $\ell$

(b) Effect of container size $C$

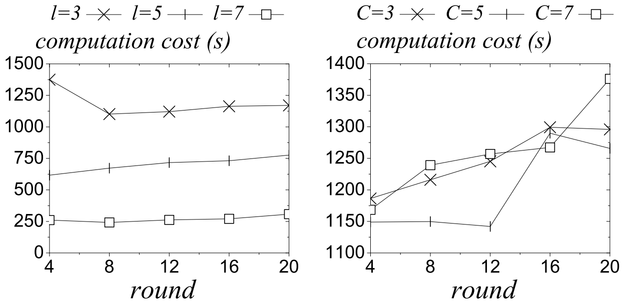
**Figure 5: Query error vs. round: HD-composition**



(a) $\ell$-diversity

(b) $m$-invariance

**Figure 6: Query error vs. round: $\ell$-diversity and $m$-invariance**



(a) Effect of $\ell$

(b) Effect of container size $C$

**Figure 7: Computation cost vs. round: HD-composition**

that, to achieve $\ell$-scarcity, normally how many individuals' sensitive values should be perturbed. Figure 4(a) shows the ratio of perturbed individuals among all individuals present in the microdata snapshot at each publishing round, where we and plot different curves to visualize the effects of different $\ell$. Figure 4(b) presents the results of a similar experiment with different container sizes $C$. From both Figure 4(a) and (b), we can find that at the first few publishing rounds, there is no perturbation at all, and the total number of perturbed tuples is tiny compared to the cardinality of microdata tables(about 0.02%).

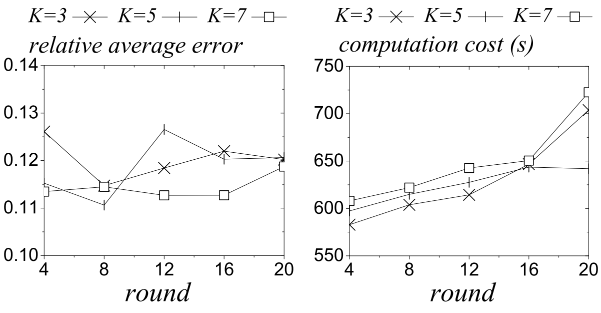
**Utility of Published Data**. In this set of experiments, we compare the query processing results on each anonymized table $T_j^*$ and its corresponding microdata table $T_j$ at each publishing round. We follow the literature conventions [19, 21, 17] to measure the error by the relative error ratio in answering an aggregate query. All the published tables are evaluated one by one. For each evaluation, we perform 5,000 randomly generated range queries which are similar to [21] on the microdata snapshot and its anonymized version, and then report the average relative error ratio. The results are plotted in Figure 5(a) and (b), in terms of different configurations of $\ell$ and container size $C$ respectively. The results show that the data utility does not downgrade much with the increase of publishing rounds, nor the container size. Besides, we also compute $\ell$-diversity and $m$-invariance generalizations for every microdata snapshot by different $\ell$ and $m$, and the results of their average relative error ratio are reported in Figure 6(a) and (b) respectively. It could be found that HD-composition has a similar performance compared with $\ell$-diversity and $m$-invariance in terms of utility.

**Computation Cost.** This experiment evaluates the efficiency of our anonymization algorithm. First, we vary $\ell$ and measure the time spent in generalization phase at each round. The effect of different $\ell$ is plotted in Figure 7(a). It could be found that the larger $\ell$ is, the faster the anonymization will be. The reason is that with larger $\ell$, the top-down specialization process can stop at an earlier stage and save a large amount of time on privacy checking. Second, we change the container size to investigate how the container size affect the time cost. The results in Figure 7(b) shows that the time costs are nearly the same for different container sizes. The reason is that the occurrence of permanent sensitive values is rare, and consequently the number of containers(for holders and decoys) are not large.

**Effect of Participation Knowledge.** The last experiment investigates how the utility and computation cost change if we need to defend against the participation knowledge. We vary the quantity of participation knowledge $\mathcal{K}$, then at each publishing round, record the relative average query error in Figure 8(a) and plot the computation cost in Figure 8(b). It could be found that different cardinalities of the participation knowledge do not have drastic influence on the results.

## 5. RELATED WORK

Researches on privacy protections for data publishing can be classified into two branches. The first branch is the one-time publication. The second branch is serial publication. For one-time publication, many privacy models are proposed to protect individual privacy. The traditional model is $k$-anonymity [14, 5, 4] which requires that each QID value either appears at least $k$ times or does not appear in the published table. However, recently, it is found that $k$-anonymity is vulnerable to homogeneity attack and background knowledge attack. One of the most influential work is $\ell$-diversity [8], which considers the inference relationship between QID values and sensitive attributes, instead of the number of occurrences of the QID values only. Since the aim of privacy protection is to guarantee the linkage probability between one and a sensitive



(a) Query error

(b) Computation cost

**Figure 8: Effect of participation knowledge vs. round**

value is no more than a threshold, $\ell$-diversity is more meaningful than $k$-anonymity. Li and Li [7] also consider the inference relationship and propose $t$-closeness which restricts that, according to the sensitive attribute, the distribution of each anonymized group is similar to the distribution of the entire published table.

Some works [8, 9, 17] for the one-time publication also consider some kinds of background knowledge. In [8], a Japanese hardly suffers from heart disease. Martin et al. [9] propose a new kind of background knowledge considering the association among individuals. For example, if an individual is linked to a sensitive value such as HIV, another individual which has a close relationship with him/her must also be linked to the same sensitive value. Wong et al. [17] consider that the background knowledge can be the principle of the anonymization algorithm. Different from previous works, we focus on the background knowledge including the participation information which denotes whether an individual is present or not in a snapshot of the microdata.

Byun et, al. [1] adopt delayed publishing to keep $\ell$-diversity dynamically, in which the anonymized data will not be published until the inserted tuples themselves could satisfy some privacy requirements. Yet the method has several drawbacks: it only considers insertions, all the publishing histories need to be maintained, and delayed durations are not bounded. Xiao and Tao [21] propose the $m$-invariance model to support both insertion and deletions. However, as we have discussed in the introduction, both of these works cannot address the problems that are identified in this paper.

Wang et al. [16] considers sequential releases where each release is on a different subset of the attributes for the same dataset. Both [11] and [2] present a privacy model to protect $k$-anonymity with serial publications that only allow insertions. Both consider $k$-anonymity and cannot be easily extended to the case where the inference relationship between QID and sensitive attributes is considered.

# 6. CONCLUSIONS

Current anonymization techniques for serial data publishing cannot support data updates on both QID attributes and the sensitive attribute and none of them consider the effect of permanent sensitive values. This paper addresses these issues by proposing a novel anonymization method by role composition. Utilizing the assumption of trusted publishers, we present an analysis showing that individual privacy can be guaranteed by our method. In the experiments on a real dataset, we find that our proposed algorithm HD-composition is efficient and also can preserve the utility of published data.

There are some promising directions for future work. It would be interesting to extend HD-composition to guard privacy against various kinds of background knowledge. One can also explore novel techniques to raise the utility for serial data publishing.

**ACKNOWLEDGEMENTS:** The research of Y. Bu, A. Fu, and R. Wong was supported in part by the RGC Earmarked Research Grant of HKSAR CUHK 4120/05E, 4118/06E, and RGC Direct Allocation 07-08. The research of Lei Chen is supported by the National Basic Research Program of China (973 Program) under grant No. 2006CB303000 and NSFC Project No. 60763001. The research of Li was supported by an ARC discovery grant DP0774450. All opinions, findings, conclusions and recommendations in this paper are those of the authors and do not necessarily reflect the views of the funding agencies.

# 7. REFERENCES

[1] J. Byun, Y. Sohn, E. Bertino, and N. Li. Secure anonymization for incremental datasets. In *Secure Data Management*, pages 48–63, 2006.

[2] B. C. M. Fung, K. Wang, A. Fu, and J. Pei. Anonymity for continuous data publishing. In *EDBT*, 2008.

[3] B. C. M. Fung, K. Wang, and P. S. Yu. Top-down specialization for information and privacy preservation. In *ICDE*, pages 205–216, 2005.

[4] K. LeFevre, D. DeWitt, , and R. Ramakrishnan. Multidimensional k-anonymity. In *M. Technical Report 1521, University of Wisconsin*, 2005.

[5] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan. Incognito: Efficient full-domain k-anonymity. In *SIGMOD Conference*, pages 49–60, 2005.

[6] J. Li, R. Wong, Ada, and J. Pei. Achieving $k$-anonymity by clustering in attribute hierarchical structures. In *DaWaK*, pages 405–416, 2006.

[7] N. Li and T. Li. $t$-closeness: Privacy beyond $k$-anonymity and $l$-diversity. In *ICDE*, 2007.

[8] A. Machanavajjhala, J. Gehrke, and D. Kifer. $\ell$-diversity: privacy beyond $k$-anonymity. In *ICDE*, 2006.

[9] D. J. Martin, D. Kifer, A. Machanavajjhala, and J. Gehrke. Worst-case background knowledge for privacy-preserving data publishing. In *ICDE*, 2007.

[10] M. Nergiz, M. Atzori, and C. W. Clifton. Hiding the presence of individuals from shared databases. In *SIGMOD*, 2007.

[11] J. Pei, J. Xu, Z. Wang, W. Wang, and K. Wang. Maintaining k-anonymity against incremental updates. In *SSDBM*, 2007.

[12] P. Samarati and L. Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression, unpublished manuscript. In *unpublished*, 1998.

[13] L. Sweeney. Weaving technology and policy together to maintain confidentiality. *Journal of Law, Medicine and Ethics*, 25(2-3):98–110, 1997.

[14] L. Sweeney. Achieving k-anonymity privacy protection using generalization and suppression. *International journal on uncertainty, Fuzziness and knowldege based systems*, 10(5):571 – 588, 2002.

[15] L. Sweeney. k-anonymity: a model for protecting privacy. *International journal on uncertainty, Fuzziness and knowldege based systems*, 10(5):557 – 570, 2002.

[16] K. Wang and B. C. M. Fung. Anonymizing sequential releases. In *KDD*, 2006.

[17] R. Wong, A. Fu, K. Wang, and J. Pei. Minimality attack in privacy preserving data publishing. In *VLDB*, 2007.

[18] R. Wong, J. Li, A. Fu, and K. Wang. (alpha, k)-anonymity: An enhanced k-anonymity model for privacy-preserving data publishing. In *SIGKDD*, 2006.

[19] X. Xiao and Y. Tao. Anatomy: Simple and effective privacy preservation. In *VLDB*, 2006.

[20] X. Xiao and Y. Tao. Personalized privacy preservation. In *SIGMOD*, 2006.

[21] X. Xiao and Y. Tao. $m$-invariance: Towards privacy preserving re-publication of dynamic datasets. In *SIGMOD*, 2007.