Journal of Engineering and Applied Sciences 13 (11): 4205-4216, 2018
ISSN: 1816-949X

Personalized Privacy Preserving Incremental Data Dissemination Through Optimal Generalization

[1]S. Ram Prasad Reddy, [2]K.V.S.V.N. Raju and [3]V. Valli Kumari
[1]Department of Computer Science and Engineering,
Vignan's of Institute of Engineering for Women, Visakhapatnam, Andhra Pradesh, India
[2]GVP College for Degree and PG Courses School of Engineering, Rushikonda,
Visakhapatnam, Andhra Pradesh, India
[3]Department of Computer Science and Systems Engineering,
Andhra University, Visakhapatnam, Andhra Pradesh, India

Abstract: A need to unveil health information for several reasons such as for health services, payment in case of insurances, health care operations, research and so on is on high demand. Personal information is to be disseminated without revealing the individual's identity in all these circumstances. Tremendous work has been carried out to provide privacy for publishing static data. Existing anonymization methods such as k-anonymity and l-diversity models have led to a number of valuable privacy-protecting techniques for static data. This very postulation implies a substantial limitation as in many applications data collection is rather a persistent process. In places where data keeps on increasing on a daily basis, the current techniques are inadequate and suffer from poor data quality and/or vulnerable to inferences. A very diminutive work has been carried out in this direction and personalized privacy for incremental datasets has not been studied. In this study, we present a solution that presents incremental data dissemination in the context of personalized privacy using optimal generalization. An algorithm in incremental mode to handle personalized privacy issues with maximum diversity and minimum anonymity is proposed. The experiments on continuously growing real world and synthetic datasets show that the proposed scheme is efficient and produces publishable data of high utility.

Key words: Privacy, personalization, incremental data dissemination, high sensitive attribute, optimal generalization, India

INTRODUCTION

Preserving individual respondent's privacy is a top priority while disseminating person-specific data. Amongst numerous approaches addressing this issue, the k-anonymity model (Sweeney, 2002a, b; Samarati, 2001) and the l-diversity model (Machanavajjhala *et al*., 2006) have drawn notable attention in the research community. In the k-anonymity model, privacy protection is achieved by ensuring that every record in a published dataset is indistinguishable from at least (k-1) other records within the dataset. Thus, the risk of record identification is guaranteed to be at most 1/k. While the k-anonymity model primarily focuses on the problem of record identification, the l-diversity model, that is built upon the k-anonymity model, addresses the risk of attribute disclosure. As attribute disclosure might occur without records being identified (e.g., due to lack of diversity in a sensitive attribute), the l-diversity model in its simplest form, additionally requires that every group of indistinguishable records contain at least l distinct sensitive attribute values, thereby, the risk of attribute disclosure is bound to be at most 1/l.

Even though, Bayardo and Agrawal (2005), Zhang *et al*. (2007), LeFevre *et al*. (2005, 2006), Fung *et al*. 2005 and Sweeney (2002a, b) have come up with a number of valuable privacy-protecting techniques, these approaches only handle static data releases. This assumption entails a significant shortcoming, as in many environs data collection is rather a continuous process. Moreover, the assumption entails "one-time" data dissemination. Obviously, this does not address today's strong demand for immediate and up-to-date information as the data cannot be released before the data collection is considered complete.

In real life scenarios, data grows/shrinks and needs to be published at consistent intervals which persuades our study. In this study, we assume that the datasets may

Corresponding Author: S. Ram Prasad Reddy, Department of Computer Science and Engineering,
Vignan's of Institute of Engineering for Women, Visakhapatnam, Andhra Pradesh, India

be updated not only by inserting new entities but also by deleting the existing entities. In all these situations, there would be a need for publishing up-to-date datasets at periodic intervals.

There is a considerable amount of study (Pei *et al*., 2007; Wang and Fung, 2006; Xiao and Tao, 2007; Bu *et al*., 2008; Byun *et al*., 2006; Fung *et al*., 2008; Wong *et al*., 2009; Cao *et al*., 2008) on incremental data publishing. Large section of the authors addressed different types of inferences and provided solutions to them. For instance, there might be a need to publish medical data of a medical centre at periodical intervals for various research purposes. Study has been carried out on the privacy protection issues for multiple data publications of multiple instances of the data (Wang and Fung, 2006). While disseminating up-to-date data, the individuals that are recorded in the data might change and the sensitive values correlated to individuals might also change. It is assumed that the sensitive values change freely.

Our objective in this study is to provide personalized privacy for publishing incremental datasets in its simplest mode minimizing the scope for inferences. In order to achieve this, we first identify the privacy requirement for incremental data dissemination. We discuss possible types of linking probabilities that an attacker may exploit by observing multiple published anonymized datasets. We develop an efficient algorithm to anonymize, check for privacy complaince and then publish the data with high assertion.

In summary, the study contributes the following: it is after examination that personalized privacy preserving incremental data dissemination has not been studied in this direction. It also ensures that publishing multiple versions of the data do not contribute too much of information to the adversary from subsequent releases. We use incremental cluster-based generalization for providing anonymity. QID sets have been considered during the process of anonymization. We conducted extensive experiments with a real world and synthetic datasets to verify with respect to privacy gain, information loss and computational time. The results indicate that our solution is elegant and promising.

Literature review: The elementary privacy apprehensive problem is publishing microdata for public use (Samarati 2001) which has been comprehensively studied. An enormous category of privacy attacks is to re-identify the individuals by merging published data with some externally available sources of information. To counter these types of attacks, the mechanism of k-anonymity was proposed (Bayardo and Agrawal, 2005; Sweeney, 2002a, b). The ideal anonymization method minimizes the information loss or maximize the utility. However, theoretical analysis (Bayardo and Agrawal, 2005) indicates that the problem of optimal anonymization under many non-trivial quality models is NP-hard.

Figure 1 depicts a partial overview of privacy preserving data publishing. The taxonomy in Fig. 1 includes various attack models, anonymity operations, information metrics, anonymity algorithms, anonymity for data mining and various data publishing scenarios. The information provided in the taxonomy is considered as the substance for providing privacy for different domains. The survey gives a glance of a variety of data publishing methods for publishing one-time data releases. Significant work has been carried on and include works on k-anonymity (Bayardo and Agrawal, 2005; Sweeney, 2002a, b) (X, Y)-anonymity (Wang and Fung, 2006), l-diversity (Machanavajjhala *et al*., 2007), confidence bounding (Wang *et al*., 2007) (X, Y)-linkability (Wang and Fung, 2006) (X, Y)-privacy (Wang and Fung, 2006), (α, k)-anonymity (Wong *et al*., 2006), LKC-privacy (Mohammed *et al*., 2009), (k, e)-anonymity (Zhang *et al*., 2007), t-closeness (Li *et al*., 2007), personalized privacy preservation (Xiao and Tao, 2006; Kumari *et al*., 2008), FF-anonymity (Wang *et al*., 2009), (c, t)-isolation (Chawla *et al*., 2005), ϵ-differential privacy (Dwork, 2006) distributional privacy (Blum *et al*., 2008) and (d, γ)-privacy (Rastogi *et al*., 2007).

This study centralizes on incremental data dissemination. Incremental data dissemination has gained importance steadily and researchers started contributing towards this. In the real world scenario, most of the datasets keep changing quite recurrently. The works on these cover sequential releases, continuous publishing, serial publishing and incremental publishing.

In sequential release, the data publisher is familiar only with a part of the raw data when a release is published. The data provider ensures that subsequent releases of data do not violate privacy requirement even when an adversary has access to all releases. The basic idea is to generalize the current release, so that, the join with previous release becomes a lossy join that hides the true join relationship to cripple a global quasi-identifier. This was addressed in sequential anonymization (Wang and Fung, 2006). The work considers sequential releases for different attribute subsets for the same dataset. The method is accomplished using top-down approach and subtree generalization. In order to the address the privacy requirement the researchers extended (X, Y)-privacy (Wang and Fung, 2006). The authors introduced lossy join an undesirable property in relational database design as a way to hide the join relationship among releases.

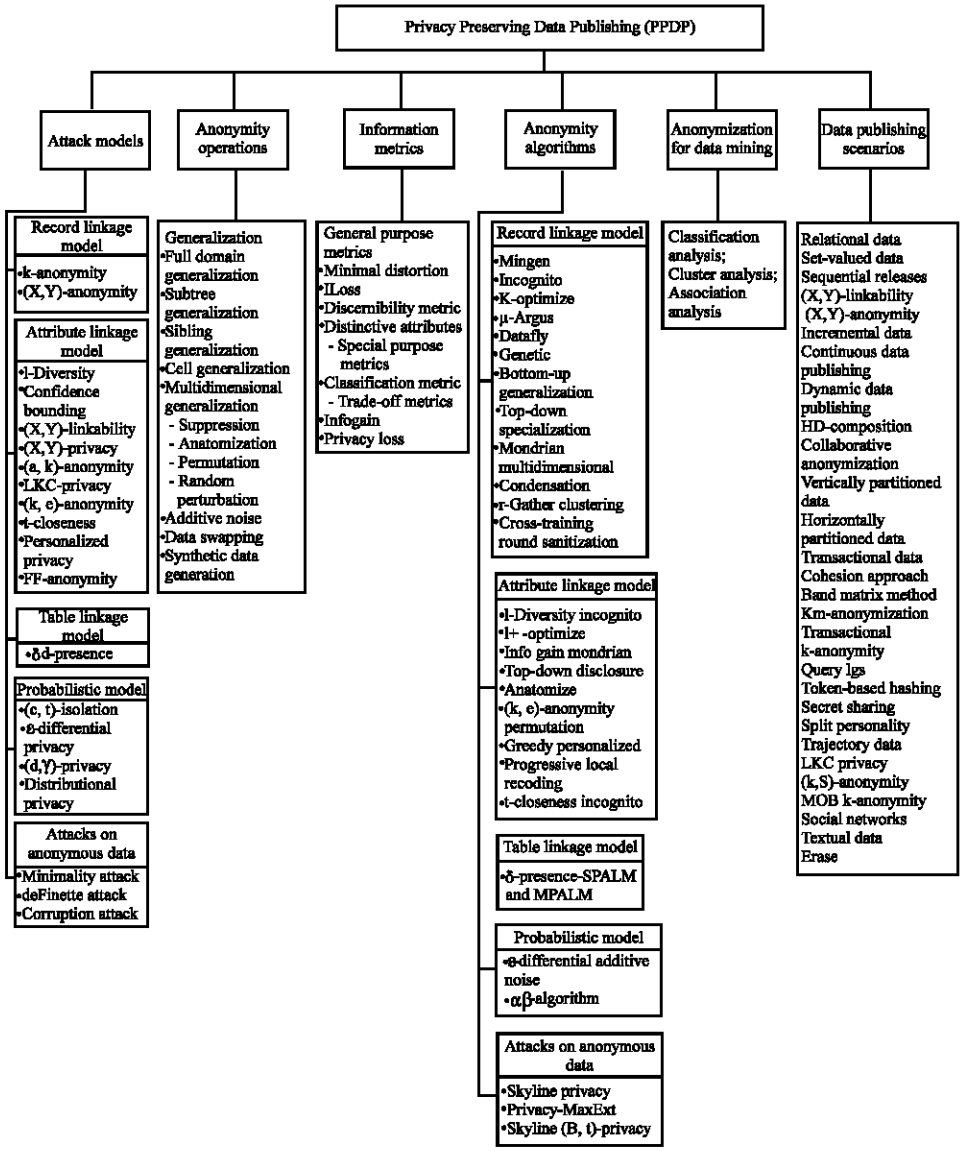

Fig. 1: Taxonomy for privacy preserving data publishing

Every release is an accumulated version of data at each instance of time in continuous publishing which contains all records collected so far. In this scenario, every data release comprises of the complete archive. It further does not allow record deletion once it is collected. Certain works include BCF-anonymity (Fung *et al*., 2008), secure anonymization (Byun *et al*., 2006), castle (Cao *et al*., 2008) and monotonic anonymization (Pei *et al*., 2007).

Fung *et al*. (2005) addressed the challenge of anonymity problem for a scenario where the data were continuously collected and published. Each release contained an accumulation of new data and the archive. Even if each release is k-anonymized, the anonymity of an individual could be compromised by cross-examining multiple releases. The authors highlighted the correspondence attacks and presented an approach to prevent such attacks. The authors proposed detection method and an anonymization algorithm namely BCF-anonymity to handle inferences. The method uses subtree generalization and does not allow deletion.

Byun *et al*. (2006) brought out an approach to securely anonymize a constantly rising dataset in an efficient manner while promising high data quality using l-diversity (Machanavajjhala *et al*., 2007). The basic idea underlying the approach is that one can efficiently anonymize a current dataset by directly inserting new records to the previously anonymized dataset. The

study further analyzed various inference channels and discussed how to avoid such inferences. The method does not handle deletions and the frequency of sensitive values was not considered. The main drawback is that if an insertion would violate any of the privacy requirements, even after generalization, the insertions are delayed until later releases. This strategy sometimes may run into starvation in which no new data could be released. It requires a very large memory buffer to store the delayed data records.

The basic idea of the proposed approach is to exploit quasi-identifier attributes to define a metric space: tuples are modeled as points in this space. CASTLE groups incoming tuples in clusters and releases all tuples belonging to the same cluster with the same generalization. CASTLE supports both numerical and categorical attributes anonymization by generalizing the latter through domain generalization hierarchies and the first through intervals. Clustering of tuples is further constrained by the need to have fresh anonymized data.

Cao *et al*. (2008) presented CASTLE (Continuously Anonymizing Streaming data via. adaptive cLustEring), a cluster-based scheme that anonymizes data streams on-the fly and at the same time, ensures the freshness of the anonymized data by satisfying specified delay constraints. The basic idea of the this approach is to exploit quasi-identifier attributes to define a metric space: tuples are modeled as points in this space. CASTLE groups arriving tuples in clusters and releases all tuples belonging to the same cluster with the same generalization. CASTLE supports both numerical and categorical attributes anonymization by generalizing the latter through domain generalization hierarchies and the first through intervals. Moreover, the authors proposed an extension of CASTLE to apply the l-diversity principle (Machanavajjhala *et al*., 2007) to data streams. Relevant features of CASTLE are the enforcement of delay constraints, its adaptability to data distributions and the use of a cluster reuse strategy that improves the performance without compromising security.

Pei *et al*. (2007) identified and reconnoitered the novel and practical problem of maintaining k-anonymity against incremental updates and proposed a solution. The authors analyzed how inferences from multiple releases may tamper the k-anonymity of data and propose the monotonic incremental anonymization property. The method uses binary anonymization tree which resembles kd-tree. It also insisted that the group contains a minimum of k tuples and a maximum of (2k-1) distinct tuples. The general idea is to progressively and consistently reduce the generalization granularity as incremental updates arrive. The approach guarantees the k-anonymity on each release and also on the inferred table using multiple releases. At the same time, the new approach utilizes the accumulated data to reduce the information loss.

Serial publishing for incremental datasets is often necessary when there are insertions, deletions and updates in the microdata. In serial publishing, the anonymization mechanism for serial publishing should provide individual-based protection. Two methods namely HD-composition (Bu *et al*., 2008) and individual privacy (Wong *et al*., 2009) were proposed.

Bu *et al*. (2008) proposed an anonymization method called HD-composition which involves two major roles, namely holder and decoy. The objective is to bind the probability of linkage between any individual and any sensitive value by a given threshold, e.g., 1/l. The authors proposed two major principles for partitioning: role-based partition and cohort-based partition. By role-based partition, in every anonymized group of the published data, for each holder of a permanent sensitive values, l-1 decoys which are not linked to s can be found. Thus, each holder is masked by l-1 decoys. By cohort-based partition, for each permanent sensitive values, construct lcohorts, one for holders and the other l-1 for decoys, restrict that decoys from the same cohort cannot be placed in the same partition, this is to imitate the properties of true holders. In total, this study presents a study of the problem of privacy preserving serial data publishing with permanent sensitive values and incremental registration lists.

Wong *et al*. (2009) proposed the privacy guarantee for transient sensitive values which was called the global guarantee. The researchers showed that the anonymized group sizes used in the data anonymization was a key factor in protecting individual privacy in serial publication. Two strategies for anonymization targeting at minimizing the average group size and the maximum group size are proposed.

Incremental publishing assumes the raw data of any previous records can be inserted, deleted and updated. A new release may contain some new records and few updated records. Every release is the data collected at each instance of time. Based on rigorous theoretical analysis, Xiao and Tao (2007) developed a new generalization principle m-invariance (Pei *et al*., 2007) that effectively limits the risk of privacy disclosure in re-publication. They accompanied the principle with an algorithm which computed privacy-guarded relations that permitted retrieval of accurate aggregate information about the original microdata.

Definition 1 (Key attribute): An attribute that uniquely identifies the tuple such as PID in Table 1 is a key attribute.

Table 1: Micro dataset

R#	PID	Job	Gender	Age	Education	Disease
1	P01	Professor	Female	21	Preschool	HIV
2	P02	Teacher	Male	48	HS-Grad	Tuberculosis
3	P03	Lawyer	Female	26	Bachelors	Pneumonia
4	P04	Police	Female	31	Some-college	Gastritis
5	P05	Engineer	Male	24	Doctorate	Flu
6	P06	Surveyor	Male	35	Bachelors	Pyrexia
7	P07	Philosopher	Female	22	Doctorate	HIV
8	P08	Clergy	Male	41	HS-Grad	Pneumonia

Definition 2 (Quasi-identifier): A set of non-sensitive attributes QID = $\{q_1, ..., q_n\}$ of a table is called a quasi-identifier if these attributes can be linked with external data to uniquely identify at least one individual in the general population.

Definition 3 (Sensitive attribute): The attributes that should not be disclosed directly to the public or may be disclosed after disassociating its value with an individual's other information.

MATERIALS AND METHODS

Privacy preserving incremental data dissemination using optimal generalization

Optimal generalization: Taxonomies are usually used to perform generalization. The QIDs are generalized using the taxonomies. In optimal generalization, instead of considering all the attributes that contribute to QIDs we focus only on those attributes that generalize the values on balancing the utility and privacy. If all the attributes of QID are considered for generalization without reducing the dimensionality, it definitely increases the computational effort. Optimal generalization slightly increases the complexity but is relatively less complex when comapred to considering all the attributes for QID and also balances the privacy and utility which is of highest concern. For the identified QID attributes, the QID-power set is generated as shown. QID-Powerset = {<Job>, <Gender>, <Age>, <Education>, <Job, Gender>, <Job, Age>, <Job, Education>, <Gender, Age>, <Gender, Education>, <Age, Education>, <Job, Gender, Age>, <Job, Gender, Education>, <Job, Age, Education>, <Gender, Age, Education>, <Job, Gender, Age, Education>.

The generalization is performed based on each combination of QID-power set. The values are recorded. The values that balance the utility and privacy is chosen as the ultimate QID generalization attribute. This is elaborated in the following section in detail.

System architecture: The simple architecture for disseminating incremental microdata is presented in Fig. 2 which publishes data only after privacy compliance verification process to ensure safe release and endorses minimum privacy breach. The privacy compliance verifier in the model addresses the linking probabilities. The clusters that do not satisfy the privacy compliance verification would be merged with one of the safe release cluster for publishing.

The data publishing considers only those tuples where the user has given his consent to release. Here, the user consent is considered as a binary value: a value 'TRUE' refers to publishing the tuple after anonymization and a value 'FALSE' implies the tuple should not be be published and hence suppressed.

Definition 4 (Safe release): This specifies the possibility of publishing the dataset with certain assurance (threshold). If the probability of linking is above the threshold parameter (ε) depending on the type of linking, then the publishable data is considered to be a safe release. This is given by the following equation:

$$\text{Safe release}\left(T'\right) \equiv \text{PL}\left(T'\right) > \varepsilon$$

Where:

T' = The anonymized table
PL = The probability of linking
ε = The user defined threshold

Definition 5 (Threshold parameter <ε>): This is an user defined parameter and is used to check the possibility of publishing the anonymized data version. This relies on the size of the dataset.

Definition 6 (High Sensitive Attribute-HSA): These are sensitive attributes that are to be disclosed only after transforming the value into a more general format using the taxonomy for all the tuples where the user's consent is "TRUE". For example, HIV is a high sensitive attribute.

Definition 7 (Low Sensitive Attribute-LSA): These are sensitive attributes that can be disclosed directly for all the tuples where the user's consent is "TRUE". For example, Pyrexia is a low sensitive attribute.

Definition 8 (Privacy breach): The knack of an attacker to gain knowledge from a recently published dataset T'_{n-1} above a threshold value ε, using the previously published datasets $\{T'_{n-1}, T'_{n-2}, ..., T'_0\}$ is known as privacy breach P_b.

Problem definition: Let $\{T'_0, T'_1, ..., T'_{n-1}\}$ be the set of released datasets. The problem of incremental data publishing is to publish $\{T'_n\}$ such that the $P_b\left(T'_{n-1}\right) \leq \lambda$ that also satisfies diversity.

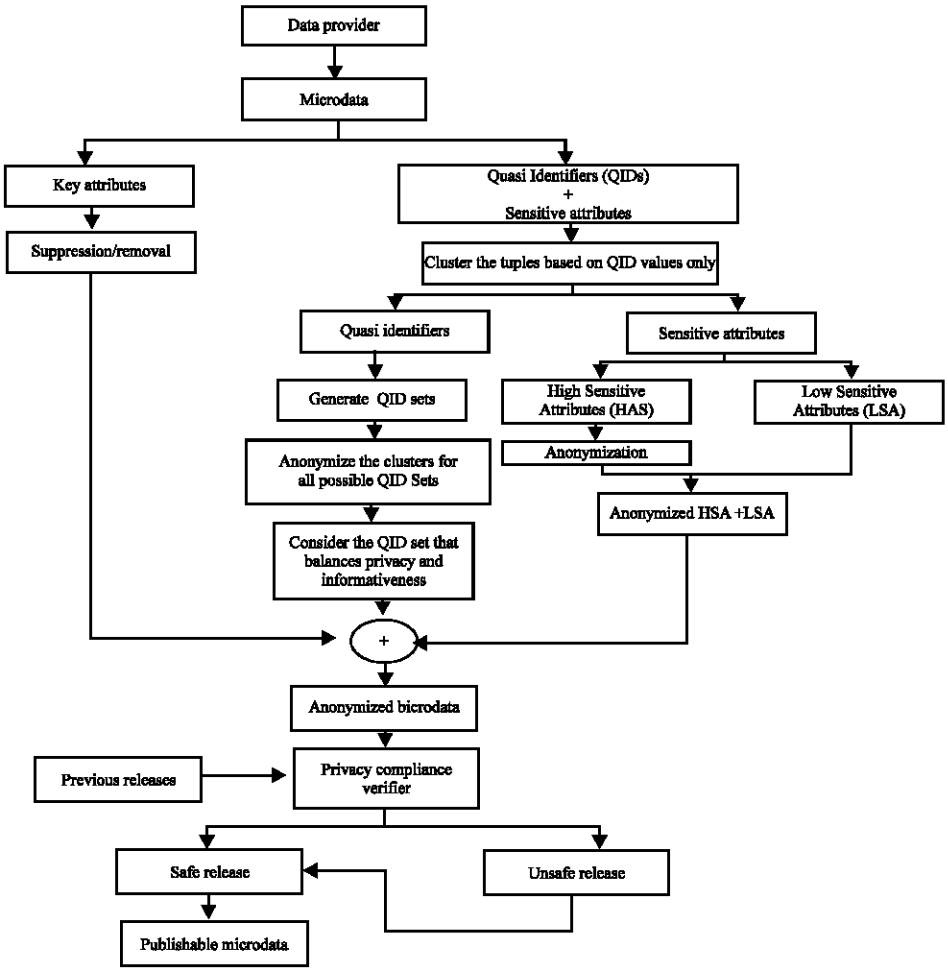

Fig. 2: Privacy preserving incremental data dissemination

Table 2: Output of phase-1

PID	PID	Job	Gender	Age	Education	Disease
P01	***	Professor	Female	21	Preschool	HIV
P02	***	Teacher	Male	48	HS-Grad	Tuberculosis
P03	***	Lawyer	Female	26	Bachelors	Pneumonia
P04	***	Police	Female	31	Some-college	Gastritis
P05	***	Engineer	Male	24	Doctorate	Flu
P06	***	Surveyor	Male	35	Bachelors	Pyrexia
P07	***	Philosopher	Female	22	Doctorate	HIV
P08	***	Clergy	Male	41	HS-Grad	Pneumonia

The method is carried out phase by phase. In phase-1, the microdata is simply split into two independent sets namely, <key identifiers> and <quasi-identifiers, sensitive attributes>. The key identifiers are either suppressed or removed. The output of phase-1 is shown in Table 2. In phase-2, the obtained dataset, ignoring key attributes is divided into 'k' clusters using the algorithm Alg_2 as depicted in Table 3.

The clusters are further split into two sub-clusters where the first subcluster consists of values related to quasi-identifiers and the second subcluster comprises of sensitive attributes in phase-3. The output of phase-3 is (2*k) subclusters. This output is presented in Table 4.

Table 3: Output of phase-2

Job	Gender	Age	Education	Disease
Professor	Female	21	Preschool	HIV
Lawyer	Female	26	Bachelors	Pneumonia
Engineer	Male	24	Doctorate	Flu
Philospher	Female	22	Doctorate	HIV
Police	Female	31	Some-College	Gastritis
Surveyor	Male	35	Bachelors	Pyrexia
Teacher	Male	48	HS-Grad	Tuberculosis
Clergy	Male	41	HS-Grad	Pneumonia

Table 4: Output of phase-3

Job	Gender	Age	Education	Disease
Professor	Female	21	Preschool	HIV
Lawyer	Female	26	Bachelors	Pneumonia
Engineer	Male	24	Doctorate	Flu
Philospher	Female	22	Doctorate	HIV
Police	Female	31	Some-college	Gastritis
Surveyor	Male	35	Bachelors	Pyrexia
Teacher	Male	48	HS-Grad	Tuberculosis
Clergy	Male	41	HS-Grad	Pneumonia

Table 5: Privacy gain and information loss for first subcluster

QID subset	Privacy gain	Information loss
<Job>	0.63	0.29
<Gender>	0.56	0.36
<Age>	0.59	0.32
<Education>	0.64	0.23
<Job, Gender>	0.65	0.25
<Job, Age>	0.69	0.40
<Job, Education>	0.68	0.41
<Gender, Age>	0.71	0.39
<Gender, Education>	0.74	0.35
<Age, Education>	0.73	0.31
<Job, Gender, Age>	0.87	0.30
<Job, Gender, Education>	0.82	0.23
<Job, Age, Education>	0.80	0.31
<Gender, Age, Education>	0.84	0.36
<Job, Gender, Age, Education>	0.88	0.40

This is followed by generation of power set for the QIDs. The basic idea behind generating subset of QIDs is to overcome the curse of dimensionality and to choose a subset that balances privacy and information loss. If QID consists of too many attributes then it would be difficult to anonymize all the attribute values as it consumes more time. So, we consider a subset of QID to handle curse of dimensionality. While choosing the subset, we also consider the amount of privacy that the subset provides and the information loss that incurs. The subset that balances both would be taken into consideration.

In phase-4, each subcluster of QID values are anonymized for each QID subset. The QID subset that shows more privacy and less information loss is chosen as the QID for that cluster. There is every scope for the QID to change for another subcluster. This might consume some time but the eventual goal of maximizing privacy and minimizing the information loss is preserved. In the similar manner, the sensitive attribute values are divided into High Sensitive Attribute values (HSA) and Low Sensitive Attribute values (LSA). The low sensitive attribute values are the values that can be directly disclosed without any anonymiziation. For example, in Table 1 where disease is a senstive attribute, Pyrexia is a low sensitive attribute and can be directly disclosed. The high sensitive attribute values are the values that needs to be anonymized using the taxonomies. The anonymized HSAs and LSAs are then merged. The output of phase-4 is presented in Table 5 and 6. Table 5 presents a scenario of values for the first subcluster. Table 6 is highlights the sensitivity of the disease, i.e., HSA and LSA.

Phase-5 concatenates suppressed key attributes, anonymized QIDs and the sensitive attributes. The output of phase-5 is depicted in Table 7-10. This forms the anonymized data which is ready to be to be published. In phase-6, the output of phase-5 dataset is sent for privacy compliance verification. The outcome of the privacy compliance verfication function would be either a safe release or an unsafe release. Phase-5 output is published if it passes the privacy compliance verification. If the outcome is a safe release then the anonymized microdata is disseminated. The unsafe releases are not published immediately and are merged with one safe releases and are then released.

Table 6: Sensitivity of disease

Disease	HSA(Y/N)
HIV	Y
Tuberculosis	Y
Pneumonia	Y
Gastritis	Y
Flu	N
Pyrexia	N

Table 7: Anonymized subcluster (QID: <Job, Gender, Age>)

PID	Job	Gender	Age	Education	Disease
***	Any job	Female	21-22	Preschool	Disease
***	Any job	Person	24-26	Bachelors	Respiratory system disease
***	Any job	Person	24-26	Doctorate	Flu
***	Any job	Female	21-22	Doctorate	Disease

Table 8: Anonymized subcluster (QID: <Gender, Age>)

PID	Job	Gender	Age	Education	Disease
***	Police	Person	31-35	Some-college	Stomach related disease
***	Surveyor	Person	31-35	Bachelors	Pyrexia

Table 9: Anonymized subcluster (QID: <Job, Age>)

PID	Job	Gender	Age	Education	Disease
***	Any job	Male	41-48	HS-Grad	Lung related disease
***	Any job	Male	41-48	Hs-Grad	Respiratory system disease

Table 10: Attributes for adult census dataset

Attribute	Types	# leaves
Age	Continuous	17-90
Work class	Categorical	8
Education	Categorical	16
Marital status	Categorical	7
Occupation	Categorical	14
Race	Categorical	5
Sex	Categorical	2
Native country	Categorical	40

For the data considered in Table 1, the number of QID attributes are four and so the possible combinations are (2^4-1). In general the number of combinations is given by 2^n-1 where 'n' is the number of quasi-identifier attributes (Algorithm 1).

Algorithms for incremental cluster-based anonymization

Algorithm 1; Algorithm for privacy preserving incremental data publishing:

Input: Dataset D

Output: Clusters C_0, C_1, C_2, ..., C_N of anonymized records

1. Split the dataset into two tables T_1 and T_2
   { T_1 contains key attributes and T_2 contains <QIDs, SAs>}
2. Suppress the key attributes in T_1
3. Cluster T_2 based on QID values only
4. for each cluster C_i

5. split the cluster into 2 sub-clusters vertically
{ C_{i1} consists of QIDs and C_{i2} consists of SAs}
6. Generate QID power set
7. for each QID subset of power set
8. Anonymize the cluster C_{i1}
9. end for
10. Consider the anonymized version that gives more privacy and less information loss
11. Split C_{i2} into HSAs and LSAs
12. Anonymize HSAs
13. Merge anonymized HSAs and LSAs
14. Concatenate output of step 8 and 13
15. Check for privacy violation
16. If privacy violation check is passed then it is considered as safe release
else the set is considered as unsafe
17. end for
18. Merge the unsafe release clusters with one of the safe release clusters and publish

Algorithm 2; Algorithm for clustering for incremental data:

Input: Dataset D` after suppressing the key identifiers
Output: Clusters $C_0, C_1, C_2, ..., C_N$
1. Tuple t_0 forms a singleton cluster C_0
2. Identify the cluster that can host the subsequent tuples (t_i)
Compute the similarity between the new tuple and the cluster C_i
Based on the value obtained from above
If there exists a cluster (Cj) that can host new tuple, then add the new tuple t_i to the cluster C_j
Otherwise, create a new cluster for t_i
3. If a cluster C_j does not have the required number of tuples, merge the cluster with
the cluster that is similar

Proposition 1: $\forall_{c_i \in C}, \cup_{i=1}^{k} C_i = T, \forall i \neq j$ and $1 \leq i, j \leq k, C_i \cap C_j = \varphi$
The algorithm uses the taxonomies for anonymization. The taxonomy for High Sensitive Attribute (HSA) is shown in Fig. 3. The numerical values in each node specify the number of descendants each node has. These values are useful in computing the privacy gain.

Privacy gain: Let, T_0 be the microdata table to be published. T_0 contains 'd' attributes: $A = \{A_1, A_2, ..., A_d\}$ and their attribute domains are $\{D[A_1], D[A_2], ..., D[A_d]\}$. A tuple $t \epsilon T$ can be represented as $t = (t[A_1], t[A_2], ..., t[A_d])$ where $t[A_i]$ $(1 = i = d)$ is the A_i value of t. Table 1 is the sample microdata considered to illustrate the notion. The tuples are distributed into diverse clusters. A tuple partition consists of several subsets of T such that each tuple belongs to exactly one subset. Specifically, let there be 'k' clusters $C_1, C_2, ..., C_k$ then $\cup_{i=1}^{k} C_i = T$ and for any $1 \leq i_1 \neq i_2 \leq k$, $C_{i_1} \cap C_{i_2} = \varnothing$.

The adversary is initially interested in finding the matching cluster. Once the matching cluster is identified, the adversary computes the conditional probability p_i (SA|QI) that measures the strength of the link between QID and SA:

$$P_i(SA|QI) = \frac{\sum_{\langle q \in QI,\ a \in SA \rangle \in T'} \langle q, a \rangle}{\sum_{\langle a \in SA \rangle \in T'} \langle a \rangle} \quad (1)$$

The privacy gain metric depends on the diversity parameter and the generalization taxonomy. The strength of privacy is given by the distribution of sensitive values for each equivalence class and the anonymization of quasi-identifier attribute values.

During data publishing, the values of the QI are modified to the higher levels in the taxonomy. Depending on the anonymity parameter λ_1, the respective taxonomy is used. The privacy gain depends on the distribution of records and the $\{\lambda_1, \lambda_2\}$ parameters.

Anonymity property 1: Given $t \epsilon B$, $PG_{(t)} = A_{(t[QI])} + PG_{(t[SA])}$ where $A_{(t[QI])}$ is the anonymization of QI in a bucket, $PG_{(t[SA])}$ is the privacy gain of a value in SA and $PG_{(t)}$ is the privacy gain of the total tuple. The privacy gain for quasi-identifier, QI is given by:

$$A_{(t[QI])} = \frac{1}{|QI|} \sum_{q_i \in QI} \frac{1}{\lambda_1} \times \frac{(\max_{subtree} - \min_{subtree})}{q_i} \quad (2)$$

Where:
$\max_{subtree}$ = The maximum value at the subtree
$\min_{subtree}$ = The minimum value at the subtree
q_i = The ith value to be anonymized

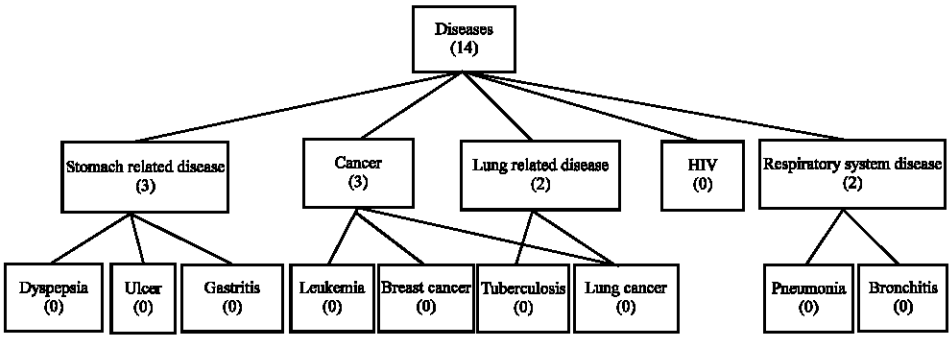

Fig. 3: Taxonomy for Sensitive Attribute Disease (HSA)

The anonymity of any $q_i \in QI$ lies in the range [0, 1]. '0' specifies that no privacy gain is being provided and '1' implies complete privacy gain. The average of the privacy gains of all QI values is the total privacy gain of the tuple. The privacy gain for sensitive attribute, SA is given by:

$$PG_{(t[SA])} = \frac{1}{|SA|}\sum_{s_j \in SA}\frac{N_{s_j}}{N_T\text{-}1} \quad (3)$$

where, N_{s_j} is the no of descendants the node has in the subtree which is given in parenthesis for each node in the taxonomy in Fig. 3 and N_T is the total number of nodes in the taxonomy which is given in parenthesis in the root of the taxonomy in Fig. 3. '0' specifies that no privacy gain is being provided and '1' implies complete privacy gain.

Anonymity property 2: $\forall_{qi \in QI,\, sj \in SA}\, PG_{(qi)} \in [0, 1]$ and $PG_{(sj)} \in [0, 1]$.

Information loss: The given tuple is classified into two parts namely QI and SA. The information loss is considered individually for all the attributes and later aggregated to compute the information loss of the total tuple. To minimize information loss, optimal anonymous tuples are generated using incremental cluster-based generalization.

Anonymity property 3: Given $t \in C$, $IL_{(t)} = IL_{(t[QI])} + IL_{(t[SA])}$ where $IL_{(t[QI])}$ is the information loss of each $qi \in QI$; $IL_{(t[SA])}$ is the information loss of each $s_j \in SA$ and $IL_{(t)}$ is the information loss of the total tuple when the tuple is ready to be published.

As incremental cluster-based generalization is engaged for anonymizing QI, the QI values undergo certain changes. This incurs certain loss of originality thus providing certain amount of information loss. So, in these terms:

$$IL_{(t[QI])} = \frac{1}{|QI|}\sum_{q_i \in QI}\frac{H_{q_i}}{H_T\text{-}1} \quad (4)$$

Where:

H_{qi} = The height of the generalized node that is, the subtree
H_T = The height of the tree
q_i = The ith value to be anonymized

For sensitive values, the values being published are the values in the next higher level of taxonomy on considering the diversity parameter. The leaf node in the taxonomy specifies the actual value of the sensitive attribute. So, $IL_{(t[SA])}$ is computed as shown as:

$$IL_{(t[SA])} = \frac{1}{|SA|}\sum_{s_j \in SA}\frac{H_{s_j}}{H_T\text{-}1} \quad (5)$$

Where:

H_{sj} = The height of the generalized node that is the subtree
H_T = The height of the tree
s_j = The jth value to be anonymized

RESULTS AND DISCUSSION

In this study, we consider the privacy breach with respect to incremental data publishing. If, F is any anonymity function and T_0 is the initial dataset to be published then the anonymized version is written as. $T_0 \xrightarrow{F} T_0'$. The experiments were performed on an Intel i5 processor machine with 8 GB of RAM. The operating system on the machine was Microsoft Windows 10. The implementation of the method was built and run in Java and the graphs were drawn in RStudio. The dataset used in our experiments was the adult census dataset from the Irvine machine learning repository (Lichman, 2013), since, this dataset was the closest to a common k-anonymization "benchmark" that, we are aware of. The actual dataset consists of 14 attributes with 48442 tuples. It has missing values also. This dataset used for result analysis consists of 8 attributes and 30,162 records. These are age, work class, education, marital status, occupation, race, gender and native country. Records with missing values are discarded because of limitations in our prototype system. The table structure is defined in Table 10. As the size of the census dataset is not too large to verify our method against incremental datasets, we created synthetic dataset consisting of {Gender, age, postal code} as the quasi-identifier and disease as the sensitive attribute. The initial synthetic dataset consisted of 1,00,000 records. The size of the synthetic dataset was gradually increased by 100%. We tested our method both on adult census data and synthetic dataset.

To analyze our model in terms of computational effort, we have implemented the model with incremental datasets. These programs have been tested by using the dataset that was taken from UCI machine learning repository and the synthetic dataset. For comparison purpose, we have taken different sizes of datasets where the size of the initial dataset is 10000. Proportionately, the dataset is increased by 100%. As, we are dealing with personalized privacy, the user consent is also taken into consideration. We go with a simple assumption for Fig. 4 and 5 that all the users have given their consent as "TRUE" and so, 100% reveal of data takes place. Figure 4 depicts the computational effort for the UCI dataset whereas presents the computational effort for

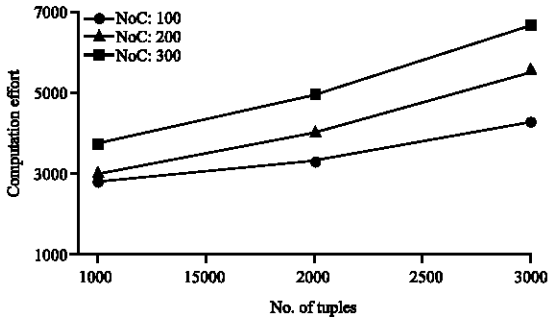

Fig. 4: Computational effort of different dataset sizes-UCI dataset

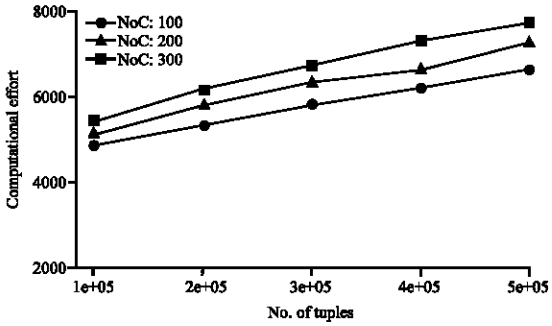

Fig. 5: Computational effort of different dataset sizes-Synthetic dataset

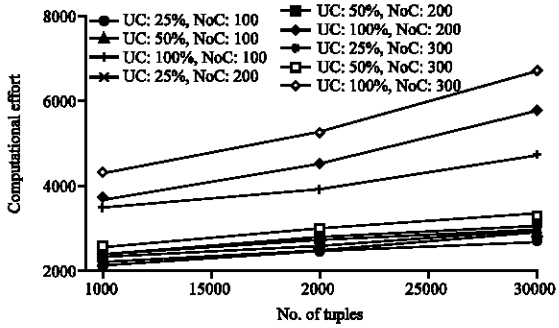

Fig. 6: Computational complexity with (25, 50 and 100%) privacy for NoC = (100, 200, 300)-UCI dataset

synthetic dataset for different number of clusters (NoC: 100, 200, 300). It is observed from the graphs that the complexity increases as the number of clusters increase.

Figure 6 and 7 highlight the computational effort for different levels of user consent. The graphs represent the user consent of 25, 50 and 100%. The number of clusters considered for analysis is 100, 200 and 300 for UCI dataset and the synthetic dataset. The computational effort increases with the number of clusters and level of user consent. Figure 8 and 9 propose the tradeoff between privacy and utility. Finally, Fig. 10 presents the

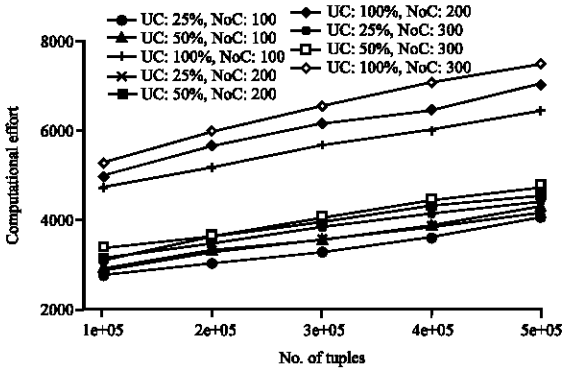

Fig. 7: Computational complexity with (25, 50 and 100%) privacy for NoC = (100, 200, 300)-Synthetic dataset

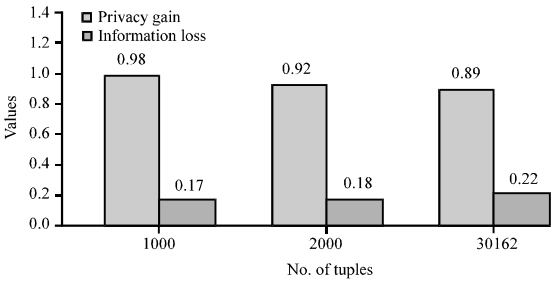

Fig. 8: Privacy gain vs. information loss for incremental datasets-UCI dataset

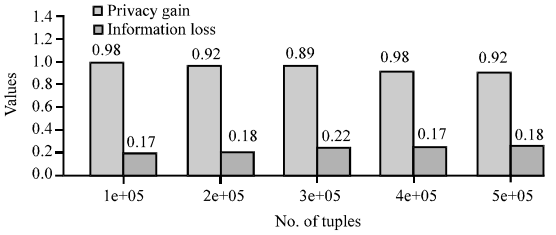

Fig. 9: Privacy gain vs. information loss for incremental datasets-Synthetic dataset

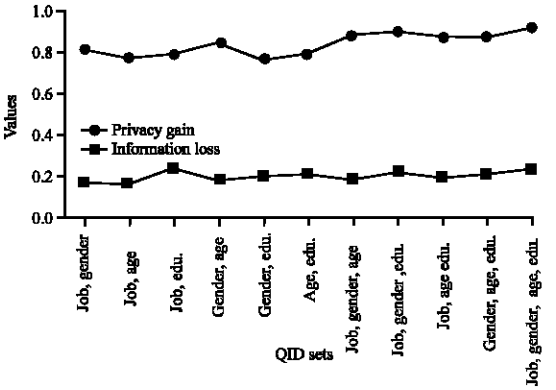

Fig. 10: QID sets-privacy gain vs. information loss (UCI dataset)

percentage of privacy gain and information loss for UCI dataset for all possible subsets QID of size greater than one.

CONCLUSION

This model brings out a practical problem of maintaining anonymity against incremental updates and proposes an effective solution. Maintaining anonymity against various types of incremental updates is an important and practical problem. Although, good progress on some scenarios have been made by Xiao and Tao (2007) and Cao *et al.* (2008) and this study, the problem at large remains open and challenging. We have provided a simple solution using incremental clustering for maintaining privacy in incremental datasets. We have also applied anonymization by considering QID subsets and choosing the one that provides more privacy and less information loss. By introducing incremental clustering, we provide an effective approach to generate anonymous datasets. We also checked for diversity in our model.

This research motivates several directions for future research. First, in this study, we consider incremental clustering based on quasi-identifiers. An extension is the notion of considering all the attributes that add to anonymity. This might provide better data utility but the privacy implications need to be carefully examined and understood. It is interesting to study the tradeoff between privacy and utility (Xu *et al.*, 2008). Second, we would like to use graph anonymization and compare our method. Third, we would like to extend clustering for handling high-dimensional data and preserve data utility. The idea can also be extended for anonymizing transaction databases which has been studied by Inan *et al.* (2009). Fourth, the incremental data publishing should be checked with respect to multiple sensitive attributes.

Finally, while a number of anonymization techniques have been designed, it remains an open problem on how to use the anonymized data. Another direction to design data mining tasks using the anonymized data (Fung *et al.*, 2005) computed by various anonymization techniques. This can be further extended to all the principles specified by Machanavajjhala (2007).

ACKNOWLEDGEMENTS

We thank the anonymous referees for their careful reading of the study and their valuable comments that significantly improved its quality.

REFERENCES

Bayardo, R.J. and R. Agrawal, 2005. Data privacy through optimal k-anonymization. Proceedings of the 21st International Conference on Data Engineering ICDE 2005, April 5-8, 2005, IEEE, London, USA., pp: 217-228.

Blum, A., K. Liggett and A. Roth, 2008. A learning theory approach to non-interactive database privacy. Proceedings of the 40th Annual ACM Symposium on Theory of Computing, May 17-20, 2008, ACM, Victoria, British, Columbia, ISBN:978-1-60558-047-0, pp: 609-618.

Bu, Y., A.W.C. Fu, R.C.W. Wong, L. Chen and J. Li, 2008. Privacy preserving serial data publishing by role composition. Proc. VLDB. Endowment, 1: 845-856.

Byun, J.W., Y. Sohn, E. Bertino and N. Li, 2006. Secure Anonymization for Incremental Datasets. In: Secure Data Management, Jonker, W. and Petkovic M. (Eds.). Springer, Berlin, Germany, ISBN:978-3-540-38984-2, pp: 48-63.

Cao, J., B. Carminati, E. Ferrari and K.L. Tan, 2008. CASTLE: A delay-constrained scheme for ks-anonymizing data streams. Proceedings of the IEEE 24th International Conference on Data Engineering (ICDE'08), April 7-12, 2008, IEEE, Cancun, Mexico, ISBN:978-1-4244-1836-7, pp: 1376-1378.

Chawla, S., C. Dwork, F. McSherry, A. Smith and H. Wee, 2005. Toward Privacy in Public Databases. In: Theory of Cryptography, Kilian, J. (Ed.). Springer, Berlin, Germany, ISBN:978-3-540-24573-5, pp: 363-385.

Dwork, C., 2006. Differential privacy. Proceedings of the 33rd International Conference on Automata, Languages and Programming, July 10-14, 2006, ACM, Venice, Italy, ISBN:3-540-35907-9 978-3-540-35907-4, pp: 1-12.

Fung, B., K. Wang, A.W.C. Fu and J. Pei, 2008. Anonymity for continuous data publishing. Proceedings of the 11th International Conference on Extending Database Technology: Advances in Database Technology, March 25-29, 2008, ACM, Nantes, France, ISBN:978-1-59593-926-5, pp: 264-275.

Fung, B.C.M., K. Wang and P.S. Yu, 2005. Top-down specialization for information and privacy preservation. Proceedings of the 21st IEEE International Conference on Data Engineering, April 5-8, 2005, Tokyo, Japan, pp: 205-216.

Inan, A., M. Kantarcioglu and E. Bertino, 2009. Using anonymized data for classification. Proceedings of the IEEE 25th International Conference on Data Engineering (ICDE'09), March 29-April 2, 2009, IEEE, Shanghai, China, ISBN:978-1-4244-3422-0, pp: 429-440.

Kumari, V.V., S.R.P. Reddy, M.A. Sowjanya, B.J. Vazram and K.V.S.V.N. Raju, 2008. A novel approach for privacy preserving publication of data. Proceedings of the 2008 International Conference on Data Mining, July 14-17, 2008, DBLP, Las Vegas, USA, pp: 506-512.

LeFevre, K., D.J. DeWitt and R. Ramakrishnan, 2005. Incognito: Efficient full-domain K-anonymity. Proceedings of the ACM SIGMOD International Conference on Management of Data, June 14-16, 2005, ACM, Baltimore, Maryland, ISBN:1-59593-060-4, pp: 49-60.

LeFevre, K., D.J. DeWitt and R. Ramakrishnan, 2006. Mondrian multidimensional K-anonymity. Proceedings of the 22nd International Conference on Data Engineering (ICDE'06), April 3-7, 2006, IEEE, Atlanta, Georgia, USA., pp: 25-25.

Li, N., T. Li and S. Venkatasubramanian, 2007. T-closeness: Privacy beyond k-anonymity and l-diversity. Proceedings of the IEEE 23rd International Conference on Data Engineering ICDE 2007, April 15-20, 2007, IEEE, Istanbul, Turkey, ISBN: 1-4244-0802-4, pp: 106-115.

Lichman, M., 2013. UCI Machine Learning Repository. University of California, Irvine, California,.

Machanavajjhala, A., J. Gehrke, D. Kifer and M. Venkitasubramaniam, 2006. L-diversity: Privacy beyond K-anonymity. Proceedings of the 22nd International Conference on Data Engineering (ICDE'06), April 3-7, 2006, IEEE, Atlanta,Georgia, USA., pp: 24-24.

Mohammed, N., B. Fung, P.C. Hung and C.K. Lee, 2009. Anonymizing healthcare data: A case study on the blood transfusion service. Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, June 28-July 01, 2009, ACM, Paris, France, ISBN:978-1-60558-495-9, pp: 1285-1294.

Pei, J., J. Xu, Z. Wang, W. Wang and K. Wang, 2007. Maintaining K-anonymity against incremental updates. Proceedings of the 19th International Conference on Scientific and Statistical Database Management (SSBDM'07), July 9-11, 2007, IEEE, Banff, Alberta, Canada, pp: 5-5.

Rastogi, V., D. Suciu and S. Hong, 2007. The boundary between privacy and utility in data publishing. Proceedings of the 33rd International Conference on Very Large Data Bases, September 23-28, 2007, VLDB Endowment, Austria, ISBN: 978-1-59593-649-3, pp: 531-542.

Samarati, P., 2001. Protecting respondents identities in microdata release. IEEE. Trans. Knowl. Data Eng., 13: 1010-1027.

Sweeney, L., 2002a. Achieving k-anonymity privacy protection using generalization and suppression. Int. J. Uncertainty Fuzziness Knowledge-Base Syst., 10: 571-588.

Sweeney, L., 2002b. k-Anonymity: A model for protecting privacy. Int. J. Uncertain. Fuzziness Knowl. Based Syst., 10: 557-570.

Wang, K. and B. Fung, 2006. Anonymizing sequential releases. Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 20-23, 2006, ACM, Philadelphia, Pennsylvania, ISBN:1-59593-339-5, pp: 414-423.

Wang, K., B.C. Fung and S.Y. Philip, 2007. Handicapping attacker's confidence: An alternative to K-anonymization. Knowl. Inf. Syst., 11: 345-368.

Wang, K., Y. Xu, A.W. Fu and R.C. Wong, 2009. FF-anonymity: When quasi-identifiers are missing. Proceedings of the IEEE 25th International Conference on Data Engineering (ICDE'09), March 29-April 2, 2009, IEEE, Shanghai, China, ISBN:978-1-4244-3422-0, pp: 1136-1139.

Wong, R.C.W., A.W.C. Fu, J. Liu, K. Wang and Y. Xu, 2009. Preserving individual privacy in serial data publishing. CoRR., 1: 1-12.

Wong, R.C.W., J. Li, A.W.C. Fu and K. Wang, 2006. The (a, k)-anonymity: An enhanced k-anonymity model for privacy preserving data publishing. Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 20-23, 2006, ACM, New York, USA., ISBN:1-59593-339-5, pp: 754-759.

Xiao, X. and Y. Tao, 2006. Personalized privacy preservation. Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data, June 27-29, 2006, ACM, Chicago, Illinois, USA., ISBN:1-59593-434-0, pp: 229-240.

Xiao, X. and Y. Tao, 2007. M-invariance: Towards privacy preserving re-publication of dynamic datasets. Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data, June 11-14, 2007, ACM, Beijing, China, ISBN:978-1-59593-686-8, pp: 689-700.

Xu, Y., K. Wang, A.W.C. Fu and P.S. Yu, 2008. Anonymizing transaction databases for publication. Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 24-27, 2008, ACM, Las Vegas, Nevada, USA., ISBN:978-1-60558-193-4, pp: 767-775.

Zhang, Q., N. Koudas, D. Srivastava and T. Yu, 2007. Aggregate query answering on anonymized tables. Proceedings of the 23rd IEEE International Conference on Data Engineering Workshops, April 15-20, 2007, Istanbul, Turkey, pp: 116-125.