Data Privacy Against Composition Attack

Muzammil M. Baig[1], Jiuyong Li[1], Jixue Liu[1], Xiaofeng Ding[1], and Hua Wang[2]

[1] School of Computer and Information Science**
University of South Australia, Mawson Lakes, SA. 5095 Australia.
{muzammil.baig, jiuyong.li, jixue.liu,
xiaofeng.ding}@unisa.edu.au
[2] Department of Maths & 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 holder. However, no method has been proposed for the anonymization scenario of multiple independent data publishing. A data holder publishes a dataset, which contains overlapping population with other datasets published by other independent data holders. No existing methods are able to protect privacy in such multiple independent data publishing. In this paper we propose a new generalization principle (ρ, α)-anonymization that effectively overcomes the privacy concerns for multiple independent data publishing. We also develop an effective algorithm to achieve the (ρ, α)-anonymization. We experimentally show that the proposed algorithm anonymizes data to satisfy the privacy requirement and preserves high quality data utility.

Keywords: Data anonymity, privacy, composition attack.

1 Introduction

Existing privacy preserving data publishing techniques focus on one-time publication [11, 8, 4] and multiple views of the same data [17]; recently address the scenario of re-publication by single data holder [14, 15]. Specifically, privacy preserving data re-publication is restricted to single data holder, and does not support overlapping population by multiple publishers. The seminal work [2] firstly identify the breach of privacy of existing anonymization methods in multiple independent data publishing, called 'composition attack'. However, the solution of [2] supports only *interactive setting* (where only data statistics and/or query results are published), and is inapplicable for *non-interactive* setting (where the data needs to be published after anonymization). Independent data publishing of overlapping subset by multiple publishers in non-interactive setting remains an open problem.

To illustrate the problem, consider Table 1(a) of Hospital-1. *Identifier attribute(s)* can directly identify individuals, such as Name, SSN etc. They should be removed in a published dataset. *Quasi identifier (QIDs) attributes* could indirectly lead to the identification of individuals in a dataset, such as Age, Zipcode and Sex etc. They are normally

** This research has been supported by ARC Discovery grants DP0774450 and DP110103142

Table 1: Patient data and its generalization at Hospital-1

(a) Original data Q_1

Identifier Attribute	***Quasi-Identifiers***		***Sensitive Attribute***
Bob	15	male	*B*
Hudson	45	male	*H*
Robi	40	female	*G*
David	20	male	*B*
Khan	25	male	*C*
Victor	50	male	*H*

(b) Generalized Q_1^*

Group ID	**Age**	**Sex**	**Disease**
1	15 – 25	male	*B*
			B
			C
2	40 – 50	*	*G*
			H
			H

generalized so that no individuals are identifiable in a generalized table. The *Sensitive attribute* contains the private information about the individuals that needs to be protected such as Disease, Income etc. A *generalized* table is considered privacy preserving, if it satisfies a privacy constraint, such as k-anonymity [11] or ℓ-diversity [8]. For example Table 1(b) is 3-anonymous and 2-diverse version of Table 1(a). In other words, 3-anonymity means that values in the QIDs have at least 3 identical copies. So one could not be distinguished from other 2 records. 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 guessed with a high confidence.

1.1 Problem Description and Motivation

Consider the patient overlapping scenario of three hospitals in figure of Table 2(c); David from Hospital-1 and Eliza from Hospital-2 were referred to Hospital-3 so the data of Hospital-3 also include the records of David and Eliza. For simplicity, we omit the overlapping scenario between Hospital-1 and Hospital-2; although our solution provides the privacy protection in any overlapping scenario.

Hospital-3 anonymized its dataset and release it as Table 3(b). Assume that an adversary knows David's QIDs (20 years old male), and the fact that David has visited both Hospital-1 and Hospital-3. The adversary would find the records of David in both hospitals since only one record matches David's QIDs and has the same disease in Table 1(b) and 3(b) respectively i.e. $\{B\}$. Therefore, David is identified in the anonymized datasets of both hospitals. The understanding remains the same for Eliza where adversary can get her disease $\{R\}$ using her QIDs in Table 2(b) and 3(b).

A patient may visit more than one hospitals of his/her area and we assume that hospitals visit information is available in public domain i.e. adversary knows about the hospitals visited by a patient. Moreover, each hospital also knows about other hospitals where it can have overlapping patients. Both are realistic assumptions. Firstly, an adversary is a person that is close to the patient (i.e. a friend, a colleague or a neighbor) and it is reasonable to believe that s/he is aware of the hospitals visited by the patient. Secondly, hospitals visit information is also part of a patient medical record so each hospital also knows about other hospitals where it can have overlapping patients. Although, each hospital knows about the overlapping with other hospitals but each hospital does not (due to internal privacy policies) or cannot (due to legal restrictions) share its original data with another organization.

Table 2: Patient data and its generalization at the Hospital-2

(a) Original data Q_2

Name	Age	Sex	Disease
Eliza	40	female	*R*
Arthur	30	male	*M*
Paul	20	male	*M*
Noreen	45	female	*S*
Mathew	15	male	*Q*
Panama	35	female	*T*

(b) Generalized Q_2^*

Age	Sex	Disease
15 – 30	male	*M*
		Q
		M
35 – 45	female	*R*
		S
		T

(c) All overlapping patients

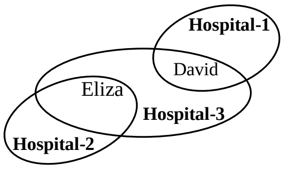


Table 3: Patient data and its generalization at the Hospital-3

(a) Original data P

Name	Age	Sex	Disease
David	20	male	*B*
Anthony	35	male	*C*
Rick	30	male	*C*
Stewart	30	male	*L*
George	28	male	*B*
Smith	38	male	*W*
Eliza	40	female	*R*

(b) Generalized P^*

Age	Sex	Disease
20 – 30	male	*B*
		C
		C
30 – 40	*	*L*
		B
		W
		R

(c) P^* with α-overlap

Age	Sex	Disease
15 – 35	male	*B*
		C
		C
		L
28 – 45 (1)	*	*B*
		W
		R
		S

The problem of overlapping data publication is not resolvable by the methods of sequential data publication, such as m-invariance [15]. m-invariance deals with two overlapping data publications of the same data holder by employing the same publication scheme. In multiple publication scenario datasets are more than two, released from different data holders, and mostly anonymized by different publication schemes. Our problem is different from sequential publication and more details are in Section 4.2.

In this paper, our proposed method (ρ, α)-anonymization (details later in Section 4) leads to the publication of Table 3(c) at Hospital-3. Now an adversary has at most 50% chance (in this simple example) to guess the sensitive value of any overlapping individual. Let us reconsider the adversary who has the precise QIDs detail of David and attempts to infer the disease of David from Tables 1(b) and 3(c). S/he can locate that the tuple of David must have been generalized in the first QID groups of Tables 1(b) and 3(c), respectively. These groups encompass the 2 common sensitive values i.e. $\{$*B*,*C*$\}$. Therefore adversary cannot get any specific disease that David has contracted. In case of Eliza, there are also two candidate diseases i.e. $\{$*R*,*S*$\}$. There is one ‘counterfeited’ tuple (shown in parentheses) in the QID group of Eliza because there was no $\{$*S*$\}$ disease in Table 3(a) (details later in Section 5).

1.2 Contributions

This paper presents the first model to prevent the composition attack in non-interactive data publishing setting by combining sampling and generalization. Our solution integrates two novel concepts: (ρ, α)-*anonymization* and *composition-based generalization*. The former is a new anonymization mechanism, which overcomes the drawbacks of generalization by combining it with sampling and provide privacy protection for composition attack. The latter is a technique that facilitates the enforcement of privacy, in the presence of overlapping population.

Secondly, we design an efficient algorithm to compute anonymous datasets that conforms to (ρ, α)-anonymization. Our algorithm aims to maximize the utility of the released data, by minimizing **(i)** the number of counterfeited tuples, and **(ii)** the amount of generalization on the QIDs. Furthermore, the algorithm is versatile, namely, it enables a data holder to produce an anonymized release, by consulting any number of already published anonymous releases of other data holders.

2 Fundamental Definitions

Let P be a dataset maintained by a data holder. There are n other published datasets $Q_1^*, Q_2^*, \ldots, Q_n^*$ which have overlapping population with P. Each published dataset Q_i^* ($i \in$ 1,2,3,$\ldots$,n) is independently anonymized from its original dataset Q_i.

We classify the columns of P and Q_i ($i \in$ 1,2,3,$\ldots$,n) into three types (already explained in Section 1): **(i)** an identifier attribute A^{id}, which is the primary key of P, **(ii)** d quasi-identifier (QIDs) attributes A_1^{qi}, A_2^{qi}, $\ldots$, A_d^{qi}, and **(iii)** a sensitive attribute A^s. The QIDs can be either numerical or categorical. For each tuple $t_p \in P$, $t_p[A]$ denotes its value on attribute A.

Definition 1 (Generalized QID group / Equivalence class). *For an anonymous dataset P^*, a generalized QID group is subset of the tuples in P. Each generalized QID group is assigned an unique ID A^g. All tuples in P^* with the same A^g have the identical values in QID attribute.*

For a tuple $t_p^* \in P^*$; the $t_p^*.QI$ denotes such generalized QID group which has t_p^* in P^*. We refer to $t_p^*.QI$ as the 'generalized QID hosting group' of the t_p^* in P^*. Next, we introduce an important notation OL.

Definition 2 (Overlapping Set). *For dataset P and each already published independent anonymous dataset $Q_i^*(i \in 1, 2, 3 \ldots, n)$, the overlapping set* (OL) *contains all those tuples in P such that:*

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

Each tuple $t \in \mathrm{OL}$ is an intersection (to be explained) of two corresponding tuples $t_i^* \in Q_i^*$ and $t_p \in P$; who satisfy the following properties:

1. *$t_i^*[A^s] = t_p[A^s]$; both tuples have same sensitive value and*
2. *$t_i^*[A_j^{qi}] \cap t_p[A_j^{qi}] \neq \emptyset, (1 \leq j \leq d)$; t_i^* and t_p have overlapping value interval in j-th QID attribute.*

Note that none of the data holder shares its original data with another data holder. Rather, before anonymizing its original data, the data holder of P gets the publicly available anonymous datasets of other data holders, i.e. Q_1^*,Q_2^*, $\ldots$, Q_n^*, and computes the overlapping set (OL) using Definition 2. After that the data holder of P applies our anonymization technique (described in detail in Section 4).

For numeric QIDs, the intersection in Definition 2 returns the overlapping value. For example, the intersection of *age* QID value 15–25 in Q_i^* and 20 in P returns 15–25 $\cap$ 20 $=$ 20. For categorical QIDs, the intersection returns the value of the closest common generalization of two values. For example, intersection of values *'male'* $\cap$ *'female'* $=$ '$\emptyset$'. If one value is the generalization of another value, the intersection returns the more specific value. For example, the intersection of '$*$'$\cap$ *'male'* $=$ '$male$'. Here '$*$' corresponds to most generalized QID value in any generalization hierarchy. In sex QID generalization hierarchy, '$*$' presents both $male$ and $female$.

3 Cases of Privacy Breach in Composition Attack

3.1 Pros and Cons of Sampling in Composition Attack

An apparent way to combat the composition attack is sampling, i.e. only publish a portion of data. After a dataset is sampled, an adversary does not know if the record is in the published dataset or not. However, sampling only reduces the chance of finding overlapping tuples, but does not reduce the confidence of an adversary for inferring the sensitive information once overlapping tuples are found.

Example 1. Let us assume that the true match is caused by the same person visiting two hospitals, and that a false match is caused by two unrelated patients who happened to have the same QIDs and disease in two datasets. Let the sample rate be 50%. The probability of a true match is 25% when a patient have visited two hospitals. The chance of two unrelated patients to have the same QIDs (false match) depends on the data distributions of two datasets. For a simple illustration, let us assume that the chance is 50%. Assume that there are 5 sensitive values and each has the same chance to associate with QIDs. The chance for two QIDs matched tuples to have the same disease is only 4% and this reduces the probability of a false match down to 2%; which is much less than the 25% probability of true match. Therefore, an adversary has a reason to be confident about true match.

3.2 Pros and Cons of Generalization in Composition Attack

Let us assume that two or more data holders achieve ℓ-diversity [8] in the overlapping set (OL); such that overlapping equivalence classes have at least ℓ overlapping patients common that are suffering from distinct diseases (although it is not trivial to achieve this, and we discuss it in the following section). Intuitively, adversary only learns that an overlapping victim suffering from one of ℓ possible diseases. However, the privacy is possibly compromised for non-overlapping victim(s).

Example 2. In published datasets P^* and Q_i^*, there are QID groups as $P^* = \{$31–35, male, (*A*,*B*,*C*)$\}$ and $Q_i^* = \{$31–35, male, (*A*,*B*)$\}$. The adversary has only 50% chance of knowing if two overlapping victims who visited both P^* and Q_i^* suffer from disease $\{A\}$ (or $\{B\}$). However, the adversary has a chance to learn the sensitive information of a victim that is not in the overlapping set (OL). For example, the adversary knows a victim (male, 31) who only visited P^*. Based on the above data publication, the adversary knows that the victim (male, 31) suffers from disease $\{C\}$.

Data publication by generalization also suffers the minimality attack [13]. An adversary can use the knowledge of an anonymization algorithm to infer the sensitive information of individuals. The same attack applies to multiple data releases.

Example 3. Assume that an algorithm follows the following procedure. If the overlapping set (OL) satisfies ℓ-diversity, publish the data section. Otherwise, generalize the data section with the adjacent tuples to make the overlapping set (OL) satisfy ℓ-diversity. If not possible, suppress tuples to make the overlapping set (OL) empty. Based on the principle, P^* is published using the information of already published $Q_i^* = \{$31–35, male, (*A*,*B*)$\}$ and $P^* = \{$35–40, male, (*A*,*C*)$\}$. The adversary knows that victim (male, 33) visited both Q_i^* and P^*. Based on the published datasets, he does not know if the victim suffers from disease $\{A\}$ or $\{B\}$. The adversary knows that the victim's record has been suppressed from the subsequent dataset P^*, but this information does not help her/him to figure out the true sensitive information of the victim either. However, s/he knows the generalization algorithm as well. S/he reasons the sensitive value of victim as disease $\{A\}$ as the following. If the victim suffers from disease $\{B\}$, the subsequent published dataset P^* should be as $P^* = \{$33–40, male, (*A*,*B*,*C*)$\}$ to maintain the 2-diversity in the overlapping set (OL). The record of the victim is suppressed from P^* because the victim does not suffer from disease $\{B\}$ and there is no possibility to generalize the data to satisfy 2-diversity in the overlapping dataset (OL). Therefore, the victim suffers disease $\{A\}$ for sure.

4 (ρ, α)-anonymization Model

ρ-sampling and α-overlapping, in short (ρ, α)-anonymization, model consists of two steps anonymization, as detailed in the following.

Definition 3 **(ρ-Sampling).** *Given a sampling probability $\rho \leq 1$, each tuple $t \in P$ is sampled with the probability of ρ without replacement, i.e. whether a tuple is included in sampled dataset P^ρ for subsequent publication is decided by tossing a coin with head probability ρ. Only if the coin heads, a tuple is included in P^ρ.*

Sampling is already a routine practice in data publishing [12], because data publishers hold gigantic data and only a subset is publicly released. As shown in previous section, an adversary infers the sensitive values of individuals in overlapping and non-overlapping datasets with different confidences. The sampling is necessary for privacy protection of non-overlapping tuples in multiple independent data releases since it reduces the confidence of locatability of an adversary. Later, in Section 4.1, we discuss in detail how sensitive value inference of a non-overlapping tuple is bounded by sampling. Next we preserve the privacy of overlapping tuple(s).

Definition 4 **(α-overlap).** *Independently published anonymous dataset P^* (formed from P^ρ) satisfies α-overlap, if for any tuple $t \in \mathrm{OL}$; its QID group in P^* contains at least α ($\alpha \geq 2$) uniformly distributed distinct sensitive values with Q_i^*.*

The overlapping set (OL) is computed by utilizing publicly available anonymous releases of other publishers using Definition 2. The rationale of α-overlap is that, if a tuple t is published by more than one publishers then all its generalized QID hosting groups must contain α common sensitive values in a way such that its sensitive values in QID hosting groups forms uniform distribution (i.e. equal number) for α common sensitive values in overlapping set (OL). The uniform distribution in overlapping set (OL) makes an adversary's confidence equally split over α sensitive values. The distributions of the sensitive values in P^* and Q_i^* can be quite different. The uniform distribution is a good trade-off between diverse distributions.

4.1 Privacy Analysis of (ρ, α)-anonymization

In this section we analyze the privacy of overlapping and non-overlapping tuples in (ρ, α)-anonymization. We start with the privacy of non-overlapping tuples.

Observation 1 *If dataset P^* satisfies (ρ, α)-anonymization, then the confidence of an adversary to derive the true sensitive value of any non-overlapping tuple t from P^* is bound by sampling probability ρ.*

Example 4. Reconsider the scenario of Example 2 with additional assumption that P^* is sampled with 50% probability. Now, adversary has maximum ρ chance that the record <male, 31–35, *C*> is the one s/he is looking for and there is no other source of information to reinforce this.

Next we reason about the privacy of overlapping tuples in (ρ, α)-anonymization.

Observation 2 *If dataset P^* satisfies (ρ, α)-anonymization with all already published anonymous datasets, then the confidence of an adversary to derive the sensitive value of any overlapping tuple $t \in \mathrm{OL}$ through the composition attack is bound by $\lceil \frac{1}{\alpha} \rceil$; where $\lceil \; \rceil$ is ceiling operator.*

An ideal situation is that the confidence of guessing a sensitive value by an adversary from an anonymous dataset is similar to the distribution of sensitive value in original data, like in t-closeness [7]. However, in the composition attack, we deal with more than one datasets which may have different distributions for sensitive values. We do not have a "standard" distribution to close to. Further, the overlapping set (OL) is a small proportion of a dataset, and may not represent the distribution of the global dataset. Uniform distribution for α common sensitive values is a good trade off. Any latter data holder, (who is at the risk of composition attack i.e. Hospital-3 in our case) can simply set α to a sufficiently larger value, for every overlapping tuple $t \in \mathrm{OL}$, to achieve the required extent of privacy preservation.

Example 5. Reconsider the scenario of Example 3, where the P^* will be $P^* = \{$31–40 (1), male, (*A*,*B*,*C*)$\}$ to maintain the 2-overlap with Q_i^*. Note that, although the adversary learns that a counterfeit exits in QID group of P^*, s/he still cannot narrow down the possible diseases of overlapping victim (male, 33). In fact, to the adversary, there is a 50% chance that either $\{A\}$ or $\{B\}$ would be the counterfeit.

Table 4: Major symbols used in different phases of composition based generalization

P	Input dataset to be anonymized	$Q^*[\ ]$	Already published overlapping datasets
ρ	Input sampling parameter for dataset P	β	Input parameter, a trade-off for efficiency and quality
k	Input parameter for k-anonymity	ℓ	Input parameter for ℓ-diversity
α	Minimum uniformly distributed distinct sensitive values to be placed in overlapping set OL		

4.2 m-invariance: similar model but not good in this scenario

m-invariance [15] model is a very typical model in serial data publication. It has certain strengths for privacy protection in multiple data publications, but it has its limitations in our problem. First, it needs a sampling process for m-invariance model in our scenario too. Second, m-invariance model requires every tuple in an overlapping dataset to persistently associate with the same set of sensitive values, called *signature* in [15]. In our scenario, the number of counterfeit or suppressed tuples must be large to satisfy the *persistent consistency* as required in m-invariance [15]. m-invariance requires every overlapping tuple in P^* to associate with m number of same sensitive values, (*signature* as defined in [15]), as the corresponding tuple in Q_i^*. In other words, m-invariance requires all sensitive values (both overlapping and non-overlapping) in QID groups of datasets P^* and Q_i^* with overlapping QID values be the same; whereas we only need to handle overlapping QID groups to combat composition attack.

Example 6. Let the sensitive values of the QID groups in $Q_1^* = \{A,B\}$, $Q_2^* = \{A,C\}$, $Q_3^* = \{A,D\}$ and the available sensitive values in $P^\rho = \{A,B,C\}$ (let $P^\rho = P$ with $\rho =$ 50%). Now to meet 2-invariance requirement we need 3 QID groups with counterfeit tuples $\{\emptyset\}$, $\{A\}$ and $\{A,D\}$ respectively. In contrast, we require one counterfeit tuple, i.e. $\{D\}$, to meet (50%, 2)-anonymization. Intuitively, m-invariance [15] principle is too strong in our scenario.

5 Composition Based Generalization

5.1 Phases

We use the running example to demonstrate the different phases of composition based anonymization to achieves (ρ,α)-anonymization; where $\rho = 50\%$, $k = 4$, $\ell = 3$, $\alpha = 2$, Q_1^*, Q_2^* and P are Tables 1(b), 2(b) and 3(a) respectively. Given already published tables Q_1^* and Q_2^* available to data holder of P, we show how to compute the anonymized version P^* from P. We perform the computation in following five phases: *sampling, division, balancing, assignment* and *generalize*. The explanation of the major symbols used in different phases of composition based generalization is shown in Table 4.

Sampling Firstly, we apply the sampling on P with input sampling probability ρ to obtain P^ρ. Each tuple of P is independently sampled. In our example, we assume sampling probability $\rho = 0.5$ and sampling function $f_\rho(t)$ returns the tuple (i.e. $f_\rho(t) = t$) if $f_\rho = 1$ and $f_x(t) = \emptyset$ if $f_\rho = 0$. In our case the probability of getting $\{0,1\}$ is 0.5. In our example, we assume that for all the tuples of P, $f_x(t) = t$, i.e. $P^\rho = P$.

Division In this phase we partition the sampled P^{ρ} into two disjoint sets i.e. *overlap tuples* $S_{\cap} = Q_i^* \cap P$ $(i \in 1, 2)$; computed as per Definition 2 and *non-overlap tuples* $S_{-} = P^{\rho} - S_{\cap}$. In case of our example the tuples with sensitive values $\{B,R\}$ and $\{C,C,L,B,W\}$ are included in $S_{\cap}$ and S_{-} respectively. For each tuple $t_{\cap} \in S_{\cap}$, we define its 'possible sensitive values' as the set of distinct sensitive values in the corresponding generalized QID hosting group in already published Q_i^*. In the running example the tuples, with sensitive value $\{B\}$ in $S_{\cap}$, has possible sensitive values as $\{B,C\}$; i.e. the distinct sensitive values in QID group-1 of Table 1(b). Whereas the set of possible sensitive values for $\{R\}$ is $\{S,R,T\}$; i.e. the distinct sensitive values in QID Group-2 of Table 2(b).

In the end of division phase, we simply divide $S_{\cap}$ into several QID groups, on the basis of their possible sensitive values. In our running example, we have two QID groups i.e. GRP_1(*B,C*) and GRP_2(*S,R,T*).

Balancing We say that a QID group GRP_i $(i \geq 1)$ is *balanced*, if it contains at least α tuples; having such distinct sensitive values that these α tuples along with their corresponding QID group(s) in already published overlapping dataset(s) comply with α-overlap principle (Definition 4). For example, the QID group GRP_1 will become balance with corresponding overlapping QID group-1 of Table 1(b) if we include two tuples (from S_{-}) having sensitive value $\{C\}$ $(\alpha = 2)$. The objective of this phase is to balance all QID groups.

Continuing our example, we cannot balance QID group GRP_2 with corresponding QID group-2 in Table 2(b) because to balance GRP_2 we need at least one tuple having either of sensitive values $\{S,T\}$ in S_{-}. As there is no such sensitive values in S_{-} so we add one counterfeit sensitive value (either of $\{S,T\}$) in the QID group GRP_2 to make it balance with corresponding QID group-2 in Table 2(b) . We add counterfeit sensitive value, in unbalanced QID group GRP_2, instead of suppressing the overlapping tuple because suppression can still breach the privacy of overlapping tuple, as shown in Example 3. A non-desiring solution can be to suppress all the tuples of P^{ρ} with the same sensitive values as of corresponding QID group in Q_i^*.

Assignment We assign remaining tuples of S_{-} (if any) in two steps. First, we include the tuples in existing QID group(s) to comply with the generalization principle(s) (k-anonymity [11], ℓ-diversity [8], t-closeness [7] etc.). Second, if necessary, new QID group(s) may be created for remaining tuples. A QID group is called *complete* if it complies with all underlaying generalization principles. The purpose of this phase is to make all QID groups complete. In running example, S_{-} has three tuples having sensitive values $\{L,B,W\}$. We have assumed that $k = 4$ and $\ell = 3$ as generalization principles. The tuple having sensitive value {L} is assigned to GRP_1 and remaining two tuples having sensitive values {B,W} are assigned to GRP_2 to make both groups complete.

The crucial part is the selection of a tuple t_{-} from S_{-} and its assignment to a QID group. We select first β tuples of S_{-} and search for the *optimal tuple* which requires least anonymization of QIDs. β is an input parameter that restricts the search of the optimal tuple within the first β tuples of S_{-}. The incorporation of β improves the per-

Algorithm 1 *overlapAnonymize*(P,$Q^*[\,]$,ρ,β,α,k,ℓ)

```
1:  $P^\rho$                                   ▷ sample each tuple of $P$ with $\rho$ probability to get $P^\rho$
2:  $S_\cap = Q^*[\,] \cap P^\rho$                        ▷ get overlap tuples (Definition 2)
3:  $S_- = P^\rho - S_\cap$                               ▷ get non-overlap tuples
4:  Divide $S_\cap$ into $GRP[\,]$ QID groups             ▷ *Division* phase
5:  $Sort(S_-)$                             ▷ sort all $S_-$ tuples on the basis of QIDs
6:  **while** $|S_-| \neq 0$ **do**                 ▷ continue till all tuples are assigned
7:      **for** $i \leftarrow 1,\ \beta$ **do**                  ▷ to access first $\beta$ tuples of $S_-$
8:          $t_- = S_-[i]$
9:          **for** $j \leftarrow 1,\ |GRP|$ **do**                       ▷ to access all groups
10:             **if** $GRP[j]$ is complete with $k$, $\ell$ and $\alpha$ **then** exclude $GRP[j]$ from $GRP[\,]$
11:             **else if** $t_-$ is optimal to $GRP[j]$ **then** assign $t_-$ to $GRP[j]$
12:             **if** $|GRP| == 0$ **then** $i = \beta$; $j = |GRP|$ ▷ set to break the loops of lines 7 and 9
13:         **end for**
14:     **end for**
15:     **if** $|GRP| == 0\ AND\ |S_-| \geq k$ **then** create new $GRP[0]$        ▷ all groups become
        *complete* but still there are more than $k$ (from $k$-anonymity) unassigned tuples
16:     **else if** $|GRP| == 0\ AND\ |S_-| < k$ **then** assign all remaining tuples to recently completed group
17:     **else if** $|GRP|\ ! = 0\ AND\ |S_-| == 0$ **then** *complete* all remaining groups by adding
        counterfeit tuples.
18: **end while**
```

formance of assignment process (as shown in experiments in Section 6) because instead of traversing all the tuples of S_-, only β tuples are searched for optimal tuple. Note that the anonymization of optimal tuple depends on the specific generalization principle(s) to be employed. Within β tuples, we calculate *Distortion* [6] caused by every tuple t_- $\in S_-$ to each QID group and assigns such t_- (also referred optimal tuple) to the QID group which has minimum distortion with t_-; as long as α-overlap (Definition 4) holds. Due to space constraint, we are omitting the calculation details of distortion between QID group and a tuple and reader is referred to the original paper [6] for further details. Algorithm 1. formally presents the assignment strategy.

In our running example we assume $k = 4$ and $\ell = 3$ and there are two QID groups; GRP_1 contains 3 tuples and GRP_2 has two tuples. We have three tuples in S_- with sensitive values $\{L,B,W\}$, as per their order in P^ρ. Due to the sorting of the S_- (Algorithm 1 line 5) on the basis of QIDs; the sorted tuples have the sensitive values as $\{B,L,W\}$. In first iteration we assign the tuple with sensitive value {L} to GRP_1 to make it complete i.e now GRP_1 has four tuples (k = 4), three distinct sensitive values (ℓ = 3) and two overlap values (α = 2). So, we exclude the GRP_1 from all subsequent iterations (Algorithm 1 line 10). Importantly, we cannot assign the tuple $\{B\}$ to GRP_1 instead of tuple $\{L\}$ (although the tuple $\{B\}$ requires less generalization for GRP_1) because after the assignment of tuple $\{B\}$ to GRP_1, the GRP_1 will not comply with α-overlap condition (Definition 4). Now, we are left with two tuples with sensitive values $\{B,W\}$ (i.e. $|S_-| \neq 0$) and GRP_2 is incomplete (as $k = 2$ instead of 4); so next two iterations assign the $\{B,W\}$ tuples to GRP_2 (Algorithm 1 line 11) and assignment algorithm finishes by breaking assignment loop (Algorithm 1 line 6).

Table 5: Attribute domain size

Attribute	Age	Sex	Education	Marital Status	Birth Place	Occupation
Domain Size	91	2	17	7	50	50

Generalize We can have two types of QID groups, i.e. *overlap QID groups* (created during division phase) and *non-overlap QID groups* (created in assignment phase). The generalization of non-overlap QID group is trivial; we get the minimum QID range that covers all the QID values and replace the original QID values with this range.

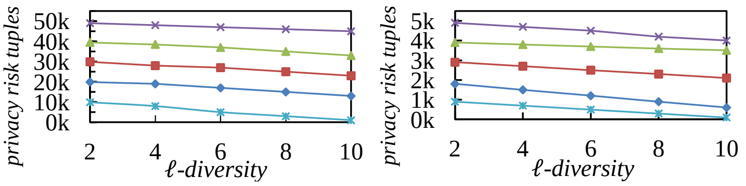
In case of overlap QID group, we get the minimum QID range that **(i)** covers all QID values in current QID group (similarly like non-overlap group) **(ii)** as well as the QID generalization range in corresponding overlap QID group of already published dataset. In our running example, the generalization range of the age in overlap QID group GRP_1 will be (15 – 35). We cannot put (20 – 35) as generalization range for age in GRP_1 because (20 – 35) only covers age QID values in GRP_1; whereas the generalization range of age in corresponding overlap group of already published Q_1^* is (15 – 25). So the minimum range for age in GRP_1 is (15 – 35) instead of (20 – 35). There is no need to generalize sex in GRP_1 because without any generalization the aforementioned both conditions of overlap group are met. In the same way, the generalization range for age in GRP_2 is (28 – 45) instead of (28 – 40). The Table 3(c) is the final outcome after generalization phase.

The complexity of the Algorithm 1 mainly depends on the computation of tuples in $S_\cap$ (Algorithm 1, line 2), sorting of S_- tuples (Algorithm 1, line 5) and searching the optimal tuple (Algorithm 1, line 11). We assume that $|P| \approx |Q_i^*|$, so major computation overhead lies on the computation of $S_\cap$ i.e. $|P| * |P|$ steps. Intuitively, the complexity of $O(m^2)$ where $m = |P|$. The more optimization of anonymization algorithm is future work we plan to pursue.

6 Experiments

All the experiments are performed on a machine running a 2.4 Ghz CPU with 3 Gigabyte memory. We deploy two real repositories BIR and OCC from United States census data downloadable from http://ipums.org. Both contain 300k and 45k tuples respectively. *BIR* includes four QID attributes, age, $gender$, $education$, $marital\ status$, and a sensitive attribute $birth\ place$. Whereas OCC contains the same QID attributes, but with different sensitive attribute $occupation$. All columns are discrete with domains size given in Table 5.

We create four disjoint sub-datasets Q_i^{bir} (Q_i^{occ}) ($n \in 1, \ldots, 4$) from $BIR(OCC)$. It suffices to clarify the generation and generalization of Q_i^{bir}, since the same method is used for Q_i^{occ}. Each sub-dataset Q_i^{bir} is of 50k tuples. Next, we form the dataset P_{bir}, also having 50k tuples. The remaining 50k tuples initiates a pool $\overset{O}{_{bir}}$. The dataset P_{bir} will be anonymized using (ρ, α)-anonymization. For OCC each sub-dataset Q_i^{occ} and P_{occ} contains 8k tuples and remaining 5k tuples goes in pool $\overset{O}{_{occ}}$. In all experiments we set sampling probability $\rho = 0.50$.



(a) P^{*}_{bir} (b) P^{*}_{occ}

Fig. 1: Successful composition attack vs. the overlap volume u

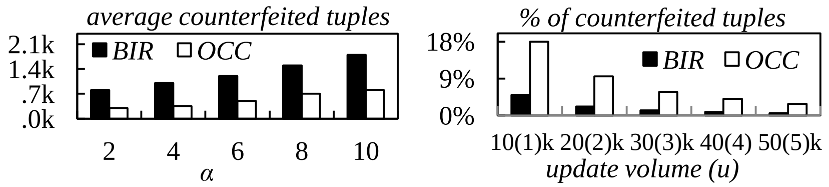
We run four sets of experiments, involving 4 sub-datasets with P_{bir}. In each set of experiment, the publication of n sub-datasets is straightforward, i.e we randomly select $u * n$ tuples from $^{O}_{bir}$ and insert separate u tuples in each Q_i^{bir}, subsequently anonymous Q_i^{bir*} (having 50k $+$ u tuples) is created that satisfies some generalization principle(s). Here u is a parameter, called *overlap volume*, controlling the overlap rate between n sub-datasets and P_{bir}. For the P_{bir}, the generalized P^{*}_{bir} is obtained by, first, inserting the same $u * n$ tuples in P_{bir}; i.e. P_{bir} has separate u overlap tuples with each Q_i^{bir*}. Consequently P_{bir}, having 50k $+$ $(u*n)$ tuples, is anonymized using (ρ, α)-anonymization that utilizes other four anonymous sub-datasets Q_1^{bir*}, $Q_2^{bir*}, \ldots, Q_4^{bir*}$. We repeat this process by increasing the u from 10k to 50k (i.e. each set of experiment includes 5 iterations on BIR). In case of OCC, we increase the overlap volume u from 1k to 5k tuples in each set of experiment, so total iterations in each set of the experiment of the OCC are also 5.

6.1 Failure of Conventional Generalization Schemes

In the first set of experiments, we aim at establishing the conjecture that the existing generalization principles may lead to severe privacy disclosure in independent data publishing. This finding was also observed in [2]. We adopt the algorithm in [5] to compute ℓ-diversity [8] as the representative generalization principle, since it is widely adopted and offers stronger privacy than k-anonymity [11]. In Fig. 1(a), we plot the number of privacy risk tuples (as explained in Section 1.1) in P^{*}_{bir} as a function of u, as this parameter changes from 10k to 50k in $^{O}_{bir}$. Regardless of u and ℓ, there are nearly 90% overlap tuples whose privacy is not preserved at all in P^{*}_{bir}. We repeat the experiments on sub-datasets Q_i^{occ} and P_{occ}. The results are illustrated in Fig. 1(b), confirming the same observations.

6.2 (ρ, α)-anonymization Evaluation

We have $n = 4$ already published sub-datasets of BIR and OCC i.e. Q_1^{x*}, $Q_2^{x*}, \ldots,$ Q_4^{x*} $(x = BIR$ or $OCC)$. We invoke the (ρ, α)-anonymization (Section 4) on P^{x} to compute the generalized version P^{*}_{x} for α-overlap publication. The computation of P^{*}_{x} utilizes already published four anonymous datasets to identify overlapping set (OL) using Definition 2. The P^{*}_{x} is characterized by two input parameters, i.e. overlap-volume u and overlap diversity α.



(a) *update volume* $u = 10(1)\text{k}$

(b) $\alpha = 6$

Fig. 2: Average and Percentage of counterfeiters in P^*_{bir} (P^*_{occ})

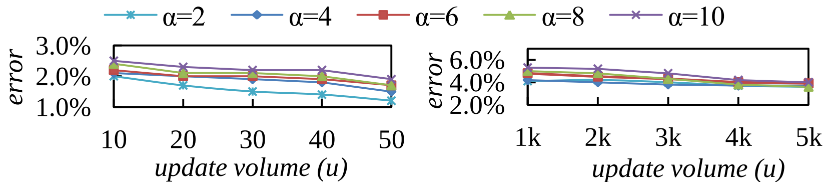
Number of Counterfeited Tuples. We start by demonstrating that only a small number of counterfeited tuples are needed to enforce (ρ, α)-overlap. In Fig. 2(a), we set u $= 10(1)$k but vary the α from 2 to 10 and measure the average counterfeited tuples in P^*_{bir} (P^*_{occ}). The average number of counterfeited tuples increase along with α because higher α requires more distinct sensitive values in each overlap QID group for balancing; failure of this causes insertion of more counterfeited tuples in QID group(s).

Next, we focus on the percentage of counterfeiters with overlap-volume u. We get the percentage of counterfeited tuples in P^*_{bir} (P^*_{occ}) for all sub-datasets of BIR (OCC). Fixing $\alpha = 6$, Fig. 2(b) shows the percentage of counterfeited tuples for both BIR and OCC as a function of u. The percentage decreases as u increases, such that (ρ, α)-overlap can utilize more overlapping tuples. This is expected, because for a fix value of α as u increases, more overlap QID groups are more likely to have same set of possible sensitive values which can accommodate larger overlap volume. Intuitively, causing less counterfeited tuples while balancing.

Utility of the Published Data. In the following set of experiments, we will use P^*_x (where $x = BIR$ or OCC) to answer queries about the original sub-dataset P_x. We concentrate on aggregate queries, since they are the basic operation for numerous mining tasks (e.g., decision tree learning, association rule mining etc.). Specifically, each query has the form:

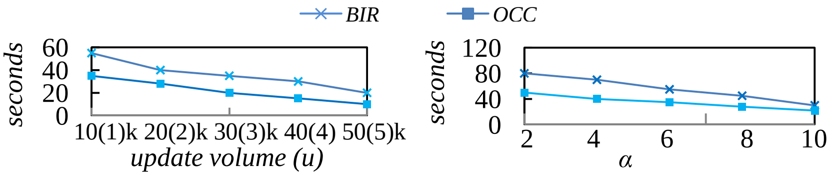
SELECT COUNT (*) FROM P^*_x WHERE $pred\{$ $t^*_x[A^{qi}_1]$ AND $\ldots$ AND $t^*_x[A^{qi}_4]$ AND $t^*_x[A^s]$ $\}$

The P^*_x is the sub-dataset generalized using (ρ, α)-overlap, $t^*_x[A^{qi}_1], \ldots, t^*_x[A^{qi}_4]$ denote the four QID attributes in P^*_x, and $t^*_x[A^s]$ is the sensitive attribute $birth\ place$ ($occupation$). For each attribute A, the condition $pred(A)$ has the form $|A|.\delta$, where $|A|$ is the domain size of A (see Table 5), and δ is a query parameter called selection range. A larger result is returned with higher δ. Our workload consists of 1000 queries with same P^*_x and $t^*_x[A^s]$. Given a query, we obtain its actual result R_{act} from original overlap sub-dataset P^x, and compute an estimated answer R_{est} from its (ρ, α)-overlap generalized version P^*_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. Adopting α $= 6$, Fig. 3 plots the workload error as a function of update volume u for $P^*_{bir}(P^*_{occ})$ respectively. In all experiments, the error is at most 2.5(5.3)% for α-overlap, indicating high utility of the (ρ, α)-overlap.



(a) P^*_{bir} $(\alpha = 6)$ (b) P^*_{occ} $(\alpha = 6)$

Fig. 3: Query error vs. update volume u



(a) $\alpha = 6$ (b) *update volume* $u = 10(1)$k

Fig. 4: Computation overhead vs. u and α

Computation Overhead The last experiment evaluates the efficiency of our overlap generalization algorithm. First in Fig. 4(a), we set $\alpha = 6$, and measure the average time of computing a generalized sub-dataset P^*_{bir} (P^*_{occ}) for different u. The cost is more expensive when u is higher, because the algorithm needs to process more tuples of overlap volume in each QID group. Then in Fig. 4(b), we fix u to 10(1)k, and got the cost as a function of α. The overhead decreases as α increases, since a larger α necessitates fewer overlap QID groups, and requires less time in generalization phase.

7 Related Work

Privacy preserving data publishing has mainly focused on taking into account other known releases, such as previous publications by the same data holder (called sequential, serial or incremental releases) [14, 15] and multiple views of the same dataset [16, 17]. Another line has considered incorporating knowledge from partitioned views of a same dataset to group individuals [16]. The sequential/multiple view release models do not fit because, in this paper, we deal with the case when there are multiple independent publishers but their release is single. A hypothetical discussion of the same problem is in [2] (driving concepts from differential privacy [1]) without the actual implementation and the test results. Although, some recent work has implemented differential privacy in data publishing [10] but it did not address the composition attack. Most relevant works to this paper are [9, 3]. But they incorporate the *coordinated model*; where all locations communicate with each other before releasing their data 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. The coordinated model also does not comply with our requirement because we are dealing with noncoordinated scenario where each location independently anonymizes its data without having any communication with other location(s). Secondly in coordinated model, all locations anonymize and publish data at the same time but in our case each location can publish its data anytime.

8 Conclusion

Existing single/serial data publishing methods do not support multiple independent data publication by different data holders having overlapping records. This paper has developed (ρ, α)-overlap anonymization model to prevent an adversary from using data releases of different data holders to infer sensitive information of overlapping individuals. We have provided an efficient algorithm for computing anonymized datasets to achieve (ρ, α)-overlap. We experimentally showed that the anonymized data adequately protects privacy and yet supports effective data analysis.

References

1. C. Dwork. Differential privacy. In *ICALP*, pages 1–12, 2006.
2. S. R. Ganta, S. P. Kasiviswanathan, and A. Smith. Composition attacks and auxiliary information in data privacy. KDD'08, pages 265–273.
3. W. Jiang and C. Clifton. A secure distributed framework for achieving k-anonymity. *JVLDB*, 15:316–333, 2006.
4. X. Jin, M. Zhang, N. Zhang, and G. Das. Versatile publishing for privacy preservation. KDD '10, pages 353–362.
5. K. LeFevre, D. J. DeWitt, and R. Ramakrishnan. Mondrian multidimensional k-anonymity. ICDE'06, pages 25–.
6. J. Li, R. C.-W. Wong, A. W.-C. Fu, and J. Pei. Anonymization by local recoding in data with attribute hierarchical taxonomies. *TKDE*, 20:1181–1194, 2008.
7. N. Li, T. Li, and S. Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and ℓ-diversity. ICDE'07, pages 106–115.
8. A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam. ℓ-diversity: Privacy beyond k-anonymity. *TKDD*, 2007.
9. B. Malin. k-unlinkability: A privacy protection model for distributed data. *DKE*, 64(1):294–311, 2008.
10. N. Mohammed, R. Chen, B. C. Fung, and P. S. Yu. Differentially private data release for data mining. KDD '11, pages 493–501, New York, NY, USA, 2011. ACM.
11. L. Sweeney. k-anonymity: a model for protecting privacy. *IJUFKS*, pages 557–570, 2002.
12. Y. Tao, X. Xiao, J. Li, and D. Zhang. On anti-corruption privacy preserving publication. ICDE'08, pages 725 –734.
13. R. C.-W. Wong, A. W.-C. Fu, K. Wang, and J. Pei. Minimality attack in privacy preserving data publishing. VLDB'07, pages 543–554.
14. R.-W. Wong, A.-C. Fu, J. Liu, K. Wang, and Y. Xu. Global privacy guarantee in serial data publishing. ICDE'10, pages 956–959.
15. X. Xiao and Y. Tao. m-invariance: towards privacy preserving re-publication of dynamic datasets. SIGMOD '07, pages 689–700.
16. B. Yang, H. Nakagawa, I. Sato, and J. Sakuma. Collusion-resistant privacy-preserving data mining. KDD'10, pages 483–492.
17. C. Yao, X. S. Wang, and S. Jajodia. Checking for k-anonymity violation by views. VLDB '05, pages 910–921.