Preservation of Utility through Hybrid k-Anonymization*

Mehmet Ercan Nergiz, Muhammed Zahit Gök, and Ufuk Özkanlı

Department of Computer Engineering, Zirve University, Gaziantep, Turkey

Abstract. Anonymization-based privacy protection ensures that published data cannot be linked back to an individual. The most common approach in this domain is to apply generalizations on the private data in order to maintain a privacy standard such as k-anonymity. While generalization-based techniques preserve truthfulness, relatively small output space of such techniques often results in unacceptable utility loss especially when privacy requirements are strict. In this paper, we introduce the *hybrid generalizations* which are formed by not only generalizations but also the *data relocation* mechanism. Data relocation involves changing certain data cells to further populate small groups of tuples that are indistinguishable with each other. This allows us to create anonymizations of finer granularity confirming to the underlying privacy standards. Data relocation serves as a tradeoff between utility and truthfulness and we provide an input parameter to control this tradeoff. Experiments on real data show that allowing a relatively small number of relocations increases utility with respect to heuristic metrics and query answering accuracy.

Keywords: Privacy, Anonymization, Privacy-preserving databases.

1 Introduction

The advance of technology along with the low cost of handling data have led service providers to collect personal information with the hope of turning this data into profit. In some cases, the potential value of such data is so great, it needs to be outsourced for analysis or it has to be published for research purposes as is the case with health related data in medical research. However, such data often contain sensitive information that needs to be kept private such as diagnosis and treatments. Thus sharing it raises every privacy concern [10]. In order to preserve the privacy of individuals, data needs to be properly anonymized before publishing meaning the link between sensitive information and individual identity should be removed. Such an anonymization must not only satisfy the

* This work was funded by The Scientific and Technological Research Council of Turkey (TUBITAK) Young Researchers Career Development Program under grant 111E047.

S. Furnell, C. Lambrinoudakis, and J. Lopez (Eds.): TrustBus 2013, LNCS 8058, pp. 97–111, 2013.

underlying privacy requirements but also preserve the utility of the data. Otherwise, it would be difficult to extract useful information from the anonymized data.

Unfortunately, just removing uniquely identifying information (e.g., SSN) from the released data is not enough to protect privacy. Works in [23] and [24] show that using publicly available sources of partially identifying information (*quasi-identifiers*) such as age, gender and zip-code, data records can be re-identified accurately even if there is no direct identifying information in the dataset. For example, in Table 1, suppose we release T as a private table. Even if T does not contain unique identifiers, an adversary that knows that her 41 years old friend Obi from USA with zip 49001 is in the dataset will be able to identify him as tuple q7.

To prevent identification, many different privacy metrics [23,24,18,16,27,21] have been introduced for various adversary models. As an example, *k-anonymity* requires that for each tuple t in the anonymization, there should be at least $k-1$ other tuples indistinguishable with t. Two individuals are said to be indistinguishable if their records agree on the set of quasi-identifier attributes. To achieve the underlying privacy standard, many algorithms have been proposed. A common feature of these algorithms is that they manipulate the data by using *generalizations* which involves replacing data values with more general values (values that include the meaning of the original value and that may also imply other atomic values, e.g., 'Italy' is changed to 'Europe') so that more tuples will express similar meanings. As an example, suppose the desired privacy standard is 3-anonymity. In Table 1, $T^*_{\mu_1}$ is a 3-anonymous generalization of T. Note that generalizations applied to T create two *equality groups* that contain similar tuples with respect to QI attributes. From the adversary's point of view, tuples with each equality group are indistinguishable from each other. If the data owner releases $T^*_{\mu_1}$ instead of T, Obi can at best be mapped to the white equality group of size 5 and to a set of salaries {18K, 35K, 14K, 25K, 29K}.

A nice feature of generalizations is that unlike perturbation techniques (that apply noise to data cells independently before publishing), generalizations preserve the truthfulness of data. However, generalizations result in information loss, thus over-generalization should be avoided as long as the privacy requirements are satisfied. To solve this problem, many heuristics have been designed, however relatively small output space of such techniques often results in huge utility loss especially when privacy requirements are strict [3]. Preservation of utility still stands as a major problem for generalization-based techniques. One of the main reasons for over-generalization is the existence of outliers in private datasets. As the neighborhood of the outliers is not heavily populated in the high dimensional domain, it becomes difficult for an anonymization algorithm to generate an equality group of sufficient size. For those algorithms that are vulnerable to outliers, a relatively large group can degrade the overall utility of the whole dataset [20].

To address the negative effects of outliers and over-generalization, in this paper, we propose the *hybrid generalization* technique which combines the

generalization technique with a *data relocation* mechanism in order to achieve more utilized anonymizations. Data relocation involves changing certain data cells (that act as outliers) to further populate small equality groups of tuples. Over relocation harms truthfulness and localized utility, thus over-relocation should be avoided as well. This can be achieved by bounding the number of relocations that the algorithm can apply, thus controlling the trade-off between truthfulness and utility. Even a small number of relocations can prevent over-generalization. As an example, in Table 1, Table $\widehat{T}$ is a relocation of Table T in which less than 10% of the data cells are relocated (see tuple q4). Table $\widehat{T}^*_{\mu_1}$ shows a 3-anonymization of $\widehat{T}$ (which we will also name as a 10%-hybrid 3-anonymization of T). $\widehat{T}^*_{\mu_1}$ is more specific than $T^*_{\mu_1}$ [1] and possibly more utilized. Our contributions in this paper are as follows:

- We introduce the hybrid k-generalization concept that allows relocation of tuples between groups to increase the overall utility at the cost of truthfulness.
- We show how one can use hybrid generalizations to achieve k-anonymity.
- We present hybrid anonymization algorithms that address three classes of adversaries.
- We empirically compare the hybrid algorithms with previously proposed algorithms and show that hybrid generalizations create better utilized anonymizations.

2 Background and Related Work

Given a dataset (table) T, $T[c][r]$ refers to the value of column c, row r of T. $T[c]$ refers to the projection of column c on T and $T[.][r]$ refers to selection of row r on T. We write $|t \in T|$ for the cardinality of tuple $t \in T$ (the number of times t occurs in T).

Although there are many ways to generalize a given value, we stick to generalizations according to domain generalization hierarchies (DGH) given in Figure 1(a).

Definition 1 (i-Gen Function). *For two data values v^* and v from some attribute A, we write $v^* = \Delta_i(v)$ if and only if v^* is the ith (grand) parent of v in the DGH for A. Similarly for tuples t, t^*; $t^* = \Delta_{i_1 \cdots i_n}(t)$ iff $t^*[c] = \Delta_{i_c} t[c]$ for all columns c. Function $\Delta(v)$ without a subscript returns all possible generalizations of a value v.*

E.g., given Figure 1(a), Δ_1(`USA`)=`N.AM`, $\Delta_{0,2,3}$(`<12,USA,47906>`)=`<12,AM, 47***>`, Δ(`USA`)={`USA`,`N. AM`,`AM`,`*`}

Definition 2 (μ-Generalization). *A generalization mapping μ is any surjective function that maps tuples from domain D to a generalized domain D^* such*

[1] 'more specific' does not necessarily mean 'more utilized'. We should take into account the cost of the relocations. We show, in Section 5, that utility gained due to lesser degrees of generalizations more than compensates the local utility loss due to relocations.

Table. 1. $\mathbf{T}$: private table; $\widehat{\mathbf{T}}$: a 10%-relocation of T; $\mathbf{T}^*_{\mu_1}$, $\widehat{\mathbf{T}}^*_{\mu_2}$: 3-anonymous single dimensional generalizations of T and $\widehat{T}$ respectively; $\mathbf{T}^*_{\mu_2}$: a single dimensional generalization of T

Id	Age	Nation	Zip	Sal.
q1	12	Greece	47906	13K
q2	19	Turkey	47907	15K
q3	17	Greece	47907	28K
q4	23	Spain	49703	14K
q5	38	Brazil	49705	18K
q6	33	Peru	49812	35K
q7	41	USA	49001	14K
q8	43	Canada	49001	25K
q9	48	Canada	49001	29K

T

Not.	Definition
T	A private table
T^*	A generalization of T
$\widehat{T}$	A relocation of T
$\widehat{T}^*$	A hybrid generalization of T
$t \in T$	A tuple in T
μ	A generalization mapping
T^*_μ	The generalization of T with mapping μ

Notations

Id	Age	Nation	Zip	Sal.
q1	12	Greece	47906	13K
q2	19	Turkey	47907	15K
q3	17	Greece	47907	28K
q4	31	Brazil	49703	14K
q5	38	Brazil	49705	18K
q6	33	Peru	49812	35K
q7	41	USA	49001	14K
q8	43	Canada	49001	25K
q9	48	Canada	49001	29K

$\widehat{T}$

Id	Age	Nation	Zip	Sal.
q1	11-30	EU	4*	13K
q2	11-30	EU	4*	15K
q3	11-30	EU	4*	28K
q4	11-30	EU	4*	14K
q5	31-50	AM	4*	18K
q6	31-50	AM	4*	35K
q7	31-50	AM	4*	14K
q8	31-50	AM	4*	25K
q9	31-50	AM	4*	29K

$T^*_{\mu_1}$

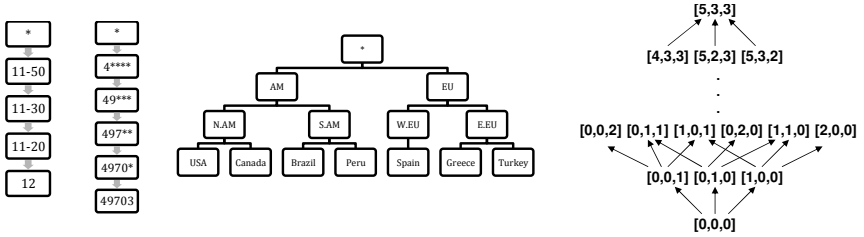
Id	Age	Nation	Zip	Sal.
q1	11-20	E. EU	47*	13K
q2	11-20	E. EU	47*	15K
q3	11-20	E. EU	47*	28K
q4	21-30	W. EU	49*	14K
q5	31-40	S. AM	49*	18K
q6	31-40	S. AM	49*	35K
q7	41-50	N. AM	49*	14K
q8	41-50	N. AM	49*	25K
q9	41-50	N. AM	49*	29K

$T^*_{\mu_2}$

Id	Age	Nation	Zip	Sal.
q1	11-20	E. EU	47*	13K
q2	11-20	E. EU	47*	15K
q3	11-20	E. EU	47*	28K
q4	31-40	S. AM	49*	14K
q5	31-40	S. AM	49*	18K
q6	31-40	S. AM	49*	35K
q7	41-50	N. AM	49*	14K
q8	41-50	N. AM	49*	25K
q9	41-50	N. AM	49*	29K

$\widehat{T}^*_{\mu_2}$

that for $t \in D$ and $t^* \in D^*$; we have $\mu(t) = t^*$ (we also use notation $\Delta_\mu(t) = \mu(t)$ for consistency) only if $t^* \in \Delta(t)$. We say a table T^* is a μ-generalization of a table T with respect to a set of attributes QI and write $\Delta_\mu(T) = T^*$, if and only if records in T^* can be ordered in such a way that $\Delta_\mu(T[QI][r]) = T^*[QI][r]$ for every row r.



(a) DGH structures (b) Generalization Lattice

Fig. 1.

In Table 1, $T^*_{\mu_1}$ and $T^*_{\mu_2}$ are two generalizations of T with mappings μ_1 and μ_2 respectively; E.g, $\Delta_{\mu_1}(T) = T^*_{\mu_1}$. Δ_{μ_1}(<41,US,49001>) = <31-50,AM,4****>

Definition 3 (Single Dimensional Generalization). *We say a mapping μ is $[i_1, \cdots, i_n]$ single dimensional iff given $\mu(t) = t^*$, we have $t^* = \Delta_{i_1 \cdots i_n}(t)$. We define in this case the level of μ as $i_1 + \cdots + i_n$.*

Each attribute in the output domain of a single dimensional mapping contains values from the same level of the corresponding DGH structure. In Table 1, $T^*_{\mu_1}$ and $T^*_{\mu_2}$ are [2,2,4] and [1,1,3] generalizations of T respectively.

Given two single dimensional mappings μ_1=$[i^1_1, \cdots, i^1_n]$ and $\mu_2 = [i^2_1, \cdots, i^2_n]$, we say μ_1 is a higher mapping than μ_2 and write $\mu_1 \subset \mu_2$ iff $\mu_1 \neq \mu_2$ and $i^1_j \geq i^2_j$ for all $j \in [1 - n]$.

We also cover multidimensional generalizations. Due to page limitations, we refer the reader to [22] for related definitions and discussion on the advantages of both approaches.

k-Anonymity privacy protection limits the linking of a record from a set of released records to a specific individual even if adversaries can link individuals to QI:

Definition 4 (k-Anonymity [23,5]). *A table T^* is k-anonymous with respect to a set of quasi-identifier attributes QI if each tuple in $T^*[QI]$ appears at least k times.*

$T^*_{\mu_1}$ is a 3-anonymous generalization of T. Note that given $T^*_{\mu_1}$, the same adversary can at best link Bob to tuples q5, q6, q7, q8, and q9.

Definition 5 (Equality group). *The equality group of tuple t in dataset T^* is the set of all tuples in T^* with identical quasi-identifiers to t.*

In dataset $T^*_{\mu_1}$, the equality group for tuple $q7$ is $\{q5, q6, q7, q8, q9\}$. We use colors to indicate equality groups in Table 1.

While k-anonymity limits identification of tuples, it fails to enforce constraints on the sensitive attributes in a given equality group, thus there is still a risk of sensitive information disclosure. We start our analysis with k-anonymity because it has a simple definition and k-anonymity and k-anonymization is still used in several domains as a privacy metric [9,27,25] and as a sub procedure [26].

In Section 4, we use the *anti-monotonicity* property of k-anonymity. Given $\mu^1 \subset \mu^2$ and a dataset T, if $\Delta_{\mu^1}(T)$ is not k-anonymous, neither is $\Delta_{\mu^2}(T)$. In Table 1, if $T^*_{\mu_2}$ is not 3-anonymous, neither is T.

There may be more than one k-anonymization of a given dataset, and the one with the most information content is desirable. Previous literature has presented many metrics to measure the utility of a given anonymization [12,20,13,6,2]. We use the LM cost metric defined in [12]. Given a is the number of attributes:

$$LM(T^*) = \frac{1}{|T^*| \cdot a} \sum_{i,j} \frac{|\Delta^{-1}(T^*[i][j])| - 1}{|\Delta^{-1}(*)| - 1}$$

Related Work. The value of utility preservation in anonymized dataset has been widely recognized by the literature since the very first works on anonymization-based privacy protection.

The first class of works on utility introduces new heuristic algorithms that generates equality groups composed of tuples that are as close to each other as possible. Grouping of close tuples achieves better utilized generalizations. [23] observes that all possible single dimensional mappings create a lattice over the subset operation. The proposed algorithm finds an optimal k-anonymous generalization (optimal in minimizing a utility cost metric) by performing a binary search over the lattice. [14] improves this technique with a bottom-up pruning approach and finds all optimal k-anonymous generalizations. [2] introduces more flexibility by relaxing the constraint that every value in the generalization should be in the same generalized domain. Works in [18,19,16,7,21] adopt previous single dimensional algorithms for other privacy notions such as ℓ-diversity, t-closeness, and δ-presence. Among other works on heterogeneous generalizations, works in [20,4,1,17] use clustering techniques to provide k-anonymity. [15] and [11] partition the multidimensional space to form k-anonymous and ℓ-diverse groups of tuples. [8] makes use of space filling curves to reduce the dimensionality of the database and provides k-anonymity and ℓ-diversity algorithms that work in one dimension. All of the above works are based on pure generalizations and are orthogonal to our approach. As will be clear in later sections, the relocation technique proposed in this paper can be used to utilize most generalizations regardless of the underlying algorithm. Even though we are not proposing standalone anonymization algorithms, our approach can be considered in this category as we aim to create better equality groups out of existing groups at the cost of truthfulness.

Another way to improve utility is by releasing more information on the equality groups without changing the groupings of the tuples. Our approach differs from such an approach as we form new groupings without specifying how we release the groups. We refer the reader to [22] for a detailed discussion on these approaches.

3 Hybrid Anonymizations

3.1 Classical Adversaries

The classical adversary is the same adversary addressed in most previous literature (see Section 2). The adversary knows the QI attributes of an individual and tries to discover, from the released dataset, the sensitive value belonging to the individual.

As mentioned in Section 1, data relocations can improve utility of released datasets. One should be careful on the number of relocations applied to the dataset as each relocation makes data less truthful. We now formally define *p%-table relocations* in which the maximum number of cell relocations is bounded by the p% of the whole dataset:

Definition 6 (p%-Table Relocations). *We say a table $\widehat{T}$ is a p%-relocation of a table T with respect to a set of attributes QI and write $\widehat{T} \frown^p T$, if and only if records in $\widehat{T}$ can be ordered in such a way that*

- *$T[c][r] \neq \widehat{T}[c][r]$ for at most p% of all possible (attribute $c \in QI$, row r) pairs and*
- *$T[c][r] = \widehat{T}[c][r]$ for all (attribute $c \notin QI$, row r) pairs.*

In Table 1, $\widehat{T}$ is a 10%-relocation of T as only two data cells (see q4) out of 27 is relocated. We now formally define hybrid anonymizations which are created by anonymizing table relocations:

Definition 7 (p%-Hybrid Generalization). *We say a table $\widehat{T}^*$ is a p%-hybrid generalization of a table T with some mapping μ if and only if there exist a $\widehat{T} \frown^p T$ such that $\widehat{T}^* = \Delta_\mu(\widehat{T})$.*

In Table 1, $\widehat{T}^*_{\mu_2}$ is a 10%-hybrid generalization of table T with mapping [1,1,3]. From now on, we assume $p = 10\%$ and do not mention p in our discussions.

Definition 8 (Hybrid k-anonymity). *We say a table $\widehat{T}^*$ is a k-anonymous hybrid of a table T if and only if $\widehat{T}^*$ is a p%-hybrid generalization of a table T and $\widehat{T}^*$ is k-anonymous.*

In Table 1, $\widehat{T}^*_{\mu_2}$ is a 3-anonymous hybrid of the table T.

As the domain of all possible generalizations of a given table T is a subset of the domain of all possible hybrid anonymizations of T, LM cost of an optimal hybrid anonymization will be at least as small as that of a generalization under the same privacy standard. For example, $T^*_{\mu_1}$ and $\widehat{T}^*_{\mu_2}$ both satisfy 3-anonymity, however, $\widehat{T}^*_{\mu_2}$ has a smaller LM cost as μ_2 is a more specific mapping. This does not necessarily mean $\widehat{T}^*_{\mu_2}$ is more utilized as LM cost does not take into account the information loss due to relocations. However, in practice, for most applications, a small number of relocations can increase the overall utility of the released dataset at the expense of decreasing utility on relocated data cells. In order to benefit from hybrid anonymizations, we now state the problem of k-anonymity in the context of hybrid anonymizations. In Section 4, given a private table T, we propose algorithms to find a k-anonymous hybrid $\widehat{T}^*$ of T that minimizes the LM cost metric. .

3.2 Statistical Adversaries

In a hybrid anonymization, the distribution of the tuples in the released data will deviate from the original distribution. If the deviation is too large, an adversary that knows about the original distribution may suspect that some groups in the released data have been artificially populated. For example, in a census dataset, if the adversary sees that there are considerably more males than females, the adversary can suspect that some females are relocated. To defend against such attack, the distance between the original distribution and relocated distribution should be bounded such that the deviation should look as if occurred by chance.

Definition 9 (α-Hybrid k-Anonymization). *Let T be a private table and X be the multinomial random variable from which the tuples are drawn from. $\widehat{T}^*$ is an α-hybrid if the hypothesis that the group sizes in $\widehat{T}^*$ are consistent with the parameters of X cannot be rejected at the significance level α.*

For significance testing, we use the Pearson's chi-squared test for multinomial distributions. Given $\widehat{T}^* = \{G_1, \cdots, G_n\}$ with mapping μ and size N, the X^2 can be approximated as follows. Let $E_i = N \cdot \sum_{t \mid \mu(t)=G_i} \mathcal{P}(X = t)$.

$$X^2 = \sum_{i}^{n} (\frac{|G_i| - E_i}{E_i})$$

Note that we assume a strong adversary that knows the exact distribution X of the tuples or sensitive values. In reality, the adversaries may only know partial information about X, such as "the number of Italians is less than Chinese". As it is difficult to predict the true background of the adversary, we assume a worst case scenario.

In addition to the above mentioned adversaries, we will also assume adversaries might know the underlying hybrid algorithm. It has been shown in [26,29] that such adversaries can reverse-engineer the anonymization algorithm and learn information that would not be allowed by the underlying privacy metric. However, ensuring a theoretical bound on the disclosure against such adversaries is not a trivial problem and can also result in a huge decrease in utility. Instead, we will make it practically hard for such an adversary to reverse-engineer the algorithm by making random decisions during the algorithm. We will employ multiple random sources within the algorithm that will generate many possible pathways for the algorithm to follow.

Algorithm 1. S-Hybrid

Require: a private table T from domain D, privacy parameter k, a utility cost metric CM, a user threshold p;

Ensure: return a minimum cost k-anonymous single dimensional hybrid generalization of T.

```
1: create lattice lat for all possible generalization mappings for D. Let n be the
   maximum level of mappings in lat.
2: for all level i from n to 0 do
3:   for all mapping μ of level i in lat do
4:     T̂* = createHybrid(Δ_μ(T), k, p)
5:     if T̂* is not k-anonymous then
6:       delete node μ and all children and grandchildren of μ from lat.
7:     else
8:       calculate cost CM(T̂*) and store the cost on the lattice node.
9: if no node exists on the lat then
10:   return null.
11: find the mapping μ with the minimum cost on the lat.
12: return createHybrid(Δ_μ(T), k, p)
```

4 Hybrid Anonymization Algorithms

In this section, we present a set of single dimensional hybrid k-anonymization algorithms each addressing a different adversary as mentioned in Section 3. All

Algorithm 2. CreateHybrid

Require: a generalization T^*, privacy parameter k, and a user threshold p;
Ensure: return $p\%-$hybrid generalization $\widehat{T}^*$ with the same mapping as that of T^*. $\widehat{T}^*$ will not contain any more non k-anonymous groups than T^*.

```
1: T̂* = T*;
2: let G be the set of equality groups in T̂*.
3: let G_sm ⊂ G be the set of groups with less than or equal to k/2 tuples.
4: let G_big ⊂ G be the set of groups with more than k/2 tuples, but less than
   k tuples.
5: for all g ∈ G_sm do
6:    let ptr be an empty group pointer to hold a target group.
7:    if G_big is not empty then
8:       find the group g' ∈ G_big that is closest to g.
9:       ptr → g'
10:   else
11:      find the group g' ∈ G − G_sm that is closest to g.
12:      ptr → g'
13:   if ptr is null then
14:      return T̂*.
15:   change all tuples in g so that the tuples are moved (relocated) into g'|ptr →
      g'
16:   remove g', update G, G_sm, G_big accordingly.
17:   if T̂* is not a p%−hybrid generalization then
18:      roll back the last change and return T̂*;
19: for all g ∈ G_big do
20:   find the group g' ∈ G − G_sm − G_big that has more than 2k − |g| tuples
      and is closest to g.
21:   if no such g' exists then
22:      return T̂*;
23:   pick any k − |g| tuples in g' and change the tuples such that they are
      moved (relocated) into g.
24:   update G, G_sm, G_big accordingly.
25:   if T* is not a p%−hybrid generalization then
26:      roll back the last change and return T̂*;
27: return T̂*;
```

algorithms trace the whole space of single dimensional mappings and returns a mapping as an approximation to the problem of hybrid anonymity. The algorithms are based on the optimal single dimensional anonymization algorithm, Incognito [14] but improve Incognito by searching the space of hybrid generalizations (Definition 7) rather than table generalizations 3.

Deterministic S-Hybrid. The pseudocode for the S-Hybrid algorithm is given in Algorithm 1. The algorithm traverses the whole space of single dimensional

mappings, applies each mapping to the private dataset, produces a hybrid generalization by calling the function createHybrid. Fortunately, the possible single dimensional mappings over a table domain form a lattice on the $\subset$ relation (see Figure 1(b)). In lines 2-8, we traverse the lattice in a top-down manner. In lines 5-6, we use the anti-monotonicity property of k-anonymity to prune the lattice, thus reduce the search space.

The pseudocode for the createHybrid function is given in Algorithm 2. The aim of the algorithm is to convert the given generalization into a k-anonymous hybrid generalization with fewest relocations as possible. We start by classifying the groups as G_{sm} (groups with less than or equal to $k/2$ tuples), G_{big} (groups with more than $k/2$ but less than k tuples) and G (all groups). The reason for such a classification is that distributing the tuples in groups of few tuples while completing groups that have almost k tuples potentially minimizes the required number of relocations. With such reasoning, the relocation of tuples are done in two phases:

Distribution: In lines 5-18, the algorithm attempts to relocate the tuples in G_{sm} first into the closest group in G_{big}. Two tuples are closest if they agree on the most number of attributes. If G_{big} is empty, the tuples are relocated into the closest k-anonymous group. After this phase, some groups in G_{big} may become k-anonymous, thus may be removed from G_{big}.

Completion: In lines 19-26, the algorithm relocates tuples from closest k-anonymous groups that has enough number of tuples into groups in G_{big}.

After each relocation, the algorithm checks if the maximum number of allowed relocations (specified with the input p) has been exceeded. If that is the case, the algorithm roll backs the last relocation and returns the non k-anonymous hybrid generalization generated so far.

As an example, in Table 1, if we use μ_2=[1,1,3] as the generalization mapping, $k = 3$, and $p = \%10$; $T^*_{\mu_2}$ will be the input to the createHybrid algorithm. The algorithm will set $G_{sm} = \{\{q4\}\}$ and $G_{big} = \{\{q5, q6\}\}$. The algorithm starts distributing the tuples in G_{sm} into groups in G_{big}. The tuple $q4$ will be sent to the only (closest) group $\{q5, q6\}$ in G_{big}. As a result, $q4$ becomes `<31,Brazil,49703>`. Note that this change only relocates 2 out of 27 data cells thus performing the change creates a 10%-table relocation. Since the resulting hybrid generalization is k-anonymous, the algorithm returns $\widehat{T}^*_{\mu_2}$.

Randomized S-Hybrid. As mentioned in Section 3, adversaries that know the underlying algorithm can attempt to reverse-engineer the algorithm and create non k-anonymous subgroups [26,29]. Such attacks pose a threat to privacy especially if the underlying algorithm is deterministic. To resist reverse-engineering attacks, we create the Randomized S-Hybrid algorithm. The algorithm makes random decisions at certain points, thus can follow multiple pathways making reverse-engineering attacks difficult. The sources of randomness can be listed as follows:

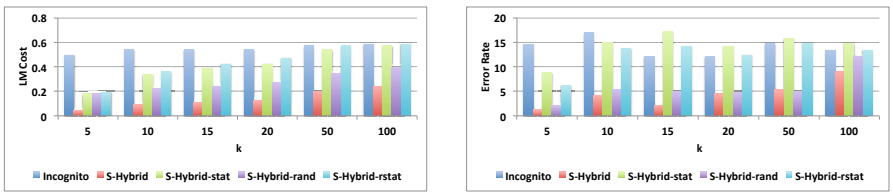
$\rightarrow$ In the distribution/completion phases, in lines 8, 11, and 20, instead of picking the closest group as the target/source group for relocation, we pick the

target group randomly from the sets G_{big} or G. Most of the time, the sizes of these sets are large enough to create a probability space of sufficient size for the flow of the algorithm.

→ In the completion phase, in line 23, instead of relocating exactly $k - |g|$ tuples, we relocate a random number of tuples such that both the target and the source group remains / becomes k-anonymous. Relocating a random number of tuples prevents the source group to contain exactly k tuples.

Statistical S-Hybrid. As mentioned in Section 3.2, statistical adversaries use the known distribution of the tuples to identify artificial relocations. α-Hybrid k-anonymization addresses such adversaries by bounding the statistical difference between the original distribution and relocated distribution. Deterministic S-Hybrid algorithm can easily be modified so that it accepts the statistical threshold α as an input and returns α-Hybrid generalizations. In Algorithm 2, in lines, 17 and 25, whenever we check if a relocation violates p%-Hybrid anonymity, we instead check if the relocation violates α-Hybrid anonymity.

It should be noted that even with low α settings, α-Hybrid anonymity is a strict privacy definition. That is, the definition allows fewer number relocations making it more difficult to create a k-anonymous hybrid from a non k-anonymous generalization. In Section 5, we empirically compare α-Hybrid anonymity with p%-Hybrid anonymity in terms of utility and show that the former allows a lower level of utility at the benefit of stronger privacy.



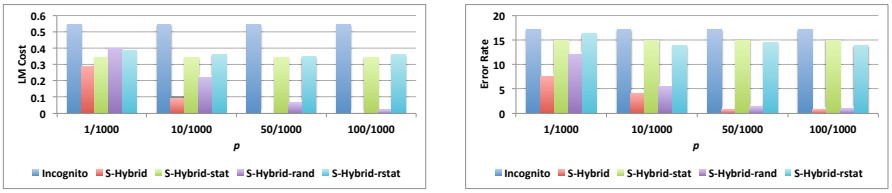
(a) Single Dimensional- LM Cost (b) Single Dimensional - Error rate

Fig. 2. Varying k - Sal

We also designed multidimensional versions of the algorithms proposed in this section (M-Hybrid). Due to space constraints, we refer the reader to [22] for details.

5 Experiments

This section presents the experimental evaluation of S-Hybrid and M-Hybrid algorithms. In addition to the three algorithms mentioned in Section 4, we also evaluate algorithm *S-Hybrid-rstat* which is randomized S-Hybrid against statistical adversaries. During our experiments, we use the real datasets 'Sal' and 'Occ' that were extracted from CENSUS data and previously used by [28,8]. Both datasets contain 100.000 tuples. As the results were similar for both dataset, we



(a) Single Dimensional- LM Cost (b) Single Dimensional - Error rate

Fig. 3. Varying Distortion Limit p - Sal

present results only on the 'Sal' dataset. Results on the 'Occ' dataset can be accessed from [22].

We used two metrics to measure utility: LM cost metric defined in Section 2 and the range query error rate used in [8,15]. Query error is calculated by issuing count queries on anonymizations and normalizing the deviation of count from the original count in the private dataset.

Varying Privacy Parameter k

We first fix the distortion limit as %1, α as %5, vary the value of k and compare S-Hybrid algorithm with the previously proposed single dimensional Incognito algorithm with respect to the LM cost and query error metric. Note that %1 distortion limit is almost a negligible sacrifice from the truthfulness of data. Figure 2 shows the results on the 'Sal' dataset. According to utility cost experiments, in nearly all cases, all Hybrid approaches give better results than the algorithm Incognito. S-Hybrid and randomized S-Hybrid perform better than statistical Hybrid algorithms. As mentioned before, this is because, compared to hybrid k-anonymity, α-hybrid k-anonymity is a strict privacy definition assuming a powerful adversary. In most cases, the number of allowed relocations for α-hybrid anonymity is much smaller than that for hybrid anonymity resulting in less utilized anonymizations. For statistical S-Hybrid and Incognito, for some cases, we observed fluctuations in error rates when we increase k. The reason is that, in these settings, the resulting mappings cannot be ordered with respect to the $\subset$ operator (see the definition of higher mappings in Section 2) and are close to each other in the generalization lattice. Any one of the mappings can be considered better utilized than the other depending on the underlying application.

Varying Distortion Limit p

In these experiments, we fix the value of k as 10, α as %5, vary the value of the distortion limit p. We present the results in Figure 3. We see that non statistical S-Hybrid algorithms increase in utility as we increase the distortion limit (e.g., as we apply more and more relocations). Statistical S-Hybrid algorithms are not very sensitive to changes in p. The reason is that significance test via the parameter α is more decisive on the number of allowed relocations than the limit via the p. Generally, significance test for further relocations fail even before

the number of relocations reach %0.1. Thus the utility of statistical S-Hybrid algorithms do not change much. The comparison of algorithms with each other is similar as mentioned in the previous section.

We also made experiments regarding the multidimensional hybrid algorithm, M-Hybrid. M-Hybrid algorithms show a similar behavior. Due to space constraints, we refer the reader to [22] for experimental results.

6 Future Work

As a possible future work, new hybrid algorithms can be designed for other privacy metrics such as ℓ-diversity, (α, k)-anonymity or δ-presence. This would be crucial in addressing different types of adversaries. There is also room for improvement for the hybrid algorithms proposed in this paper. For example, one can design hybrid algorithms that would theoretically bound the probability of identification against algorithm-aware adversaries. Hybrid techniques can also be evaluated with respect to different cost metrics and real applications so that utility gain can better be quantified under different scenarios.

References

1. Agrawal, G., Feder, T., Kenthapadi, K., Khuller, S., Panigrahy, R., Thomas, D., Zhu, A.: Achieving anonymity via clustering. In: PODS 2006: Proceedings of the 25th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, Chicago, IL, USA, June 26-28, pp. 153–162 (2006)
2. Bayardo, R.J., Agrawal, R.: Data privacy through optimal k-anonymization. In: ICDE 2005: Proceedings of the 21st International Conference on Data Engineering, pp. 217–228. IEEE Computer Society, Washington, DC (2005)
3. Brickell, J., Shmatikov, V.: The cost of privacy: destruction of data-mining utility in anonymized data publishing. In: Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2008, pp. 70–78. ACM, New York (2008)
4. Byun, J.-W., Kamra, A., Bertino, E., Li, N.: Efficient k-anonymization using clustering techniques. In: Kotagiri, R., Radha Krishna, P., Mohania, M., Nantajeewarawat, E. (eds.) DASFAA 2007. LNCS, vol. 4443, pp. 188–200. Springer, Heidelberg (2007)
5. Ciriani, V., De Capitani di Vimercati, S., Foresti, S., Samarati, P.: k-anonymity. In: Secure Data Management in Decentralized Systems, pp. 323–353 (2007)
6. Domingo-Ferrer, J., Torra, V.: Ordinal, continuous and heterogeneous k-anonymity through microaggregation. Data Mining and Knowledge Discovery 11(2), 195–212 (2005)
7. Fung, B.C.M., Wang, K., Yu, P.S.: Top-down specialization for information and privacy preservation. In: ICDE 2005: Proceedings of the 21st International Conference on Data Engineering, pp. 205–216. IEEE Computer Society, Washington, DC (2005)
8. Ghinita, G., Karras, P., Kalnis, P., Mamoulis, N.: Fast data anonymization with low information loss. In: VLDB 2007: Proceedings of the 33rd International Conference on Very Large Data Bases, pp. 758–769. VLDB Endowment (2007)

9. Gionis, A., Mazza, A., Tassa, T.: k-anonymization revisited. In: IEEE 24th International Conference on Data Engineering, ICDE 2008, pp. 744–753 (April 2008)
10. Standard for privacy of individually identifiable health information. Federal Register, 66(40) (February 28, 2001)
11. Hore, B., Ch, R., Jammalamadaka, R., Mehrotra, S.: Flexible anonymization for privacy preserving data publishing: A systematic search based approach. In: Proceedings of the 2007 SIAM International Conference on Data Mining (2007)
12. Iyengar, V.S.: Transforming data to satisfy privacy constraints. In: KDD 2002: Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 279–288. ACM, New York (2002)
13. Kifer, D., Gehrke, J.: Injecting utility into anonymized datasets. In: SIGMOD 2006: Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data, pp. 217–228. ACM, New York (2006)
14. LeFevre, K., DeWitt, D.J., Ramakrishnan, R.: Incognito: Efficient full-domain k-anonymity. In: SIGMOD 2005: Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data, pp. 49–60. ACM, New York (2005)
15. LeFevre, K., DeWitt, D.J., Ramakrishnan, R.: Mondrian multidimensional k-anonymity. In: ICDE 2006: Proceedings of the 22nd International Conference on Data Engineering, Atlanta, GA, April 3-7, pp. 25–35 (2006)
16. Li, N., Li, T.: t-closeness: Privacy beyond k-anonymity and l-diversity. In: ICDE 2007: Proceedings of the 23rd International Conference on Data Engineering, Istanbul, Turkey, April 16-20 (2007)
17. Lin, J.-L., Wei, M.-C., Li, C.-W., Hsieh, K.-C.: A hybrid method for k-anonymization. In: Asia-Pacific Services Computing Conference, APSCC 2008, pp. 385–390. IEEE (2008)
18. Machanavajjhala, A., Gehrke, J., Kifer, D., Venkitasubramaniam, M.: ℓ-diversity: Privacy beyond k-anonymity. In: ICDE 2006: Proceedings of the 22nd IEEE International Conference on Data Engineering, Atlanta Georgia (April 2006)
19. Nergiz, M.E., Atzori, M., Clifton, C.: Hiding the presence of individuals in shared databases. In: SIGMOD 2007: Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data, Beijing, China, June 11-14 (2007)
20. Nergiz, M.E., Clifton, C.: Thoughts on k-anonymization. Data and Knowledge Engineering 63(3), 622–645 (2007)
21. Nergiz, M.E., Clifton, C.: δ-Presence without complete world knowledge. IEEE Transactions on Knowledge and Data Engineering, 868–883 (2009)
22. Nergiz, M.E., Gok, M.Z., Ozkanli, U.: Preservation of utility through hybrid k-anonymization. Technical Report TR 2013-001, Department of Computer Engineering, Zirve University (2013)
23. Samarati, P.: Protecting respondent's identities in microdata release. IEEE Transactions on Knowledge and Data Engineering 13(6), 1010–1027 (2001)
24. Samarati, P., Sweeney, L.: Generalizing data to provide anonymity when disclosing information (abstract). In: PODS 1998: Proceedings of the Seventeenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, p. 188. ACM, New York (1998)
25. Tamersoy, A., Loukides, G., Nergiz, M.E., Saygin, Y., Malin, B.: Anonymization of longitudinal electronic medical records. IEEE Transactions on Information Technology in Biomedicine 16(3), 413–423 (2012)

26. Wong, R.C.-W., Fu, A.W.-C., Wang, K., Pei, J.: Minimality attack in privacy preserving data publishing. In: VLDB 2007: Proceedings of the 33rd International Conference on Very Large Data Bases, pp. 543–554. VLDB Endowment (2007)
27. Wong, R.C.-W., Li, J., Fu, A.W.-C., Wang, K. (α, k)-anonymity: An enhanced k-anonymity model for privacy preserving data publishing. In: KDD 2006: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 754–759. ACM, New York (2006)
28. Xiao, X., Tao, Y.: Anatomy: Simple and effective privacy preservation. In: VLDB 2006: Proceedings of 32nd International Conference on Very Large Data Bases, Seoul, Korea, September 12-15, pp. 139–150 (2006)
29. Zhang, L., Jajodia, S., Brodsky, A.: Information disclosure under realistic assumptions: privacy versus optimality. In: CCS 2007: Proceedings of the 14th ACM Conference on Computer and Communications Security, pp. 573–583. ACM, New York (2007)