Efficient Algorithms for Masking and Finding Quasi-Identifiers [*]

Rajeev Motwani [†] Ying Xu [‡]

Abstract

A quasi-identifier refers to a subset of attributes that can uniquely identify most tuples in a table. Incautious publication of quasi-identifiers will lead to privacy leakage. In this paper we consider the problems of finding and masking quasi-identifiers. Both problems are provably hard with severe time and space requirements. We focus on designing efficient approximation algorithms for large data sets.

We first propose two natural measures for quantifying quasi-identifiers: distinct ratio and separation ratio. We develop efficient algorithms that find small quasi-identifiers with provable size and separation/distinct ratio guarantees, with space and time requirements sublinear in the number of tuples. We also propose efficient algorithms for masking quasi-identifiers, where we use a random sampling technique to greatly reduce the space and time requirements, without much sacrifice in the quality of the results. Our algorithms for masking and finding quasi-identifiers naturally apply to stream databases. Extensive experimental results on real world data sets confirm efficiency and accuracy of our algorithms.

1 Introduction

A quasi-identifier (also called a semi-key) is a subset of attributes which uniquely identifies most entities in the real world or tuples in a table. A well-known example is that the combination of gender, date of birth, and zipcode can uniquely determine about 87% of the population in United States. Quasi-identifiers play an important role in many aspects of data management, including privacy, data cleaning, and query optimization.

As pointed out in the seminal paper of Sweeney [25], publishing data with quasi-identifiers leaves open attacks that combine the data with other publicly available information to identify represented individuals. To avoid such linking attacks via quasi-identifiers, the concept of k-anonymity was proposed [25, 24] and many algorithms for k-anonymity have been developed [23, 2, 4]. In this paper we consider the problem of masking quasi-identifiers: we want to publish a subset of attributes (we either publish the exact value of every tuple on an attribute, or not publish the attribute at all), so that no quasi-identifier is revealed in the published data. This can be viewed as a variant of k-anonymity where the suppression is only allowed at the attribute level. While this approach is admittedly too restrictive in some applications, there are two reasons we consider it. First, the traditional tuple-level suppression may distort the distribution of the original data and the association between attributes, so sometimes it might be desirable to publish fewer attributes with complete and accurate information. Second, as noted in [15], the traditional k-anonymity algorithms are expensive and do not scale well to large data sets; by restricting the suppression to a coarser level we are able to design more efficient algorithms.

We also consider the problem of finding small keys and quasi-identifiers, which can be used by adversaries to perform linking attacks. When a table which is not properly anonymized is published, an adversary would be interested in finding keys or quasi-identifiers in the table so that once he collects other persons' information on those attributes, he will be able to link the records to real world entities. Collecting information on each attribute incurs certain cost to the adversary (for example, he needs to look up yellow pages to collect the area code of phone numbers, to get party affiliation information from the voter list, etc), so the adversary wishes to find a subset of attributes with a small size or weight that is a key or almost a key to minimize the attack cost.

Finding quasi-identifiers also has other important applications besides privacy. One application is data cleaning. Integration of heterogeneous databases sometimes causes the same real-world entity to be represented by multiple records in the integrated database due to spelling mistakes, inconsistent conventions, etc. A critical task in data cleaning is to identify and remove such fuzzy duplicates [3, 6]. We can estimate the ratio of fuzzy duplicates, for example by checking some samples manually or plotting the distribution of pairwise similarity; now if we can find a quasi-

[*] P3DM'08, April 26, 2008, Atlanta, Georgia, USA.

[†] Stanford University. rajeev@cs.stanford.edu. Supported in part by NSF Grant ITR-0331640, and a grant from Media-X.

[‡] Stanford University. xuying@cs.stanford.edu. Supported in part by Stanford Graduate Fellowship and NSF Grant ITR-0331640.

identifier whose "quasiness" is similar to the fuzzy duplicate ratio, then those tuples which collide on the quasi-identifier are likely to be fuzzy duplicates. Finally, quasi-identifiers are a special case of approximate functional dependency [13, 22], and their automatic discovery is valuable to query optimization and indexing [9].

In this paper, we study the problems of finding and masking quasi-identifiers in given tables. Both problems are provably hard with severe time and space requirements, so we focus on designing efficient approximation algorithms for large data sets. First we define measures for quantifying the "quasiness" of quasi-identifiers. We propose two natural measures – separation ratio and distinct ratio.

Then we consider the problem of finding the minimum key. The problem is NP-hard and the best-known approximation algorithm is a greedy algorithm with approximation ratio $O(\ln n)$ (n is the number of tuples); however, even this greedy algorithm requires multiple scans of the table, which are expensive for large databases that cannot reside in main memory and prohibitive for stream databases. To enable more efficient algorithms, we sacrifice accuracy by allowing approximate answers (quasi-identifiers). We develop efficient algorithms that find small quasi-identifiers with provable size and separation/distinct ratio guarantees, with both space and time complexities sublinear in the number of input tuples.

Finally we present efficient algorithms for masking quasi-identifiers. We use a random sampling technique to greatly reduce the space and time requirements, without sacrificing much in the quality of the results.

Our algorithms for masking and finding minimum quasi-identifiers naturally apply to stream databases: we only require one pass over the table to get a random sample of the tuples and the space complexity is sublinear in the number of input tuples (at the cost of only providing approximate solutions).

1.1 Definitions and Overview of Results

A *key* is a subset of attributes that uniquely identifies each tuple in a table. A *quasi-identifier* is a subset of attributes that can distinguish almost all tuples. We propose two natural measures for quantifying a quasi-identifier. Since keys are a special case of functional dependencies, our measures for quasi-identifiers also conform with the measures of approximate functional dependencies proposed in earlier work [13, 22, 11, 8].

> (1) An *α-distinct quasi-identifier* is a subset of attributes which becomes a key in the table remaining after the removal of at most a $1-\alpha$ fraction of tuples in the original table.
>
> (2) We say that a subset of attributes *separates* a pair of tuples x and y if x and y have different values on at least one attribute in the subset. An *α-separation quasi-identifier* is a subset of attributes which separates at least an α fraction of all possible tuple pairs.

	age	sex	state
1	20	Female	CA
2	30	Female	CA
3	40	Female	TX
4	20	Male	NY
5	40	Male	CA

Table 1. An example table. The first column labels the tuples for future references and is not part of the table.

We illustrate the notions with an example (Table 1). The example table has 3 attributes. The attribute *age* is a 0.6-distinct quasi-identifier because it has 3 distinct values in a total of 5 tuples; it is a 0.8-separation quasi-identifier because 8 out of 10 tuple pairs can be separated by *age*. $\{sex, state\}$ is 0.8-distinct and 0.9-separation.

The separation ratio of a quasi-identifier is always larger than its distinct ratio, but there is no one-to-one mapping. Let us consider a 0.5-distinct quasi-identifier in a table of 100 tuples. One possible scenario is that projected on the quasi-identifier there are 50 distinct values and each value corresponds to 2 tuples, so its separation ratio is $1 - \frac{50}{\binom{100}{2}} \approx 0.99$; another possible scenario is that for 49 of the 50 distinct values there is only one tuple for each value, and all the other 51 tuples have the same value, and then this quasi-identifier is 0.75-separation. Indeed, an α-distinct quasi-identifier can be an α'-separation quasi-identifier where α' can be as small as $2\alpha - \alpha^2$, or as large as $1 - \frac{2(1-\alpha)}{n}$. Both distinct ratio and separation ratio are very natural measures for quasi-identifiers and have different applications as noted in the literature on approximate functional dependency. In this paper we study quasi-identifiers using both measures.

Given a table with n tuples and m attributes, we consider the following problems. The *size* of a key (quasi-identifier) refers to the number of attributes in the key.

> *Minimum Key Problem:* find a key of the minimum size. This problem is provably hard so we also consider its relaxed version:
>
> *(ϵ,δ)-Separation or -Distinct Minimum Key Problem:* look for a quasi-identifier with a small size such that, with probability at least $1-\delta$, the output quasi-identifier has separation or distinct ratio at least $1-\epsilon$.
>
> *β-Separation or -Distinct Quasi-identifier Masking Problem:* delete a minimum number of attributes such that there is no quasi-identifier with separation or distinct ratio greater than β in the remaining attributes.

In the example of Table 1, $\{age, state\}$ is a minimum key, with size 2; the optimal solution to 0.8-distinct quasi-identifier masking problem is $\{sex, state\}$; the optimal solution to 0.8-separation quasi-identifier masking problem is $\{age\}$, $\{sex\}$ or $\{state\}$, all of size 1.

The result data after quasi-identifier masking can be viewed as an approximation to k-anonymity. For example, after 0.2-distinct quasi-identifier masking, the result data is approximately 5-anonymous, in the sense that on average each tuple is indistinguishable from another 4 tuples. It does not provide perfect privacy as there may still exist some tuple with a unique value, nevertheless it provides anonymity for the majority of the tuples. The k-anonymity problem is NP-hard [17, 2]; further, Lodha and Thomas [15] note that there is no efficient approximation algorithm known that scale well for large data sets, and they also aim at preserving privacy for majority. We hope to provide scalable anonymizing algorithm by relaxing the privacy constraints. Finally we would like to maximize the utility of published data, and we measure utility in terms of the number of attributes published (our solution can be generalized to the case where attributes have different weights and utility is the weighted sum of published attributes).

We summarize below the contributions of this paper.

1. We propose greedy algorithms for the (ϵ, δ)-separation and distinct minimum key problems, which find small quasi-identifiers with provable size and separation (distinct) ratio guarantees, with space and time requirements sublinear in n. In particular, the space complexity is $O(m^2)$ for the (ϵ, δ)-separation minimum key problem, and $O(m\sqrt{mn})$ for (ϵ, δ)-distinct. The algorithms are particularly useful when $n \gg m$, which is typical of database applications where a large table may consist of millions of tuples, but only a relatively small number of attributes. We also extend the algorithms to find the approximate minimum β-separation quasi-identifiers. (Section 2)
2. We present greedy algorithms for β-separation and β-distinct quasi-identifier masking. The algorithms are slow on large data sets, and we use a random sampling technique to greatly reduce the space and time requirements, without much sacrifice in the utility of the published data. (Section 3)
3. We have implemented all the above algorithms and conducted extensive experiments using real data sets. The experimental results confirm the efficiency and accuracy of our algorithms. (Section 4)

2 Finding Minimum Keys

In this section we consider the Minimum Key problem. First we show the problem is NP-hard (Section 2.1) and the best approximation algorithm is a greedy algorithm which gives $O(\ln n)$-approximate solution (Section 2.2). The greedy algorithm requires multiple scans of the table, which is expensive for large tables and inhibitive for stream databases. To enable more efficient algorithms, we relax the problem by allowing approximate answers, i.e. the (ϵ, δ)-Separation (Distinct) Minimum Key problem. We develop random sampling based algorithms with approximation guarantees and sublinear space (Section 2.3, 2.4).

2.1 Hardness Result

The Minimum Key problem is NP-Hard, which follows easily from the NP-hardness of the *Minimum Test Collection* problem.

> *Minimum Test Collection:* Given a set S of elements and a collection C of subsets of S, a test collection is a subcollection of C such that for each pair of distinct elements there is some set that contains exactly one of the two elements. The Minimum Test Collection problem is to find a test collection with the smallest cardinality.

Minimum Test Collection is equivalent to a special case of the Minimum Key problem where each attribute is boolean: let S be the set of tuples and C be all the attributes; each subset in C corresponds to an attribute and contains all the tuples whose values are *true* in this attribute, then a test collection is equivalent to a key in the table. Minimum Test Collection is known to be NP-hard [7], therefore the Minimum Key problem is also NP-hard.

2.2 A Greedy Approximation Algorithm

The best known approximation algorithm for Minimum Test Collection is a greedy algorithm with approximation ratio $1 + 2\ln|S|$ [18], i.e. it finds a test collection with size at most $1 + 2\ln|S|$ times the smallest test collection size. The algorithm can be extended to the more general Minimum Key problem, where each attribute can be from an arbitrary domain, not just boolean.

Before presenting the algorithm, let us consider a naive greedy algorithm: compute the separation (or distinct) ratio of each attribute in advance; each time pick the attribute with the highest separation ratio in the remaining attributes, until we get a key. The algorithm is fast and easy to implement, but unfortunately it does not perform well when the attributes are correlated. For example if there are many attributes pairwise highly correlated and each has a high separation ratio, then the optimal solution probably includes only one of these attributes while the above greedy algorithm is likely to pick all of them. The approximation ratio of this algorithm can be arbitrarily bad.

A fix to the naive algorithm is to pick each time the attribute which separates the largest number of tuple pairs not yet separated. To prove the approximation ratio of the algorithm, we reduce Minimum Key to the Minimum Set Cover problem. The reduction plays an important role in designing algorithms for finding and masking quasi-identifiers in later sections.

> *Minimum Set Cover:* Given a finite set S (called the *ground set*) and a collection C of subsets of S, a *set cover* I is a subcollection of C such that every element in S belongs to at least one member of I. *Minimum Set Cover* problem asks for a set cover with the smallest size.

Given an instance of Minimum Key with n tuples and m attributes, we reduce it to a set cover instance as follows: the ground set S consists of all distinct unordered pairs of tuples ($|S| = \binom{n}{2}$); each attribute c in the table is mapped to a subset containing all pairs of tuples separated by attribute c. Now a collection of subsets covers S if and only if the corresponding attributes can separate all pairs of tuples, i.e., those attributes form a key, therefore there is a one-to-one map between minimum set covers and minimum keys.

Consider the example of Table 1. The ground set of the corresponding set cover instance contains 10 elements where each element is a pair of tuples. The column *age* is mapped to a subset c_{age} with 8 pairs: $\{(1,2),(1,3),(1,5),(2,3),(2,4),(2,5),(3,4),(4,5)\}$; the column *sex* is mapped to a subset c_{sex} with 6 pairs, and *state* 7 pairs. The attribute set $\{age, sex\}$ is a key; correspondingly the collection $\{c_{age}, c_{sex}\}$ is a set cover.

The *Greedy Set Cover Algorithm* starts with an empty collection (of subsets) and adds subsets one by one until every element in S has been covered; each time it chooses the subset covering the largest number of uncovered elements. It is well known that this greedy algorithm is a $1 + \ln |S|$ approximation algorithm for Minimum Set Cover.

LEMMA 2.1. *[12] The Greedy Set Cover Algorithm outputs a set cover of size at most* $1 + \ln |S|$ *times the minimum set cover size.*

The *Greedy Minimum Key Algorithm* mimics the greedy set cover algorithm: start with an empty set of attributes and add attributes one by one until all tuple pairs are separated; each time chooses an attribute separating the largest number of tuple pairs not yet separated. The running time of the algorithm is $O(m^3 n)$. It is easy to infer the approximation ratio of this algorithm from Lemma 2.1:

THEOREM 2.1. *Greedy Minimum Key Algorithm outputs a key of size at most* $1 + 2\ln n$ *times the minimum key size.*

The greedy algorithms are optimal because neither problem is approximable within $c \ln |S|$ for some $c > 0$ [10]. Note that this is the worst case bound and in practice the algorithms usually find much smaller set covers or keys.

2.3 (ϵ, δ)-Separation Minimum Key

The greedy algorithm in the last section is optimal in terms of approximation ratio, however, it requires multiple scans ($O(m^2)$ scans indeed) of the table, which is expensive for large data sets. In this and next section, we relax the minimum key problem by allowing quasi-identifiers and design efficient algorithms with approximate guarantees.

We use the standard (ϵ, δ) formulation: with probability at least $1 - \delta$, we allow an "error" of at most ϵ, i.e. we output a quasi-identifier with separation (distinct) ratio at least $1 - \epsilon$. The (ϵ, δ) Minimum Set Cover Problem is defined similarly and requires the output set cover covering at least a $1 - \epsilon$ fraction of all elements.

Our algorithms are based on random sampling. We first randomly sample k elements (tuples), and reduce the input set cover (key) instance to a smaller set cover (key) instance containing only the sampled elements (tuples). We then solve the exact minimum set cover (key) problem in the smaller instance (which is again a hard problem but has much smaller size, so we can afford to apply the greedy algorithms in Section 2.2), and output the solution as an approximate solution to the original problem. The number of samples k is carefully chosen so that the error probability is bounded. We present in detail the algorithm for (ϵ, δ)-set cover in Section 2.3.1; the (ϵ, δ)-Separation Minimum Key problem can be solved by reducing to (ϵ, δ) Minimum Set Cover (Section 2.3); we discuss (ϵ, δ)-Distinct Minimum Key in Section 2.4.

2.3.1 (ϵ, δ) Minimum Set Cover The key observation underlying our algorithm is that to check whether a given collection of subsets is a set cover, we only need to check some randomly sampled elements if we allow approximate solutions. If the collection only covers part of S, then it will fail the check after enough random samples. The idea is formalized as the following lemma.

LEMMA 2.2. $s_1, s_2, \ldots, s_k$ *are* k *elements independently randomly chosen from* S*. If a subset* S' *satisfies* $|S'| < \alpha|S|$*, then* $Pr[s_i \in S', \forall i] < \alpha^k$*.*

The proof is straightforward. The probability that a random element of S belongs to S' is $|S'|/|S| < \alpha$, therefore the probability of all k random elements belonging to S' is at most α^k.

Now we combine the idea of random sample checking with the greedy algorithm for the exact set cover. Our *Greedy Approximate Set Cover algorithm* is as follows:

1. Choose k elements uniformly at random from S (k is defined later);
2. Reduce the problem to a smaller set cover instance: the ground set $\tilde{S}$ consists of the k chosen elements; each subset in the original problem maps to a subset which is the intersection of $\tilde{S}$ and the original subset;
3. Apply Greedy Set Cover Algorithm to find an exact set cover for $\tilde{S}$, and output the solution as an approximate set cover to S.

Let n be the size of the ground set S, and m be the number of subsets. We say a collection of subsets is an α-*set cover* if it covers at least an α fraction of the elements.

THEOREM 2.2. *With probability* $1-\delta$*, the above algorithm with* $k = \log_{\frac{1}{1-\epsilon}} \frac{2^m}{\delta}$ *outputs a* $(1-\epsilon)$*-set cover whose cardinality is at most* $(1 + \ln \log_{\frac{1}{1-\epsilon}} \frac{2^m}{\delta})|I^*|$*, where* I^* *is the optimal exact set cover.*

Proof. Denote by $\tilde{S}$ the ground set of the reduced instance ($|\tilde{S}| = k$); by $\tilde{I}^*$ the minimum set cover of $\tilde{S}$. The greedy algorithm outputs a subcollection of subsets covering all k elements of $\tilde{S}$, denoted by $\tilde{I}$. By Lemma 2.1, $|\tilde{I}| \le (1 + \ln |\tilde{S}|)|\tilde{I}^*|$. Note that I^*, the minimum set cover of the original set S, corresponds to a set cover of $\tilde{S}$, so $|\tilde{I}^*| \le |I^*|$, and hence $|\tilde{I}| \le (1 + \ln k)|I^*|$.

We map $\tilde{I}$ back to a subcollection I of the original problem. We have

$$|I| = |\tilde{I}| \le (1 + \ln k)|I^*| = (1 + \ln \log_{\frac{1}{1-\epsilon}} \frac{2^m}{\delta})|I^*|.$$

Now bound the probability that I is not a $1-\epsilon$-set cover. By Lemma 2.2, the probability that a subcollection covering less than a $1-\epsilon$ fraction of S covers all k chosen elements of $\tilde{S}$ is at most

$$(1-\epsilon)^k = (1-\epsilon)^{\log_{\frac{1}{1-\epsilon}} \frac{2^m}{\delta}} = (1-\epsilon)^{\log_{1-\epsilon} \frac{\delta}{2^m}} = \frac{\delta}{2^m}.$$

There are 2^m possible subcollections; by union bound, the overall error probability, i.e. the probability that any subcollection is not a $(1-\epsilon)$-cover of S but is an exact cover of $\tilde{S}$, is at most δ. Hence, with probability at least $1-\delta$, I is a $(1-\epsilon)$-set cover for S.

If we take ϵ and δ as constants, the approximation ratio is essentially $\ln m + O(1)$, which is smaller than $1 + \ln n$ when $n \gg m$. The space requirement of the above algorithm is $mk = O(m^2)$ and running time is $O(m^4)$.

2.3.2 (ϵ, δ)-Separation Minimum Key The reduction from Minimum Key to Minimum Set Cover preserves the separation ratio: an α-separation quasi-identifier separates at least an α fraction of all pairs of tuples, so its corresponding subcollection is an α-set cover; and vice versa. Therefore, we can reduce the (ϵ, δ)-Separation Minimum Key problem to the (ϵ, δ)-Set Cover problem where $|S| = O(n^2)$. The complete algorithm is as follows.

1. Randomly choose $k = \log_{\frac{1}{1-\epsilon}} \frac{2^m}{\delta}$ pairs of tuples;
2. Reduce the problem to a set cover instance where the ground set $\tilde{S}$ is the set of those k pairs and each attribute maps to a subset of the k pairs separated by this attribute;
3. Apply Greedy Set Cover Algorithm to find an exact set cover for $\tilde{S}$, and output the corresponding attributes as a quasi-identifier to the original table.

THEOREM 2.3. *With probability* $1-\delta$*, the above algorithm outputs a* $(1-\epsilon)$*-separation quasi-identifier whose size is at most* $(1 + \ln \log_{\frac{1}{1-\epsilon}} \frac{2^m}{\delta})|I^*|$*, where* I^* *is the smallest key.*

The proof directly follows Theorem 2.2. The approximation ratio is essentially $\ln m + O(1)$. The space requirement of the above algorithm is $mk = O(m^2)$, which significantly improves upon the input size mn.

2.4 (ϵ, δ)-Distinct Minimum Key

Unfortunately, the reduction to set cover does not necessarily map an α-distinct quasi-identifier to an α-set cover. As pointed out in Section 1.1, an α-distinct quasi-identifier corresponds to an α'-separation quasi-identifier, and thus reduces to an α'-set cover, where α' can be as small as $2\alpha - \alpha^2$, or as large as $1 - \frac{2(1-\alpha)}{n}$. Therefore reducing this problem directly to set cover gives too loose bound, and a new algorithm is desired.

Our algorithm for finding distinct quasi-identifiers is again based on random sampling. We reduce the input (ϵ, δ)-Distinct Minimum Key instance to a smaller (exact) Minimum Key instance by randomly choosing k tuples and keeping all m attributes. The following lemma bounds the probability that a subset of attributes is an (exact) key in the sample table, but not an α-distinct quasi-identifier in the original table.

LEMMA 2.3. *Randomly choose* k *tuples from input table* T *to form table* T_1*. Let* p *be the probability that an (exact) key of* T_1 *is not an* α*-distinct quasi-identifier in* T*. Then*

$$p < e^{-\frac{(\frac{1}{\alpha}-1)k(k-1)}{2n}}$$

Proof: Suppose we have n balls distributed in $d = \alpha n$ distinct bins. Randomly choose k balls without replacement, and the probability that the k balls are all from different bins is exactly p. Let $x_1, x_2, \ldots, x_d$ be the number of balls in the d bins ($\sum_{i=1}^{d} x_i = n, x_i > 0$), then

$$p = \frac{\sum_{all\{i_1,i_2,\ldots,i_k\}} x_{i_1} x_{i_2} \ldots x_{i_k}}{\binom{n}{k}}.$$

p is maximized when all x_is are equal, i.e. each bin has $\frac{1}{\alpha}$ balls. Next we compute p for this case. The first ball can be from any bin; to choose the second ball, we have $n-1$ choices, but it cannot be from the same bin as the first one, so $\frac{1}{\alpha} - 1$ of the $n-1$ choices are infeasible; similar arguments hold for the remaining balls. Summing up, the probability that all k balls are from distinct bins is

$$\begin{aligned} p &= 1(1 - \frac{\frac{1}{\alpha}-1}{n-1})(1 - \frac{2(\frac{1}{\alpha}-1)}{n-2}) \ldots (1 - \frac{(k-1)(\frac{1}{\alpha}-1)}{n-(k-1)}) \\ &\le e^{-(\frac{\frac{1}{\alpha}-1}{n-1} + \frac{2(\frac{1}{\alpha}-1)}{n-2} + \frac{(k-1)(\frac{1}{\alpha}-1)}{n-(k-1)})} \\ &< e^{-\frac{(\frac{1}{\alpha}-1)k(k-1)}{2n}} \square \end{aligned}$$

The Greedy (ϵ, δ)-Distinct Minimum Key Algorithm is as follows:

1. Randomly choose $k = \sqrt{\frac{2(1-\epsilon)}{\epsilon} n \ln \frac{2^m}{\delta}}$ tuples and keep all attributes to form table T_1;
2. Apply Greedy Minimum Key Algorithm to find an exact key in T_1, and output it as a quasi-identifier to the original table.

THEOREM 2.4. *With probability $1-\delta$, the above algorithm outputs a $(1-\epsilon)$-distinct quasi-identifier whose size is at most $(1+2\ln k)|I^*|$, where I^* is the smallest exact key.*

The proof is similar to Theorem 2.2, substituting Lemma 2.2 with Lemma 2.3. k is chosen such that $p \le \frac{\delta}{2^m}$ to guarantee that the overall error probability is less than δ. The approximation ratio is essentially $\ln m + \ln n + O(1)$, which improves the $1 + 2\ln n$ result for the exact key. The space requirement is $mk = O(m\sqrt{mn})$, sublinear in the number of tuples of the original table.

2.5 Minimum β-Separation Quasi-identifier

In previous sections, our goal is to find a small quasi-identifier that is almost a key. Note that ϵ indicates our "error tolerance", not our goal. For (ϵ, δ)-Separation Minimum Key problem, our algorithm is likely to output quasi-identifiers whose separation ratios are far greater than $1-\epsilon$. For example, suppose the minimum key of a given table consists of 100 attributes, while the minimum 0.9-separation quasi-identifier has 10 attributes, then our $(0.1, 0.01)$-separation algorithm may output a quasi-identifier that has say 98 attributes and is 0.999-separation. However, sometimes we may be interested in finding 0.9-separation quasi-identifiers which have much smaller sizes. For this purpose we consider the *Minimum β-Separation Quasi-identifier Problem*: find a quasi-identifier with the minimum size and separation ratio at least β.

The Minimum β-Separation Quasi-identifier Problem is at least as hard as Minimum Key since the latter is a special case where $\beta = 1$. So again we consider the approximate version by relaxing the separation ratio: we require the algorithm to output a quasi-identifier with separation ratio at least $(1-\epsilon)\beta$ with probability at least $1-\delta$.

We present the algorithm for approximate β-set cover; the β-separation quasi-identifier problem can be reduced to β-set cover as before.

The *Greedy Minimum β-Set Cover algorithm* works as follows: first randomly sample $k = \frac{16}{\beta\epsilon^2} \ln \frac{2^m}{\delta}$ elements from the ground set S, and construct a smaller set cover instance defined on the k chosen elements; run the greedy algorithm on the smaller set cover instance until get a subcollection covering at least $(2-\epsilon)\beta k/2$ elements (start with an empty subcollection; each time add to the subcollection a subset covering the largest number of uncovered elements).

THEOREM 2.5. *The Greedy Minimum β-Set Cover algorithm runs in space $mk = O(m^2)$, and with probability at least $1-\delta$, outputs a $(1-\epsilon)\beta$-set cover with size at most $(1 + \ln \frac{(2-\epsilon)\beta k}{2})|I^*|$, where I^* is the minimum β-set cover of S.*

The proof can be found in our technical report. This algorithm also applies to the minimum exact set cover problem (the special case where $\beta = 1$), but the bound is worse than Theorem 2.2; see our technical report for detailed comparison.

The minimum β-separation quasi-identifier problem can be solved by reducing to β-set cover problem and applying the above greedy algorithm. Unfortunately, we cannot provide similar algorithms for β-distinct quasi-identifiers; the main difficulty is that it is hard to give a tight bound to the distinct ratio of the original table by only looking at a small sample of tuples. The negative results on distinct ratio estimation can be found in [5].

3 Masking Quasi-Identifiers

In this section we consider the quasi-identifier masking problem: when we release a table, we want to publish a subset of the attributes subject to the privacy constraint that no β-separation (or β-distinct) quasi-identifier is published; on the other hand we want to maximize the utility, which is measured by the number of published attributes. For each problem, we first present a greedy algorithm which generates good results but runs slow for large tables, and then show how to accelerate the algorithms using random sampling. (The algorithms can be easily extended to the case where the attributes have weights and the utility is the sum of attribute weights.)

3.1 Masking β-Separation Quasi-identifiers

As in Section 2.2, we can reduce the problem to a set cover type problem: let the ground set S be the set of all pairs of tuples, and let each attribute correspond to a subset of tuple pairs separated by this attribute, then the problem of Masking β-Separation Quasi-identifier is equivalent to finding a maximum number of subsets such that at most a β fraction of elements in S is covered by the selected subsets. We refer to this problem as *Maximum Non-Set Cover problem*. Unfortunately, the Maximum Non-Set Cover problem is NP-hard by a reduction from the Dense Subgraph problem. (See our technical report for the hardness proof.)

We propose a greedy heuristic for masking β-separation quasi-identifiers: start with an empty set of attributes, and add attributes to the set one by one as long as the separation ratio is below β; each time pick the attribute separating the least number of tuple pairs not yet separated.

The algorithm produces a subset of attributes satisfying the privacy constraint and with good utility in practice, however it suffers from the same efficiency issue as the greedy algorithm in Section 2.2: it requires $O(m^2)$ scans of the table and is thus slow for large data sets. We again use random sampling technique to accelerate the algorithm: the following lemma gives a necessary condition for a β-separation quasi-identifier in the sample table (with high probability), so only looking at the sample table and pruning all attribute sets satisfying the necessary condition will guarantee the privacy constraint. The proof of the lemma is omit for lack of space.

LEMMA 3.1. *Randomly sample k pairs of tuples, then a β-separation quasi-identifier separates at least $\alpha\beta$ of the k pairs, with probability at least $1 - e^{-(1-\alpha)^2\beta k/2}$.*

The *Greedy Approximate β-Separation Masking Algorithm* is as follows:

1. Randomly choose k pairs of tuples;
2. Let $\beta' = (1 - \sqrt{\frac{2\ln(2^m/\delta)}{\beta k}})\beta$. Run the following greedy algorithm on the selected pairs: start with an empty set C of attributes, and add attributes to the set C one by one as long as the number of separated pairs is below $\beta' k$; each time pick the attribute separating the least number of tuple pairs not yet separated;
3. Publish the set of attributes C.

By the nature of the algorithm the published attributes C do not contain quasi-identifiers with separation greater than β' in the sample pairs; by Lemma 3.1, this ensures that with probability at least $1 - 2^m e^{-(1-\beta'/\beta)^2\beta k/2} = 1 - \delta$, C do not contain any β-separation quasi-identifier in the original table. Therefore the attributes published by the above algorithm satisfies the privacy constraint.

THEOREM 3.1. *With probability at least $1 - \delta$, the above algorithm outputs an attribute set with separation ratio at most β.*

We may over-prune because the condition in Lemma 3.1 is not a sufficient condition, which means we may lose some utility. The parameter k in the algorithm offers a tradeoff between the time/space complexity and the utility. Obviously both the running time and the space increase linearly with k; on the other hand, the utility (the number of published attributes) also increases with k because the pruning condition becomes tighter as k increases. Our experiment results show that the algorithm is able to dramatically reduce the running time and space complexity, without much sacrifice in the utility (see Section 4).

3.2 Masking β-Distinct Quasi-identifiers

For masking β-distinct quasi-identifiers, we can use a similar greedy heuristic: start with an empty set of attributes, and each time pick the attribute adding the least number of distinct values, as long as the distinct ratio is below β. And similarly we can use a sample table to trade off utility for efficiency.

1. Randomly choose k tuples and keep all the columns to form a sample table T_1;
2. Let $\beta' = (1 - \sqrt{\frac{2\ln(2^m/\delta)}{\beta k}})\beta$. Run the following greedy algorithm on T_1: start with an empty set C of attributes, and add attributes to the set C one by one as long as the distinct ratio is below β'; each time pick the attribute adding the least number of distinct values;
3. Publish the set of attributes C.

Lemma 3.2 and Theorem 3.2 state the privacy guarantee of the above algorithm.

LEMMA 3.2. *Randomly sample k tuples from the input table T into a small table T_1 ($k \ll n$, where n is the number of tuples in T). A β-distinct quasi-identifier of T is an $\alpha\beta$-distinct quasi-identifier of T_1 with probability at least $1 - e^{-(1-\alpha)^2\beta k/2}$.*

Proof. By the definition of β-distinct quasi-identifier, the tuples has at least βn distinct values projected on the quasi-identifier. Take (any) one tuple from each distinct value, and call those representing tuples "good tuples". There are at least βn good tuples in T.

Let k_1 be the number of distinct values in T_1 projected on the quasi-identifier, and k' be the number of good tuples in T_1. We have $k_1 \geq k'$ because all good tuples are distinct. (The probability that any good tuple is chosen more than once is negligible when $k \ll n$.) Next we bound the probability $Pr[k' \leq \alpha\beta k]$. Since each random tuple has a probability at least β of being good, and each sample are chosen independently, we can use Chernoff bound (see [19] Ch. 4) and get

$$Pr[k' \leq \alpha\beta k] \leq e^{-(1-\alpha)^2\beta k/2}$$

Since $k_1 \geq k'$, we have

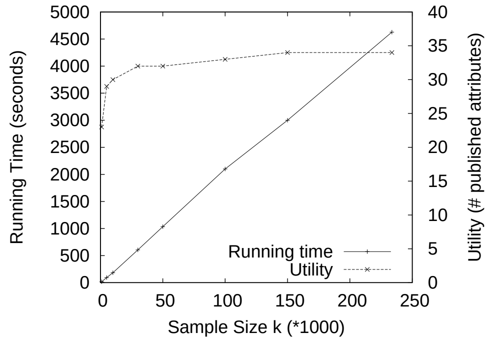
$$Pr[k_1 \leq \alpha\beta k] \leq Pr[k' \leq \alpha\beta k] \leq e^{-(1-\alpha)^2\beta k/2}$$

Hence with probability at least $1 - e^{-(1-\alpha)^2\beta k/2}$, the quasi-identifier has distinct ratio at least $\alpha\beta$ in T_1.

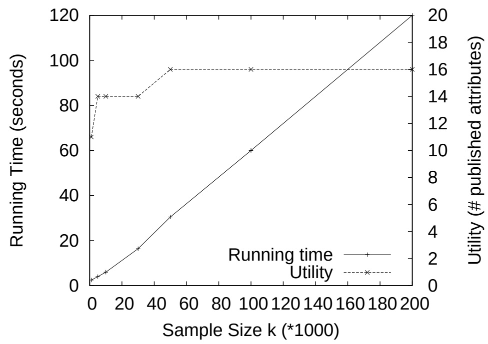
THEOREM 3.2. *With probability at least $1-\delta$, the attribute set published by the algorithm has distinct ratio at most β.*

4 Experiments

We have implemented all algorithms for finding and masking quasi-identifiers, and conducted extensive experiments using real data sets. All experiments were run on a 2.4GHz Pentium PC with 1GB memory.



(a) Masking 0.5-distinct quasi-identifiers



(b) Masking 0.8-separation quasi-identifiers

Figure 1. Performance of masking quasi-identifier algorithms with different sample sizes on table *california*. Figures (a) and (b) show how the running time (the left y axis) and the utility (the right y axis) change with the sample size (the parameter k) in Greedy Approximate algorithms for masking 0.5-distinct and 0.8-separation quasi-identifiers.

4.1 Data Sets

One source of data sets is the census microdata "Public-Use Microdata Samples (PUMS)" [1], provided by US Census Bureau. We gather the 5 percent samples of Census 2000 data from all states and put into a table "census". To study the performance of our algorithms on tables with different sizes, we also extract 1 percent samples of state-level data and select 4 states with different population sizes – Idaho, Washington, Texas and California. We extract 41 attributes including age, sex, race, education level, salary etc. We only use adult records (age ≥ 20) because many children are indistinguishable even with all 41 attributes. The table *census* has 10 million distinct adults, and the sizes of *Idaho, Washington, Texas* and *California* are 8867, 41784, 141130 and 233687 respectively.

We also use two data sets *adult* and *covtype* provided by UCI Machine Learning Repository [21]. The *covtype* table has 581012 rows and 54 attributes. We use 14 attributes of *adult* including age, education level, marital status; the number of records in *adult* is around 30000.

4.2 Masking Quasi-identifiers

The greedy approximate algorithms for masking quasi-identifiers are randomized algorithms that guarantee to satisfy the privacy constraints with probability $1 - \delta$. We set $\delta = 0.01$, and the privacy constraint are satisfied in all experiments, which confirms the accuracy of our algorithms.

Figure 1 shows the tradeoff between the running time and the utility (the number of attributes published), using the *california* data set. Both the running time and the utility decrease as the sample size k decreases; however, the running time decreases linearly with k while the utility degrades very slowly. For example, running the greedy algorithm for masking 0.5-distinct quasi-identifiers on the entire table (without random sampling) takes 80 minutes and publishes 34 attributes (the rightmost point in Figure a); using a sample of 30000 tuples the greedy algorithm takes only 10 minutes and outputs 32 attributes. Figure b shows the impact of k on the masking separation quasi-identifier algorithm. To run the greedy algorithm for masking 0.8-separation quasi-identifier on the entire table takes 728 seconds (not shown in the figure); using a sample of 50000 pairs offers the same utility and only takes 30 seconds. The results show that our random sampling technique can greatly improve time and space complexity (space is also linear in k), with only minor sacrifice on the utility.

Data Sets	Greedy		Greedy Approximate	
	time	utility	time	utility
adult	36s	12	-	-
covtype	-	-	2000s	46
idaho	172s	33	-	-
wa	880s	34	620s	33
texas	3017s	35	630s	33
ca	4628s	34	606s	32
census	-	-	755s	30

Table 2. Algorithms for masking 0.5-distinct quasi-identifiers. The column "Greedy" represents the greedy algorithm on the entire table; the column "Greedy Approximate" represents running greedy algorithm on a random sample of 30000 tuples. We compare the running time and the utility (the number of published attributes) of the two algorithms on different data sets. The results of Greedy on *census* and *covtype* are not available because the algorithm does not terminate in 10 hours; the results of Greedy Approximate on *adult* and *Idaho* are not available because the input tuple number is less than 30000.

Data Sets	Greedy		Greedy Approximate	
	time	utility	time	utility
adult	19s	5	2s	5
covtype	2 hours	38	104s	37
idaho	147s	24	30s	23
wa	646s	23	35s	23
texas	1149s	19	34s	19
ca	728s	16	30s	16
census	-	-	170s	17

Table 3. Algorithms for masking 0.8-separation quasi-identifiers. The column "Greedy" represents the greedy algorithm on the entire table, and the column "Greedy Approximate" represents running greedy algorithm on a random sample of 50000 pairs of tuples. We compare the running time and the utility of the two algorithms on different data sets. The result of Greedy on *census* is unavailable because the algorithm does not terminate in 10 hours.

Table 2 and 3 compare the running time and the utility (the number of published attributes) of running the greedy algorithm on the entire table versus on a random sample (we use a sample of 30000 tuples in Table 2 and a sample of 50000 pairs of tuples in Table 3). Results on all data sets confirm that the random sampling technique is able to reduce the running time dramatically especially for large tables, with only minor impact on the utility. For the largest data set *census*, running the greedy algorithm on the entire table does not terminate in 10 hours, while with random sampling it only takes no more than 13 minutes for masking 0.5-distinct quasi-identifier and 3 minutes for masking 0.8-separation quasi-identifier.

4.3 Approximate Minimum Key Algorithms

Finally we examine the greedy algorithms for finding minimum key and (ϵ, δ)-separation or -distinct minimum key in Section 2. Table 4 shows the experimental results of the Greedy Minimum Key, Greedy $(0.1, 0.01)$-Distinct Minimum Key, and Greedy $(0.001, 0.01)$-Separation Minimum Key algorithms on different data sets.

The Greedy Minimum Key algorithm (applying greedy algorithm directly on the entire table) works well for small data sets such as *adult, idaho*, but becomes unaffordable as the data size increases. The approximate algorithms for separation or distinct minimum key are much faster. For the table *California*, the greedy minimum key algorithm takes almost one hour, while the greedy distinct algorithm takes 2.5 minutes, and greedy separation algorithm merely seconds; for the largest table *census*, the greedy minimum key algorithm takes more than 10 hours, while the approximate algorithms take no more than 15 minutes. The space and time requirements of our approximate minimum key algorithms are sublinear in the number of input tuples, and we expect the algorithms to scale well on even larger data sets.

We measure the distinct and separation ratios of the output quasi-identifiers, and find the ratios always within error ϵ. This confirms the accuracy of our algorithms.

Theorem 2.3 and 2.4 provide the theoretical bounds on the size of the quasi-identifiers found by our algorithms ($\ln m$ or $\ln mn$ times the minimum key size). Those bounds are worst case bounds, and in practice we usually get much smaller quasi-identifiers. For example, we find that the minimum key size of *adult* is 13 by exhaustive search, and the greedy algorithm for both distinct and separation minimum key find quasi-identifiers no larger than the minimum key. (For other data sets in Table 4, computing the minimum key exactly takes prohibitively long time, so we are not able to verify the approximation ratio of our algorithms.) We also generate synthetic tables with known minimum key sizes, then apply the greedy distinct minimum key algorithm (with $\epsilon = 0.1$) on those tables and are always able to find quasi-identifiers no larger than the minimum key size. Those experiments show that in practice our approximate minimum key algorithms usually perform much better than the theoretical worst case bounds, and are often able to find quasi-identifiers with high separation (distinct) ratio and size close to the minimum.

5 Related Work

The implication of quasi-identifiers to privacy is first formally studied by Sweeney, who also proposed the k-anonymity framework as a solution to this problem [25, 24]. Afterwards there is numerous work which studies the complexity of this problem [17, 2], designs and implements algorithms to achieve k-anonymity [23, 4], or extends upon the framework [16, 14]. Our algorithm for masking quasi-identifiers can be viewed as an approximation to k-anonymity where the suppression must be conducted at the attribute level. Also it is an "on average" k-anonymity because it does not provide perfect anonymity for every individual but does so for the majority; a similar idea is used in [15]. On the other side, our algorithms for finding keys/quasi-identifiers attempt to attack the privacy of published data from the adversary's point of view, when the publish data is not k-anonymized. To the best of our knowledge, there is no existing work addressing this problem.

Our algorithms exploit the idea of using random samples to trade off between accuracy and space complexity, and can be viewed as streaming algorithms. Streaming algorithms emerged as a hot research topic in the last decade; see [20] for a survey of this area.

Keys are special cases of functional dependencies, and quasi-identifiers are a special case of approximate functional dependency. Our definitions of separation and dis-

Data Sets	Greedy		distinct Greedy ($\epsilon = 0.1$)			separation Greedy ($\epsilon = 0.001$)		
	time	key size	time	key size	distinct ratio	time	key size	separation ratio
adult	35.5s	13	8.8s	13	1.0	3.11s	5	0.99995
covtype	964s	5	78.1s	3	0.9997	27.1s	2	0.999996
idaho	50.4s	14	15.2s	8	0.997	1.07s	3	0.9999
wa	490s	22	34.1s	8	0.995	7.14s	3	0.99993
texas	2032s	29	120s	14	0.995	13.2s	4	0.99995
ca	3307s	29	145s	13	0.994	16.3s	4	0.99998
census	-	-	808s	17	0.993	120s	3	0.99998

Table 4. Running time and output key sizes of the Greedy Minimum Key, Greedy $(0.1, 0.01)$-Distinct Minimum Key, and Greedy $(0.001, 0.01)$-Separation Minimum Key algorithms. The result of Greedy Minimum Key on *census* is not available because the algorithm does not terminate in 10 hours.

tinct ratios for quasi-identifiers are adapted from the measures for quantifying approximations of functional dependencies proposed in [13, 22].

6 Conclusions and Future Work

In this paper, we designed efficient algorithms for discovering and masking quasi-identifiers in large tables. We developed efficient algorithms that find small quasi-identifiers with provable size and separation/distinct ratio guarantees, with space and time complexity sublinear in the number of input tuples. We also designed efficient algorithms for masking quasi-identifiers in large tables.

All algorithms in the paper can be extended to the weighted case, where each attribute is associated with a weight and the size/utility of a set of attributes is defined as the sum of their weights. The idea of using random samples to trade off between accuracy and space complexity can potentially be explored in other problems on large tables.

References

[1] Public-use microdata samples (pums). *http://www.census.gov/main/www/pums.html*.

[2] G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D.Thomas, and A. Zhu. Anonymizing tables. In *ICDT*, 2005.

[3] R. Ananthakrishna, S. Chaudhuri, and V. Ganti. Eliminating fuzzy duplicates in data warehouses. In *VLDB*, 2002.

[4] R. Bayardo and R. Agrawal. Data privacy through optimal k-anonymization. In *ICDE*, 2005.

[5] M. Charikar, S. Chaudhuri, R. Motwani, and V. Narasayya. Towards estimation error guarantees for distinct values. In *PODS*, 2000.

[6] S. Chaudhuri, K. Ganjam, V. Ganti, and R. Motwani. Robust and efficient fuzzy match for online data cleaning. In *SIGMOD*, 2003.

[7] M. R. Garey and D. S. Johnson. Computers and intractability. 1979.

[8] C. Giannella and E. Robertson. On approximation measures for functional dependencies. *Information Systems*, 2004.

[9] C. M. Giannella, M. M. Dalkilic, D. P. Groth, and E. L. Robertson. Using horizontal-vertical decompositions to improve query evaluation. *LNCS 2405*.

[10] B. Halldorsson, M. Halldorsson, and R. Ravi. Approximability of the minimum test collection problem. In *ESA*, 2001.

[11] Y. Huhtala, J. Karkkainen, P. Porkka, and H. Toivonen. Discovery of functional and approximate dependencies using partitions. In *ICDE*, 1998.

[12] D. Johnson. Approximation algorithms for combinatorial problems. In *J. Comput. System Sci.*, 1974.

[13] J. Kivinen and H. Mannila. Approximate dependency inference from relations. In *Theoretical Computer Science*, 1995.

[14] N. Li, T. Li, and S. Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *ICDE*, 2007.

[15] S. Lodha and D. Thomas. Probabilistic anonymity. *Technical Report*.

[16] Machanavajjhala, J. Gehrke, and D. Kifer. l-diversity: privacy beyond k-anonymity. In *ICDE*, 2006.

[17] A. Meyerson and R. Williams. On the complexity of optimal k-anonymity. In *PODS*, 2004.

[18] B. Moret and H. Shapiro. On minimizing a set of tests. In *SIAM Journal on Scientific and Statistical Computing*, 1985.

[19] R. Motwani and P. Raghavan. Randomized algorithm. 1995.

[20] S. Muthukrishnan. Data streams: Algorithms and applications. 2005.

[21] D. Newman, S. Hettich, C. Blake, and C. Merz. Uci repository of machine learning databases. *http://www.ics.uci.edu/∼mlearn/MLRepository.html*.

[22] B. Pfahringer and S. Kramer. Compression-based evaluation of partial determinations. In *SIGKDD*, 1995.

[23] P. Samarati and L. Sweeney. Generalizing data to provide anonymity when disclosing information. In *PODS*, 1998.

[24] L. Sweeney. Achieving k-anonymity privacy protection using generalization and suppression. In *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 2002.

[25] L. Sweeney. k-anonymity: a model for protecting privacy. In *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 2002.