iTcs English

清华大学理论计算机科学研究中心

PRIVACY PRESERVATION FOR MULTIPLE SENSITIVE ATTRIBUTES

Yang Ye

Institute forTheoretical Computer Science, Tsinghua University, Beijing 100084, China

yey05@mails.tsinghua.edu.cn

Dapeng Lv, Yu Liu, Jianhua Feng

Dept. of Computer Science, Tsinghua University, Beijing 100084, China

{liuyu-05,lvdp05}@mails.tsinghua.edu.cn, fengjh@tsinghua.edu.cn

Ⅰ.Background:

1. **Three types of attributes in the mcrodata:**

(1) Identifying Attributes：
Name, Social Security Number

(2) Quasi-identifying (QI) Attributes:
Date of birth, Zip code and Gender

(3) Sensitive Attributes:
Disease, Salary

2. **Objectives of Privacy Preservation:**

(1) Prevent direct or indirect disclosure of sensitive values.

(2) Enable the researcher to effectively investigate the relationship between sensitive attributes and other attributes.

3. **The prevalent privacy preservation technique: anonymization**

(1) Eliminate Identifying Attributes.

(2) Generalization on QI Attributes to form QI groups.

(3) Anonymization principles on QI groups:
k-anonymity, l-diversity, etc.

4. **Anonymization principles: bound the strength of privacy preservation**

(1) K-anonymity(against link attack):
Each QI group with size at least k .

(2) L-diversity(against homogenous and background attack):
Each QI group contains at least l "well-represented" sensitive values.

Ⅱ.Two Different Cases of Privacy Preservation:

1. **The SSA Case(the work of the state of the art):**
Each tuple contains one single sensitive attribute
2. **The MSA Case(our Work):**
Each tuple contains multiple sensitive attributes

Ⅲ. The Running Example

TABLE I
THE MICRODATA TABLE

Tuple#	Sex	Zip	Birth.	Occ.	Sal.[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 II
PART OF A VOTE REGISTER LIST

Name	Sex	Zip	Birth.
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 III
THE GENERALIZED TABLE

#	Sex	Zip	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

1. Here the value i for **salary** means monthly income is **1000i----1000(i+1)** dollars.
2. In this example, "**Sex**", "**Zip**" and "**Birth.**" are treated as QI attributes while "**Occ**." and "**Salary**" are sensitive attributes.
3. Through **generalization**, a generalized table, TABLE Ⅲ, is formed, Which contains two QI-groups.
4. Each group contains 4 tuples, therefore, TABLE Ⅲ satisfies **4-anonymity.**
5. Besides, for either sensitive attribute "**Occ.**" or "**Salary**", the first group contains at least 3 different values. **Group 1** satisfies **3-diversity** for "Occ." and "Salary" respectively.

Ⅳ. New Privacy Risks in the MSA cases

1. **Chain Attack:**
If an adversary locate **Carl** in the first group and he previously knows Carl' **occupation** is "police", he will obtain Diana's "**Salary**" information of "**8000-9000**" with full confidence. This kind of attack is termed "**link attack**".
2. **Exclusion Attack:**
If an adversary locates **Carl** in the first group and although he does not know Carl's occupation, but he can conclude Carl is not a **nurse** because nurse is an job for women. So, the adversary knows Carl corresponds to the 3rd or the 4th tuple, so he can deduce Carl's "**Salary**" information is "**8000-10000**". This kind of attack is termed "**exclusion attack**".
3. **The Mechanism of new Privacy Risks in Anonymization in the MSA cases.**
In each group, the distribution of "Salary" attribute for each value of "Occupation" l lacks diversity.
4. A QI group like group **2** may be satisfactory, in which each occupation value corresponds to 2 different salary values. However, we can prove to obtain such groups to cover the whole table is indeed impractical.

Ⅴ. Our solution: Decompose

1. Decompose in the **SSA** case:
Resembles "**Anatomy**" and "**Permutation**" Table will be partitioned into "**SA-groups**"

DEFINITION 1 (SA-Group)
A SA-group contains tuples with untransformed QI values and each tuple is associated with the union of these tuples' original sensitive values.

2. Choose one Sensitive attribute as "**Primary Sensitive Attribute**" and form SA-groups with this attribute
The **Group-Forming Method**: To Be Discussed Latter.

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.

3. Extending to Multiple Sensitive Attributes: By possibly **adding noise** in other attributes.
The **Noise-Adding Method**: To Be Discussed Latter.

ⅵ. Decompose in the SSA case:

1. The **Diversity Parameter l:** resembles the concept in l-diversity
2. The **Group-Forming Method : Largest---l M**ethod

(1) Bucketization :
Tuples with identical sensitive attributes are placed in the same bucket;
Bi: the i-th largest buckets
| *Bi* | =ni

(2) Group Forming:
In each iteration of group forming, one tuple is removed from each of the l largest buckets to form a new SA-group

after one iteration, the size of some buckets will be changed. So in the beginning of every iteration, the buckets are sorted according to their sizes

The result of decompose on **Table Ⅱ** is depicted in **Table Ⅳ**

3. Intensive Study on **Largest—l M**ethod:

Theorem 1: The **Largest-l** group forming method creates as many groups as possible

DEFINITION 3 (l-Property) : The original data distribution satisfies **l-Property** iff

(1) ni / n ≤ 1 / l
(2) n = k · l

Theorem 2: If the original data distribution satisfies **l-Property**, then after the **Largest-l** Group Forming method, no tuple will be left.

Corollary 1 : If the original data distribution satisfies **l-Property (1)** while does not satisfy **l-Property (2),** then after the **Largest-l** Group Forming procedure , there will be only one tuple in the non-empty buckets.

Corollary 2: The optimal assignment of diversity parameter l is **| n/n1** ⌋

TABLE IV
THE DECOMPOSED TABLE FOR SINGLE SENSITIVE ATTRIBUTE

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

TABLE V
THE DECOMPOSED TABLE FOR TWO SENSITIVE ATTRIBUTES

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

Ⅶ. Extending Decompose to the MSA case.

1. Form SA-groups according to the **Primary Sensitive Attribute**
2. For each other group and each other sensitive attribute, unite up the original values, reduplicated values are counted just once, because multiple counts just increase privacy disclosure risk, as shown in **Group 1** of **Table Ⅴ** (salary value 8)

3. Possibly in combination with **Adding Noise**

(1) In **Group 1** of **Table Ⅴ**, there are only 3 distinct values for "salary". However, the optimal assignment of diversity parameter for "salary" is **l=8/2=4**.

(2) Noise values are not arbitrarily chosen, in fact, **4** and **7** are allowed noise values while **9** is not**.**

(3) Because by linking **Group 1** 's **Occupation** values with the **sensitive table**, the adversary can deduce **9** cannot be a salary value of Group 1 . *Detailed implementation of **Adding Noise** is neglected for lack of space*

(4) The final publishing of Decompose is shown in Table Ⅵ, where we choose 4 as the noise value

TABLE VI
THE FINAL PUBLISHING OF DECOMPOSE

The Sensitive Table

Occ.	Sal.
nurse	1
nurse	4
police	8
cook	9
actor	2
actor	7
clerk	8
clerk	2

The Decomposed Table after Adding Noise

Grp#	Sex	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	4
	M	10076	1985/03/01	clerk	8
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

Ⅷ. Experiments

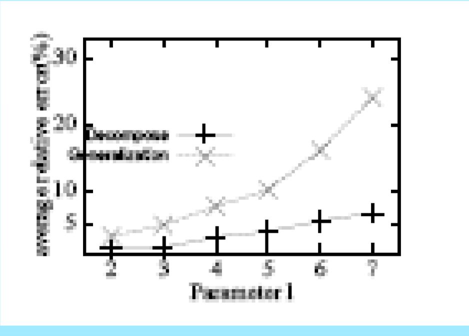
1. Setting
Real Database : **Adult (Downloaded at :**http://www.ics.uci.edu/m~learn/mlrepository.html)

2. The measurement : *average relative error* in answering aggregate query
relative error = |act – est|/act
act = actual result derived from the microdata
est = the estimate computed from the published table.

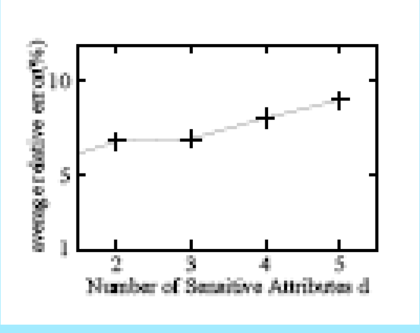
3. Aggregate Query:
SELECT COUNT(*) FROM Unknown-Microdata
WHERE pred(A 1) AND ... AND pred(A q) AND pred(S 1) ... AND pred(S d)

pred(A) of the form (A = x1 OR A = x2 OR ... OR A = xb)

4 Comparison between Decompose in SSA and Generalization
(Parameter l)



5 . Decompose for multiple sensitive attributes
(Parameter d: number of)



Ⅸ. Acknowledgement

This work was Supported in part by the National Natural Science Foundation of China Grant 60553001, 60573094, the National Basic Research Program of China Grant 2007CB807900, 2007CB807901, the National High Technology Development 863 Program of China under Grant No.2007AA01Z152 and 2006AA01A101, the National Grand Fundamental Research 973 Program of China under Grant No.2006CB303103, and Basic Research Foundation of Tsinghua National Laboratory for Information Science and Technology (TNList).

清華大學 Tsinghua University