

On Multidimensional k -Anonymity with Local Recoding Generalization

Yang Du*

Tian Xia*

Yufei Tao[†]

Donghui Zhang*[‡]

Feng Zhu*

Abstract

This paper presents the first theoretical study, on using local-recoding generalization (LRG) to compute a k -anonymous table with quality guarantee. First, we prove that it is NP-hard both to find the table with the maximum quality, and to discover a solution with an approximation ratio at most $5/4$. Then, we develop an algorithm with good balance between the approximation ratio and time complexity. The quality of our solution is verified by experiments.

1 Introduction

Privacy preserving data publication has received considerable attention from the database community in the past few years. Specifically, given a relation R containing the sensitive information of individuals, the objective is to publish a set of data that allows accurate mining of data patterns in R , and prevents the privacy of each individual from being revealed.

The hardness of this problem lies in the existence of certain attributes in R called *quasi-identifiers*, which can be used in combination to restore the identity of each individual. Consider, for example, Table 1, which contains the payroll records of the employees in a company. In particular, *Salary* is sensitive, because no employee is willing to disclose her/his salary. Simply publishing the projection on $\{Age, Start-year, Salary\}$ is not feasible. This is because if a person knows the age and start-year of, for instance, Alice, s/he can still obtain the precise salary of Alice. In this case, *Age* and *Start-year* are the *quasi-identifier* attributes.

One remedy is *k-anonymity*, which replaces each quasi-identifier value with a more general form, such that the quasi-identifier values of each tuple are identical to those of at least $k - 1$ other tuples in the published table. Table 2 illustrates a possible 3-anonymous relation for Table 1. Here, each tuple has transformed age and start-year as same as

Name	Age	Start-year	Salary
Alice	25	2001	7k
Bob	30	2004	1k
Christina	35	1990	2k
Daniel	40	1995	3k
Emily	45	2000	6k
William	55	1985	3k

Table 1. The original payroll table (microdata).

Age	Start-year	Salary
[25, 45]	[2000, 2004]	7k
[25, 45]	[2000, 2004]	1k
[35, 55]	[1985, 1995]	2k
[35, 55]	[1985, 1995]	3k
[25, 45]	[2000, 2004]	6k
[35, 55]	[1985, 1995]	3k

Table 2. A 3-anonymous generalization of Table 1.

those of 2 other tuples. Such transformation is called *generalization*. In general, a k -anonymous generalization with a larger k offers a higher degree of privacy protection.

The purpose of publishing generalized data is to allow users to do further analysis on top of it. Therefore, given a certain requirement on the degree of privacy protection, it is important to make the information lost in the generalization as less as possible. In this paper, we aim to tackle the problem of computing k -anonymous tables with quality guarantees under certain scheme called LRG[3]. Our contributions can be summarized as follows.

- First, we prove that this problem is NP-hard. Furthermore, we show that, unfortunately, it cannot be solved even approximately in polynomial time, if the approximation ratio has to be very small (lower than 1.25).
- Second, we present the **NNG** algorithm with good balance between approximation ratio and time complexity. Given a dataset S with cardinality n and dimensionality d , the algorithm runs in $O(d \cdot n^2)$ time and has an approximation ratio of $6d$.

* College of Computer and Information Science, Northeastern University, Boston, MA 02115. {duy, tianxia, donghui, zhufeng}@ccs.neu.edu

[†] Department of Computer Science, City University of Hong Kong, Kowloon, HONG KONG

[‡] Supported by NSF CAREER Award IIS-0347600.

The rest of the paper is organized as follows. Section 2 formally defines the problem. Section 3 presents our hardness results. Section 4 propose the approximate algorithm NNG. Finally, Section 5 reviews some related work and concludes the paper with directions for future work.

2 Problem Definition

To study this problem formally, we map it into a partitioning problem in d -dimensional space. In particular, each tuple in R is mapped into a point in a d -dimensional space, where each dimension corresponds to a quasi-identifier attribute¹. The generalization of k or more tuples is mapped into a rectangle bounding all points mapped from those tuples. As a result, the problem is naturally mapped into finding a good k -anonymous partition defined as follows.

Definition 1 (k -Anonymous Partition) Given a set S of d -dimensional points, an **k -anonymous partition (k -AP)** is a disjoint partition P of S that includes a set of clusters $\{S_1, S_2, \dots, S_m\}$ satisfying

1. $\bigcup_{1 \leq i \leq m} S_i = S$;
2. for any $1 \leq i \neq j \leq m$, $S_i \cap S_j = \emptyset$;
3. for any $1 \leq i \leq m$, $|S_i| \geq k$.

Definition 2 (Local Recoding Generalization [3]) A k -AP uniquely decides a k -anonymous table under the **local-recoding generalization (LRG)** scheme, where, for each cluster S_i ($1 \leq i \leq m$) in P , the generalized form of all tuples in S_i corresponds to the MBR (minimum bounding rectangle) of S_i .

To retain information, good generalization should produce MBRs with small size, which motivates the following metric for measuring the quality of generalization:

Definition 3 (MAXSIZE Metric) Given a k -anonymous partition $P = \{S_1, S_2, \dots, S_m\}$ of dataset S , the **MAXSIZE** of P is defined as $\max_{1 \leq i \leq m} \{PR(S_i)\}$, where $PR(S_i)$ is the perimeter of the MBR of S_i .

Here we use perimeter to measure the size of MBR. Without ambiguity, in the sequel, we use the “size” and “perimeter” of a rectangle interchangeably. Based on this metric, computing the best k -anonymous generalization is to find the k -AP with the smallest MAXSIZE:

Problem 1 (Optimal k -Anonymous Partition) Given a value of k and a set S of d -dimensional points, the **optimal k -anonymous partition (OPT- k -AP)** is the k -anonymous partition P of S that has the smallest MAXSIZE, among all possible k -AP of S .

¹Unless specifically stated, we consider that each quasi-identifier attribute is numeric.

3 Hardness of Our Problem

This section shows the hardness of the problem. The main results are Theorem 1 and Theorem 2, which show that the OPT - k -AP problem is NP-hard, and that there does not exist any polynomial algorithm with approximation ratio smaller than $5/4$, respectively. To prove the NP-hardness, we reduce from the planar-3SAT problem which is known to be NP-complete [5]. While our proof is inspired by the framework proposed in [2], the details are non-trivial and significantly different from theirs. Due to space limitations, the details of the proofs are omitted.

Theorem 1 The OPT - k -AP problem is NP-hard for $k \geq 3$.

Theorem 2 If $P \neq NP$, there is no polynomial-time algorithm for the OPT - k -AP problem with approximation ratio less than $\frac{5}{4}$.

4 The NNG Algorithm

This section proposes the **Nearest-Neighbor-Group (NNG)** algorithm. As its name suggests, the algorithm groups the objects according to the proximity to others.

The NNG algorithm (Figure 2) first finds the $k - 1$ nearest objects to a given object and takes the MBR of these k points (step 1). Then it picks a subset of non-overlapping MBRs from the results of first step (Step 2). Note that this step greedily picks MBRs until, as indicated by the requirements, that any non-picked MBR must overlap with some picked one. Then it creates a set of initial partitions, one corresponding to a picked MBR (Step 3). Next, Step 4 assigns each object not partitioned yet to the closest picked MBR. To improve the quality of the result (without affecting the worst-case approximation ratio), Step 5 further divides each partition that contains $2k$ or more points, if any, to sub-partitions.

Theorem 3 For the OPT - k -AP problem on d -dimensional space, the NNG algorithm has an approximation ratio of $6d$ and time complexity $O(d \cdot n^2)$.

The quality of the NNG algorithm is bounded by $6d$. However, our experiments reveals it is much better in real cases. The dataset we used in our experiments is the adult census dataset from the UC Irvine Machine Learning Repository[7], which has become a commonly used benchmark for k -anonymity. We select the numerical attributes (up to 5) as the quasi-identifier attributes and follow the process of [4] to remove the tuples with missing values. For each attribute, we normalize its domain to $[0,1]$. We compare NNG with the OPT algorithm, which returns the lower bound of the optimal MAXSIZE by finding the minimum rectangle that contains at least k objects. Figure 1 is

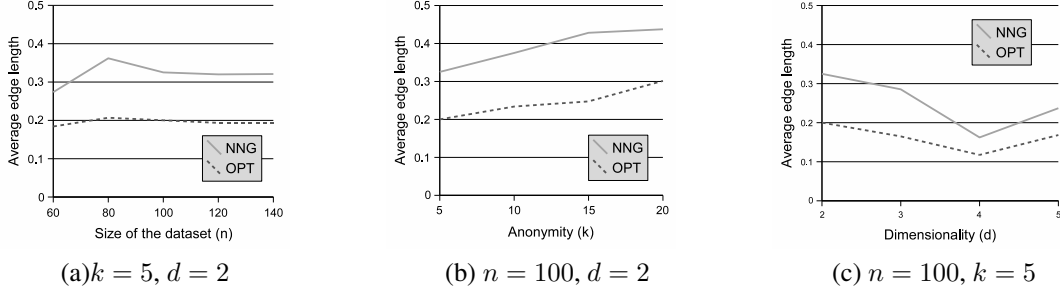


Figure 1. The Quality of NNG Algorithm vs. OPT.

Algorithm NNG

Input: Dataset S .

Return: A k -anonymity partition.

1. For each object $s \in S$, find its $k - 1$ nearest neighbors, and takes the MBR of this k objects.
2. Greedily pick a subset R of MBRs found in previous step, such that:
 - $\forall r_i, r_j \in R, r_i$ and r_j do not overlap;
 - $\forall r_k \notin R, \exists r_i$ found in step 1 which overlaps with r_k .
3. Let the objects covered by each MBR in R be a separate partition. Let P represent the set of partitions.
4. **for** each point $s \in S$ which does not belong to any partition in P , assign s to the partition corresponding to the nearest partition in P to s .
5. Make sure no partition has more than $2k - 1$ objects.
6. **return** P .

Figure 2. The NNG Algorithm.

the results when we vary the cardinality (size of the dataset), anonymity, and dimensionality. It shows that the NNG algorithm achieves approximation ratio about 1.5.

5 Related Work and Conclusions

Many algorithms have been proposed for computing a k -anonymous table. Based on various heuristics, these “engineering approaches” are not concerned about the theoretical aspects of the problem. In the worst case, they require computation time exponential to the dataset size[4], whereas the results do not have quality guarantees. The existing “theoretical algorithms” [1, 4, 6] with good complexities assume either *global recoding generalization* (GRG) or *suppression*, both of which are special case of LRG. Therefore, LRG is naturally superior to them.

This paper is the first work that considers k -anonymous generalization with quality guarantees under the LRG

scheme. We addressed a crucial problem named the *OPT- k -AP* problem. This paper aims to find an optimal partitioning of a set of objects, where every partition has at least k objects. Here *optimality* means to minimize the perimeter of the largest partition. We proved that the problem is NP-hard and proposed an algorithm with good balance between running time and approximation ratio. In particular, it has an approximation ratio of $6d$ and complexity of $O(dn^2)$.

We plan to continue our work in two directions. The first direction is to discover new algorithms with better approximation ratio and/or time complexity. The second direction is a variation of the problem, where the optimality is defined as minimizing the average perimeter of the partitions. We expect this new metric will lead to algorithms with less information loss.

References

- [1] G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D. Thomas, and A. Zhu. Anonymizing tables. In *ICDT*, pages 246–258, 2005.
- [2] S. Khanna, S. Muthukrishnan, and M. Paterson. On Approximating Rectangle Tiling and Packing. In *Proc. of the Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 384–393, 1998.
- [3] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan. Incognito: Efficient full-domain k -anonymity. In *SIGMOD*, pages 49–60, 2005.
- [4] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan. Mondrian multidimensional k -anonymity. In *ICDE*, 2006.
- [5] D. Lichtenstein. Planar formulae and their uses. *SIAM Journal on Computing*, 11(2):423–454, 1982.
- [6] A. Meyerson and R. Williams. On the complexity of optimal k -anonymity. In *PODS*, pages 223–228, 2004.
- [7] D.J. Newman, S. Hettich, C. L. Blake, and C.J. Merz. UCI Repository of machine learning databases, 1998.