Utility-Based k-Anonymization

Qingming Tang
Dept. of Computer Science
Fuzhou University
Fuzhou, China
tqm2004@gmail.com

Yinjie Wu
Dept. of Computer Science
Fuzhou University
Fuzhou, China
yjwu@fzu.edu.cn

Shangbin Liao
Dept. of Computer Science
Fuzhou University
Fuzhou, China
liaoshangbin@gmail.com

Xiaodong Wang
Dept. of Computer Science
Fuzhou University
Fuzhou, China
wangxd@fzu.edu.cn

***Abstract*—k-Anonymity is a well-researched mechanism for protecting private information released in Web. It requires that each tuple of a public released table must be indistinguishable from at least other $k-1$ tuples. Subject to this constraint, how to release data as useful as possible is challenge. Most previous works try to develop flexible anonymization method to reduce information loss, however, utility of released data is ignored.**

This paper studies utility-based k-anonymization. We first analyze deficiencies of previous global recoding model and the state-of-the-art local recoding model from utility view. Best of our knowledge, this is the first paper to evaluate the two models from such a view. Effectively combing both global and local recoding, we then propose a hybrid algorithm for utility based k-anonymization. The algorithm greedily partitions original table into non-overlapping sub-tables in global recoding phase, and then employ local recoding to go on divide each sub-table into smaller ones if possible. Experiments on famous adult data set show the utility and also information loss advantage of our algorithm towards the advanced algorithms proposed in recent related literatures.

***Keywords*-Privacy ; k-Anonymity; Utility; Network; Data Publishing**

I. Introduction

Lots of corporations, governments, institutes or even individuals release data tables which contain person-specific information to support knowledge discovery and various decision making. Directly releasing private data, especially that contains sensitive information(e.g. disease), violates personal privacy. Thus, before such kind of data is released, attributes such as name, social security number must be removed. These attributes are called *identifiers* for they can be used to explicitly identify a target tuple. However, just removing identifers is not safe because attackers can use other attributes, combining with external data tables, to re-identify the target tuple[1], [2], [3] or to infer useful information about the target [4], [5], [6].

For example, Table 1a is a medical table released by a hospital without name information while Table 1b is a voter list released by government. Voter list has the basic(unsensitive) information of each candidate. Identities of the tuples in Table 1a can be discovered or narrowed down to a small scope by joining Table 1a and Table 1b on $\{age,zipcode\}$. As shown in Table 1c, half of the tuples are re-identified and others suffer more serious privacy risk. Typically, just like $\{age,zipcode\}$, a minimal set of attributes of a table that can be used to be joint with external information to (partly) recover the identities of individual records is called *quasi-identifier*(QI).

To prevent such kind of problem, which is also called *linking attack* [1], [7], Sweeney and Samarati propose the k-anonymity principle[1], [2], [3]. This privacy principle requires that each tuple of a released table can not be identified with a probability higher than $\frac{1}{k}$, that is each tuple is indistinguishable from at least other $k-1$ tuples. Table 2a and 2b show two 2-anonymous tables of Table 1a. In the two tables, each tuple can not be identified with a probability higher than $1/2$ through *linking attack*. Note that, disease attribute is not changed because it belongs to personal privacy and thus is not possible to be published with an explicit identifier; also, to protect the integrity and security of such information is the goal of research works in this area.

The process to achieve k-anonymity is called *anoymization*. According to literatures, existing operations of anonymization includes *generalization* [1], [2], [7], *perturbation* [7], *suppression* [1], [2], [7] and *anatomization* [7]. *Generalization* is the most widely used one among the existing operations. This operation replaces(or recodes) a value with another less specific but semantically consistent value to hide details. For example, a writer may be replaced with artist, and a specific number may be replaced with a segment, just as shown in Table 2a and 2b. As anonymization process unfortunately leads to information loss, a common trend of previous works is to reduce information loss as much as possible. Thus, many metrics to effectively measure the information loss, such as *discernibility penalty*(DM) [7], [8], are proposed and many flexible recoding model for finding lower information loss results are developed. The ideal anonymization process is to find the optimal result under given information loss metric, however, finding optimal result is NP-hard in most recoding model under any non-trivial metric [7], [9], [8]. Thus, most efficient algorithms are based on greedy strategy or incomplete search.

Unlike most previous works, this paper studies k-anonymization from utility view. We show the interesting fact that lower information loss does not always lead to

Table I
EXAMPLE FOR LINKING ATTACK:(A) IS THE A TABLE WITHOUT NAME INFORMATION, (B) IS A VOTER LIST AND (C) IS CREATED BY JOINING (A) AND (B)

(a) Medical Table

Age	Zipcode	Disease
20	101	H1N1
20	101	HIV
20	101	flu
30	102	Pneumonia
30	103	HBV
40	102	HIV
40	103	dyspepsia
50	104	flu
50	104	flu
50	104	HIV

(b) Voter List

Name	Age	Zipcode
Linda	20	101
Bill	20	101
Sam	20	101
Hans	25	102
Sarah	30	102
Mary	30	103
Jacky	40	102
Tom	40	103
Ana	50	104
Bob	50	104
Jane	50	104

(c) Joint Table

Name	Age	Zipcode	Disease
Linda	20	101	H1N1, HIV or flu
Bill	20	101	H1N1, HIV or flu
Sam	20	101	H1N1, HIV or flu
Sarah	30	102	Pneumonia
Mary	30	103	HBV
Jacky	40	102	HIV
Tom	40	103	dyspepsia
Ana	50	104	HIV or flu
Bob	50	104	HIV or flu
Jane	50	104	HIV or flu

higher utility, and the state-of-the-art algorithms for k-anonymization violate the data utility though information loss is greatly reduced. We propose a hybrid algorithm for utility-based data publishing and show that our algorithm finds better anonymization results than the state-of-the-art algorithms from utility view.

In the end of this section, we briefly introduce the organization of this paper. In the next section, we introduce previous two recoding models: *global recoding* and *local recoding*. In section 3, we discuss what is utility based k-anonymization. In section 4, we show the drawbacks of the two previous recoding models and propose our own hybrid algorithm for data publishing. In section 5, we runs our algorithm over famous adult dataset. The experimental results show that our algorithm is better than previous related works. In section 6, we conclude the whole paper.

II. GLOBAL AND LOCAL RECODING

A public released table can be divided into several "anonymization groups", tuples in each group are indistinguishable by *linking attack*. In Table 2b, there are four groups and each group contains two tuples. Famous *global recoding* model requires that groups of a public released table must not be "overlapping" with each other. More specifically, given a table T, if we use D to denote the *Cartesian product* of all its attribute domains, then *global recoding* requires that each element of D can just be mapped to one generalized form.

A large number of previous algorithms, such as algorithms based on Sweeney and Samarati's framework [1], [2], [3], Bayardo and Agrawal's *K-Optimize* [8], Iyengar's *Genetic Algorithm* [10], Wang et al.'s *Bottom-Up Generalization* [11], Fung et al.'s *Top-Down Specialization* [12] and LeFevre et al.'s *Mondrian* [9] use *global recoding* to find reasonable or optimal anonymization result under some additional constraint. Among all these algorithms, LeFevre et al.'s *Mondrian* is a successful one both in information loss and runtime. This algorithm is based on greedy strategy, and it always find anonymization results with lower information loss than most other algorithms that also use *global recoding*.

In contrast to *global recoding*, *local recoding* is more flexible as it allows different anonymization groups of a released table to be overlapping. Generally speaking, this flexibility always leads to smaller anonymization groups comparing with *global recoding*, and thus significantly reduce the information loss according to previous information loss metrics. The state-of-the-art algorithms for k-anonymization using *generalization* employ *local recoding*, such as *Relaxed Mondrian* [9] and Xu et al.'s *Bottom-up Greedy Algorithm* [13].

III. UTILITY-BASED ANONYMIZATION

Most previous algorithms concern the information loss of the released table. These algorithms try to optimize the solution according to some information loss metrics. Low information loss algorithms always produce small anonymization groups. However, information loss metrics just concern the size of an anonymization group, that is the number of indistinguishable tuples within the group. But, in fact, when generalization is applied, a small anonymization group may have a too vague representation [13]. We can briefly explain this fact from geometric view. If total orders are set over each attribute domain, then all tuples are points in a multidimensional space and each anonymization group(with generalized form) is a multidimensional rectangular box. It is possible that points within a box distribute sparsely though the total number of points is small. In this case, the multidimensional box is huge(vague representation) though there is just a few points within it(small size).

Table II
GLOBAL RECODING AND LOCAL RECODING FOR TABLE 1A

(a) Global 2-anonymous Table

Age	Zipcode	Disease
20	101	H1N1
20	101	HIV
20	101	flu
30	[102-103]	Pneumonia
30	[102-103]	HBV
40	[102-103]	HIV
40	[102-103]	dyspepsia
50	104	flu
50	104	H1N1
50	104	HIV

(b) Local 2-anonymous Table

Age	Zipcode	Disease
20	101	H1N1
20	101	HIV
[20-30]	[101-102]	flu
[20-30]	[101-102]	Pneumonia
[30-40]	[102-103]	HBV
[30-40]	[102-103]	HIV
[40-50]	[103-104]	dyspepsia
[40-50]	[103-104]	flu
50	104	H1N1
50	104	HIV

In contrast with information loss metrics, there are some metrics that focus on the utility of the data. These metrics concern whether a released table is useful for queries or data mining. According to previous literatures, there are two easy and powerful utility metrics. One is the *normalized certainty penalty*(NCP) [13]. This metric takes the final representation of the anonymization group into account. That is, it concerns whether the "box" is small. Another metric is query answerability. Query answerability measures how precisely the generalized(anonymized) table can respond for queries. Original(un-generalized) data can answer queries very precisely. For an generalized table, the better query answerability, the data is closer to original dataset. Simply speaking, utility anonymization is to find high-utility anonymized dataset subject to some privacy requirements. In this paper, we focus on k-anonymity. We use NCP and query answerability to measure the utility of dataset.

IV. HYBRID ALGORITHM

A. Global vs Local

It has been believed by many that *local recoding* is much more powerful than *global recoding*. In turns of information loss, *local recoding* really is better because its flexibility always leads to anonymization groups with smaller size. If we set orders over each attribute domain of a given table, and treat each tuple as a point in a multidimensional space. Then a *global recoded* solution can be viewed as a strict partition of a multidimensional rectangular box, while a *local recoding* solution is a relaxed partition. That is, local recoding allows boxes(anonymization groups) intersect with its neighbors, just as Figure 1 shows. [9] has proven that, subject to k-anonymity, the average anonymization group size of a global recoded solution may exceed $2(k-1)d$ in worst case. Here, d is the attribute number of the given table. In contrast, the worst case for local recoded data is just $2k-1$.

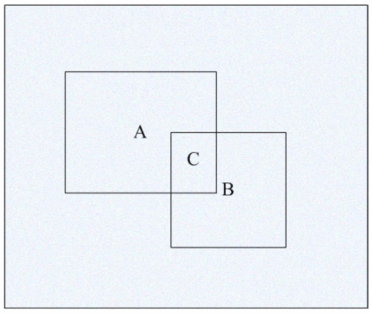

Figure 1. Two Intersecting Boxes(Groups) A and B. Here, C is there intersection

Thus, one trend of recent works is to employ *local recoding* to obtain an anonymized result with low information loss. However, most these works ignore one thing that the powerful flexibility of *local recoding* brings both lower information loss and deeper confusion. For example, Table 2a and Table 2b are optimal 2-anonymous tables for Table 1a. They uses global and local recoding respectively. We query the two anonymized tables eight times respectively using *age*=20, *age*=30, *age*=40, *age*=50 and *zipcode*=101, *zipcode*=102, *zipcode*=103, *zipcode*=104. Table 2a answers $3+2+2+3+3+4+4+3=24$ instances while Table 2b answers $4+4+4+4+4+4+4+4=32$ instances. Thus we find an interesting case: though the optimal 2-anonymous table created using *local recoding* has lower information loss, it answers less precise results than the optimal 2-anonymous table created using *global recoding*.

Now we briefly explain why such an interesting case will happen. Given a table T that just contains two QI attributes. After generalization(using *local recoding*), A and B are two anonymization groups of the k-anonymous table. C is the intersection of the two groups, just as Figure 1 shows. Use $Att1$ and $Att2$ to denote two attributes respectively, and use $|Att1|$, $|Att2|$ to denote the two domains. Assume A matches to queries: $Att1=X$, $Att2=Y$ and $(Att1=X, Att2=Y)$. Here, X and Y are subset of $|Att1|$ and $|Att2|$ respectively. Also, assume B matches to quires: $Att1=X'$, $Att2=Y'$ and $(Att1=X', Att2=Y')$. Thus, C matches to queries: $Att1=X\bigcap X'$, $Att2=Y\bigcap Y'$ and $(Att1=X\bigcap X', Att2=Y\bigcap Y')$.

Obviously, for queries matched to C, the anonymized table will answer all elements of $A\bigcup B$ and leads to a "very rough" answer. When the number and average size of such intersection grow, the answerability of the table will be greatly weakened. That is why Table 2a is better than

Table 2b in terms of answerability.

B. Hybrid Algorithm

In previous section, we have shown that *global recoding* is less flexible but less confusing, while *local recoding* is more flexible but more confusing. This motivates us to combines the two models together to overcome both their deficiencies. We try to use *global recoding* to obtain many un-overlapping anonymization groups first, then we use *local recoding* to go on divide each group into smaller ones.

We call such model as *hybrid recoding*. As *local recoding* is employed, huge anonymization group will not exist. And all intersection will occur just when some anonymization groups have come to the worst case of *global recoding*. In fact, our experiments in the next section show that *hybrid recoding* is effective. Now we describe our algorithm. Our algorithm can be divided into two phases: *global phase* and *local phase*. In *global phase*, we employ greedy strategy to partition the table into small sub-tables according to the requirement of *global recoding*. Then, we go on partition the sub-tables whose size are greater than or equal to $2k$ in *local phase*. In this paper, we use Xu et al's *Bottom Up Algorithm* [13] to implement the *local phase*.

Now we discuss how to use greedy strategy to strictly partition the original table in *global phase*. We adopt the geometric view. That is, total orders are set over each attribute domain when a table T is given, and each tuple can be treated as a point in a multidimensional space. We use $\Omega(T)$ to denote the minimum multidimensional box that contains all tuples of T(note that in the case of 2 dimension, the box is a rectangle, and in the case of 3 dimension, the box is a cuboid). The following is the pseudo-code description of the *global phase*.

Algorithm 1 : Hybrid Recoding Algorithm-Global Phase(HRGP)

1: **Input**: A table T, k, a metric χ(such as DM,NCP)
2: $Temp \leftarrow \Omega(T)$.
3: **if** $|Temp| \geq 2k$ **then**
4: **if** $Temp$ can be strictly partitioned into two rectangular boxes S_1,S_2 such that $|S_1| \geq k \wedge |S_2| \geq k$ **then**
5: Find a partition to minimize $|\chi(S_1) - \chi(S_2)|$
6: $T_1 \leftarrow$ all elements in S_1; $T_2 \leftarrow$ all elements in S_2
7: Run HRGP(T_1,k,χ) and HRGP(T_2,k,χ) respectively.
8: **end if**
9: **end if**

This algorithm recursively partitions the rectangular box $\Omega(T)$. In each run, the algorithm examines whether the input box contains at least $2k$ tuples. If the condition is met, the algorithm tries to greedily partitions it into two small boxes subject to the requirement of *global recoding*. Here, the algorithm eventually partitions the box according to given metric. When no boxes can be strictly partitioned, the *local phase* is executed to go on partition those large boxes(which contains no less than $2k$ tuples) into overlapping boxes. Note that the *Hybrid Recoding Algorithm* allows a data publisher to use different metrics. Thus a data publisher can use a proper metric for the publishing.

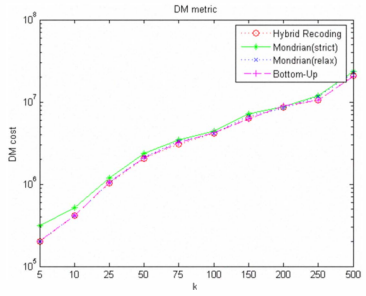

Figure 2. Quality-Penalties of the four algorithms according to DM when k grows

V. Experiments

Our experiments evaluate the utility(and quality) of the anonymized solution of our algorithm by comparing with those produced by Xu et al' *Bottom-Up Algorithm* and *Mondrian Algorithm*. Here, *Bottom-Up Algorithm* is one of the state-of-the-art algorithm which uses *local recoding*. *Mondrian Algorithm* is very famous and has been widely used. It has two versions: one version finds the anonymized solution subject to the requirement of *global recoding* while the other employs *local recoding*. All algorithms are run in the same experimental environment: a Pentium 4, dual 2.2 GHZ processor machine with 2 GB RAM. The dataset used is famous "adult dataset". This dataset is used in most of previous related works [1], [2], [8], [9], [14], [15]. Removing all the uncomplete records in this dataset, there are totally 30162 records. We choose 7 regular attributes as our QI(age, workclass, education, marital-status, occupation, race,sex,native-country), and use attribute "salary" as classification identifier for *classification penalty metric*(CM) [10], [7].

The main purpose of the experiment is to evaluate the utility(NCP Cost and Query Answerability) of the anonymized solution of our algorithm, however, we also compares the information loss and time cost. According to Figure 2, 3 and 4, our anonymized solution has less CM and DM cost, and produces more anonymization groups, if comparing with both strict and relaxed Mondrian. This result demonstrates our previous conclusion that *hybrid recoding* suffers less information loss comparing with *global recoding*. Also, our algorithm is not inferior than powerful *Bottom-Up Algorithm* in terms of information loss, according to the three figures.

Figure 5 and Figure 6 concern NCP cost and query

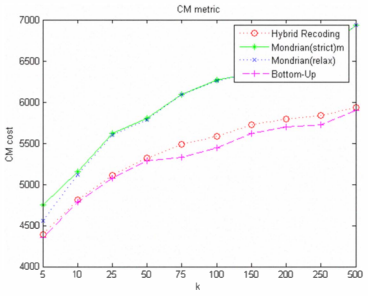

Figure 3. Quality-Penalties of the four algorithms according to CM when k grows

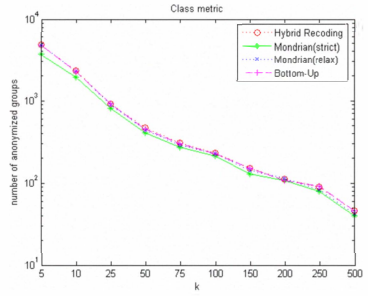

Figure 4. Quality-The number of groups of the solutions resulting from the four algorithms with respect to growing k

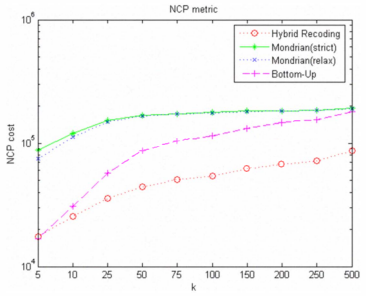

Figure 5. Utility-Penalties of the four algorithms according to NCP with respect to growing k

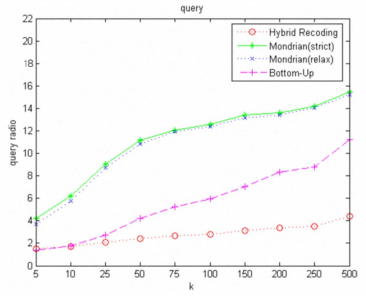

Figure 6. Utility-Answering ratio/Query answerability of the four algorithms, when k grows

answerability respectively. From the two figures, we find that our algorithm has significantly less NCP cost and much lower query ratio than other three algorithms. This Result shows that our algorithm(and also the *hybrid recoding*) is powerful in terms of utility view. Note that *query ratio* is used to evaluate the precision of answers to given queries. It shows the answerability of a table. To calculate query ratio, we first randomly generate 7000 queries(with different length), then we use the 7000 queries to query the anonymized solution resulting from the four algorithms. We count the total number of matched tuples of each solution(anonymized table) for the 7000 queries, and divide the four numbers by 7000×30162.

We also evaluate the time cost of the four algorithms. It is obvious that our algorithm runs much faster than the *Bottom-Up Algorithm* according to Figure 7.

VI. Relatedwork

[1], [2]introduce the notion of k-anonymity. *Generalization* [7], [1], [2], [3] is the most widely used operation for achieving this privacy principle. All generalization based algorithms can be divided into two categories: *global recoding* and *local recoding*. Various metrics of information loss [7] are proposed for evaluating anonymized data, and quite a lot of algorithms are designed according to these information loss metrics. As finding optimal k-anonymity was proved to be NP-hard even in easy cases [7], [9], most these algorithms are approximate or focusing on a limited solution space. For example, Samarati gives an algorithm based on binary search in a limited solution space, his algorithm is relatively efficient than other similar search algorithms [2]. Bayardo designs an efficient ordered portioning schemes [8]. [9] uses *hierarchical partition* to achieve k-anonymity. Soon [15] propose an optimal algorithm based on *hierarchical partition*. Xu et al. firstly show in their paper that low information loss does not equal high utility and propose a new algorithm[13], then, more and more works does not only take information loss into account..

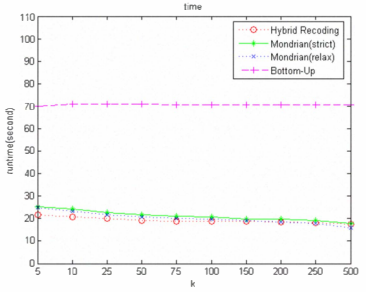

Figure 7. Time cost of the four algorithms with respect to growing k

VII. Conclusion

In this paper, we focus on utility-based k-anonymization. We show that the flexibility of powerful *local recoding* may lead to higher confusion, which weakens the utility of the anonymized data. We use an example to demonstrate the possibility of this situation and give a brief but proper analysis. We design our hybrid algorithm to overcome both the deficiencies of *local recoding* and *global recoding*. Experimental results on adult dataset shows that our algorithm is very powerful. It has lower information loss than advanced global recoding algorithm-Mondrian. Its NCP cost and query ratio is lower than two powerful local recoding algorithms. Thus, it is reasonable to say that our algorithm can produce high-utility anonymized data comparing with recent related works. What's more, it is faster, at least comparing with powerful *Bottom-Up Algorithm.*

Acknowledgment

The research is supported by the Natural Science Foundations of Fujian Province under Grant No. 2009J01295 and No. 2010J01330, and by the Foundation of Fujian Education Department under Grant No. JA09004.

References

[1] L. SWEENEY, "k-anonymity: a model for protecting privacy," *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, vol. 557-570, Oct. 2002.

[2] ——, "Achieving k-anonymity privacy protection using generalization and suppression," *International Journal on Uncertainty, Fuzziness, and Knowledge-based Systems*, vol. 571-588, Oct. 2002.

[3] P. SAMARATI and L. SWEENEY, "Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression," SRI International, Tech. Rep., Mar. 1998.

[4] A. MACHANAVAJJHALA, J. GEHRKE, D. KIFER, and M. VENKITASUBRAMANIAM, "l-diversity: Privacy beyond k-anonymity." in *Proc. of the 22nd IEEE International Conference on Data Engineering(ICDE), Atlanta, GA*, 2006.

[5] N. LI, T. LI, and S. VENKATASUBRAMANIAN, "t-closeness: Privacy beyond k-anonymity and l-diversity," in *Proc. of the 23rd IEEE International Conference on Data Engineering (ICDE). Istanbul, Turkey*, 2007.

[6] T. Truta, A. Campan, and P.Meyer, "Generating micro data with p-sensitive k-anonymity property." in *Proc. of SDM*, 2007.

[7] B. Fung, K. Wang, R. Chen, and P. Yu, *Privacy-preserving data publishing: A survey on recent developments*. ACM Computing Surveys.

[8] R. BAYARDO and R. AGRAWAL, "Data privacy through optimal k-anonymization," in *Proc. of the 21st IEEE International Conference on Data Engineering (ICDE)*, 2005, pp. 217–228.

[9] K. LEFEVRE, D. DEWITT, and R. RAMAKRISHNAN, "Mondrian multidimensional k-anonymity," in *Proc. of the 22nd IEEE International Conference on Data Engineering (ICDE)*, 2006.

[10] V. IYENGAR, "Transforming data to satisfy privacy constraints," in *Proc. of the 8th ACM SIGKDD*, 2002, pp. 279–288.

[11] K. WANG, P. YU, and CHAKRABORTY, "Bottom-up generalization: a data mining solution to privacy protection," in *Proc. of the 4th IEEE International Conference on Data Mining (ICDM)*, 2004.

[12] B. FUNG, K. WANG, and P. YU, "Top-down specialization for information and privacy preservation," in *Proc. of the 21st IEEE International Conference on Data Engineering (ICDE)*, 2005.

[13] J. XU, W. WANG, J. PEI, X. WANG, B. SHI, and A. FU, "Utility-based anonymization using local recoding," in *Proc. of the 12th ACM SIGKDD. Philadelphia, PA*, 2006.

[14] K. LEFEVRE, D. DEWITT, and R. RAMAKRISHNAN, "Incognito: Efficient full-domain k-anonymity," in *Proc. of the 25th ACM SIGMOD*, 2006, pp. 49–60.

[15] S. B. Hore and R. Jammalamadaka, "Flexible anonymization for privacy preserving data publishing: A systematic search based approach," in *Proc. of the Seventh SIAM International Conference on Data Mining*, 2007.