Enhanced Algorithm for Data Privacy Preservation using Data Anonymization with Low Information Loss in Public cloud

Amalraj Irudayasamy
Periyar University, Salem, T.Nadu, India

Arockiam Lawrence
St. Joseph's College, Trichy, T. Nadu, India

Abstract

Data are scattered in public cloud to share among the stake holders which produces much concern over the protection of individual privacy. Without revealing the delicate information, publishing data in a public cloud is a challenging aspect. To reveal data to the public leads to the introduction of many models like k-anonymity and l-diversity. These methods safeguard the data against the adversary by attaining the knowledge about the sensitive information. But, in these approaches there exists many shortcomings such as the information loss. The research focuses the issues only with the k-anonymity. An efficient framework is proposed for privacy preservation. Considering the information loss metrics, one-dimensional quasi identifiers are focused and studied the properties of optimal solutions for k-anonymity. To generalize, an efficient multi-dimensional quasi identifiers using space mapping methods are proposed. These experimental evaluations prove that the proposed approaches are more efficient in terms of execution time and information loss.

1. Introduction

The difficult of privacy preserving data publishing has presently gathered great attentions and the data owner bothers about the security concerns. Several organizations have practice of publishing and sharing data with other stake holders for analysis [1]. Sensitive information may be disclosed to the adversaries and chances are there to misuse the information. Generally, recognizing attributes are not unveiled to defend the private data. But, still many studies have revealed that mere unveiling the data is not sufficient to preserve the confidentiality. It is because of the existence of the quasi identifiers in the published data [2] [3] [4] [5] [6].

Quasi identifiers are those can be combined with information acquired from unlike sources in order to render the private records. Records with identical quasi-identifier values establish an equivalence class [7]. *K-anonymity* is commonly accomplished by generalization or suppression, which accidentally lead to information loss. Still, the data should remain to be helpful as possible. Hence a trade-off among privacy and information loss happens [8].

Freshly, several concepts were projected to address the shortcomings of *k*-anonymity. *K-anonymity* could disclose information which is more sensitive, when there are various matching *Sensitive Attribute (SA)* values inside an equivalence class. *K*-anonymity algorithm can be modified by adjustable equivalence class authentication complaint [9]. However, such a method may arrive at unnecessary information loss.

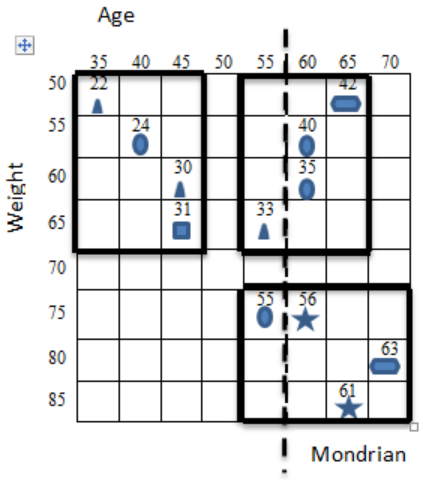
Consider the micro data in figure 1(a), where the combinations of age and weight are the quasi-identifiers and disease is the sensitive attribute. Let the required degree of anonymity be $k=4$. The current state-of-the-art *k*-anonymity algorithm i.e., Mondrian [10], categories the data points along each dimensions separately. Then the data points are segregated across the dimension with the broadest normalized range of values. In the given example, the normalized ranges for both dimensions are the similar. Mondrian selects the first one i.e., Age and splits into segments 35−55 and 60−70 which is shown in figure 1(b). Further segregating is not potential, because any division would result in groups with less than 4 records. A different method is proposed. Mapping of the multi-dimensional quasi-identifier to a *1-D* value is proposed. In the given example an 8 × 8 Hilbert space filling curve [11] is used. The resulting sorted *1-D* values are shown in Figure 1(a) (column *1D*). Next, the *1-D* space is partitioned. It is proved that the optimal *1-D* partitions are non-overlapping and contain between k and $2k - 1$ records.

Three groups which correspond to *1-D* ranges 22 through 31, 33 through 42 and 55 through 63 are obtained. The resulting *2-D* partitions are enclosed by three rectangles in figure 1(b). Note that the proposed method causes less information loss. For instance, there is a 1/12 chances for a person who weighs 65kg and is 45 years old, to suffer from pneumonia. According to Mondrian, the probability is only 1/40 which proves the proposed method of partitioning is more accurate.

However, there exists optimal partitioning consisting of only consecutive ranges with respect to each individual value of the sensitive attribute [12]. Based on the sensitive attribute property, a heuristic method of grouping records which are close to each other in the *1-D* space is proposed. But these groupings have different sensitive attribute values. From the consequence it is concluded for instance, that no person younger than 55 suffers from Alzheimer's. Obviously the resulting information loss is unacceptable [13]. Moreover, while the proposed technique resembles clustering, experiments show that existing clustering-based anonymization techniques are worse in terms of information loss and they are considerably slower [14].

Age	*Weight*	*Disease*	*1-D*
35	50	Gastritis	22
40	55	Diabetes	24
45	60	Gastritis	30
45	65	Pneumonia	31
55	65	Gastritis	33
60	60	Diabetes	35
60	55	Diabetes	40
65	50	Alzheimer	42
55	75	Diabetes	55
60	75	Flu	56
65	85	Flu	61
70	80	Alzheimer	63

(a) Original Micro data



(b) K-anonymous groups

Figure 1. k – anonymity

The rest of the paper is organized as follows: Section 2 comprises important definitions and reviews the related work. Section 3 presents the proposed solutions for the *k*-anonymity problem with respect to *1-D*. In section 4, extension of a multi-dimensional case using quasi identifiers are presented. Experimental evaluation is presented in Section 5 and the conclusion along with the future work is given in section 7.

2. Background and Related work

The terminologies and its related work are presented as follows:

2.1. Definition 1 (Quasi – identifier)

Given a database table T(A1,A2, . . . ,An), a quasi-identifier attribute set QT = {A1,A2, . . . ,Ad} $\subseteq$ {A1,A2, . . . ,An} is a minimal set of attributes, which can be joined with external information in order to reveal the personal identity of individual records [17].

A set of tuples which are fuzzy in the projection of T on Q_T is called *equivalence class*. Two commonly employed techniques to preserve privacy are generalization and suppression [15]. Generalization describes equivalence classes for tuples as multi-dimensional ranges in the Q_T space, and substitutes their actual Q_T values with a representative value of the entire range of the equivalent class (e.g., replaces the city with the state). Generalization ranges are usually specified by a generalization hierarchy, or taxonomy tree e.g., city, state, country. Suppression excludes some Q_T attributes or entire records (known as outliers) from the micro data, overall.

The privacy-preserving conversion of the micro data is stated to as recoding. Local recoding allows the same detailed value to be mapped to different generalized values in each equivalence class [16]. Local recoding is more flexible and has the prospective to attain inferior information loss [17]. The recoding method can also be categorized into *1-D*, where the mapping is accomplished for each attribute independently. In Multi-dimensional method, the cartesian products of multiple attributes are mapped. Multi-dimensional mappings are more accurate. A local recoding, multi-dimensional transformations is developed to support the research [18]. All privacy-preserving transformations cause information loss, which must be reduced in order to preserve the ability to extract meaningful information from the published data.

2.2 Information loss metrics

There are many metrics proposed on information loss. The *Height Metric* (*HM*) measures information loss centered on the summation of the generalization levels functional to all quasi-identifiers, which is autonomous of the actual input dataset [7]. However, it is not clear how *HM* can be extended to support general purpose applications. The *Precision Metric* (*PM*), on the other hand, measures information loss based on the average of the normalized generalization levels applied to all quasi-identifiers [3]. It is independent of the actual input dataset. , *PM* does not capture the distribution of records in the Q_T space. *Average Equivalence Class Size* and *Discernibility Metric (DM) metrics* in which the information loss based on the size of the equivalence classes resulting from a transformation is measured. The actual values of the quasi-identifiers in the input dataset are not considered. *Non-uniform Entropy* metric measures information loss based on the loss of entropy, i.e., information content. More accurate is the *Generalized Loss Metric* and the similar *Normalized Certainty Penalty (NCP)* [7] [10] [16].

For numerical attributes, the *NCP* of an equivalence class *G* is defined as

$$NCP_{ANum}\ (G) = \frac{max^{G}_{ANum} - min^{G}_{ANum}}{max_{ANum} - min_{ANum}}$$

where the numerator and denominator represent the ranges of attribute *ANum* for the class *G* and the entire table, respectively. In the case of categorical attributes, where no total order or distance function exists, NCP is defined with respect to the taxonomy tree of the attribute:

$$NCP_{ACat}(G) = \begin{cases} 0, & card(u) = \\ Card(u)/|A_{Cat}|, & otherwise \end{cases}$$

where u is the lowest common ancestor of all *ACat* values included in G, *card(u)* is the number of leaves (i.e., attribute values) in the sub-tree of u, and |*ACat*| is the total number of distinct *ACat* values. The *NCP* of class *G* over all quasi-identifier attributes is:

$$NCP(G) = \sum_{i=1}^{d}. NCP_{Ai}\ (G) \qquad (1)$$

where d is the number of attributes in Q_T (i.e., dimensionality). Ai is either a numerical or categorical attribute and has a weight w_i, where $\Sigma w_i = 1$. *NCP* measures information loss for a single equivalence class. Based on the above new metric, called Global Certainty Penalty *(GCP*) is introduced, which measures the information loss of the entire anonymized table. Let P be the set of all equivalence classes in the released anonymized table. GCP is defined as:

$$GCP\ (P) = \frac{d.N}{d.N}$$

where *N* denotes the number of records in the original table (i.e., micro data) and d is the dimensionality of Q_T. The advantage of the above formulation is its ability to measure information loss among tables with varying cardinality and dimensionality. Furthermore, *GCP* is between 0 and 1, where 0 signifies no information loss and 1 corresponds to total information loss.

2.3 k-anonymity

DEFINITION 2 (*K-ANONYMITY*) A database *T* with a quasi-identifier attribute set Q_T conforms to the *k*-anonymity property, if and only if each unique tuple in the projection of *T* on Q_T occurs at least *k* times [17]. An optimal solution to the k-anonymity problem should minimize information loss [18]. An optimal solution to the *k*-anonymity problem should minimize information loss. Formally:

Problem 1. Given a table *T,* a quasi-identifier set Q_T and a privacy bound expressed as the degree of anonymity *k*, determine a partitioning *P* of *T* such that each partition $G \in P$ has at least *k* records, and *GCP(P)* is minimized.

Meyerson and Williams proved that optimal k-anonymity for multi-dimensional quasi-identifiers is *NP-hard*, under both the generalization and suppression models [19]. For the latter, they proposed an approximate algorithm that minimizes the number of suppressed values. The approximation bound is *O(k·logk)*. Aggarwal et al. improved approximation bound to *O(k)* [2]. Several approaches limit the search space by considering only global recoding proposes an optimal algorithm for single-dimensional recoding with respect to the CM and DM metrics [16]. Incognito introduces a dynamic programming approach, which finds an optimal solution for any metric by considering all possible generalizations [9].

To address the inflexibility of single-dimensional recoding, Mondrian employs multi-dimensional global recoding, which achieves finer granularity [10]. Similar to *kd-trees*, Mondrian partitions the space recursively across the dimension with the widest normalized range of values. Mondrian can also support a limited version of local recoding. If many points fall on the boundary of two groups, they

may be divided between the two groups. Since Mondrian uses space partitioning, the data points within a group are not necessarily close to each other in the Q_T space e.g., points 22 and 55 in figure 1(b), causing high information loss.

Another family of multi-dimensional local recoding methods is based on clustering. *k*-anonymity is treated as a special clustering problem, called *r*-cellular clustering [17] [20]. A constant factor approximation of the optimal solution is proposed, but the bound only holds for the Euclidean distance metric. Furthermore, the computation and I/O cost may be high in practice. Many have proposed an agglomerative and divisive recursive clustering algorithms, which attempt to minimize the *NCP* metric [19] [20]. The latter called Top Down in the following is the better of the two. Top Down performs a two-step clustering: first, all records are in one cluster, which is recursively divided as long as there are at least *2k* records in each cluster. In the second step, the clusters with less than *k* members are either grouped together, or they borrow records from clusters with more than *k* records. The complexity of Top Down is $O(N^2)$. The research shows that Top Down is inefficient in terms of information loss and computational cost. The research shows that the proposed metric is efficient than the existing one in terms of information loss and computational cost.

3. Optimal 1d k-anonymity

In this section, an optimal solution to the k-anonymity problem with minimal information loss is presented. A quasi-identifier hash based approach is formulated in detail. At first, large-volume data sets are partitioned into a variety of relatively small data sets which are then stored in cloud data nodes [20]. Original generalized data sets are partitioned according to *QI-groups.* Similar *QI-groups* are mapped into the same data nodes. These *QI-groups* are then hashed by domain values in their quasi-identifiers in the current generalization level. Checking is done, whether *k*-anonymous status is violated and whether anonymized data sets are over-generalized.

Description: Dynamically maintain anonymity of an anonymized data set and overcomes the problem of *k*-anonymity for one-dimensional quasi-identifiers along with minimal information loss.

Input: A data set *D* and its already anonymized data set *D**, where *D** satisfies *k*-anonymity.

Output: Anonymized dataset with minimum information loss.

Step: 1 Hash generalized data set *D**

1.1 Partition *D** and *D* into *DNodes* = $\{DN_1, DN_2, \ldots, DNn\}$.

1.2 Establish quasi-identifier hash for *QI*-groups by domain value q ∈ $\bigcup_{i=1}^{m} Cuti$.

1.3 Count statistical information for links of *QI*-groups and sort *QI*-groups in each links in terms of their anonymity.

1.3.1 Set *D* in ascending order of *1-D* Q_T

for *i :=k* to 2*k*-1
Opt(i) = Opt_I([1,i])
prev(i) = *NIL* /* used to reconstruct solution */
for *i :=k* to 2*k* to *N*
for *j := max*{ *k,i-2k* + 1 } to *i- k*
Opt(i) = $\min_j$ **{** *Opt (j)* **+** *Opt ([j* + 1, *I])}*
prev(i) = *j* value that minimizes *Opt(i)*
i = N /* output *k –anonymized groups* */
while (*prev(i)* ≠ *NIL)*
output group with boundaries [*prev(i)* + 1 , *i*]
i =prev(i)
output group [1 , *i*]

Although the problem is *NP-hard* in the general case, the complexity is linear to the size of the input table for *1-D* quasi-identifiers [14]. Let D = $\{d_i\}$ 1≤ *I* ≤ *N* be the set of records in table *T*, where *N* = |*T*|. *D* is a totally ordered set according to the 1-D quasi-identifier Q_T. The goal is to compute a partitioning of *R* that minimizes *GCP* and satisfies the *k*-anonymity property.

An algorithm that computes the 1-D optimal *k*-anonymity partitioning of *D* needs only to consider groups with records that are consecutive in the Q_T space. If two groups with at least *k* records each overlap, the records can be swapped between them such that, the number of records in each group remains the same and the overlap is eliminated, without increasing *GCP*.

4. General multi-dimensional case

Here, the *1-D k*-anonymity algorithms are extended to multi-dimensional quasi-identifiers [22] [23]. Let Q_T be a quasi-identifier with *d* attributes i.e., d dimensions. The *d*-dimensional Q_T is mapped to one dimension and execute the *1-D* algorithms on the transformed data. Recall that optimal k-anonymity is *NP-hard* [13] [14] in the multi-dimensional case. The solutions obtained through mapping are not optimal. However, due to the good locality properties of the

space mapping techniques, information loss is low, as demonstrated experimentally in the forthcoming sections. The information loss of each *k*-anonymous group is measured using *NCP* and the information loss over the entire partitioning using *GCP*.

A well-known space-mapping technique, the Hilbert space-filling curve is employed. The Hilbert curve is a continuous fractal which maps each region of the space to an integer. With high probability, if two points are close in the multi-dimensional space, they will also be close in the Hilbert transformation [11]. Figure 10(a), for instance, shows the transformation from *2-D* to *1-D* for the 8 × 8 grid shown in the previous example the granularity of the regions can be arbitrarily small. The data set is totally ordered with respect to the *1-D* Hilbert value.

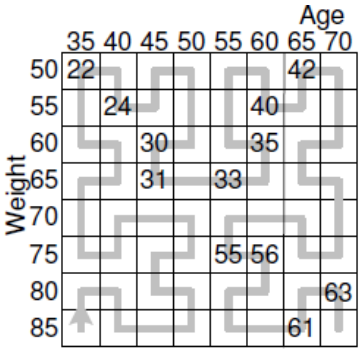


Figure 2. Hilbert Curve

The cluster centers are ordered according to any method (e.g., Hilbert ordering). Each data point is then assigned to its closest cluster center according to Euclidean distance. The data set is totally ordered with respect to the *1-D* Hilbert value.

Table 1. Census Dataset Characteristics

Attribute	Cardinality	Type
Age	79	Numerical
Gender	2	Hierarchical(2)
Educational Level	17	Numerical
Marital Status	6	Hierarchical(3)
Race	9	Hierarchical(2)
Work Class	10	Hierarchical(4)
Country	83	Hierarchical(3)
Occupation	50	Sensitive Value
Salary Class	50	Sensitive Value

Regardless of the technique, in order to perform the data mapping, each attribute value must be assigned to a number. For numerical attributes, the values are used directly. Furthermore, the semantic distance between two numeric attribute values can be measured as the difference between the two values. For categorical attributes and their associated taxonomy tree, labeling approach [3, 9] is adopted, where each attribute value is assigned to a distinct integer according to the in-order traversal of the taxonomy tree. If an equivalence class spans across different sub-trees, it is penalized according to *NCP*.

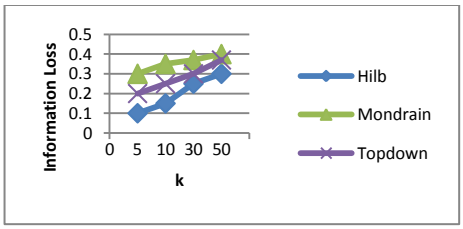
The overhead of the Hilbert mapping is *O(d)* per record, hence the method is efficient. *1-D* k-anonymity algorithms require the input to be sorted on Q_T , the cost is *O(N logN)*. Assuming sorted input, the proposed methods need to scan the data only once. Therefore the I/O cost is linear.

5. Experimental evaluation

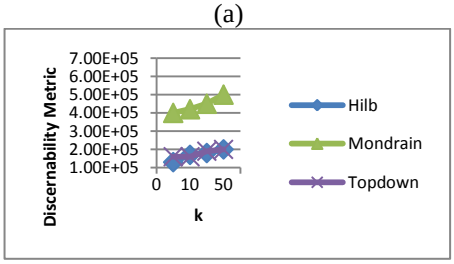
Evaluation of the proposed techniques against the existing methodology is carried out here. All algorithms are implemented in JAVA and the experiments were run on a leased cloud environment from amazon EC2. The workload consists mainly of the *CENSUS* [24] dataset, containing information of 500,000 persons. The schema is summarized in Table 1. There are nine attributes. The first seven represent the quasi-identifier Q_T, whereas the last two (i.e., Occupation and Salary) are the sensitive attributes *(SA)*. Two of the attributes are numerical and the rest are hierarchical. The number of levels in the taxonomy trees is shown in parentheses. Input tables with 50,000 to 400,000 records are generated by randomly selecting tuples from the entire dataset. Due to the small size of *ADULT*, the larger which is more realistic *CENSUS* dataset are used for most of the experiments. The *GCP* metric is used to measure the information loss. Recall that the values of *GCP* are in the range from 0 and 1. 0 is the best score i.e., no information loss.

5.1. k- Anonymity Evaluation

In the following experiments, the *1-D* optimal *k*-anonymity algorithm is compared with the existing techniques like Mondrian *k*- anonymity [10], and the Top Down clustering-based technique [20]. For optimal *1-D* algorithm, we consider the Hilbert with 12 bits per dimension. In the base version, partitioning is guided by accurate cost estimation at the original multi-dimensional space. The amortized complexity for calculating the cost is *O(d)*, where *d* is the dimensionality of Q_T . The algorithm estimates the cost at the *1-D* space in *O(1)* time. Since the algorithm is used only to estimate the real cost, the resulting information loss is expected to be higher.



(a)



(b)

Figure 3. Adult Dataset, variable k

First, the *ADULT* dataset which varies *k* between 5 and 50 are considered. The information loss is shown in Figure 3(a) *Hilb* outperform the existing methods. In Figure 3(b) the experiment is repeated using the *DM* metric, which was also used in the original Mondrian paper. *DM* is not particularly accurate to characterize information loss. Since *DM* considers only the partition size, these methods behave similarly, although they are considerably different in terms of information loss, which is also a symptom, that *DM* is not a suitable metric.

In the next experiment the *CENSUS* dataset [8] are used. The input size is set to N = 200, 000 records. K varies from 10 to 100. Figure 4 presents the results. Hilb achieved lower information loss, compared to Top Down and Mondrian, in all cases. Subsequently, for fixed k = 50, the records N varies from 50, 000 to 400, 000. Figure 5 shows the results. All methods manage to reduce information loss when the size of the input increases, since the data density becomes higher and the probability of finding good partitions increases. Hilb is better than Mondrian and Top-Down in all cases. The proposed method is also very fast when compared to the other methods. Hilb method is also suitable for real-life high dimensional data.

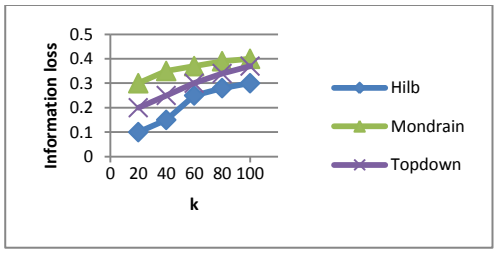


Figure 4. Census Dataset variable k

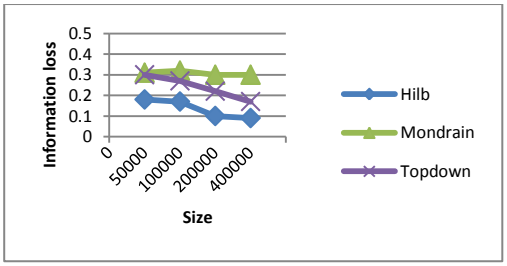


Figure 5. Census Dataset variable size

6. Conclusion and future work

It is demonstrated that, for k-anonymity, the proposed algorithms are superior to existing techniques in terms of information loss. Hilb is the best, but is a bit slower than Mondrian. It is by far superior in terms of information loss. For specific applications, other multi-dimensional to *1-D* mappings may be more appropriate. Any such mapping can be used in the proposed framework. Lastly, note that the projected methods scale well with the input size, since the computational complexity is linear, the required memory is constant and only one scan of the data is necessary.

In the research, a framework is developed for solving the k-anonymity problems, by mapping the quasi-identifiers to one dimension. However, a set of properties have been identified for the optimal *1-D* solution. Guided by these properties, efficient algorithms are developed at the *1-D* space. Popular transformations namely the Hilbert curve is used to solve the anonymity problem. In future, other transformations can easily be incorporated in the projected framework. The experiments demonstrate that the proposed methods clearly outperform the existing methods in terms of information loss. Moreover, the projected algorithms are linear to the input size. Therefore they are applicable to very large datasets. In the future, the dual problem shall be investigated, when given a maximum allowable information loss and identify the best possible degree of privacy.

7. References

[1] G. Aggarwal, T. Feder, K. Kenthapadi, S. Khuller, R. Panigrahy, D. Thomas, and A. Zhu, "Achieving Anonymity via Clustering", in Proc. of ACM PODS, pp 153–162, 2006.

[2] G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D. Thomas, and A. Zhu, "Approximation Algorithms for k-Anonymity", Journal of Privacy Technology, 2005.

[3] R. J. Bayardo and R. Agrawal, "Data Privacy through Optimal k-Anonymization", in Proc. of ICDE, pp 217–228, 2005.

[4] J. H. Friedman, J. L. Bentley, and R. A. Finkel, "An Algorithm for Finding Best Matches in Logarithmic Expected Time", ACM Transactions on Math. Software. 3(3):209–226, 1977.

[5] A. Froomkin, "The Death of Privacy. Stanford Law" Review, 52(5):1461–1543, 2000.

[6] V. Harinarayan, A. Rajaraman, and J. D. Ullman, "Implementing Data Cubes Efficiently", in Proc. Of ACM SIGMOD, pp 205–216, 1996.

[7] V. S. Iyengar, "Transforming Data to Satisfy Privacy Constraints" in Proc. of SIGKDD, pp 279–288, 2002.

[8] D. Kifer and J. Gehrke, "Injecting Utility into Anonymized Datasets",in Proc. of ACM SIGMOD, pp 217–228, 2006.

[9] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Incognito: Efficient Full-domain k-Anonymity", in Proc. of ACM SIGMOD, pp 49–60, 2005.

[10] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Mondrian Multidimensional k-Anonymity", in Proc. Of ICDE, 2006.

[11] B. Moon, H. Jagadish, and C. Faloutsos, "Analysis of the Clustering Properties of the Hilbert Space-Filling Curve" IEEE TKDE, 13(1):124–141, 2001.

[12] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Workload-aware Anonymization", in Proc. of KDD, pp 277–286, 2006.

[13] N. Li, T. Li, and S. Venkatasubramanian, "t-Closeness: Privacy Beyond k-Anonymity and l-Diversity", in Proc. of ICDE, pp 106–115, 2007.

[14] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "*L*-Diversity: Privacy Beyond k-Anonymity" , in Proc. of ICDE, 2006.

[15] P. Samarati and L. Sweeney, "Generalizing Data to Provide Anonymity when Disclosing Information" (abstract), in PODS (see also Technical Report SRI-CSL-98-04), 1998.

[16] J. Xu, W. Wang, J. Pei, X. Wang, B. Shi, and A. Fu, "Utility-Based Anonymization Using Local Recoding", in Proc. of SIGKDD, pp 785–790, 2006.

[17] L. Sweeney, "k-Anonymity: A Model for Protecting Privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems", 10(5):557–570, 2002.

[18] Y. Tao and X. Xiao," Personalized Privacy Preservation", in Proc. of ACM SIGMOD, pp 229–240, 2006.

[19] A. Meyerson and R. Williams, "On the Complexity of Optimal K-anonymity", in Proc. of ACM PODS, pp 223–228, 2004.

[20] R. Wong, A. Fu, J. Pei, K. Wang, S. Wan, and C. Lo, "Multidimensional k-anonymization by Linear Clustering Using Space-filling Curves", TR 2006-27, Simon Fraser University, March 2006.

[21] X. Xiao and Y. Tao, "Anatomy: Simple and Effective Privacy Preservation", in Proc. of VLDB, pp 139–150, 2006.

[22] Q. Zhang, N. Koudas, D. Srivastava, and T. Yu, "Aggregate Query Answering on Anonymized Tables", in Proc. of ICDE, pp 116–125, 2007.

[23] R. Zhang, P. Kalnis, B. C. Ooi, and K.-L. Tan, "Generalized Multidimensional Data Mapping and Query Processing", ACM TODS, 30(3):661–697, 2005.

[24] http://ww.ipums.org, "Census Dataset Characteristics", accessed on 14, Dec 2014.