Decomposition: Privacy Preservation for Multiple Sensitive Attributes

Yang Ye[1], Yu Liu[2], Chi Wang[2], Dapeng Lv[2], and Jianhua Feng[2]

[1] Institute for Theoretical Computer Science, Tsinghua University
Beijing, 100084, P.R.China
yey05@mails.tsinghua.edu.cn
[2] Department of Computer Science, Tsinghua University
Beijing, 100084, P.R.China
{liuyu-05,wangchi05,lvdp05}@mails.tsinghua.edu.cn
fengjh@tsinghua.edu.cn

Abstract. Aiming at ensuring privacy preservation in personal data publishing, the topic of anonymization has been intensively studied in recent years. However, existing anonymization techniques all assume each tuple in the microdata table contains one single sensitive attribute (the *SSA* case), while none paid attention to the case of multiple sensitive attributes in a tuple (the *MSA* case). In this paper, we conduct the pioneering study on the MSA case, and propose a new framework, decomposition, to tackle privacy preservation in the MSA case.

1 Introduction

Anonymization[1] is the most popularly adopted approach for privacy-preserving data publishing. Anonymization techniques typically perform *generalization*[1,2] on QI attributes, as depicted in Table 3. Principles such as k-anonymity[1] put constraints on each QI-group. The widely adopted principle l-diversity[3] requires each group contains at least l "*well-represented*" sensitive values, and reduces the risk of *sensitive attribute disclosure* to no higher than *1/l*.

Current researches on anonymization all assume there is one single sensitive attributes (the SSA case) in the microdata table. This assumption is arbitrary. In the running example, two attributes, *Occupation* and *Salary* are sensitive attributes. Consider an adversary who obtains the QI values $\{M, 10076, 1985/03/01\}$ of Carl. Given the published Table 3, s/he can locate Carl in the first QI-group. However, since the first two tuples of Group 1 have "*nurse*" as the occupation value and according to common sense, nurse is generally a female occupation, thereby the adversary can locate Carl in the last two tuples. S/he will be able to reveal with high confidence that Carl's monthly salary is 8000-10000 dollars (In tables of this paper, integer i in "salary" column means the monthly salary is between the range of $1000i - 1000(i + 1)$ dollars).

This paper provides the first study towards privacy preservation in the MSA case. We propose a new publishing methodology, decomposition, to achieve privacy preservation in the MSA case. Instead of performing generalization on QI

X. Zhou et al. (Eds.): DASFAA 2009, LNCS 5463, pp. 486–490, 2009.

Table 1. The Microdata Table

Tuple#	Gender	ZipCode	Birthday	Occupation	Salary [1]
1(Alice)	F	10078	1988/04/17	nurse	1
2(Betty)	F	10077	1984/03/21	nurse	4
3(Carl)	M	10076	1985/03/01	police	8
4(Diana)	F	10075	1983/02/14	cook	9
5(Ella)	F	10085	1962/10/03	actor	2
6(Finch)	M	10085	1988/11/04	actor	7
7(Gavin)	M	20086	1958/06/06	clerk	8
8(Helen)	F	20087	1960/07/11	clerk	2

Table 2. Part of a Vote Register List

Name	Gender	ZipCode	Birthday
Alice	F	10078	1988/04/17
Betty	F	10077	1984/03/21
Carl	M	10076	1985/03/01
Diana	F	10075	1983/02/14
Ella	F	10085	1962/10/03
Finch	M	10085	1988/11/04
Gavin	M	20086	1958/06/06
Helen	F	20087	1960/07/11

Table 3. The Generalized Table

#	Gender	ZipCode	Birth.	Occ.	Sal.
1	*	1007*	1983-88	nurse	1
2	*	1007*	1983-88	nurse	4
3	*	1007*	1983-88	police	8
4	*	1007*	1983-88	cook	9
5	*	*008*	1958-88	actor	2
6	*	*008*	1958-88	actor	7
7	*	*008*	1958-88	clerk	8
8	*	*008*	1958-88	clerk	2

attributes and forming QI-groups, our technique decomposes the table into so-called *SA-groups*. To retain valuable information lost in the transformed sensitive attributes, the original *sensitive table* is also published without privacy leakage.

2 General Idea of Decomposition

We term our methodology "*decomposition*". Firstly, it publishes the decomposed sensitive table. Secondly, instead generalized on QI attributes, tuples are grouped properly. Their QI values remain unchanged while tuples within a group share the union of their sensitive values, as shown in Table 4 and Table 5.

Definition 1. *(*SA-group*) A SA-group G contains tuples with their original, non-transformed QI values and for each S^i, each tuple in G is associated with the set of $G.S^i$ values.*

We first assume there is one single sensitive attribute S and aim at achieving l-diversity. We shall research, given a diversity parameter l, how to decompose the table into SA-groups so that: (i) each group had better contains exactly l distinct sensitive values. (ii) the number of such SA-groups should be maximized. Following **Largest-l group forming Procedure** is adopted: first place tuples with identical sensitive values into a same "*bucket*". Let B_i denote the i^{th} largest bucket and $\mathscr{B} = \{B_1, B_2, \ldots, B_m\}$ denote the set of buckets. We have:

Table 4. The Decomposed Table for Single Sensitive Attribute

#	Gender	Zip.	Birth.	Occ.
1	F	10078	1988/04/17	police
	F	10085	1962/10/03	nurse
	M	20086	1958/06/06	actor
	M	10076	1985/03/01	clerk
2	F	10077	1984/03/21	nurse
	M	10085	1988/11/04	actor
	F	10075	1983/02/14	cook
	F	20087	1960/07/11	clerk

Table 5. The Decomposed Table for Two Sensitive Attributes

#	Gender	Zip	Birth.	Occ.	Sal.
1	F	10078	1988/04/17	police	1
	F	10085	1962/10/03	nurse	2
	M	20086	1958/06/06	actor	8
	M	10076	1985/03/01	clerk	
2	F	10077	1984/03/21	nurse	2
	M	10085	1988/11/04	actor	4
	F	10075	1983/02/14	cook	7
	F	20087	1960/07/11	clerk	9

$n_i = |B_i|,\ n_1 \geq n_2 \geq \cdots \geq n_m$. In each iteration, one tuple is removed from each of the l largest buckets to form a new SA-group. Similar to [5], we can prove:

Theorem 1. *The Largest-l group forming procedure creates as many groups as possible.*

We shall also investigate: (*i*) in which case there will be not tuples left after the procedure; and (*ii*) what is the property of residual tuples, if any.

Theorem 2. *When the Largest-l group forming procedure terminates, there will be no residual tuples if and only if the buckets formed after the bucketizing step satisfy the following properties (we term it l-Property):*
(i) $\frac{n_i}{n} \leq \frac{1}{l},\ i = 1, 2, \ldots, m$ *(Use the same notation:* n_i, m, n *as in Theorem 1);*
(ii) $n = kl$ *for some integer* k.

When the buckets formed through bucketization satisfy the first condition while do not satisfy the second condition of l-Property, we have following conclusion:

Corollary 1. *If the buckets satisfy:* $\frac{n_i}{n} \leq \frac{1}{l}$*, then when the Largest-l group forming terminates, each non-empty bucket contains just one tuple.*

Corollary 2. *The largest permissable assignment to the diversity parameter l is* $l_{per} = \lfloor \frac{n}{n_1} \rfloor$

The extension of Decomposition to the MSA case is intuitive. First, the sensitive table T^S is published. Next, one sensitive attribute (denoted S^{pri}), is chosen as the "*primary sensitive attribute*" and largest-l procedure is exerted on S^{pri} to form SA-groups.

Definition 2. *(*Primary Sensitive Attribute*) In the MSA case, the* primary sensitive attribute *is the sensitive attribute chosen by the publisher, according to which SA-groups are formed.*

Third, for each SA-group and each non-primary sensitive attribute, the original values are united up, as depicted in Table 5. Reduplicated values are counted once because multiple counts just increase the privacy disclosure risk. We should not assign a uniform l for all S^i. Instead, each S^i should have its own l_i.

Table 6. The Final Publishing of Decomposition

The Sensitive Table		The Decomposed Table after Adding Noise					
Occupation	**Salary**	**Group#**	**Gender**	**ZipCode**	**Birthday**	**Occupation**	**Salary**
nurse	1	1	F	10078	1988/04/17	police	1
nurse	4		F	10085	1962/10/03	nurse	2
police	8		M	20086	1958/06/06	actor	4
cook	9		M	10076	1985/03/01	clerk	8
actor	2	2	F	10077	1984/03/21	nurse	2
actor	7		M	10085	1988/11/04	actor	4
clerk	8		F	10075	1983/02/14	cook	7
clerk	2		F	20087	1960/07/11	clerk	9

Definition 3. *($(l_1, l_2, \ldots, l_d)$-*diversity*) A decomposed table is said to satisfy $(l_1, l_2, \ldots, l_d)$-diversity, if for each of its SA-group G and each $i \in \{1, 2, \ldots, d\}$, $G.S^i$ contains at least l_i distinct sensitive values.*

As for some non-primary sensitive attribute S^i, there may be groups with less than l_i distinct S_i values, like in Group 1 of Table 5, $l_{per}(Salary) = \frac{8}{2} = 4$. For Group 1 to satisfy the privacy goal, some "*noise*" is added. In sum, the final publishing of *decomposition* is shown in Table 6.

3 Experiments

In the experiments, we utilized the "Adult" database from the UCI Machine Learning Repository (http://www.ics.uci.edu/mlearn/mlrepository.html) and the *KL-divergence* metric to measure data quality.

First we treat *Work-class* as the sensitive attribute and develop 4 tables from Adult: q-QI-Adult ($5 \leq q \leq 8$). q-QI-Adult takes the first d of other attributes as QI. We compare decomposition against the widely-adopted multi-dimensional generalization algorithm Mondrian[4] when achieving l-diversity. Figure 1 through Figure 4 depicts the KL-divergence of the anonymized datasets created by two algorithms. We also compare the execution time of both techniques. For lack of space, only the result on 8-QI-Adult is in Figure 5. Decomposition greatly outperforms generalization in both data quality and efficiency.

For the MSA case, we develop 4 tables: d-SA-Adult ($1 \leq d \leq 4$). d-SA-Adult uses the first 5 attributes as QI attributes and the subsequent d attributes as sensitive attributes. *Work-Class* is treated as primary sensitive attribute. Figure 6 depicts the KL-divergence of decompositiond d-SA-Adult tables where l_{pri} is set from 3 to $l_{per}(work\text{-}class) = 7$. For each non-primary sensitive attribute S^i, l_i is set to $l_{per}(S^i)$. In Figure 6, the experimental result is quite close to the theoretical estimation of $\log(\prod_i l_i)$. Figure 7 depicts the execution time of decomposition on d-SA-Adult tables. Again, each non-primary sensitive attribute is set to its largest permissible diversity parameter while l_{pri} varies from 3 to 7.

We conduct a separate experiment to measure the number of noises in the MSA case. This experiment is on 2-SA-Adult, which takes *Work-Class* as the

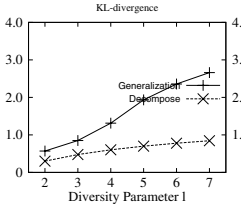

Fig. 1. 5-QI (SSA)

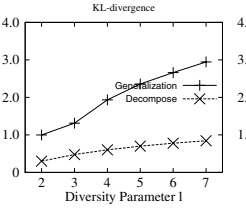

Fig. 2. 6-QI (SSA)

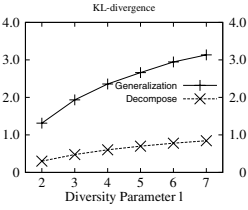

Fig. 3. 7-QI (SSA)

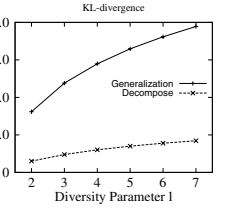

Fig. 4. 8-QI (SSA)

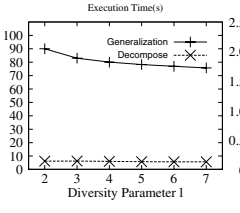

Fig. 5. Time (SSA)

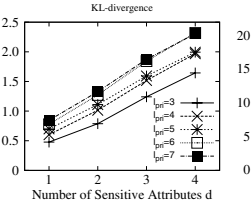

Fig. 6. MSA

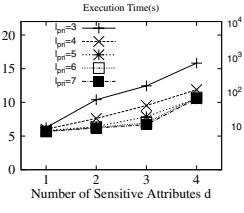

Fig. 7. Time (MSA)

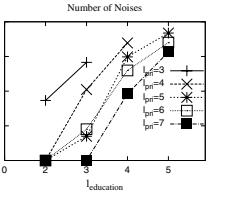

Fig. 8. Noise (MSA)

primary sensitive attribute and *Education* as the non-primary sensitive attribute. Figure 8 depicts the number of noises as the function of l_{pri} and $l_{Education}$.

4 Conclusions

This paper conducts the pioneering research towards privacy preservation in the MSA case, and lays down a foundation for future works, including combining categorical and numerical sensitive attributes, working on dynamic dataset, etc.

Acknowledgments. This work is partly supported by the National Natural Science Foundation of China Grant 60873065, 60553001, the National Basic Research Program of China Grant 2007CB807900, 2007CB807901, and the National High Technology Development 863 Program of China Grant 2007AA01Z152.

References

1. Sweeney, L.: Achieving k-anonymity privacy protection using generalization and suppression. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems 10(5), 571–588 (2002)
2. LeFevre, K., DeWitt, D.J., Ramakrishnan, R.: Incognito: efficient full-domain k-anonymity. In: SIGMOD, pp. 49–60 (2005)
3. Machanavajjhala, A., Gehrke, J., Kifer, D., Venkitasubramaniam, M.: l-diversity: privacy beyond k-anonymity. In: ICDE, pp. 24–26 (2006)
4. LeFevre, K., DeWitt, D.J., Ramakrishnan, R.: Mondrian: multidimensional k-anonymity. In: ICDE, p. 25 (2006)
5. Ye, Y., Deng, Q., Wang, C., Lv, D., Liu, Y., Feng, J.-H.: BSGI: An Effective Algorithm towards Stronger l-Diversity. In: Bhowmick, S.S., Küng, J., Wagner, R. (eds.) DEXA 2008. LNCS, vol. 5181, pp. 19–32. Springer, Heidelberg (2008)