Dynamic Anonymization for Marginal Publication

Xianmang He*, Yanghua Xiao, Yujia Li, Qing Wang, Wei Wang, and Baile Shi

Fudan University, Shanghai 200433, China
{071021057,shawyh,wangqing,wangwei1,bshi}@fudan.edu.cn

Abstract. Marginal publication is one of important techniques to help researchers to improve the understanding about correlation between published attributes. However, without careful treatment, it's of high risk of privacy leakage for marginal publications. Solution like ANGEL has been available to eliminate such risks of privacy leakage. But, unfortunately, query accuracy has been paid as the cost for the privacy-safety of ANGEL. To improve the data utility of marginal publication while ensuring privacy-safety, we propose a new technique called dynamic anonymization. We present the detail of the technique and theoretical properties of the proposed approach. Extensive experiments on real data show that our technique allows highly effective data analysis, while offering strong privacy guarantees.

Keywords: privacy preservation, marginal publication, dynamic anonymization, m-invariance.

1 Introduction

In recent year, we have witnessed the tremendous growth of the demand to publish personal data, which posed great challenges for protecting the privacy in these data. For example, medical records of patients may be released by a hospital to aid the medical study. Suppose that a hospital wants to publish records of Table 1, called microdata (T). Since attribute *Disease* is sensitive, we need to ensure that no adversary can accurately infer the disease of any patient from the published data. For this purpose, unique identifiers of patients, such as *Name* should be anonymized or excluded from the published data. However, it is still possible for the privacy leakage if adversaries have certain background knowledge about patients. For example, if an adversary knows that Bob is of age 20, Zipcode 12k and Sex M, s/he can infer that Bob's disease is bronchitis since the combination of Age, Zipcode and Sex uniquely identify each patient in Table 1. The attribute set that uniquely identify each record in a table is usually referred to as a quasi-identifier (QI for short) of the table.

* This work was supported in part by the National Natural Science Foundation of China (No.61003001, No.61033010 and NO.90818023) and Specialized Research Fund for the Doctoral Program of Higher Education (No.20100071120032).

J.B. Cushing, J. French, and S. Bowers (Eds.): SSDBM 2011, LNCS 6809, pp. 451–460, 2011.

Table 1. Microdata

	Age	Zip	Sex	Disease
Bob	20	12k	M	bronchitis
Alex	19	20k	M	flu
Jane	20	13k	F	pneumonia
Lily	24	16k	F	gastritis
Jame	29	21k	F	flu
Linda	34	24k	F	gastritis
Sarah	39	19k	M	bronchitis
Mary	45	14k	M	flu
Andy	34	21k	F	pneumonia

Table 2. Generalization T^*

GID	Age	Zip	Sex	Disease
1	[19-20]	[12k-20k]	M	bronchitis
1	[19-20]	[12k-20k]	M	flu
2	[20-24]	[13k-16k]	F	pneumonia
2	[20-24]	[13k-16k]	F	gastritis
3	[29-34]	[21k-24k]	F	flu
3	[29-34]	[21k-24k]	F	gastritis
4	[34-45]	[14k-21k]	*	bronchitis
4	[34-45]	[14k-21k]	*	flu
4	[34-45]	[14k-21k]	*	pneumonia

Table 3. Marginal $\langle Zip, Disease \rangle$

Zip	Disease
[12k-13k]	bronchitis
[12k-13k]	pneumonia
[14k-16k]	gastritis
[14k-16k]	flu
[19k-20k]	flu
[19k-20k]	bronchitis
[21k-24k]	gastritis
[21k-24k]	flu
[21k-24k]	pneumonia

Table 4. GT

GID	Zip	Batch-ID
1	[12k-13k]	1
1	[12k-13k]	2
2	[14k-16k]	4
2	[14k-16k]	2
3	[19k-20k]	1
3	[19k-20k]	4
4	[21k-24k]	3
4	[21k-24k]	3
4	[21k-24k]	4

Table 5. BT

Batch-ID	Disease	Count
1	bronchitis	1
1	flu	1
2	pneumonia	1
2	gastritis	1
3	flu	1
3	gastritis	1
4	bronchitis	1
4	flu	1
4	pneumonia	1

To protect privacy against attack guided by background knowledge, generalization has been widely used in privets anonymization solutions [1, 2, 3, 4, 5]. In a typical generalization solution, tuples are first divided into subsets (each subset is referred to as a QI-group). Then, QI-values of each QI-group are generalized into less specific forms so that tuples in the same QI-group cannot be distinguished from each other by their respective QI-values. As an example, we generalize Table 1 into Table 2 such that there exists at least two records in each QI-group. After generalization, the age(=20) of Bob has been replaced by an interval [19-20]. As a result, even if an adversary has the exact QI values of Bob, s/he can not exactly figure out the tuple of Bob from the first QI-group.

Motivation 1: Privacy leakage of marginal publication. Privacy preservation of generalization comes at the cost of information loss. Furthermore, generalization generally loses less information when the number of QI attributes is smaller [6]. Hence, to enhance the understanding about the underlying correlations among attributes, the publisher may further release a refined generalization of the projection on attributes of interest. This approach is referred to as *marginal publication*. For example, a researcher may request refined correlations of Zipcode and Disease several weeks later after the publication of Table 2. To satisfy the request, the publisher further publish Table 3, which is a more accurate generalization of $\langle Zipcode, Disease \rangle$ compared to that in Table 2, hence capturing the correlations between Zipcode and Disease better.

However, it is of possible risk of privacy leakage in solutions of marginal publication. Continue the above example. Suppose an adversary knows Bob's QI-values. Then, by Table 2 s/he infers that Bob's disease is in the set {bronchitis,

flu}. By Table 3, s/he infers that Bob has contracted the disease either pneumonia or bronchitis. By combining the above knowledge, the adversary makes sure that Bob have contracted bronchitis.

Motivation 2: Information loss of existing solutions. To overcome the privacy leakage of marginal publication, Tao et al. [5] propose an anonymization technique ANGEL (illustrated in Example 1), which releases each marginal with strong privacy guarantees. Many QI-groups of the anonymized table released by ANGEL may contain a large number of sensitive values. The number of these values in the worst case is quadratic to the number of tuples in the QI-group. As a result, there will exist significant average error when answering aggregate queries. To give a clear explanation, assume that a researcher wants to derive an estimation for the following query:

Select Count(∗) From Table GT and BT Where ZipCode ∈ [12k, 24k] And Disease='Penumonia'.

By estimating from Table 4 and 5, we can only get an approximate answer 4, which is much larger than the actual query result 2(see Table 1). Then, we may wonder whether there exists an approach that can protect privacy for marginal publication while ensuring the data utility of the published data. This issue is addressed in this paper.

Example 1. Suppose that the publisher need to release a marginal containing $\langle Zipcode, Disease\rangle$. If the privacy principle is 2-unique, the parameter k of ANGEL will be 2. After running ANGEL under this parameter, the result will be two tables GT (shown in Table 4 which is 2-anonymity) and BT (shown in Table 5 which is 2-unique).

Related Work. Although improving the data utility of marginal publication is desired, rare works can be found to solve this problem. We give a brief review on the previous works about marginal publication. It was shown in that when the set of marginals overlap with each other in an arbitrarily complex manner, evaluating the privacy risk is NP-hard [7, 8]. The work of [7], on the other hand, is applicable only if all the marginals to be published form a decomposable graph. The method in [8] requires that, except the first marginal, no subsequent marginal released can have the sensitive attribute. The work of [8] shows that, checking whether a set of marginals violates k-anonymity is a computationally hard problem. The method in the paper [9] requires that, except the first marginal, no subsequent marginal released can have the sensitive attribute. For example, after publishing Table 3(Marginal $\langle Zip, Disease\rangle$), the publisher immediately loses the option of releasing any marginal which contains the attribute Disease. This is a severe drawback since the sensitive attribute is very important for data analysis. The work that is closest to ours is ANGEL that is proposed by Tao et al. [5]. ANGEL can release any marginals with strong privacy guarantees, which however comes at the cost of information loss. Please refer to Section 1 for details.

Contributions and paper organization. To reduce the information loss of marginal publication, we propose a dynamic anonymization technique, whose effectiveness is verified by extensive experiments. We systematically explored the theoretic properties of marginal publication, and proved that the generalization principle m-invariance can be employed to ensure the privacy safety of marginal publication.

The rest of the paper is organized as follows. In Section 2, we give the preliminary concepts and formalize the problem addressed in this paper. In Section 3, we present the dynamic anonymization technique as our major solution. In Section 4, experimental results are evaluated. Finally, the paper is concluded in Section 5.

2 Problem Definition

In this section, we will formalize the problem addressed in this paper and explore some theoretic property of marginal publication.

Marginal Publication. Let T be a microdata table, which has d QI-attributes $A_1, ..., A_d$, and a sensitive attribute (SA) S. We consider that S is categorical, and every QI-attribute $A_i (1 \leq i \leq d)$ can be either numerical or categorical. For each tuple $t \in T, t.A_i (1 \leq i \leq d)$ denotes its value on A_i, and $t.A_s$ represents its SA value. We first give the fundamental concepts. A *QI-group* of T is a subset of the tuples in T. A *partition of* T is a set of disjoint QI-groups whose union equals T.

Now, we will formalize the key concepts in marginal publication. In the following texts, without loss of generality, we assume that all marginals released contain the sensitive attribute. A marginal published without the sensitive attribute is worthless for the QI-conscious adversary.

Definition 1 (Marginal). *Marginal M_j is a generalized version of certain projection on microdata T. The correspondent schema of the projection is referred to as* the schema of the marginal. *A* trivial marginal *is the marginal that contains no QI-attributes. Any other marginals are non-trivial.*

Given a microdata T, the number of its possible non-trivial marginals can be quantified by the following lemma. In following texts, without explicit statement, a marginal is always non-trivial.

Lemma 1. *There are $2^d - 1$ different non-trivial marginals, where d is the number of QI-attributes.*

Given multiple marginals of a microdata, it is possible for an adversary to infer privacy of a victim by combining knowledge obtained from different marginals. This implies that *the intersection of the sensitive sets obtained from different marginals must cover sufficiently large number of values so that the adversaries can not accurately infer the sensitive information of a victim.* One

principle achieving this objective is m-invariance. We first give the definition of m-invariance and two basic concepts to define m-invariance: signature and m-unique. The privacy guarantee of m-invariance is established by Lemma 3 in the paper [4].

Definition 2 (Signature, m-Unique [4,10]). *Let P be a partition of T, and t be a tuple in a QI-group $G \in P$. The signature of t in P is the set of distinct sensitive values in G. An anonymized version T^* is m-unique, if T^* is generated from a partition, where each QI-group contains at least m tuples, each with a different sensitive value.*

Definition 3 (m-Invariance [4,10]). *A set S of partitions is m-invariant if (1) Each partition in S is m-unique; (2) For any partitions $P_1, P_2 \in S$, and any tuple $t \in T$, t has the same signature in P_1 and P_2.*

In this paper, we adopt the normalized certainty penalty (NCP [3]) to measure the information loss. Now, we are ready to give the formal definition about the problem that will be addressed in this paper.

Definition 4 (Problem Definition). *Given a table T and an integer m, we need to anonymize it to be a set of marginals $M_j (1 \leq j \leq r)$ such that (1)Existence: these marginals are m-invariant; (2)Optimality: and the information loss measured by NCP is minimized.*

Existence of m-Invariant marginals. Given a table T and an integer m, is it possible to generate a set of marginals $M_j (1 \leq j \leq r)$ that is m-invariant? The answer is positive. We will show in Theorem 1 that if a table T is *m-eligible*, there exists a set of marginals $M_r (1 \leq j \leq r)$ that is m-invariant. A table T is *m-eligible* if it has at least one m-unique generalization. Then, to determine the existence of $m-$invariant marginals for a table, we only need to find a sufficient and necessary condition to characterize m-eligibility of a table, which is given in Theorem 2.

Theorem 1. *If a table T is m-eligible, then there exists a set of marginals $\{M_1, M_2, \cdots, M_r\}$ that is m-invariant.*

Theorem 2. *A table T is m-eligible, if and only if the number of tuples that have the same sensitive attribute values is at most $\frac{|T|}{m}$, where $|T|$ is the number of tuples in table T.*

3 Dynamic Anonymization Technique

In this section, we will present our the detail of our solution: dynamic anonymization, which contains three steps: partition, assign and decomposition. Each step will be elaborated in following texts.

The Partitioning Step. The partitioning step aims to partition tuples of T into disjoint sub-tables T_i such that each T_i is m-eligible. The detailed procedure

is presented in Figure 1. Initially, S contains T itself (line 1); then, each $G \in S$ is divided into two generalizable subsets G_1 and G_2 such that $G_1 \cup G_2 = G,\ G_1 \cap G_2 = \emptyset$ (line 5-7). Then for each new subset, we check whether $G_1(G_2)$ satisfies m-eligible (line 8). If both are generalizable, we remove G from S, and add $G_1,\ G_2$ to S; otherwise G is retained in S. The attempts to partition G are tried k times and tuples of G are randomly shuffled for each time (line 3-4). Our experimental results show that most of G can be partitioned into two m-eligible sub-tables by up to $k = 5$ tries. The algorithm stops when no sub-tables in S can be further partitioned.

In the above procedure, the way that we partition G into two subsets G_1 and G_2 is influential on the information loss of the resulting solution. To reduce information loss, we *distribute tuples sharing the same or quite similar QI-attributes into the same sub-tables*. For this purpose, we artificially construct two tuples $t_1, t_2 \in G$ with each attribute taking the maximal/minimal value of the corresponding domains, and then insert them G_1 and G_2 separately (line 6). After this step, for each tuple $w \in G$ we compute $\Delta_1 = NCP(G_1 \cup w) - NCP(G_1)$ and $\Delta_2 = NCP(G_2 \cup w) - NCP(G_2)$, and add tuple w to the group that leads to lower penalty (line 7). After successfully partitioning G, remove the artificial tuples from G_1 and G_2 (line 8).

Input: A microdata T, integers k and m
Output: A set S consisting of sub-tables of T;
/* the parameter k is number of rounds to partition G*/
1. $S = \{T\}$;
2. While($\exists G \in S$ that has not been partitioned)
3. For $i = 1$ to k
4. Randomly shuffle the tuples of G;
5. Set $G_1 = G_2 = \emptyset$;
6. Add tuple t_1 (t_2) of extremely maximal (minimal) value to G_1 (G_2);
7. For any tuple w
compute Δ_1 and Δ_2.
If($\Delta_1 < \Delta_2$) then Add w to G_1, else add w to G_2;
8. If both G_1 and G_2 are $m-$eligible
remove G from S, and add $G_1 - \{t_1\}, G_2 - \{t_2\}$ to S, **break**;
9.Return S;

Fig. 1. The partitioning step

The Assigning Step. After the partitioning step, we enter into the assigning step, which is accomplished by the assign algorithm proposed in paper [4]. Given a set of sub-tables T_i passed from the previous phase, the assigning step is to divide each T_i into buckets such that each bucket constitutes a bucketization. The concepts about bucket and bucketization are given by following definitions.

Definition 5 (Bucket [10],Bucketization). *Given T and T^*, a bucket B is a set of tuples in T whose signatures in T^* are identical. The signature of a bucket B is the set of sensitive values that appear in B. A bucketization U is a set of disjoint buckets, such that the union of all buckets equals T.*

The Decomposition Step. In real applications, a publisher may want to release a set of marginals $\{M_1, M_2, \cdots, M_r\}$ that overlap with each other in an arbitrary manner. To help publishers accomplish this, we use the third step: decomposition step to produce a set of marginals $\{M_1, M_2, \cdots, M_r\}$ that are m-invariant. Depending on marginals of different attribute sets, the bucketization U is decomposed differently. Each decomposition of U is a partition of the microdata T. All the partitions constitute an m-invariant set while offering strong privacy guarantees.

Definition 6 (Decomposition [10]). *Let B be a bucket with signature K. A decomposition of B contains $\frac{|B|}{|K|}$ disjoint QI-groups whose union is B, and all of them have the same signature K.*

The decomposition algorithm runs iteratively and maintains a set bukSet of buckets. Let $B \in U$ be a bucket with a signature containing $s \geq m$ sensitive values $\{v_1, v_2, ..., v_s\}$. The decomposition phase starts by initializing a set bukSet= $\{B\}$. Then, we recursively decompose each bucket B_i in bukSet that contains more than s tuples into two buckets B_1 and B_2 until each bucket in bukSet contain exactly s tuples. The final bukSet is returned as the QI-groups for generalization. The resulting decomposition is guaranteed to be m-invariant, which is stated in Theorem 3.

Now, we elaborate the detailed procedure to decompose a single bucket $B \in U$ with the signature $K = \{ v_1, v_2, \cdots, v_s\}$ into B_1, B_2. Suppose the schema of marginal $M_j (1 \leq j \leq r)$ is $\langle A_1, A_2, ..., A_t \rangle$. We first organize B into s groups such that the i-th $(1 \leq i \leq s)$ group denoted by Q_i contains only the tuples with the sensitive value v_i. Then, by one attribute A_i, we can sort the tuples in each group into the ascending order of their A_i values. After sorting by A_i, we assign the first $\frac{|Q_i|}{2} (1 \leq i \leq s)$ tuples to B_1, and the remaining tuples to B_2. In this way, we get a decomposition of B by A_i. Similarly, we can get another $t - 1$ decompositions by A_j with $j \neq i$. Among all the t decompositions, we pick the one that minimizes the sum of $NCP(B_1)$ and $NCP(B_2)$ as the final decomposition.

Theorem 3. *Given the bucketization U, the marginals $M_j (1 \leq j \leq r)$ produced by the decomposing algorithm are m-invariant.*

Since our marginals $M_j (1 \leq j \leq r)$ enforce m-invariance, we can guarantee the privacy preservation of marginals produced by above decomposition, which is given in following corollary. The computation of above decomposition is also efficient enough (see Theorem 4).

Corollary 1. *A QI-conscious adversary has at most $\frac{1}{m}$ confidence in inferring the sensitive value of any individual in $M_j (1 \leq j \leq r)$, even if s/he is allowed to obtain all versions of $M_j (1 \leq j \leq r)$.*

Theorem 4. *For a single bucket B, the decomposition algorithm can be accomplished in $O(|B| \cdot log^2(|B|) \cdot t)$, where t is the size of attributes of the required marginal, and $|B|$ is the cardinality of bucket B.*

4 Experimental Evaluation

In this section, we experimentally evaluate the effectiveness and efficiency of the proposed technique. We utilize a real data set CENSUS (http://ipums.org) that is widely used in the related literatures. We examine five marginals M_1, M_2 ... M_5, whose dimensionalities are 2, 3, ..., 6, respectively. Specifically, M_1 includes attributes Age and Occupation, M_2 contains attributes of M_1 and Gender(for simplicity, we denote $M_2 = M_1 \cup \{Gender\}$), $M_3 = M_2 \cup \{Education\}$, $M_4 = M_3 \cup \{Marital\}$, and $M_5 = M_4 \cup \{Race\}$. We run all experiments on a PC with 1.9 GHz CPU and 1 GB memory. For comparisons, the results of the state-of-art approach: ANGEL will also be given.

Utility of the Published Data. Since data utility of ANGEL can't be measured by NCP, for the fairness of comparison, we evaluate the utility of the published data by summarizing the accuracy of answering various aggregate queries on the data, and has been widely accepted in the literature [3, 4, 10, 11]. Specifically, each query has the form: *select count(*) from* M_j *where* $A_1 \in b_1$ *and* $A_2 \in b_2$ *and* $\cdots$ *and* $A_w = b_w$, where w is query dimensionality. $A_1, ..., A_{w-1}$ are $w-1$ arbitrary distinct QI-attributes in M_j, but A_w is always Occupation. Each $b_i (1 \leq i \leq w)$ is a random interval in the domain of A_i. The generation of $b_1, \cdots, b_w$ is governed by a real number $s \in [0, 1]$, which determines the length of range $b_i (1 \leq i \leq w)$ as $\lfloor |A_i| \cdot s^{1/w} \rfloor$. We derive the estimated answer of a query using the approach explained in [11]. The accuracy of an estimation is measured by its relative error, which is measure by $|act - est|/act$ where act and est denote the actual and estimated results, respectively.

We conduct the first set of experiments to explore the influence of m on data utility. Towards this, we vary m from 4 to 10. Figure 2 plots the error as a function m. Compared to ANGEL that produces about 50%-200% average error, the published data produced by our algorithm is significantly more useful. It is quite impressive to see that the error of our algorithm is consistently below 11 % despite of the growth of m. Figure 3 shows the error as a function of $M_i (2 \leq i \leq 5)$(ANGEL is omitted due to its high query error). We can see that the accuracy increases as the number of QI-attributes of the marginal decreases, which can be attributed to the fact that M_i is more accurate than $M_{i+1} (2 \leq i \leq 4)$.

We evaluate the influence of parameter s on the query error. The result is shown in Figure 4. Evidently, the accuracy is improved for larger s, which can be naturally explained since larger s implies larger query intervals, which in return reduce the query error.

Efficiency. We evaluate the overhead of performing marginal publication. Figure 5 shows the cost of computing marginals $(M_1, \cdots, M_5)$ for m varying from 4 to 10. We can see that the cost drops as m grows. However, the advantages of our method in data utility do not come for free. From figure 5, we can see that the time cost of our algorithm is 115% to 143% of ANGEL, however, which is acceptable especially for those cases where query accuracy is the critical concern.

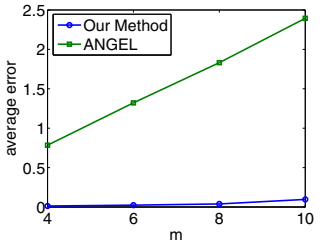

Fig. 2. Accuracy vs. $m(M_5)$

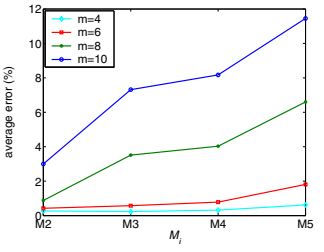

Fig. 3. Accuracy vs. $M_i(s = 0.1)$

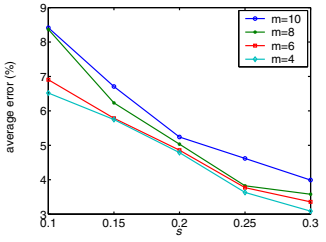

Fig. 4. Accuracy vs. $s(M_5)$

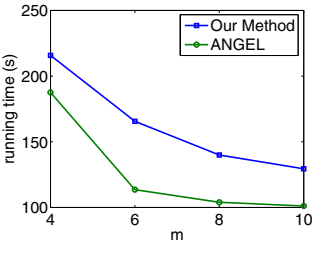

Fig. 5. Running time

5 Conclusion

In this paper, we systematically investigate characteristics of marginal publications. We propose a technique called dynamic anonymization to produce a set of anonymized marginals for a given schema of marginals. As verified by extensive experiments, the marginals produced by our approach not only guarantees the privacy safety of published data but also allows high actuary of query estimation.

References

1. Sweeney, L.: k-anonymity: a model for protecting privacy. Int. J. Uncertain. Fuzziness Knowl. -Based Syst. 10(5), 557–570 (2002)
2. Samarati, P.: Protecting respondents' identities in microdata release. IEEE Trans. on Knowl. and Data Eng. 13(6), 1010–1027 (2001)
3. Xu, J., Wang, W., Pei, J., Wang, X., Shi, B., Fu, A.W.-C.: Utility-based anonymization using local recoding. In: KDD 2006, New York, pp. 785–790 (2006)
4. Xiao, X., Tao, Y.: M-invariance: towards privacy preserving re-publication of dynamic datasets. In: SIGMOD 2007, pp. 689–700. ACM, New York (2007)
5. Tao, Y., Chen, H., Xiao, X., Zhou, S., Zhang, D.: Angel: Enhancing the utility of generalization for privacy preserving publication. TKDE, 1073–1087 (2009)
6. Aggarwal, C.C.: On k-anonymity and the curse of dimensionality. In: VLDB 2005, pp. 901–909. VLDB Endowment (2005)

7. Kifer, D., Gehrke, J.: Injecting utility into anonymized datasets. In: SIGMOD 2006, New York, NY, USA, pp. 217–228 (2006)
8. Yao, C., Wang, X.S., Jajodia, S.: Checking for k-anonymity violation by views. In: VLDB 2005, pp. 910–921. VLDB Endowment (2005)
9. Wang, K., Fung, B.C.M.: Anonymizing sequential releases. In: KDD 2006, New York, NY, USA, pp. 414–423 (2006)
10. Xiao, X., Tao, Y.: Dynamic anonymization: accurate statistical analysis with privacy preservation. In: SIGMOD 2008, pp. 107–120 (2008)
11. Zhang, Q., Koudas, N., Srivastava, D., Yu, T.: Aggregate query answering on anonymized tables. In: ICDE 2006, pp. 116–125 (2007)