Privacy Protection on Multiple Sensitive Attributes

Zhen Li and Xiaojun Ye

Key Laboratory for Information System Security, Ministry of Education
School of Software, Tsinghua University, 100084 Beijing, China
li-zhen05@mails.tsinghua.edu.cn, yexj@tsinghua.edu.cn

Abstract. In recent years, a privacy model called *k-anonymity* has gained popularity in the microdata releasing. As the microdata may contain multiple sensitive attributes about an individual, the protection of multiple sensitive attributes has become an important problem. Different from the existing models of single sensitive attribute, extra associations among multiple sensitive attributes should be invested. Two kinds of disclosure scenarios may happen because of *logical associations*. The *Q&S Diversity* is checked to prevent the foregoing disclosure risks, with an α *Requirement* definition used to ensure the diversity requirement. At last, a two-step greedy generalization algorithm is used to carry out the multiple sensitive attributes processing which deal with quasi-identifiers and sensitive attributes respectively. We reduce the overall distortion by the measure of *Masking SA*.

1 Introduction

In this information growing society, there has been a tremendous growth in the amount of personal data that will be collected and analyzed. In many scenarios, access to large amounts of personal data is essential for accurate inferences to be drawn. As a original form of information, microdata is a valuable source of data for the allocation of public funds, medical research, and trend analysis. For example, a hospital may release patient's diagnosis records so that researchers can analyze the characteristics of various diseases or use them to produce various statistical reports. The data providers must be careful when providing outside users access to such data, because they have obligations towards the individuals to which the data refer to make sure that it is (almost) impossible for a user to use the data to disclose confidential information about these individuals[11,14]. In order to use personal data without disclose confidential information, efficient anonymization techniques should be adopted. First, some uniquely identifying attributes like *names* and *social security numbers* are removed from the table. However, sets of other attributes (like *age*, *gender*, and *zipcode*) can be linked with external data to uniquely identify individuals. These attributes are called *quasi-identifier*[1]. To prevent linking attacks using quasi-identifiers, Samarati and Sweeney proposed a model of privacy called *k-anonymity*[9]. A release of microdata is said to adhere to k-anonymity if each

S. Qing, H. Imai, and G. Wang (Eds.): ICICS 2007, LNCS 4861, pp. 141–152, 2007.

record in the released dataset has at least (k-1) other records with respect to the set of quasi-identifier attributes[2,7,12]. However, the k-anonymity model has drawbacks itself, because lack of diversity in sensitive attributes and adversaries' background knowledge may lead to additional disclosure risks. Two improved k-anonymity model *l-diversity*[8] and *(α,* k*)-anonymity*[16] are proposed to solve this problem. However, both models only focus on dealing with microdata with single sensitive attribute (single-SA). When there comes the situation that attackers can do some inference disclosures based on multiple sensitive attributes, data providers should consider all the possible conditions that may happen before releasing the microdata with multiple sensitive attributes (multi-SA).

The existing methods concerning about multi-SA microdata publishing are discussed on[8]. The idea is described as follows: Suppose S and V are two sensitive attributes of one microdata set. Only if we treated S as part of the quasi-identifier when checking for diversity in V (and vice versa), we can ensure the diversity principle held for the entire dataset. Effectively protect the privacy of microdata with multi-SA, and at the same time with considerable utilities of the data. This is the original intention of this paper. The main contributions of this paper include:

(1) We set out by analyzing the problems of multiple sensitive publishing and the disclosure scenarios which may happen because of *logical associations* existing between multiple sensitive attributes. Then a *Q&S Diversity* requirement is proposed to prevent attacks in the foregoing disclosure scenarios. And finally, an α *Requirement* definition is given to ensure the diversity requirement.
(2) We propose an effective multiple sensitive attributes processing framework integrating different generalization algorithms on quasi-identifiers and sensitive attributes respectively. In order to evaluate the performance of our framework, the corresponding information loss metrics are subsequently defined. And we experimentally show the effectness of overall distortion reduction based on our proposed measure implemented in the framework.

The rest of the paper is organized as follows. We start by discussing related work (Section 2). Section 3 analyzes the problems existing in the multiple sensitive attributes publishing, and takes measures to prevent the disclosure risk caused by logical associations. Section 4 explains the whole generalization framework for the multiple sensitive attributes processing. Section 5 experimentally evaluates the effectiveness of our solutions, and we conclude finally in Section 6 our future work.

2 Related Work

At present, many k-anonymity models have been proposed in the literature to prevent re-identification risks caused by external information linking with quasi-identifiers [3,6]. However, these k-anonymity models have drawbacks themselves, because they do not consider problems existing in the diversity of sensitive attributes. Two improved k-anonymity model l-diversity[8] and (α,k)-anonymity[16]

are proposed to solve this problem. The parameter l should be "well-represented". We should ensure the l-diversity requirement on sensitive attribute at the same time with k-anonymity requirement on quasi-identifiers. A more practical approach is not to consider every value in the sensitive attribute as sensitive. If we only have a small number of sensitive vales, a reasonable protection is that the inference confidence from a group of k-anonymous records to a sensitive value is below a threshold. This is the basic idea of the (α,k)-anonymity model. Most of the existing k-anonymity methods focus only on dealing with single sensitive attribute. However, as inference disclosure may be deduced on multiple attributes, the multiple sensitive attributes privacy protection should be supported in k-anonymity. This is the motivation of this work.

3 Multiple Sensitive Attributes

3.1 Basic Definitions

Definition 1(Equivalence Class). *Let T be a dataset with quasi-identifier attributes $Q_1, \ldots, Q_d$. An equivalence class for T is the set of all tuples in T containing identical values $(q_1, \ldots, q_d)$ for the quasi-identifiers.*

Definition 2(K-Anonymity Property). *T is said to satisfy the k-anonymity property with respect to the quasi-identifiers if each tuple $(q_1, \ldots, q_d)$ on $Q_1, \ldots, Q_d$ occurs at least k times.*

To prevent the re-identification risk, if we make sure that each record is in an *equivalence class* containing at least k members, we can guarantee that each record relates to at least k individuals even if the released records are linked to external information. This is the basic idea of k-anonymity.

Definition 3(α Requirement). *Given an equivalence class E and a value s in the domain of sensitive attribute S from dataset T. Let (E, s) be the set of tuples in equivalence class E containing s for S and α be a user-specified threshold, where $0<\alpha<1$. T satisfies α Requirement if for each sensitive value s in S, the relative frequency of s in every equivalence class is less than or equal to α.*

This definition presents another way to ensure diversity of sensitive values. The basic idea is similar with[16]. If this α Requirement is satisfied, we would have at least $1/\alpha$ diversity ensured, because if each frequency of s does not exceed the threshold $1/\alpha$, there will be at least $1/\alpha$ different values in an *equivalence class*. Beacause frequency of s in every *equivalence class* is less than or equal to α, by setting α,individuals can control the frequency of s. This is heavily needed beacause if one sensitive value s appears too frequently, intruders may disclose s with high confidence.

3.2 Disclosure Risks on Multi-SA

If the microdata is treated only as one sensitive attribute, we just need to consider this attribute's diversity in each *equivalence class*. This is called single diversity

(SD). But for the microdata with multi-SA consideration, associations among sensitive attributes should also be considered. Because these associations can lead to additional disclosure scenarios. Most of the time, one sensitive attribute do play a part in the statistical analyzing as the identifiers of other ones. For example,*Disease* and *Household Disease* are two sensitive attributes of Medical microdata. As shown in Table1, if users want to stat. *"the probability of Disease coming from Household Disease"*, data providers have to publish multi-SA at the same time. In this situation, in addition to satisfy the requirement of single sensitive attribute disclosure controlled, associations existing between multiple sensitive attributes should also be considered. Otherwise, some disclosure scenarios may happen because of these associations. There are two types of association. One can be regarded as the semantic association, the other one the logical association. The semantic association is just another say of dependency. Dependencies in the microdata can be of a logical nature or of a statistical nature[13]. Ignoring such dependencies may lead to underestimation of the disclosure risk. However, current disclosure risk measures we adopt do not take into account them which might exist between variables or records, because they might complicate the analysis considerably. A proper treatment of such data may require tailor-made models[13], which can be time-consuming and complicated. Therefore, they are not taken into account when assessing the disclosure risk or when modifying the data in an attempt to increase their safety. By now, we just come across one work which considers hiding strong associations between some attributes and sensitive classes and combines k-anonymity with association hiding. This model is called the template-based model [13]. This model is good for users who know exactly what inferences are damaging, but is not suitable for users who do not know. So it can not gain most of the popularity. Another type of association is the logical association which will be mostly considered in the following.

Definition 4(Logical Association). *Suppose* $S_1, \ldots, S_m$ *are* m *sensitive attributes of dataset* T*,* t *denoted a tuple of* T*. For the publishing of microdata with multi-SA, in each tuple* t*, the values* $t.S_1, \ldots, t.S_m$ *should not be disordered in order to keep statistics property. Each time a disclosure or elimination of* $t.S_i$ *(*$1 \leq i \leq m$*) means the disclosure or elimination of the other* $t.S_j$ *(*$1 \leq j \leq m, i \neq j$*). This is caused by the logical association among multi-SA.*

As the above definition shows, the logical association happens because of the position corresponding, i.e. all the sensitive values of each record are in the same line and the sensitive values of each column should not be disordered in order to keep the original statistical characteristic.

If the logical association exists among multi-SA and disclosure happens on one attribute, the other ones will be disclosed correspondingly. Two kinds of disclosure risks will happen on each single attribute, *positive disclosure* and *negative disclosure*. After ensuring the single l-diversity of each sensitive attribute, intruders need to eliminate at least l-1 possible values of S in order to infer a positive disclosure. If the positive disclosure of one attribute successes, the corresponding values of other sensitive attributes will be disclosed at the same

time. Another kind of disclosure is the negative disclosure. If intruders can eliminate one sensitive value with high confidence, the logical corresponding values of the other sensitive attributes can also be eliminated. If the left values in each sensitive attribute lack diversity, disclosure risk happens. This disclosure risk happens on one sensitive attribute because of negative disclosure on another attribute. For example, as Table 1 shows, *Disease* and *Household Disease* are two sensitive attributes, and consider the set of quasi-identifier attributes in the first *equivalence class*. This *equivalence class* is 3-diverse for attribute *Disease*, and also the same with attribute *Household Disease*. However, if intruders have the background knowledge that the value for *Disease* is not *Albinism*, they can make sure that the value for attribute *Household Disease* cannot be *Albinism* or *No*, and therefore must be *Asthma*. Thus we see that an equivalence class with diversity in each sensitive attribute separately in a multi-SA microdata may still violate the principle of diversity.

For multiple sensitive attributes, besides the requirement of single l-diversity on each attribute, additional measures should be taken in order to prevent the attacks happened by the lack of diversity between sensitive attributes.

Table 1. Medical database with sensitive attribute Disease &Household Disease

Age	Sex	Zipcode	Disease	Household Disease
[10,30]	M	[15001,20000]	Albinism	Albinism
[10,30]	M	[15001,20000]	Albinism	No
[10,30]	M	[15001,20000]	Albinism	No
[10,30]	M	[15001,20000]	Asthma	Asthma
[10,30]	M	[15001,20000]	Pneumonia	Asthma
[10,30]	M	[15001,20000]	Pneumonia	Asthma
[30,60]	F	[30000,60000]	Haemophilia	Hepatitis
[30,60]	F	[30000,60000]	Cold	No
[30,60]	F	[30000,60000]	Liver cancer	Pneumonia
[30,60]	F	[30000,60000]	Liver cancer	Hepatitis
[30,60]	F	[30000,60000]	Liver cancer	Hepatitis
[30,60]	F	[30000,60000]	Cold	No

Definition 5 (Q&S Diversity). *Let T be a dataset with non-sensitive attributes $Q_1, \ldots, Q_d$ and sensitive attributes $S_1, \ldots, S_m$. S_i is treated as the sole sensitive attribute and $Q_1, \ldots, Q_d$, $S_1, \ldots, S_{i-1}$, $S_{i+1}, \ldots, S_m$ is treated as the quasi-identifier. If the value of S_i is diverse according to $Q_1, \ldots, Q_d$, $S_1, \ldots, S_{i-1}$, $S_{i+1}, \ldots, S_m$, We say S_i is Q&S Diverse.*

Definition 6 (Multi-Diversity (MD)). *Let T be a dataset with non-sensitive attributes $Q_1, \ldots, Q_d$ and sensitive attributes $S_1, \ldots, S_m$. If for each sensitive attribute S_i (i = 1,... ,m), the Q&S Diversity is satisfied, we say T is satisfied Multi-Diversity.*

Only the released dataset which satisfies *MD* requirement can be regarded as the proper result of multiple sensitive attribute processing. See the example in

Table 1 again, the *MD* is satisfied only if the *Q&S Diversity* are satisfied in both following situations: regarding *Age,Sex,Zipcode,Disease* as quasi-identifiers with *Household Disease* sensitive and regarding *Age,Sex,Zipcode,Household Disease* as quasi-identifiers with *Disease* sensitive. However, this requirement is a little too strict which may lead to over-distortion of quasi-identifiers and correspondingly cause too much information loss. The masking of sensitive attribute values with more general patterns may help to alleviate this problem.

Definition 7 (Masking SA). *Suppose s_{child} is one sensitive value in relation T, s_{parent} is the parent node of s_{child} in the more general domain with N leaf nodes in the sensitive attribute tree. If we replace s_{child} with s_{parent}, the possibility of disclosing s_{child} is reduced to 1/N. This measure which can reduce disclosure risks is called Masking SA.*

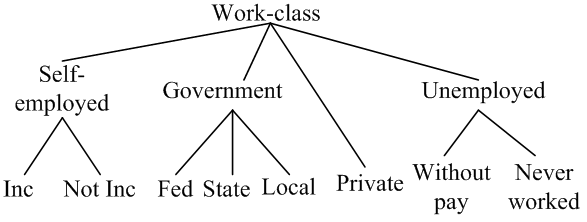

Fig. 1. Taxonomy tree for the attribute Education

For example, as in Fig. 1 shows, *Local Government* is one leaf node of *Government*. If we mask *Local Government* with *Government*, the possibility of disclosing *Local Government* is reduced to 1/3. The implementing of *Masking SA* and an effective generalization algorithm framework will be proposed in the following.

4 The Generalization Framework

4.1 Two-Step Generalization Algorithm

As the number of sensitive attributes grows, it is not hard to see that we will inevitably need larger and larger *equivalence class* to ensure the diversity of sensitive attributes[10,17]. To avoid over-distortion of quasi-identifiers, we implement *Masking SA* on sensitive attributes in company with the quasi-identifiers generalization. Although making sensitive values more general may result in less accurate values on sensitive attributes, it retains more information on the quasi-identifiers. Generally, the diversity requirement of sensitive attribute according to each quasi-identifier *equivalence class* and each sensitive *equivalence class* should not be the same. This can be proved by the experiments in the next section. We define α-*QI* and α-*SA* with respect to QI and SA to show this different diversity requirement.

Our generalization scheme is composed by a *two-step generalization* [18] as shown in Fig.2. The first phase applies quasi-identifiers generalization on microdata, and we choose the *top-down specialization greedy algorithm* for it [5]. Then, the second step produces the final microdata by performing *Masking SA* on the foregoing result quasi-identifiers *equivalence class*, employing a *bottom-up local recoding algorithm* for each equivalence class [13]. The execution proceeds in rounds. In each iteration,

- The top-down specialization greedy algorithm slightly refines one of $f_1, \ldots, f_d$ and lead to a new T^* with lower information loss. We choose the "best" attribute for the refining function. The "best" attribute means that the refine of that can involve the largest number's tuples. The core of the greedy algorithm is to make the largest extent specialization in each round, so as to find the anonymity result with the least information loss quickly.
- For each *equivalence class*, the bottom-up local recoding algorithm finds the corresponding value of sensitive *equivalence class*, in which the *α-SA requirement* is not satisfied. Then, we carry on the local recoding which also adopt the greedy algorithm, i.e. finding the value with the least number's tuples from the same generalization hierarchy with the former value. Afterwards, we impose the generalization function on the tuples which contains these two values. The generalization is only done in this special *equivalence class*, therefore, this is a local recoding algorithm.

The current round finishes after executing the two-step generalization algorithm. As our generalization framework is devoting itself into finding the optimum result with the least information loss, we should measure the amount of information gain by implying the top-down specialization on quasi-identifiers and the information loss by implying the bottom-up local recoding on sensitive attributes. If we get more information gain than the information loss, we will carry on the next iteration. Or else, the current result is regarded as the optimum one and we finish the iteration.

4.2 Information Loss Metrics

In order to evaluate the effectiveness of our two-generalization algorithm, the corresponding information loss metrics are subsequently defined. Based on the general loss metric (LM)[6], information in all the potentially identifying attributes will be assumed to be described equally important in LM. So the total information loss due to generalizations will be computed by summing up a normalization information loss for each of these attributes.

Definition 8 (Distortion Ratio). *Given a microdata set T, after the processing of generalization, a T^* is obtained. We compute all the tuple's information loss of T^*, compared with the overall tuple's absolute information loss by making all the attribute values to the most generalized domain. The result is called Distortion Ratio.*

The Greedy Two-step Generalization Algorithm

Input: microdata T, generalization hierarchies of all attributes, value of α-QI and α-SA
Output: publishable relation T^*
Body:
for each QI attribute $Q_i(1 \leq i \leq d)$
initialize a generalization function f_i with a single partition covering the entire domain of Q_i
T^*= the relation after applying QI-generalization on T according to $F=f_1, \ldots, f_d$
While(true)
T^*_{best}= T^*
for each QI-group
check whether the α-QI and α-SA are satisfied;
if(true)
finding the best $F'=f'_1, \ldots, f'_d$ obtain from F with "single partition"(specialization)
T^*'=the relation after applying F' on the quasi-identifiers *equivalence class* of T^*
else if(α-QI is not satisfied)
withdraw the last "single partition" on quasi-identifiers
else if(α-SA is not satisfied)
s= the value of corresponding sensitive *equivalence class*
s-company= the value with the least number tuples in the same generalization hierarchy with s
do the *Masking SA* on s and s-Company of this quasi-identifiers *equivalence class*
end if
if (Distortion Ratio(T^*')< Distortion Ratio (T^*_{best}))
T^*_{best}= T^*
else
return T^*_{best}
End While

Fig. 2. Algorithm for the Greedy Two-step Generalization

Let v be a value in the domain of attribute A. We use *InfoLoss(v^*)* to capture the amount of information loss in generalizing v to v^*. The number of values in v^* is expressed by *value.number(v^*)* and the number of values in the domain of A by *value.number(domain A)*. Formally,

$$InfoLoss(v^*) = \frac{value.number(v^*) - 1}{value.number(domainA)} \quad (1)$$

For instance, in Fig. 1, the taxonomy of *Work-class* has 8 leaves, generalizing *Local Government* to *Government* results in *InfoLoss(Government)*=(3-1)/8=1/4, where 3 is the number of leaves under *Government*. Obviously, if v is not generalized, *InfoLoss(v^*)* equals 0, i.e., no information is lost. The

overall information loss $InfoLoss_{tuple}(t^*_{qi})$ and $InfoLoss_{tuple}(t^*_{sa})$ of a generalized tuple t^* respectively equals the follows,

$$InfoLoss_{tuple}(t^*_{qi}) = \sum_{i=1}^{d} InfoLoss(t^*.A^{qi}_i) \tag{2}$$

$$InfoLoss_{tuple}(t^*_{sa}) = \sum_{i=1}^{m} InfoLoss(t^*.A^{sa}_i) \tag{3}$$

The total information loss of the entire relation T^* is computed following the definition 3, given out respectively for quasi-identifiers and sensitive attributes as the *QI Distortion Ratio* and *SA Distortion Ratio*,

$$QI.DistortionRatio(T^*) = \frac{\sum_{\forall t^* in T^*} InfoLoss_{tuple}(t^*_{qi})}{\sum_{\forall t^* in T^*} \sum_{i=1}^{d} 1} \tag{4}$$

$$SA.DistortionRatio(T^*) = \frac{\sum_{\forall t^* in T^*} InfoLoss_{tuple}(t^*_{sa})}{\sum_{\forall t^* in T^*} \sum_{i=1}^{m} 1} \tag{5}$$

The overall Distortion Ratio is the sum of *QI Distortion Ratio* and *SA Distortion Ratio*.

5 Experiments

This section experimentally evaluates the effectiveness of our approach using the Adult Database from the UCI Machine Learning Repository[4]. We select about 48000 tuples from the Adult Database. The microdata has 8 attributes: *Salary, Marital-status, Family-status, Race, Gender, Education, Occupation and Employment*. All the columns are categorical. In our experiments, we used *Occupation* and *Employment* as two sensitive attributes, *Education*, *Occupation* and *Employment* for three-sensitive attributes, *Family-status*, *Education*, *Occupation* and *Employment* for four-sensitive attributes. The left columns are quasi-identifiers. All experiments were run on a Celeron(R) PC with CPU 2.40GHz and RAM 512MB.

Intuitively, the dataset *Distortion Ratio* with *MD* is higher than that with *SD*, because the *MD* requirement is much stricter than the *SD* requirement. In fact, as illustrated in Fig. 3, the contrast is not so far. We regard the sensitive attribute *Education* as quasi-identifier, obtaining the curse "1-attribute", and ensure the *MD* on both *Education* and *Occupation*, obtaining the curse "2-attribute", showing in Fig. 3(a). This result perfectly proves the good performance of *Masking SA*. The conclusion is same with Fig. 3(b). It is this *Masking SA* which prevents quasi-identifiers from over-distortion.

Fig. 4 shows the *QI Distortion Ratio* and *SA Distortion Ratio* and Execution Time. As we see, in Fig. 4(a), the *QI Distortion Ratio* decreases when α increases, opposite for the curse of *SA Distortion Ratio*. We can also see that the extent of *SA Distortion Ratio* is not large, however, decreasing the *QI Distortion Ratio* all the

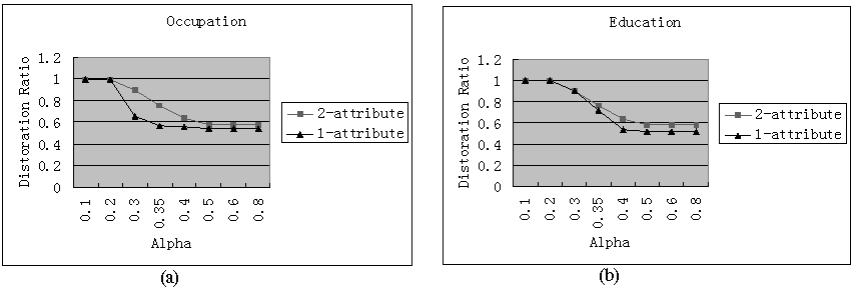

Fig. 3. Distortion Ratio Comparison between SD and MD

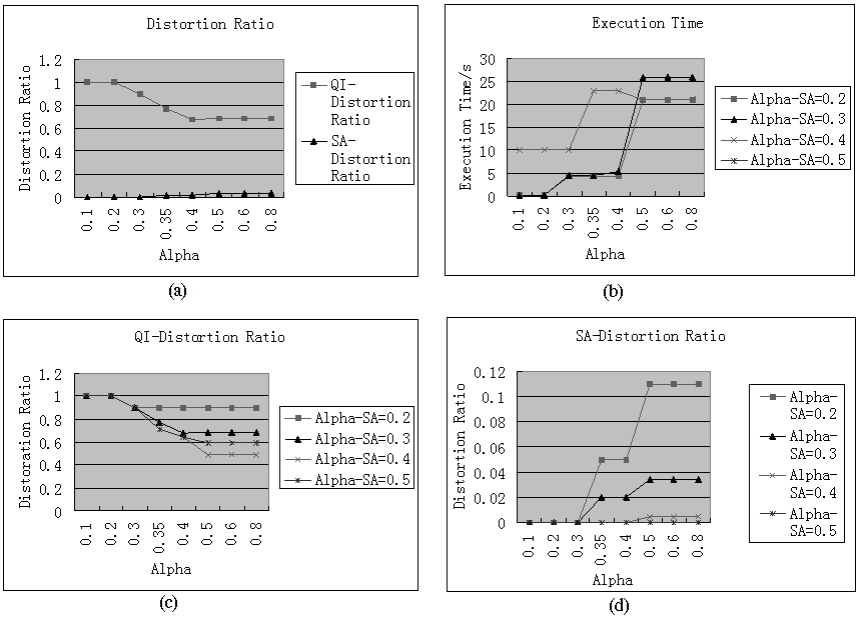

Fig. 4. QI and SA Distortion Ratio and Execution Time Versus α-QI andα-SA

same. The higher diversity requirement, the more information loss. Therefore, we should choose α between 0.3 and 0.5. The execution time in Fig. 4(b) displays the computation cost of different α-SA parameter. Fig. 4(c) and Fig. 4(d) respectively shows the *QI Distortion Ratio* and *SA Distortion Ratio* according to different diversity requirement of (α-SA). As we see, the requirement should not be too strict. Otherwise, we would get the terrible result with both *QI Distortion Ratio* and *SA Distortion Ratio* so high, e.g., when α-SA =0.2.

Fig. 5 shows the performance of different number's sensitive attributes. As the number increases, both the *QI Distortion Ratio* and *SA Distortion Ratio* become higher. This result illustrates that the more sensitive attributes, the

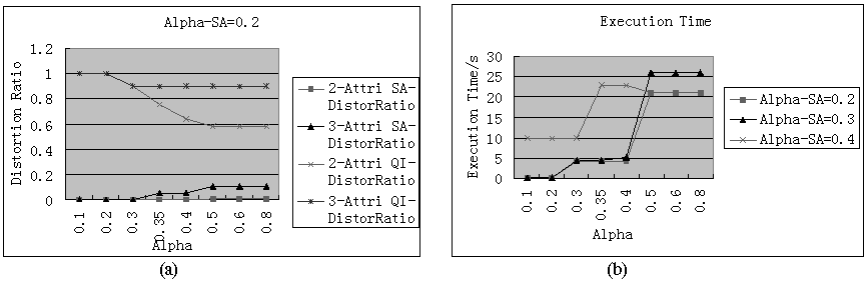

Fig. 5. Distortion Ratio and Execution Time of Different Number Sensitive Attributes

heavier information loss we get and the more Execution time it costs. We should reduce the diversity requirement of *MD* as the number of sensitive attributes increases, otherwise, the released dataset will be useless. Moreover, in the view of statistics analysis, if the number of sensitive attributes is large enough, we should not impose *MD* on them again, because attackers impossibly has so much background knowledge as to reduce identifying possibility by detecting association between sensitive attributes. In conclusion, experiments have proved the feasibility and advantage of our processing method for microdata anonymity with multiple sensitive attributes and illustrated the effectiveness of *Masking SA*. Through these experiments, we also obtain the principle of how to set proper parameters so as to make good performance. Moreover, we have discovered one important principle that whether imposing *MD* requirement on a dataset should be according to the number of sensitive attributes of the releasing microdata.

6 Conclusion

Most of the existing k-anonymity methods focus only on dealing with single-SA. For microdata with multi-SA publishing, the disclosure scenarios may happen because of *logical associations* existing between multi-SA. A *Q&S Diversity* requirement is proposed to prevent inference attacks in the foregoing disclosure scenarios. The *MD definition* is proposed to ensure the diversity among multi-SA. We make sure the diversity by ensuring the frequency of each sensitive value below the threshold α in each *equivalence class*. Additionally, we reduce the overall distortion by the measure of *Masking SA*. We propose a multiple sensitive attributes processing framework implementing top-down specialization on quasi-identifiers and local recoding bottom-up generalization on sensitive attributes. The experiment proves the feasibility and advantage of our method, and we also get additional knowledge from the experiment results. For future work, in order to retain more information of a released dataset, we will consider the local suppression as the supplement of our generalization techniques to integrate with this framework and validate our solution by more real experiments.

Acknowledgements

This work is supported by National Natural Science Foundation(60673140) and National High-Tech R&D Program(2007AA01Z156).

References

1. Bettini, C., Wang, X.S., Jajodia, S.: The role of quasi-identifiers in k-anonymity revisited. Technical Report N. RT-11-06 DICo - University of Milan, Italy (2006)
2. Bayardo, R., Agrawal, R.: Data privacy through optimal k-anonymity. In: Proc of the ICDE (2005)
3. Ciriani, V., De Capitani di Vimercati, S., Foresti, S., Samarati, P.: Samarati. k-Anonymity. In: Secure Data Management in Decentralized Systems (2007)
4. `ftp://ftp.ics.uci.edu/pub/machine-learning-databases/adult`
5. Fung, B., Wang, K., Yu, P.: Top-down specialization for information and privacy preservation. In: Proc of ICDE (April 2005)
6. Iyengar, V.: Transforming data to satisfy privacy constraints. In: Proc of SIGKDD (2002)
7. LeFevre, K., DeWitt, D.J., Ramakrishnan, R.: Mondrian multidimensional k-anonymit. In: Proc of ICDE (2006)
8. Machanavajjhala, A., Gehrke, J., Kifer, D., Venkitasubramaniam, M.: l-diversity: Privacy beyond k-anonymity. In: Proc of ICDE (2006)
9. Sweeney, L.: K-anonymity: A model for protecting privacy. International Journal on Uncertainty, Fuzziness, and Knowledge-based Systems (2002)
10. Sweeney, L.: Achieving k-anonymity privacy protection using generalization and suppression. Int'l Journal on Uncertainty, Fuzziness, and Knowledge-based Systems (2002)
11. Samarati, P.: Protecting respondents' identities in microdata release. IEEE Trans. on Knowledge and Data Engineering (2001)
12. Samarati, P., Sweeney, L.: Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Technical report, CMU, SRI (1998)
13. Wang, K., Yu, P., Chakraborty, S.: Bottom-up generalization: A data mining solution to privacy protection. In: Perner, P. (ed.) ICDM 2004. LNCS (LNAI), vol. 3275, Springer, Heidelberg (2004)
14. Willenborg, L., deWaal, T.: Elements of Statistical Disclosure Control. Lecture Notes in Statistics. Springer, Heidelberg (2000)
15. Wang, K., Fung, B.C.M, Yu, P.S.: Template-based privacy preservation in classifica-tion problems. In: Proc of ICDM 2005 (2005)
16. Wong, R.C.-W., Li, J., Fu, A.W.-C., Wang, K.: α,k-Anonymity: An Enhanced k-Anonymity Model for Privacy-Preserving Data Publishing. In: Proc of KDD 2006 (2006)
17. Xu, J., Wang, W., Pei, J., Wang, X., Shi, B., Fu, A.W.-C.: Utility-Based Anonymization Using Local Recoding. In: Proc of KDD 2006 (2006)
18. Xiao, X., Tao, Y.: Personalized Privacy Preservation. In: Proc of the SIGMOD (2006)