Cloning for Privacy Protection in Multiple Independent Data Publications

Muzammil M. Baig, Jiuyong Li, Jixue Liu
School of Computer and Information Science
University of South Australia, Mawson Lakes,
SA. 5095 Australia
{muzammil.baig, jiuyong.li,
jixue.liu}@unisa.edu.au

Hua Wang
Department of Maths and Computing
University of Southern Queensland,
Toowoomba, Queensland, 4350 Australia.
wang@usq.edu.au

ABSTRACT

Data anonymization has become a major technique in privacy preserving data publishing. Many methods have been proposed to anonymize one dataset and a series of datasets of a data owner. However, no method has been proposed for the anonymization of data of multiple independent data publications. A data owner publishes a dataset, which contains overlapping population with other datasets published by other independent data owners. In this paper we analyze the privacy risk in the such scenario and vulnerability of partitioned based anonymization methods. We show that no partitioned based anonymization methods can protect privacy in arbitrary data distributions, and identify a case that the privacy can be protected in the scenario. We propose a new generalization principle ϵ-cloning to protect privacy for multiple independent data publications. We also develop an effective algorithm to achieve the ϵ-cloning. We experimentally show that the proposed algorithm anonymizes data to satisfy the privacy requirement and preserves good data utility.

Categories and Subject Descriptors

H.2.0 [**Database Management**]: General

General Terms

Management, Security, Theory

1. INTRODUCTION

Private individual-specific information such as customer data, employee data etc. are maintained and shared for various purposes. The advantages of such sharing are well documented but in the recent past several instances of data privacy breaches [2], due to data sharing, have resulted in financial and reputational losses for enterprises. Partition-based privacy preserving data publishing techniques address this problem by anonymizing data such that individual privacy is preserved when data is shared or released. The basic idea behind these techniques is *one-in-crowd* which guarantees that an individual cannot be distinguished from a minimum number of other people. Partition-based anonymization techniques are widely discussed in literature and well known schemes include k-anonymity [25], (α,k)-anonymity [28], ℓ-diversity [20], t-closeness [19] and anatomy [32]. Existing partition-based techniques focus on one-time publication [25, 20], multiple views of the same data [34, 33]; and the series of datasets by single data owner [29, 10, 31]. Multiple data publications are restricted to single owner, and does not support privacy preserving data publications of overlapping records by multiple publishers. Ganta et al. [11] firstly identified the privacy breach of overlapping population within multiple published datasets, and called this privacy breach *composition attack*. However, the solution of [11] supports only *interactive setting* (where only data statistics and/or query results are released), and is inapplicable for *non-interactive setting* (where the data is published after anonymization). Privacy preserving data anonymization against known and unknown overlapping population remains an open problem. To the best of our knowledge no solution exists for these scenarios.

	Hospital-1	Hospital-2	Hospital-3
Nancy	●	●	
David	●		●
Eliza		●	●

Figure 1: Full overlapping scenario among three hospitals

Before we illustrate the problem, firstly let us consider patient overlapping scenario of three hospitals in Figure 1. There are three overlapping patients i.e. Nancy, David and Eliza. Nancy has visited Hospitals 1 and 2, Eliza, has visited the Hospitals 2 and 3 and David has visited the Hospitals 1 and 3. All hospitals independently anonymize their datasets and are unaware of their overlapping records with other hospitals.

Table 1(a) is the original data of Hospital-1, which has three type of attributes. *Identifier attributes* can directly identify individuals, such as Name, SSN etc. They should be removed in a published dataset. *Quasi identifier (QI) attributes* could indirectly lead to the identification of individuals in a dataset, such as Age, Zipcode and Sex etc. They are normally generalized so that no individuals are identifiable in a generalized dataset. The *Sensitive attribute* contains the private information about the individuals that needs to be protected such as Disease, Income etc. There is another type of attribute called *non-sensitive attribute*, which do not fall under aforementioned types but is useful for some data analysis. In our case, we do not consider them since they are unrelated to

CIKM'11, October 24–28, 2011, Glasgow, Scotland, UK.

Table 1: Hospital-1

(a) Original Table

<Marry, 15, F> C
<Vince, 18, M> C
<Rabi, 25, F> D
<Nana, 30, F> D
<Nancy, 50, F>D
<Bob, 23, M> G
<Steve, 40, M> G
<David, 15, M>J
<Paul, 35, M> J

(b) 4-anonymous release

Age	Sex	Disease
15 – 40	M	C (1) G (2) J (2)
15 – 50	F	C (1) D (3)

(c) Privacy clone release

Age	Sex	Disease
15 – 50	*	J (2) G (2) C (2) D (3)

(d) ϵ-clone release with micro-statistics

Age	Sex	Disease
~19 15 – 25	M (2) F (2)	C (1) D (1) G (1) J (1)
~35 18 – 50	M (3) F (2)	C (1) D (2) G (1) J (1)

Table 2: Hospital-2

(a) Original Table

<Alex, 15, M > A
<Mary, 22, F > A
<Tom, 38, M > A
<Linda, 10, F> D
<Khan, 18, M> D
<Nancy, 50, F> D
<Andy, 25, M> G
<Eliza, 30, F> G
<Maxi, 45, F> G
<Hussy, 23, M> J
<Tina, 35, F> J
<Paul, 35, M> J

(b) 6-anonymous release

Age	Sex	Disease
10 – 38	M	A (2) D (1) G (1) J (2)
10 – 50	F	A (1) D (2) G (2) J (1)

(c) Privacy clone release

Age	Sex	Disease
10 – 50	*	D (3) J (3) A (3) G (3)

(d) ϵ-clone release with micro-statistics

Age	Sex	Disease
~19 19 – 23 (2)	M (4) F (2)	A (2) D (2) G (1) J (1)
10 – 25		
~39 35 – 39 (2)	M (2) F (4)	A (1) D (1) G (2) J (2)
30 – 50		

data anonymization. A *generalized* table is considered privacy preserved, if it satisfies a privacy requirement, such as k-anonymity [25], (α,k)-anonymity [28], ℓ-diversity [20], t-closeness [19] etc. Hospital-1 releases Table 1(b) as a 4-*anonymous* and 2-*diverse* version of the original Table 1(a). 4-anonymity means that values in the quasi-identifier have at least 4 identical copies. So one could not be distinguished from other 3 records. Such a group is called an *equivalence class* (formal definition in Section 4). 2-diversity means that each of such a group has at least 2 distinct values in the sensitive attribute. So the sensitive value of each individual could not be inferred with a high confidence.

2. PROBLEM DESCRIPTION AND MOTIVATION

All existing partition-based privacy preserving data publishing techniques, such as [20, 32, 19, 22, 6, 16, 26, 14, 30] focus on forming equivalence classes (also called partitions or QI groups) based upon some generalization principles. To illustrate the problem, let us assume that Hospital-2 releases Table 2(b) as a 6-anonymous and 3-diverse version of the original Table 2(a) and Hospital-3 releases Table 3(b) as 4-anonymous and 3-diverse version of the original Table 3(a). The Nancy's equivalence class in Table 1(b) (Hospital-1's anonymous release) and in Table 2(b) (Hospital-2's anonymous release) has only 1 common sensitive value i.e. $\{D\}$. So an adversary knowing the QIs of Nancy (50 years old female) and the fact she has visited two hospitals can derive her sensitive value from both tables. Such utilization of more than one individually anonymized datasets to infer the privacy of overlapping individual(s) is called a 'composition attack' [11]. Using the composition attack adversary can get the diseases of David and Eliza in Tables 1(b), 3(b) and Tables 2(b), 3(b) respectively.

Individually all three anonymous datasets, Table 1(b), Table 2(b) and Table 3(b) pose low privacy risk but collectively compromise the privacy of overlapping patients due to the composition attack [11, 3]. In other words independently anonymized datasets do not retain the privacy properties under the composition attack. Note that there are other possible generalized tables of original tables too. However, they suffer the same type of privacy disclosures.

Our method proposed in this paper leads to the publication of Table 1(d), 2(d) and 3(c). Specifically, each equivalence class of these tables contains all the sensitive values of its respective original dataset. Moreover, unlike a definite QI range (as done in previous releases of these tables), each equivalence class of Table 1(d), 2(d) and 3(c) has *micro-statistics* values (to be discussed in detail in Section 6). The purpose of releasing such statistics is for better privacy preserving and retaining statistical properties.

Let adversary now attempt to infer the diseases of Nancy, David and Eliza from Table 1(d), 2(d) and 3(c). The adversary can locate that the equivalence classes of Nancy has 3, and David/Eliza have 2 sensitive values common i.e. $\{D,G,J\}$ and $\{G,J\}$ respectively. Therefore adversary cannot get a specific disease that Nancy, David or Eliza has contracted.

One straightforward approach that can tackle the composition attack is trivial sanitizer [4] which simply suppresses all quasi-identifiers (QIs) or all sensitive attributes and publish all *non-sensitive* attributes intact. In our case, such sanitization makes data less useful since all information of QIs and sensitive value is lost. Our motivation is to achieve privacy against composition attack and also to maintain, as much as possible, the high data utility of anonymous releases.

Table 3: Hospital-3

(a) Original Table

	Name	Age	Sex	Marital Status	Disease
David	t_1	**15**	**M**	**Never-married**	**J**
	t_2	18	F	Never-married	J
	t_3	25	M	CIV-spouse	J
	t_4	35	F	AF-spouse	J
	t_5	38	M	Divorced	J
	t_6	10	M	Never-married	B
	t_7	18	M	Never-married	B
	t_8	20	M	Never-married	B
	t_9	25	F	Never-married	B
	t_{10}	30	F	Separated	B
	t_{11}	35	F	Divorced	B
	t_{12}	45	M	Separated	B
	t_{13}	20	M	AF-spouse	G
Eliza	t_{14}	**30**	**F**	**CIV-spouse**	**G**
	t_{15}	40	F	CIV-spouse	G
	t_{16}	10	F	Never-married	L
	t_{17}	17	F	Never-married	L
	t_{18}	22	M	Separated	L
	t_{19}	28	F	Widowed	L

(b) 4-anonymous release

Age	Sex	Marital Status	Disease
10 – 18	*	Never-married	B (2) J (2) L (1)
17 – 25	*	*	B (2) J (1) L (1)
20 – 40	*	*	B (1) G (2) L (1)
25 – 28	*	*	B (1) J (2) L (1)

(c) ϵ-clone release with micro-statistics

Age	Sex	Marital Status	Disease
~15 14 – 16 (1)	M (3)	AF-spouse (1) Never-married (2)	G (1) J (2) B (2) L (1)
~21 19 – 21 (1)	M (2) F (1)	Never-married (2) CIV-spouse (1)	
10 – 20			
~22 18 – 22 (1)	M (1) F (2)	CIV-spouse (1) Never-married (2)	G (1) J (2) B (2) L (1)
~27 25 – 27 (1)	M (1) F (2)	Separated (1) AF-spouse (1) Never-married (1)	
18 – 35			
~38 37 – 39 (1)	M (1) F (2)	CIV-spouse (1) Divorced (2)	G (1) J (2) B (2) L (1)
~29 28 – 28 (1)	F (2) * (1)	Widowed (1) * (1) Separated (1)	
28 – 40			

3. CONTRIBUTIONS

The problem of overlapping data publications is not resolvable by the methods of sequential data publishing, such as [29, 31, 10] and multiple views of the same dataset [33, 34]. Both, sequential data and multiple view, deals with two overlapping data publications of the same data owner. The data owner *knows* the previously released data/view(s) of the dataset and *uses* these data/view(s) during the anonymization process of the next publication. In our scenario of multiple publications, a publisher is *unaware* of other datasets that have overlapping records with it. Therefore, methods of sequential and multiple view data publication are inapplicable to multiple independent data publications. The sequential/multiple views inference is between the two different releases of same location; whereas the composition attack works between the independent releases of mutually unknown locations.

Firstly, In this paper we analyze the privacy risk in multiple independent data publications situations and vulnerability of partition based anonymization methods in such an environment. We show that no partition based anonymization methods can protect privacy in arbitrary data distributions, and identify a case that the privacy can be protected in.

Secondly, this paper proposes a multiple data publications model to protect a dataset from the composition attack when different locations independently release the anonymous data of overlapping population. The core of our solution is the integration of two novel concepts: ϵ-*cloning* and *micro-statistics based anonymization.* The former is a new model, whose satisfaction ensures the association of a tuple with all sensitive values of data; hence providing the maximum privacy protection. The latter is an anonymization technique that facilitates the implementation of ϵ-cloning.

4. BASIC DEFINITIONS

Let P be a dataset maintained by a publisher. There are n other publishers which have overlapping subset with P. We assume that all n publishers and their published datasets are unknown to the publisher of P. Each published dataset Q_i^* ($i \in$ 1,2,3, …, n) is independently anonymized from its original dataset Q_i.

We classify the columns of P and Q_i into three types (already explained in Section 1): **(i)** an identifier attribute A^{id}, which is the primary key of P and Q_i, **(ii)** the d quasi-identifier (QI) attributes $A_1^{qi}, A_2^{qi}, \ldots, A_d^{qi}$, and **(iii)** a sensitive attribute S. The QI attributes can be either numerical or categorical. For each tuple $t \in P$, $t[A]$ denotes its value on attribute A.

Definition 1 (Generalized QI group / Equivalence class) *For anonymous dataset P^*, an equivalence class (E) is set of the tuples in P^* with the same values in QI attributes. Each equivalence class is assigned an ID A^g.*

For a tuple $t \in P^*$; we refer $t.E$ as the **hosting equivalence class** of the tuple t in P^*. When the publisher P releases P^*, the adversary can use P^* and any Q_i^* to intrude the privacy of overlapping subset by the composition attack. To formalize the attack, we first introduce a notation $O(P^*, Q_i^*)$.

Definition 2 (Overlapping set) *Let publisher P releases P^*, anonymous version of P, for each independent release $Q_i^*(i \in 1, 2, 3 \ldots, n)$ the overlapping set $O(P^*, Q_i^*)$ contains all those tuples in P^* and Q_i^* such that:*

$$O(P^*, Q_i^*) = \bigcup_{i=1}^{n} (P^* \cap Q_i^*) \;\; (i \in 1, 2, 3 \ldots, n)$$

Each tuple in $O(P^*, Q_i^*)$ is an intersection (to be explained) of two corresponding tuples $t_i \in Q_i^*$ and $t \in P^*$; which satisfy the following requirements:

1. *$t[S] = t_i[S]$; both tuples have the same sensitive value*
2. *$\forall_j : t[A_j^{qi}] \cap t_i[A_j^{qi}] \neq \emptyset$; t and t_i have overlapping value intervals in every QI attribute.*

For two value intervals (or values), the intersection returns the overlapping range of two intervals. For example, for age QI intervals (15–25) $\cap$ (20–30) = (20–25). For categorical QI values in generalization taxonomies, the intersection returns the most specific QI value of two QI values if the one is a generalization of another otherwise returns $\emptyset$. For example, for sex QI values ($* \cap male = male$), ($* \cap female = female$) and ($female \cap male = \emptyset$); '$*$' corresponds to the most generalized QI value in any generalization hierarchy. In sex QI generalization hierarchy, '$*$' presents both $male$ and $female$.

Assume that a tuple t occurs in two generalized tables P^* and Q_i^*. As per Definition 2, tuple $t \in O(P^*, Q_i^*)$ and it has two corresponding tuples; each in P^* and Q_i^*. The probability for inferring

the sensitive value of overlapping tuple t from tables P^* and Q_i^* is $\text{prob}(t[S] \mid (P^*, Q_i^*))$:

$$\text{prob}\left(t[S]|(P^*, Q_i^*)\right) = \text{prob}\left(t[S]|O(P^*, Q_i^*)\right) \quad (1)$$

Equation (1) reveals the reason of the failure of partition based generalization schemes like k-anonymity [25] , ℓ-diversity [20], t-closeness [19] etc. for the composition attack . After proper anonymization, the corresponding generalized tuples of overlapping tuple t are protected in each of P^* and Q_i^*. However, composition attack inference has nothing to do with their respective individual probabilities in P^* and Q_i^*. Composition attack inference is only determined by their overlapping set $O(P^*, Q_i^*)$. An overlapping set $O(P^*, Q_i^*)$ may not satisfy either generalization scheme of individual anonymized datasets. In the worst case:

$$\text{prob}\left(t[S]|O(P^*, Q_i^*)\right) = 1 \quad (2)$$

The worst case (2) occurs when only one record in $O(P^*, Q_i^*)$ matches overlapping tuple t or all tuples have the same sensitive value.

Example 1. Let $t =$ <Nancy, 50, F, D> and Q_1^* and P^* as Table 1(b) and 2(b) respectively. The overlapping set $O(P^*, Q_1^*)$ contains the tuple <15–50, F, D>. So the probability for inferring the true sensitive value of Nancy from tables Q_1^* and P^* is $\text{prob}\left(t[S] \mid O(P^*, Q_1^*)\right) = 1$. All matching records in $O(P^*, Q_1^*)$ have the same sensitive value.

5. PRIVACY PRESERVING IN MULTIPLE INDEPENDENT PUBLISHING

In the previous section, we have shown that the privacy of an individual in the overlapping dataset can be inferred with 100% accuracy. Now we discuss when we can prevent such inference and how to do it.

We start with the inference of the probabilities of sensitive values shared by two different datasets with overlapping QI values. Let S be a set of all possible sensitive values of two datasets D_1 and D_2. Let $f_{D_1}(s)$ be the frequency of sensitive values in dataset D_1 and $f_{D_2}(s)$ for D_2. Note when a sensitive value, say s_j, is not present in a dataset, for example D_1, $f_{D_1}(s_j) = 0$.

The probability for inferring $s_i \in C$ from both datasets based on their sensitive values distribution is given as the following. Note that $\alpha = \frac{|D_1|}{|D_2|}$ and $|D_1|$ and $|D_2|$ are sizes of datasets D_1 and D_2 respectively. Let us assume that $|D_1| \leq |D_2|$ (we can swap the datasets if otherwise) and $0 < \alpha \leq 1$.

$$\begin{aligned} \text{prob}(s_i|(D_1, D_2)) &= \text{prob}(s_i|O(D_1, D_2)) \\ &= \frac{\min(f_{D_1}(s_i) \cdot |D_1|, f_{D_2}(s_i) \cdot |D_2|)}{\sum_{r=1}^{|S|} \min(f_{D_1}(s_r) \cdot |D_1|, f_{D_2}(s_r) \cdot |D_2|)} \\ &= \frac{\min(\alpha f_{D_1}(s_i), f_{D_2}(s_i))}{\sum_{r=1}^{|S|} \min(\alpha f_{D_1}(s_r), f_{D_2}(s_r))} \end{aligned} \quad (3)$$

The probability for inferring an overlapping sensitive value from two overlapping datasets relies on **1)** the number of overlapping sensitive values, **2)** their frequencies, especially in the smaller dataset of two datasets. In other words, it depends on what have been remained in the intersection. Next, we show how to use equation (3) in the following example.

Example 2. Let D_1 be Table 1(c) and D_2 be Table 2(c). $S = \{A, C, D, G, J\}$. We infer Nancy's sensitive value knowing her record is in both datasets. The sensitive values in $O = \{D(3), G(2), J(2)\}$ where the numbers in parentheses are counts. $\text{prob}(D \mid O(D_1, D_2)) = 3/(3+2+2) = 3/7$.

If we use the distributional information to infer the probability, we use the following table.

frequency	A	C	D	G	J
$\lvert D_1\rvert = 9$	0	2/9	1/3	2/9	2/9
$\lvert D_2\rvert = 12$	1/4	0	1/4	1/4	1/4

We obtain $\text{prob}(s = D|(D_1, D_2))$:

$$= \frac{(3/4) * (1/3)}{(3/4) * (1/3) + (3/4) * (2/9) + (3/4) * (2/9)} = 3/7$$

Now we present the objective of privacy preservation in multiple independent publications in the following.

$$\forall i : \text{prob}(s_i|(D_1, D_2)) - \max(\text{prob}(s_i|D_1), \text{prob}(s_i|D_2)) < \theta \quad (4)$$

$\text{prob}(s_i|D_1)$ and $\text{prob}(s_i|D_2)$ are the confidences of the adversary inferring the sensitive value s_i individually from two published datasets D_1 and D_2 respectively. θ is a small positive number. Let us recall what an adversary knows. Based on the background knowledge and the quasi-identifier values, an adversary knows that the victim's record is in both datasets D_1 and D_2. $\max(\text{prob}(s_i \mid D_1), \text{prob}(s_i \mid D_2))$ means the maximum probability that the adversary would know the sensitive value from the two published datasets individually. The objective (4) is to bound the improvement of the confidence of the adversary (caused by the composition attack) by θ. If $\theta = 0$, then the adversary gets no advantage by using two datasets collectively. $\text{prob}(s_i|D_1)$ and $\text{prob}(s_i|D_2)$ can be estimated as $f_{D_1}(s_i)$ and $f_{D_2}(s_i)$ if two datasets have been anonymized properly.

Now, we discuss when objective (4) can be achieved. We firstly study how difference in sensitive value distributions in D_1 and D_2 affect the inference probability. Let us fix $\alpha = 1$.

$$\text{prob}(s_i|(D_1, D_2)) = \frac{\min(f_{D_1}(s_i), f_{D_2}(s_i))}{\sum_{r=1}^{|S|} \min(f_{D_1}(s_r), f_{D_2}(s_r))} \quad (5)$$

Let $\beta = \frac{1}{\sum_{r=1}^{|S|} \min(f_{D_1}(s_r), f_{D_2}(s_r))}$. When distributions of sensitive values in both D_1 and D_2 are similar, $\beta \approx 1$. When the distributions are significantly different, β is a large number.

Example 3. Let $S = \{s_1, s_2, s_3, s_4, s_5\}$, datasets D_1 and D_2 contain some of the sensitive values of S. We give some distributions of sensitive values in D_1 and D_2 and their corresponding β values as the following.

frequency	s_1	s_2	s_3	s_4	s_5
D_1	0.1	0.2	0.3	0.4	0
D_2	0.1	0.2	0.3	0.4	0

$\beta = 1.$

frequency	s_1	s_2	s_3	s_4	s_5
D_1	0.1	0.2	0.3	0.4	0
D_2	0	0.2	0.3	0.4	0.1

$\beta = 1.11$

frequency	s_1	s_2	s_3	s_4	s_5
D_1	0.1	0.2	0.3	0.4	0
D_2	0	0.1	0.2	0.3	0.4

$\beta = 1.67$

In an ideal situation when $\beta = 1$, the overlapping dataset does not reveal more privacy of an overlapping individual than each individual dataset does. Therefore, the privacy of overlapping individuals in multiple datasets is preserved.

In the worst situation, when β is a large number, distributions of two datasets are almost distinct and have very little overlap. In such situation, $\text{prob}(s_i|(D_1, D_2)) \approx 1$ and an adversary can obtain the probability of an overlapping sensitive value with the nearly certain confidence using just the distribution information.

From the above examples, we see that distributions dominate the inference probability in the composition attack. Thus, we have the following observation.

Observation 1 *There is not a general solution for the composition attack for any distribution of two datasets.*

PROOF. Given a dataset D_1 with a distribution. We can always find a distribution of another dataset D_2 so that β is a large number. As a result, $\text{prob}(s_i|(D_1, D_2)) \approx 1$, and the objective (4) will be dissatisfied for some i's. □

For example, in the worst case, the privacy will be revealed just based on distributional information of two datasets, such as two datasets have only a single sensitive value in common. Such a disclosure does not even need QI information.

In the following, we consider a solution in a restricted form. We firstly define the *closed community*. In a closed community, the dataset of each data owner (i.e hospital in our case) is drawn from the same population and their data distributions are similar. In our words, $\beta \approx 1$. For example, different hospitals take different sets of patients, but in most cases, the disease distributions of the hospitals are similar. There are some exceptions that the distributions of some specialized hospitals are different from those of general hospitals. However, in such exceptions, the knowledge of a patient visiting a specialized hospital gives an adversary large chance for guessing the disease anyway. Let us assume that a number of general hospitals plan to publish their datasets independently in anonymized form (for research purposes etc.) and want to ensure privacy against the composition attack. So, a solution in the assumption of the closed community, where hospitals have similar distributions, is still useful.

In the following discussions, we assume that $\beta \approx 1$ for a solution in a closed community. When $\beta \approx 1$, α does not make much difference because both distributions are similar. An ideal solution can be achieved if sensitive values in every equivalence class of P^* and Q_i^* has the same distribution as of P and Q_i respectively. To formalize such best solution scenario, we introduce the following concept.

Definition 3 (Clone equivalence class) *Given the distribution of sensitive values of dataset P as $f(S)$. An equivalence class $E \subset P^*$ is called a clone equivalence class if it has all the sensitive values S and the distribution of sensitive values in E is the same as $f(S)$.*

Example 4. A clone equivalence class of Table 2(a) is Table 2(c). Apparently, the utility of this clone equivalence class is low. Table 2(d) has two clone equivalence classes generalized by our *micro-statistics method* (to be discussed later).

Observation 2 (Property of clone equivalence classes) *Let all equivalence classes of anonymized datasets P^* and Q_i^* be clone equivalence classes. The objective (4) can be achieved when $\theta \geq \max\big(\beta - \min(\text{prob}(s_j \mid P^*), \text{prob}(s_j \mid Q_i^*)) - \max(\text{prob}(s_j \mid P^*), \text{prob}(s_j \mid Q_i^*))\big)$ for all $j \in$ 1,2,3, …, $|S|$.*

PROOF. If all the equivalence classes of anonymized datasets P^* and Q_i^* are clone equivalence classes (Definition 3), the probability for inferencing a sensitive value in an equivalence class is the same as that for inferencing the sensitive value in whole dataset. Therefore, we can replace D_1 and D_2 with P^* and Q_i^* in equations (3) and (4) and the lower bound of θ is obtained. □

Since $\max(\text{prob}(s_i|P^*), \text{prob}(s_i|Q_i^*)) \geq \min(\text{prob}(s_i|P^*), \text{prob}(s_i|Q_i^*))$ and $\beta \approx 1$, the θ can be very small.

Clone equivalence classes are ideal but inappropriate for real world situations. We may lose a lot of utility if we require datasets to be anonymized to such an ideal situation (as we see in Tables 1(c) and 2(c)). We need a practical goal. We relax the requirement of clone equivalence classes as the following ϵ-clone equivalence classes.

Definition 4 (ϵ-clone equivalence class) *Given the distribution of sensitive values of dataset P as $f(S)$. An equivalence class $E \subset P^*$ is called ϵ-clone if it has all the sensitive values S and for every sensitive value in S, the difference of probabilities of the sensitive value in P and in E is bound by a small number ϵ. More specifically, $|\text{prob}(s_i|P) - \text{prob}(s_i|E)| \leq \epsilon$.*

The magnitude of ϵ is determined by the data distribution of a dataset. We will analyze the upper bound of the ϵ in Section 6.2 after the algorithm is presented. Finally, we present the problem of creating ϵ-clone equivalence classes.

Definition 5 (ϵ-clone publication) *The objective of privacy preserving ϵ-clone publication is to compute such anonymous P^* that each equivalence class E in P^* is ϵ-clone equivalence class and retains as much information in anonymous P^* as possible.*

The requirement of ϵ-clone equivalence class looks like t-closeness [19]. However, they are different in the following way. The t-closeness bounds the total difference of the distribution between an equivalence class and the dataset, but ϵ-cloning bounds the frequency difference of each sensitive value of an equivalence class and the dataset. When the overall difference is bounded in two distributions (like in t-closeness), the frequency differences of a sensitive value between dataset and an equivalence class can still be large. Therefore, t-closeness [19] is also vulnerable to the composition attack as shown in [11].

6. MICRO-STATISTICS BASED ANONYMIZATION

There is a practical problem for anonymizing a table to comply with ϵ-clone publication. The size of an ϵ-clone equivalence class can be big when the distribution of sensitive values is skewed. The generalization of such a large equivalence class makes a dataset useless, as we saw in Table 1(c) and 2(c) where most QI values have been suppressed. In this section, we will introduce a *micro-statistic* technique to make such a super equivalence class more useful than the generalization.

6.1 Phases

We aim at achieving two intuitive goals. First, each equivalence class has all the sensitive values of the anonymized dataset. Second, we attempt to reduce, as much as possible, the frequency difference for each sensitive value between original dataset and the equivalence class i.e. ϵ. We do not pre-specify ϵ but adaptively find it out in the process of anonymization. The detail discussion on ϵ is given in Section 6.2.

6.1.1 Division

This phase divides all the tuples of P into $|S|$ ($|S|$ is the number of distinct sensitive values in P) sub-datasets i.e. W_1, W_2, …, $W_{|S|}$, such that each sub-dataset contains only the tuples with same sensitive value. In the end of this phase, we calculate the frequency of each sub-dataset $f(W_j)$ in original dataset P. This phase is different from the *bucket partition* phase of [5] because [5] starts out with buckets of more than one *closely related* distinct sensitive values in one bucket but we strictly allocate only one sensitive value to each sub-dataset.

As an example, we consider Table 3(a) as P, $u = 4$ because P (Table 3(a)) has 4 distinct sensitive values i.e, {J,B,G,L}. We create 4 sub-datasets i.e. $W_1(J)$, $W_2(B)$, $W_3(G)$, $W_4(L)$. We calculate the ratio of each sub-dataset to the whole dataset; for example the sensitive value J has 5 tuples and its ratio in Table 2(a) is 0.26.

6.1.2 Assignment

First, we create b empty equivalence classes (E_1, E_2, E_3 …, E_b); where $b = \min(|W_j|)$, i.e. the sub-dataset with the minimum tuples. We call such sub-dataset as the *key sub-dataset*, denoted as W^{κ}. Next, we decide how many tuples to be assigned to each empty equivalence class E_r ($r = 1, 2, 3, \ldots, b$).

The assignment algorithm takes the $|S|$ sub-datasets, already created in the division phase, and calculates a separate γ_j value for each sub-dataset; γ_j is the number of tuples to be assigned to each equivalence class E_r from each sub-dataset W_j. To calculate the γ_j, we first make a temporary calculation of $\eta_j = \frac{|W_j|}{b}$. Where b is the number of equivalence classes created earlier in this phase i.e. $b = |W^{\kappa}|$. Now we calculate γ_j for j-th sub-dataset as:

$$\gamma_j = round(\eta_j) \tag{6}$$

Where $round$ is a function that returns the nearest integer from η_j. For example, $round(2.3) = 2$, $round(2.8) = 3$ and $round(2) = 2$. In an ideal situation, the $\gamma_j = \eta_j$. Such ideal situation exists for all sub-datasets of Table 2(a). If η_j is not an integer then there is a *deficiency* of some tuples in W_j; there are not enough tuples in W_j that can equally be allocated to all b equivalence classes. In such case, we need to handle this *deficiency* by suppressing/adding $\hbar_j$ tuples from/to sub-dataset W_j, where:

$$\hbar_j = \begin{cases} (\gamma_j * b) - |W_j| & \text{if } (\eta_j - \lfloor \eta_j \rfloor) > 0.5 \\ |W_j| - (\gamma_j * b) & \text{if } (\eta_j - \lfloor \eta_j \rfloor) < 0.5 \end{cases} \tag{7}$$

We *add* $\hbar_j$ counterfeited tuples to sub-dataset W_j if deficiency ($\eta_j - \lfloor \eta_j \rfloor$) > 0.5 because in such case suppression will cause more utility lost. The counterfeit tuples only has sensitive value s_i in sub-dataset W_j and all QIs have '$*$' values. In other case, we *suppress* the $\hbar$ tuples from sub-dataset W_j if deficiency ($\eta_j - \lfloor \eta_j \rfloor$) < 0.5 because in such case counterfeit tuples will cause more utility lost. If ($\eta_j - \lfloor \eta_j \rfloor$) $= 0.50$ then suppress and counterfeit equal number of sub-datasets.

Once γ_j information is ready for all sub-datasets, the assignment algorithm selects a sub-dataset W_j and calculates the *Distortion* (a distance metric) [18] of first ρ ($\gamma_j \leq \rho \leq |W_j|$) tuples with equivalence class E_r. Next, out of ρ tuples it assigns γ_j *closest* (defined in next paragraph) tuples to equivalence class E_r. This process is repeated for all equivalence classes. The ρ is a parameter, called *distortion control*, to improve the performance of the assignment algorithm and it restricts the calculation of *Distortion* to the first ρ tuples of the sub-dataset, instead of calculating distortion for all tuples of the sub-dataset.

Now we explain the *closest* tuples in detail. Alongside complying with the ϵ-clone publication objective (Definition 5), we also need to make sure that anonymous dataset P^* maintains a reasonable utility and it has been done through putting *closest* tuples of each sub-dataset W_j into an equivalence class E_r. We use *Distortion* [18] to measure the closeness between a tuple t and equivalence class E_r. The distortion is the sum of the generalization steps in the generalization hierarchies if t is assigned to E_r. Still, the problem of forming equivalence classes with the minimum distortions is similar to other optimal k-anonymity problems [1, 23]. So, we *randomly* pick ρ tuples from each sub-dataset W_j and calculate the *Distortion* for each of ρ tuples with an equivalence class E_r. We assign that tuple t out of ρ tuples to an equivalence class E_r that has the minimum distortion with it. Due to the space constraint, we omit the calculation details of distortion and readers are referred to the original paper [18] for further details.

In the running example where we assume Table 3(a) as P, the b $= |W_3(G)|$ i.e. 3 and we create 3 empty equivalence classes. We have already created 4 sub-datasets in division phase. For all four sub-datasets, the $\gamma_1(J) = 2$, $\gamma_2(B) = 2$, $\gamma_3(G) = 1$, $\gamma_4(L) = 1$ and deficiency values ($\eta_j - \lfloor \eta_j \rfloor$) are 0.66, 0.33, 0, 0.33 respectively. We need to add 1 counterfeit tuple in $W_1(J)$ and need to suppress 1 tuple from sub-datasets $W_2(B)$ and $W_4(L)$. In the assignment phase, we do not suppress any tuple and once the $(\gamma_j * b)$ closest tuples are assigned to b equivalence classes; we suppress the remaining ($|W_j| - (\gamma_j * b)$) tuples in the sub-dataset W_j.

Assignment operation accesses each sub-dataset and assigns 'closest' γ_j tuples of each sub-dataset W_j to each of three equivalence classes.

6.1.3 Statistics.

In the beginning of this phase, we suppress any remaining tuples in all sub-datasets and divide/merge or leave intact each equiva-

Algorithm 1 *Micro-Statistic based Anonymization*

Input:
P ▷ Input dataset to be anonymized
k ▷ Input parameter for k-anonymity
ρ ▷ Input parameter to be utilized in *Assignment* phase
λ ▷ Input parameter to be utilized in *Statistics* phase

Output:
P^* ▷ Anonymized dataset

1: Create w ($w = \frac{P}{|S|}$) empty sub-datasets $W[\,]$ ▷ *Division* phase
2: Populate sub-datasets $W[\,]$, each sub-dataset has tuples with same sensitive value in P ▷ *Division* phase
3: $b = \min(|W[j]|)(j \in 1, 2, 3, \ldots, |W[\,]|)$ ▷ Get number of tuples in *key sub-dataset*
4: Create b empty equivalence classes $E[\,]$ ▷ *Assignment* phase
5: $\gamma[j] = round(\frac{|W[j]|}{b})$ ▷ get γ for each sub-dataset
6: **for** $j \leftarrow 1$, $|W|$ **do** ▷ Access all sub-datasets one-by-one
7: **for** $r \leftarrow 1$, $|E|$ **do** ▷ Access all equivalence classes
8: Calculate *Distortion Distance* for first ρ tuples of $W[j]$ with $E[r]$
9: $E[r] = \gamma[j]$ of ρ in $W[j]$ ▷ out of ρ tuples in $W[j]$, assign γ tuples with minimum *distortion* to $E[r]$
10: **end for**
11: **end for**
12: Merge/divide equivalence classes in $E[\,]$ as per k
13: Generate 'micro-statistics' for each equivalence class in $E[\,]$ using input parameter λ ▷ *Statistics* phase
14: Combine all $E[\,]$ to form P^* ▷ Anonymized dataset created

lence class as per the input parameter k from k-anonymity [25]. We divide the QIs of any equivalence class into sub-equivalence classes if $|E_r| > k$ or merge an equivalence class with some another one if $|E_r| < k$. We only divide the QIs into sub-equivalence classes and leave the sensitive values intact because we have to maintain the association of each tuple of sub-equivalence class with all the sensitive values of the equivalence class. We merge those equivalence classes that have minimum distortion with each other.

Next, we calculate the *micro-statistics* for all QIs within each sub/equivalence class. We have two types of QIs i.e. categorical and numerical. For categorical QI, we include the number of distinct QI values in each sub-equivalence class. For numerical QI, first we calculate the 'average' of QI values within sub/equivalence class and next select an indicative range that covers λ QI values of the same sub/equivalence class; where $\lambda \leq k$ and is an input parameter. Algorithm 1 formally presents the anonymization operation.

In our running example, we have two remaining tuples t_{12} = <45, M, Separated> in $W_2(B)$ and t_{18} = <22, M, Separated> in $W_4(L)$; we suppress these tuples first. Now, we consider k = 3 and $\lambda = 1$ and divide each equivalence class into two sub-equivalence classes, each of 3 tuples. As last step, we generate the micro-statistics for all the sub-equivalence classes. We have 3 QIs, the Age is numerical whereas Sex and $Marital\ Status$ are categorical. We replace the Age QI values in all sub-equivalence classes with the average Age of each sub-equivalence class and also include an indicative range that covers 1 ($\lambda = 1$) Age value nearest to the average value in each sub-equivalence class. For Sex and $Marital\ Status$ QIs, we put the number of all distinct QI values in all sub-equivalence classes. The result is Table 3(c).

The Table 2(a) is case of equivalence class merger. Ideally, there should be 3 equivalence classes (having 4 tuples each) but we assume $k = 6$ for Table 2(a); so one equivalence class is split and merged into others two.

6.2 Discussion

The ϵ-clone equivalence classes formed by our algorithm makes use of *micro-statistics* instead of *local recoding* [17] and *global recoding* [25] generalization schemes often adopted in literature. We employ the micro-statistics scheme because it provides high degree of privacy and utility in anonymous dataset (shown in experiments of Section 7). The concept of ϵ-cloning can also be implemented by local/global-recoding generalization. Although it will result in same privacy guarantee but with lower utility (shown in Tables 1(c), 2(c)).

Now, we discuss the effectiveness of our scheme (micro-statistics based anonymization) in achieving the ϵ-clone publication. The major concern of the ϵ-clone publication is the magnitude of ϵ. We aim at the maximum possible value of ϵ in our scheme. A large ϵ means high distribution difference between the original and the anonymized dataset.

We now compute the maximum value of ϵ, for all distinct sensitive values $s_1, s_2, \ldots, s_u$ of a dataset P. As dataset P has u distinct sensitive values so there will be u sensitive sub-datasets as $W_1, W_2, \ldots, W_u$. We also assume the ratio of these u sub-datasets as: $f(W_1) \leq f(W_2) \leq \ldots \leq f(W_u)$. As the ratio $f(W_1)$ of sensitive value s_1 is the smallest, the sub-dataset W_1 will be the 'key sub-dataset' and the number of initial empty equivalence classes $b = |W_1|$ (i.e. the number of tuples in key sub-dataset). In an ideal situation, if $\hbar_j = 0$ ($j = 2, 3, \ldots, n$) for all n sub-datasets (it occurs when $\gamma_j = \eta_j$ in (6)) then $\epsilon_j = 0$ for all sub-datasets. In other case, where $\hbar_j > 0$ (i.e. $\gamma_j \neq \eta_j$), we need to *distort* the sub-dataset(s) by suppressing and/or adding counterfeiting $\hbar_j$ tuples. In such case, we need to find the upper bound of the ϵ_j. We denote such maximum value of ϵ_j as $\epsilon_j^{\max}$.

Table 4: Attribute domain size

Attribute	Age	Sex	Education	Birth Place	Occupation	Income
Domain Size	79	2	17	57	50	50

We consider the worst case and assume that, excluding the key sub-dataset W_1, for all other sub-datasets $\gamma_j \neq \eta_j$ and $(\eta_j - \lfloor \eta_j \rfloor)$ = 0.50. Now we need to compute the $\hbar_j$ suppress and counterfeit tuples for equal number of sub-datasets. The values of $\hbar_j$ are calculated as per (7) and in any case the maximum number of distorted (suppressed and/or counterfeit) tuples, denoted as $\hbar_j^{\max}$, are:

$$\hbar_j^{\max} \leq |W^{\kappa}|/2 \tag{8}$$

The W^{κ} is the key sub-dataset (W_1 in our case). So as per (8), the maximum number of distorted (suppress and/or counterfeit) tuples in any sub-dataset are not more than the half number of the tuples in the key sub-dataset. Now if the property (8) holds for all sub-datasets then following property will also hold for any sub-dataset W_j:

$$\epsilon_j^{\max} = f(W^{\kappa})/2 \tag{9}$$

Now for any dataset P having at least two distinct sensitive values, the maximum frequency of key sub-dataset can be $f(W^{\kappa}) = 0.40$ to get worst case scenario when $(\eta - \lfloor \eta \rfloor) = 0.50$ for non-key sub-dataset i.e. we have to suppress/counterfeit $(0.40/2 * |P|)$ tuples from/to non-key sub-dataset. Even in such extreme case, the $\epsilon_j^{\max} = 0.10$ for suppression and $\epsilon_j^{\max} = 0.07$ for counterfeiting. If we have 3 sensitive values the $\epsilon_j^{\max} = 0.125$. In real life datasets, we have moderate number of distinct sensitive values; causing even lower $\epsilon_j^{\max}$. For Table 3(c) the $\epsilon_j^{\max}$ is 0.07.

ϵ-cloning is also not vulnerable to other privacy threats like minimality attack [27] and deFinett attack [15]. The minimality attack [27] exploits the *minimum generalization* required for anonymization and it cannot be successful on our scheme due to incorporation of *micro-statistics* instead of (local or global) generalization. Adversary cannot get the exact QIs range for any equivalence class from its micro-statistics and adversary's confidence remains split about the inclusion of a tuple in an equivalence class or not. The deFinett attack [15] can be successful when the frequency difference of a sensitive value between original and anonymous dataset is 'large'; which is not in the case of ϵ-clone equivalence classes because each ϵ-clone equivalence class has all the sensitive values of the original dataset with a small ϵ threshold. So, adversary gets very little information from ϵ-cloned dataset after deFinett attack [15].

7. EXPERIMENTS

Experiments are performed on a machine running a 2.4 Ghz CPU with 3 Giga-byte memory. We deploy two real world datasets OCC and SAL from United States census data downloadable from http://ipums.org. Both contain 600k tuples and each tuple contains the information of an American adult. OCC includes four QI attributes, age, sex, $education$, $birth\ place$, and a sensitive attribute $occupation$. SAL contains the same set of QI attributes, but different sensitive attribute $income$. All columns, except age, are discrete and the sizes of their domains are given in Table 4.

We create five disjoint sub-datasets Q_i^{occ} (Q_i^{sal}) ($i \in 1, \ldots, 5$) from $OCC(SAL)$. It suffices to clarify the generation and generalization of Q_i^{occ}, since the same method is applied for Q_i^{sal}. We

assign 100k tuples to each sub-dataset Q_i^{occ}. The remaining 100k tuples initiates an overlap pool. All sub-datasets and overlapping pool have same set of sensitive values.

In the publication of 5 sub-datasets, we randomly select specific *overlapping tuples* from overlap pool and insert them in each Q_i^{occ}, subsequently each anonymous Q_i^{occ*} (having 100k+ overlapping tuples) is created that satisfies ϵ-cloning. Here overlapping tuples control the overlap rate among 5 sub-datasets. We repeat this process by increasing the overlapping tuples as 20k, 40k, 60k, 80k, 100k, i.e. our experiments include 5 groups of sub-datasets each with size of 120k, 140k, 160k, 180k, 200k. In each group, we the same overlapping rate. We refer to these sub-datasets Q_i^{occ} (Q_i^{sal}) as independent publishers. Broadly, we have five separate publishers for OCC and SAL.

7.1 Failure of Conventional Generalization

In the first set of experiments, we show that the existing generalization methods lead to severe privacy disclosure in independent data publications. The findings were also observed previously in [11, 3]. We increase the overlapping tuples from 20k to 100k and adopt the algorithm in [16] to compute ℓ-diverse [20] versions Q_1^{occ*}, Q_2^{occ*}, ..., Q_n^{occ*}. Then, we identify all the overlapping tuples that appear in any Q_i^{occ*} and whose sensitive values will definitely be revealed (called *privacy risk* tuples) using any another sub-dataset. These privacy risk tuples are extracted using Definition 2; where we use Q_1^{occ*} as P^* and all other 4 sub-datasets are treated as Q_i^*. In Fig. 2(a), we plot the maximum number of privacy risk tuples in any sub-dataset as a function of ℓ. We repeat this process by increasing ℓ from 2 to 10. The ℓ-diversity [20] fails to support independent publications of overlapping data, because it results in a large number of privacy risk tuples. Although fewer privacy risk tuples exist as ℓ increases, privacy risk tuples still cannot be completely prevented with larger ℓ.

As all sub-datasets Q_i^{x*} ($x = occ$ or sal) and overlap pool have the same set of sensitive values (Table 4) so ϵ-cloning preserves the privacy because in ϵ-cloning the overlapping set (Definition 2) always has all the sensitive values of the whole dataset and privacy compromise scenario ((2) in Section 4) does not occur. We repeat these experiments on sub-datasets Q_i^{sal} and the results are illustrated in Fig. 2(b), confirming the same observations.

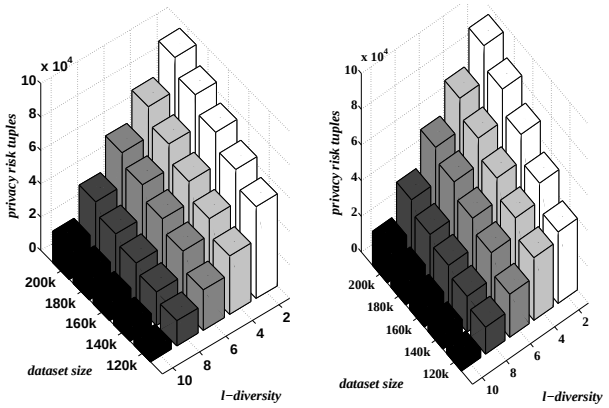
7.2 ϵ-Cloning Evaluation

In following experiments, we examine the effectiveness of ϵ-cloning. We invoke the micro-statistics based anonymization on $Q_1^x, Q_2^x, \ldots, Q_5^x$ ($x = occ$ or sal) to compute generalized versions $Q_1^{x*}, Q_2^{x*}, \ldots, Q_5^{x*}$. Each Q_i^{x*} is characterized by overlapping tuples and distortion control parameter ρ (ρ from Section 6.1.2). We use 20% of dataset size as ρ, unless otherwise mentioned.

We cannot benchmark our experiments with existing partition-based generalization schemes i.e. k-anonymity [25], ℓ-diversity [20], t-closeness [19] etc. because all of these schemes are prone to composition attack [11, 3].

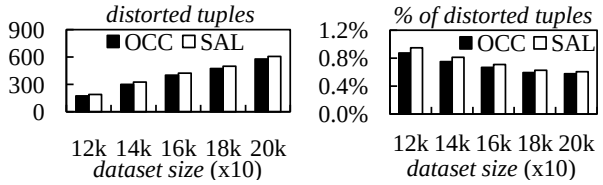
7.2.1 Number of Distorted Tuples

Our algorithm needs to suppress and/or counterfeit some tuples in the Assignment phase. We demonstrate that only a small number of distorted tuples (both suppressed and counterfeited) are needed to enforce ϵ-cloning. In Fig. 3(a), we vary the dataset size from 120k to 200k and measure the number of distorted tuples in OCC (SAL). The number of distorted tuples increases along with dataset size because higher dataset size requires more tuples in each equivalence class for assignment; resulting possibility of more distortion. The maximum number of distorted tuples are only 579(607) for dataset size of 200k in OCC (SAL).



(a) OCC (b) SAL

Figure 2: Successful composition attack with increasing overlapping tuples and ℓ-diversity)



(a) Distorted tuples vs. dataset size

(b) Percentage of distortion vs. dataset size

Figure 3: Average and percentage of distorted (suppressed and counterfeited) tuples in OCC and SAL with increasing dataset size

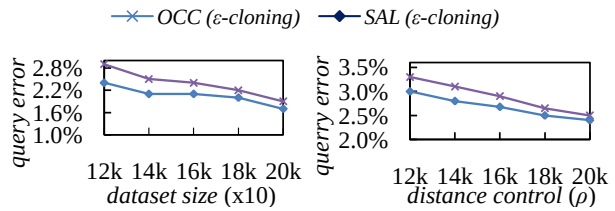
In Fig. 3(b), we show the percentage of distorted tuples in OCC (SAL) versus the dataset size. The percentage decreases as dataset size increases. As a result, that micro-statistics based anonymization can utilize more tuples. As dataset size increases, the total number of distorted tuples also increases (Fig 3(a)) but overall the percentage of distorted tuples with dataset size decreases (Fig 3(b)). In any case the percentage of distorted tuples is less than 1% of dataset size.

7.2.2 Utility of the Published Data.

In the following set of experiments, we will use Q_i^{x*} ($x = occ$ or sal) to answer queries about the original sub-dataset Q_i^x. We use aggregate queries, since they are the basic operation for numerous data mining tasks (e.g., decision tree learning, association rule mining, etc.). Specifically, each query has the form:

SELECT COUNT (*) FROM Q_i^{x*} WHERE $pred(t[A_1^{qi}]$ AND ... AND $t[A_4^{qi}]$ AND $t[S])$

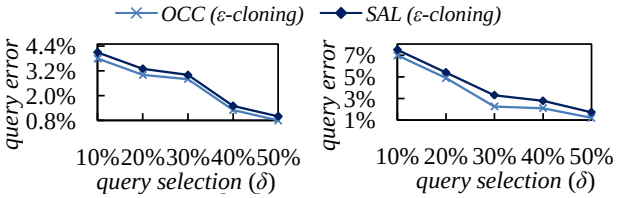
The Q_i^{x*} is the sub-dataset generalized using ϵ-cloning, ($t[A_1^{qi}]$, ..., $t[A_4^{qi}]$) denote the four QI attributes, and $t[S]$ is the sensitive attribute $occupation$ ($income$). For each attribute A, the condition $pred(A)$ has the form $A \in \delta$. Here δ is a query parameter called *selection range*, and has length $|A|.\delta$, where $|A|$ is the domain size



(a) Median relative error vs. dataset size, ($\delta = 25\%$)

(b) Median relative error vs. ρ (dataset size = 120k, $\delta = 25\%$)

Figure 4: Query error with fixed query selection range (δ) and increasing dataset size and increasing distance control parameter (ρ)



(a) Workload error vs. δ (dataset size = 120k)

(b) Workload error vs. δ (dataset size = 200k)

Figure 5: Average workload error with increasing query selection range (δ) and increasing dataset size

of attribute A (see Table 4). A larger result is returned with higher δ. Our workload consists of 10000 queries with same sub-dataset Q_i^{x*} and sensitive value $t[S]$.

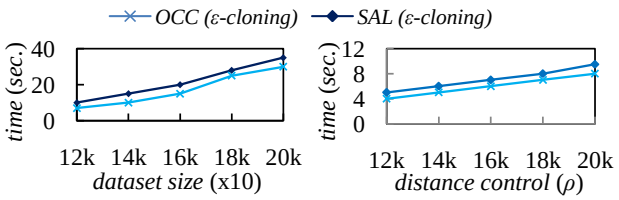
Given a query, we obtain its actual result R_{act} from the original sub-dataset Q_i^x, and compute an estimated answer R_{est} from its ϵ-clone generalized version Q_i^{x*}. The relative error of a query equals $|R_{act} - R_{est}|/R_{act}$. We measure the workload error as the median relative error of all the queries of all sub-datasets.

Fig. 4(a), plots the workload error as a function of dataset size for OCC and SAL respectively. In all experiments, the median error is at most 2.5%. In the experiments of Fig. 4(b), we set dataset size to 120k and measure the workload error as the function of distortion control parameter ρ. The error decreases (accuracy improves) with ρ because higher ρ results the more search in finding the closest tuple during the assignment phase. Our experiments show that error does not vary significantly with dataset size, and is not very sensitive to the distortion control parameter ρ.

In the experiments of Fig. 5, we set dataset size to 120k and 200k (i.e. minimum and maximum values of dataset in Fig. 3) and study workload error in Q_i^{x*} ($x = occ$ or sal) as the function of query selection range δ. The accuracy improves i.e. workload error decreases with increase in δ. This is expected because higher δ leads to larger query results, whereas, in general, aggregate analysis is effective for sizable queries.

7.2.3 Computation Overhead.

The last experiment evaluates the efficiency of our micro-statistics based generalization algorithm. In Fig. 6(a), we measure the average time of computing a generalized sub-dataset Q_i^{occ*} (Q_i^{sal*}) as function of dataset size. The cost is more expensive when dataset size is higher, because the algorithm needs to process more tuples in a dataset. In Fig. 6(b), we fix dataset size to 120k, and get the computation time as a function of distortion control parameter ρ. The overhead increases as ρ increases, since a larger ρ results the involvement of more tuples in finding the closest tuple during the assignment phase. In comparison Figure 6(b) and Figure 4(b), we see that ρ is a trade off for the efficiency and utility.



(a) Computation time vs. dataset size

(b) Computation time vs. ρ (dataset size = 120k)

Figure 6: Computation overhead with increasing dataset size and increasing distance control parameter (ρ)

8. RELATED WORK

It is important to point out that the partition-based schemes in the literature were not designed to be used in contexts where independent releases of overlapping population are available from different locations. Thus, essentially the composition attack is not a flaw in these schemes, but rather it directs the community's attention to an important direction of research.

There has already been made substantial progress in privacy preserving data publishing. One line has focused on taking into account other, known releases, such as previous publications by the same organization (called sequential, serial or incremental releases) [29, 31, 10] and multiple views of the same dataset [33, 34]. Another line has considered incorporating knowledge from partitioned views to group individuals [33]. In our case, each publisher is independent and unaware from the dataset of other publisher(s).

Some other works have sought to model unknown background knowledge of adversary [22, 6]. Martin et al. [22] and Chen et al. [6] provide complexity measures for an adversary's side information (roughly, they measure the size of the smallest formula within a CNF-like class that can encode the side information). Both works design schemes that provably resist attacks based on side information whose complexity is below a given threshold. A hypothetical discussion of the same problem is in [11], driving concepts from differential privacy [7, 9]. Differential privacy guarantees that the attacker should not be able to distinguish between two possibilities, i.e. a specific person's record is in or not in a statistical database, thus preserving the privacy. Most relevant to this paper are works that elaborate the privacy risk due to the anonymous data release of overlapping population by multiple locations [21, 13, 12]. All of these incorporate the *co-ordinated model*; where all locations communicate with each other to calculate the privacy risk of overlapping population and subsequently release dataset that is *k-linkable* i.e. each overlapping record is minimum linked to *k* records in each release.

Independent and asynchronous release by multiple locations (and hence composition attacks) fall outside the models proposed by these works. The sequential release models do not fit because they assume the multiple synchronous releases from the single location. In this paper, we deal with the case when there are multiple independent publishers. The complexity-based measures do not fit because independent releases appear to have complexity that is linear in the size of the datasets. The differential privacy is ideally suitable for interactive-setting (where there is no public release of anonymous data) and solution purposed in [11] lacks the actual im-

plementation and test results. Moreover recent test results show that no differently private algorithm can have meaningful utility unless the privacy requirement is very low [8, 24]. The co-ordinated model also does not satisfy our requirement because we are dealing with a non co-ordinated scenario where each location independently anonymizes its data without having any communication with other location(s).

9. CONCLUSION

Existing data publishing and serial data publishing methods do not support multiple independent data publication by different publishers where there are overlapping individual records. This paper has developed ϵ-cloning model to prevent an adversary from using independent data releases of multiple data owners to infer sensitive information of overlapping individuals. We have provided an efficient algorithm for computing anonymized dataset to achieve ϵ-cloning. We experimentally showed that the anonymized data adequately protects privacy and yet supports effective data analysis.

10. ACKNOWLEDGEMENT

The authors would like to thank the anonymous reviewers for their constructive suggestions. This research has been supported by ARC Discovery grants DP0774450 and DP110103142.

11. REFERENCES

[1] AGGARWAL, G., FEDER, T., KENTHAPADI, K., MOTWANI, R., PANIGRAHY, R., THOMAS, D., AND ZHU, A. Anonymizing tables. In *ICDT'05* (2005), pp. 246–258.

[2] AGRAWAL, R., KIERNAN, J., SRIKANT, R., AND XU, Y. Hippocratic databases. VLDB '02, pp. 143–154.

[3] BAIG, M. M., AND JIUYONG LI, JIXUE LIU, H. W. Studying genotype-phenotype attack on k-anonymized medical and genomic data. *AusDM* (2009), 159–166.

[4] BRICKELL, J., AND SHMATIKOV, V. The cost of privacy: destruction of data-mining utility in anonymized data publishing. KDD '08, pp. 70–78.

[5] CAO, J., KARRAS, P., KALNIS, P., AND TAN, K.-L. Sabre: a sensitive attribute bucketization and redistribution framework for t-closeness. *JVLDB 20* (2011), 59–81.

[6] CHEN, B.-C., LEFEVRE, K., AND RAMAKRISHNAN, R. Privacy skyline: privacy with multidimensional adversarial knowledge. VLDB '07, pp. 770–781.

[7] DWORK, C. Differential privacy. In *ICALP* (2006), pp. 1–12.

[8] DWORK, C., MCSHERRY, F., NISSIM, K., AND SMITH, A. Calibrating noise to sensitivity in private data analysis. 3rd Theory of Cryptography Conference, Springer, pp. 265–284.

[9] FRIEDMAN, A., AND SCHUSTER, A. Data mining with differential privacy. KDD'10, ACM, pp. 493–502.

[10] FUNG, B. C. M., WANG, K., FU, A. W.-C., AND PEI, J. Anonymity for continuous data publishing. EDBT '08, pp. 264–275.

[11] GANTA, S. R., KASIVISWANATHAN, S. P., AND SMITH, A. Composition attacks and auxiliary information in data privacy. KDD'08, pp. 265–273.

[12] JIANG, W., AND CLIFTON, C. Privacy-preserving distributed k-anonymity. In *Data and Applications Security XIX* (2005), vol. 3654 of *LNCS*, Springer, pp. 924–924.

[13] JIANG, W., AND CLIFTON, C. A secure distributed framework for achieving k-anonymity. *JVLDB 15* (2006), 316–333.

[14] JIN, X., ZHANG, M., ZHANG, N., AND DAS, G. Versatile publishing for privacy preservation. KDD '10, pp. 353–362.

[15] KIFER, D. Attacks on privacy and definetti's theorem. SIGMOD'09, pp. 127–138.

[16] LEFEVRE, K., DEWITT, D. J., AND RAMAKRISHNAN, R. Mondrian multidimensional k-anonymity. ICDE'06, pp. 25–.

[17] LEFEVRE, K., DEWITT, D. J., AND RAMAKRISHNAN, R. Incognito: efficient full-domain k-anonymity. SIGMOD'05, pp. 49–60.

[18] LI, J., WONG, R. C.-W., FU, A. W.-C., AND PEI, J. Anonymization by local recoding in data with attribute hierarchical taxonomies. *TKDE 20* (2008), 1181–1194.

[19] LI, N., LI, T., AND VENKATASUBRAMANIAN, S. t-closeness: Privacy beyond k-anonymity and ℓ-diversity. ICDE'07, pp. 106–115.

[20] MACHANAVAJJHALA, A., KIFER, D., GEHRKE, J., AND VENKITASUBRAMANIAM, M. ℓ-diversity: Privacy beyond k-anonymity. *TKDD* (2007).

[21] MALIN, B. k-unlinkability: A privacy protection model for distributed data. *DKE 64*, 1 (2008), 294–311.

[22] MARTIN, D., KIFER, D., MACHANAVAJJHALA, A., GEHRKE, J., AND HALPERN, J. Worst-case background knowledge for privacy-preserving data publishing. ICDE'07, pp. 126 –135.

[23] MEYERSON, A., AND WILLIAMS, R. On the complexity of optimal k-anonymity. PODS '04, ACM, pp. 223–228.

[24] MURALIDHAR, K., AND SARATHY, R. Does differential privacy protect terry gross' privacy? In *Privacy in Statistical Databases*, vol. 6344 of *LNCS*. Springer, 2011, pp. 200–209.

[25] SWEENEY, L. k-anonymity: a model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.* (2002), 557–570.

[26] TAO, Y., XIAO, X., LI, J., AND ZHANG, D. On anti-corruption privacy preserving publication. ICDE'08, pp. 725 –734.

[27] WONG, R. C.-W., FU, A. W.-C., WANG, K., AND PEI, J. Minimality attack in privacy preserving data publishing. VLDB'07, pp. 543–554.

[28] WONG, R. C.-W., LI, J., FU, A. W.-C., AND WANG, K. (α, k)-anonymity: an enhanced k-anonymity model for privacy preserving data publishing. 754–759.

[29] WONG, R.-W., FU, A.-C., LIU, J., WANG, K., AND XU, Y. Global privacy guarantee in serial data publishing. ICDE'10, pp. 956–959.

[30] WONG, W. K., MAMOULIS, N., AND CHEUNG, D. W. L. Non-homogeneous generalization in privacy preserving data publishing. SIGMOD'10, pp. 747–758.

[31] XIAO, X., AND TAO, Y. m-invariance: towards privacy preserving re-publication of dynamic datasets. SIGMOD '07, pp. 689–700.

[32] XIAO, X., AND TAO, Y. Anatomy: simple and effective privacy preservation. VLDB'06, pp. 139–150.

[33] YANG, B., NAKAGAWA, H., SATO, I., AND SAKUMA, J. Collusion-resistant privacy-preserving data mining. KDD'10, pp. 483–492.

[34] YAO, C., WANG, X. S., AND JAJODIA, S. Checking for k-anonymity violation by views. VLDB '05, pp. 910–921.