CHEMICAL ENGINEERING TRANSACTIONS

VOL. 46, 2015

A publication of
AIDIC
The Italian Association
of Chemical Engineering
Online at www.aidic.it/cet

Guest Editors: Peiyu Ren, Yancang Li, Huiping Song
Copyright © 2015, AIDIC Servizi S.r.l.,
ISBN 978-88-95608-37-2; **ISSN** 2283-9216

DOI: 10.3303/CET1546031

The *K*-Anonymization Method Satisfying Personalized Privacy Preservation

Jinling Song*, Liming Huang, Gang Wang, Yan Kang, Haibin Liu

Hebei Normal University of Science & Technology, Qinhuangdao 066004, China.
songjinling99@126.com

Even if *k*-anonymity model can prevent publishing data from disclosing privacy effectively and efficiently, due to the uneven distribution of the sensitive data, ordinary *k*-anonymization method cannot guarantee each tuple satisfying the personalized privacy requirement of it's data owner although the publishing table has been satisfied *k*-anonymity constraint. The reason which *k*-anonymity table fails to satisfy personalized privacy requirement is analyzed firstly, then Correlate degree of Sensitive Values, Leakage Collection, privacy disclosure metric and data quality metric are presented. At last an anonymization method satisfying personalized privacy requirements is presented, in which a utility-driven adaptive clustering method is proposed to partition tuples with similar best data quality.

1. Introduction

The general process of privacy preserving data publishing is: firstly, the data owners offer the data publisher their individual data containing sensitive information, and then, the data publisher treats the collected data uniformly to meet certain privacy model, finally, the data publisher publishes the data satisfying privacy requirement to the data receiver for statistical analysis. In the data release process, although the data owner has authorized the data publisher to deal with his individual data for protecting privacy, the uniform treatment of data may not meet the privacy requirement of each individual. So, this leads to the problem of personalized privacy preservation, that is, each data owner will specify the privacy preservation level for his data independently and the publishing data set should satisfy the privacy requirement of each individual.

K-anonymity is a typical privacy model, which reduces privacy leakage based on generalizing the values on quasi-identifier attributes (called *k*-anonymizaiton). Due to the sensitive value do not distribute symmetrically, ordinary *k*-anonymization method may lead to the same sensitive value appears frequently in a *QI*-group. Even the generalizing data set meet the *k*-anonymity constraint, some tuples may violate the privacy requirements of their owner, which implicates that the disclosing risk is still exist. Table 1 is the diagnosis records published by a hospital, in which the sensitive attribute is "disease". To resist the linking attack, we can *k*-anonymize Table 1 before publishing. Assume the privacy request of each data owner is 50%, Table 1 should be 2-anonymous from the principle of the *k*-anonymity model. Table 3 is 2-anonymous table of Table 1 on {Age, Postcode}. But a closer look reveals that the risk of privacy leakage still exists in it: attacker can deduce that Zhao is related to the fourth tuple or fifth tuple in Table 3 according to his Age (12) and Postcode ("22000"), so he can infer that Zhao suffers from "pneumonia" because the disease is same in the two tuples.

What we can deduce from the example is that the uniform *k*-anonymization of publishing data set can't preserve the privacy of each individual even the privacy requirement of each one is consistent. The reason is the personal privacy requirements are neglected in the *k*-anonymization process and result in the slope of sensitive value. So, the *k*-anonymization method need to consider the privacy requirement of each individual adequately, we present a *k*-anonymization method which aims at personalized privacy preservation.

Please cite this article as: Song J.L., Huang L.M., Wang G., Kang Y., Liu H.B., 2015, The k-anonymization method satisfying personalized privacy preservation, Chemical Engineering Transactions, 46, 181-186 DOI:10.3303/CET1546031

Table 1: Medicine diagnosis records

Age	Zipcode	Diease
5	12000	gastric ulcer
9	14000	dyspepsia
8	19000	bronchitis
12	22000	pneumonia
19	24000	pneumonia

Table 2: Voter information list

Name	Age	Zipcode
Li	5	12000
Wang	9	14000
Wang	6	18000
Liu	8	19000
Chen	7	17000
Zhao	12	22000

Table 3: 2-anonymous table

Age	Zipcode	Diease
[5-10]	[10001-20000]	gastric ulcer
[5-10]	[10001-20000]	dyspepsia
[5-10]	[10001-20000]	bronchitis
[11-20]	[20001-25000]	pneumonia
[11-20]	[20001-25000]	pneumonia

2. Related work

K-anonymity privacy protection model (L. Sweeney (2002)) got the wide attention of experts and scholars when it was presented by L. Sweeney. Previous studies mostly focus on *k*-anonymization algorithm under different scenarios. *Datafly* algorithm was adopted by L. Sweeney (1997), which have promoted the generation of *k*-anonymity model. To improve the data precision of the generated table, *Mingen* algorithm was adopted in (L. Sweeney (2002)). The global *Incognito* algorithm was proposed by K. Lefvre et al (2005), which generalize all the domain values of attributes. *Multi-dimensional* algorithm was proposed by Kristen LeFevre et al (2006), which generalize multi-attributes at the same time. Fung B C M et al presented the *TDS* (top-down specialization) algorithm which achieves the *k*-anonymity by gradual specialization from the most generalization state (attribute values are represented by the root nodes in classification tree). To preserve the clustering information of anonymous data, Fung B C M et al (2009) extended *TDS* algorithm. Bo WANG and Jing YANG (2012) proposed a local coding anonymous algorithm was proposed based on the attribute hierarchy. Although the above algorithms are all excellent in their considering scenarios, no one consider the personalized privacy requirement. Zakerzadeh H and Osborn SL (2011), Cao JM et al. (2011) presented the anonymization method for numerical streaming data. HYUNJI L and JAE-WOO C (2013) present Density-based *k*-anonymization scheme in location-based services. Yuan MX et al (2011), M.E.S karkala et al. (2011) utilized the *k*-anonymity model in social network. Jinling SONG et al (2014) presented a multi-objective optimization method for selecting the value of *k* in the *k*-anonymity Model.

3. Basic definitions

Definition 1 (Quasi-identifier, *QI*) If a dataset *T* (A^{QI}, A^{S}) can be connected with other dataset by attributes A^{QI}={A_i,…,A_j} and re-identify the privacy of some individuals. Then the attribute set A^{QI} is called Quasi-identifier.

Definition 2 (Generalization Tree, Gtree) Let *D* is a finite domain of the attribute A_i, then the Generalization Tree of A_i is a tree whose leaves are attribute values in *D* and each non-leaf node is a value summarizing it's children. Gtrees of the numeric attribute *Age* and Categorical attribute *Disease* are shown in Figure 1.

Definition 3 (Generalization) Let the Gtree of attribute A_i is GT_{Ai}, the generalization of A_i is the process of mapping a value *v* to its ancestor in in GT_{Ai}. The value has been generalized is called generalization value.

Definition 4 (*k*-anonymization) For the dataset *T* (A^{QI}, A^S), if we generalize the values on A^{QI} and get a dataset *T** in which each tuple has at least *k*-1(*K*≥2) other tuples same with it on A^{QI}, then *T** is the *k*-anonymized dataset of *T*, the generalization process from *T* to *T** is called *k*-anonymization.

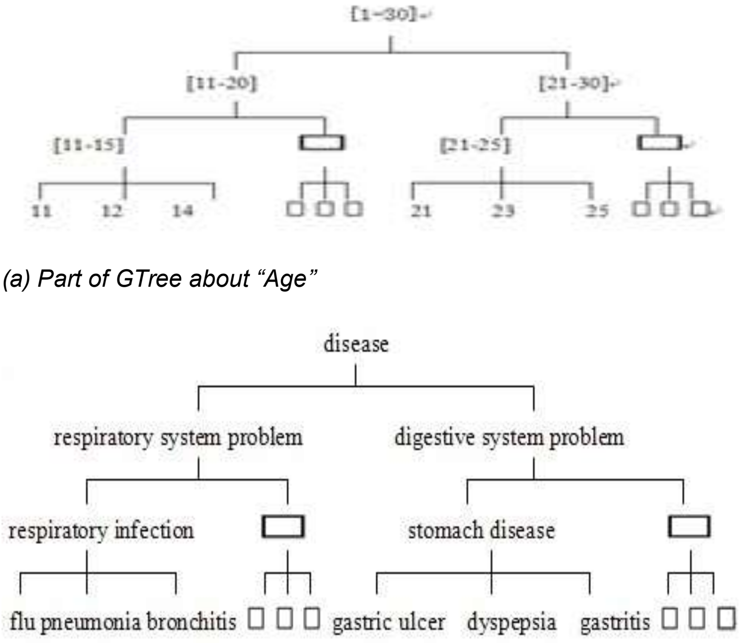
Definition 5 (*QI* Group) For a *k*-anonymous dataset *T**(A^{QI}, A^S), the generalized tuples with the same value in $T^*[A^{QI}]$ are called a *QI* group, i.e. *QG*. The *QI* groups in *T** are denoted as *QG* (*T**) = {QG_1, QG_2, QG_m}, where $|QG_i|$≥*k*,and $|QG_1|+|QG_2|+\ldots+|QG_m|= |T^*|$.

4. The metric for privacy leakage

Definition 6 (Correlate degree of sensitive values) Let t^* and t^*_i are generalizing tuples in the *k*-anonymous table, *t* is the original tuple of t^*, the correlate degree of the sensitive value of tuple t^* and t^*_i is denoted as $D_{Correlate}$ (t^*, t^*_i). $D_{Correlate}$ (t^*, t^*_i)=1/*Nleaf*(t^*_i) when $t[A^S]$ is a leaf of the tree rooted as $t^*_i [A^S]$, where *Nleaf*(t^*_i) is the count of the leaves in the tree $t^*_i [A^S]$. $D_{Correlate}$ (t^*, t^*_i)=1 if $t[A^S]$ is same as $t^*_i [A^S]$, else $D_{Correlate}$ (t^*, t^*_i)=0.

Definition 7(Leakage Collection) Let t^* is a tuple in the anonymous table, all the tuples relevant with the privacy leakage of t^* in it's *QI* group is called the Leakage collection of t^*, denoted as *LC*(t^*). *LC*(t^*) expresses the tuple set composed by t^*_i making $D_{Correlate}$ (t^*, t^*_i)>0 in the *QI* group containing t^*.

Definition 8 (Privacy leakage probability of t^*) Let the *QI* group including t^* is QG_t, then the privacy leakage probability of t^* (denoted as *Pleak* (t^*)) only associated with other tuples in the QG_t. Let the cardinality of *QI* is *b*, then the probability that an attacker matches t^* to any t^*_i is 1/*b*, so the privacy leakage probability that t^* correlates to t^*_i is $D_{Correlate}$ (t^*,t^*_i)/*b*. Then $Pleak(t^*) = \sum_{i=1}^{b} D_{Correlate}(t^*, t^*_i)/b$, where t^*_i represents the *i*th tuple in QG_t.



(a) Part of GTree about "Age"

(b) Part of GTree about "Disease"

Figure 1: Gtree about attribute Age and Disease

5. The Metric of Information Left Degree

Definition 9 (Information Left Degree of value) After the attribute value *v* is generalized to v^*, the size of information left in v^* is called information left degree of v^*, denoted by *ILeft_value*(v^*), $ILeft\_value(v^*) = 1 - \frac{\text{The number of leaf nodes in SubT}(v^*)\text{-1}}{\text{The number of all leaf nodes in GTree}}$, *SubT*($v^*$) is the child tree rooted by v^* in *GTree*.

Definition 10 (Information Left Degree of tuple) Let the generalized tuple of *t* is t^*, and *d* is the sum of dimensions of A^{QI} and A^S in the tuple *t*, then the information left degree of tuple t^* (denoted as *ILeft_tuple*(t^*)) can be expressed as *ILeft_tuple*(t^*)= (*ILeft_value*($t^*[A^S]$)+ *ILeft_value*($t^*[A^{QI}]$))/*d*.

Definition 11 (Information Left Degree of Anonymous Table) Let T^* is the anonymous table of original table T, the information left degree of T^* can be expressed as $ILeft\_table(T^*)=\sum_{i=1}^{|T^*|} ILeft\_tuple(t_i^*)/|T^*|$.

6. The Anonymization Method Satisfying Personalized Privacy Preservation

Considering the limitation and insufficiency in traditional *k-anony*mization methods, our method will not only satisfy the privacy requirements of each data owner but also persist as more as information. The algorithm contains the following steps: firstly, we *k-anony*mize the original table by a utility-driven adaptive clustering method which can partition tuples with best data quality similarly, Secondly, check each *QI* group and call *SA_Generalization* algorithm to ensure each tuple satisfy personalized privacy requirement.

To partition tuples in a table with less information loss (that is with best data quality), the intuition is that the tuples in a partitioned group must be more similar and the tuples in two groups must be have more difference. If we sort the tuples on *QI* attributes, the adjacent tuples will be more similar than other tuples which are far away. So, we can partition the front k tuples to a group and the group will have least information loss naturally. To the (k+1)th tuple, it can belong to this group or the next group as the first tuple, we decide it by judging the information left degree change which can make the best data quality of (k+1)th tuple greeedly, so it can make the anonymity table with best data quality similarly.

Theorem 1. Let the k-anonymous table of the table T is T^*, and the *QI* group of T^* that tuple t^* locates in is QG_t, and b is the cardinality of QG_t (that is $b=|QG_t|$), then the minimum privacy leakage probability of tuple t^* is $1/b$.

By Theorem 1, if the privacy requirement of data owner is less than $1/b$, then the tuple t^* will not satisfy the privacy requirement whatever to do for it, that is the anonymous table fails to satisfy privacy requirement.

Theorem 2. For any tuple t^* in a *QI* group, if there is a tuple t^*_i ($t^*_i \neq t^*$) in the *QI* group satisfy $D_{\text{Correlate}}(t^*,t^*_i)>0$, then generalizing the sensitive attribute value of the t^*_i will decrease $Pleak(t^*)$.

***Anonymity_Generalization* (T, A^{QI}, *pleak*$_j$ ($1 \le j \le |T|$))**

Input: publishing table T, quasi-identifier attributes A^{QI}, the privacy requirements (maximum privacy leakage probability) of each data owner *pleak*$_j$ ($1 \le j \le |T|$);

Output: the anonymous table T^* satisfying personalized privacy requirements

```
$T^*=\varnothing$; /*$T^*$ stores the anonymous table */
flag=false; /* identify whether the anonymous table satisfy the personalized privacy requirements */
$T'$ =Sort table $T$ on $A^{QI}$;
$QG_i$ =the front $k$ tuples of $T'$to an anonymized group;
while ($l<|T'|$) /* $|T'|$ is the number of the tuples in publishing table $T$*/
    {if ($ILeft$_table($QG_i \cup t_l$)-$ILeft$_table ($QG_i$)$\le$ $ILeft$_tuple ($t_l \in${ the $k$ tuples of $T'$ from $t_l$ }))
        $QG_i = QG_i \cup t_l$;
    else
        {$T' = T' - QG_i$;
        $QG_i$= generalizing $QG_i$ on $A^{QI}$;
        $T^* = T^* \cup QG^*_i$;
        $QG_i$ =the front $k$ tuples of $T'$;
    }}
for each $QG_i \in T^*$
    {for each ($t^*_j$ in $QG_i$)
        if($pleak_j<1/|QG_i|$) then
            {flag=true;
            return; }
    if (not flag)
        $QG_i$ =$SA\_Generalization$($QG_i$, $pleak_l$($1 \le j \le |QG_i|$)); }
return $T^*$;
```

SA_Generalization algorithm is called to dispose the group QG_i to satisfy the privacy requirements given by data owners. It examines each tuple in the QG_i, if the privacy leakage probability of a tuple t^* is greater than the privacy requirement, then the sensitive value that has the largest information left degree in the leakage collection LC (t^*) will be generalized to make it achieve the requirement.

***SA_Generalization* (QG_i, *pleak*$_j$ ($1 \le j \le$b))**

Input: QG_i (containing tuple $t^*_1, t^*_2, \ldots, t^*_b$), the privacy requirements *pleak*$_j$ ($1 \le j \le$b) of each data owner in QG_i

Output: QG_i satisfying the privacy requirements of each data owner

```
For each tuple $t^*_j \in QG_i$ ($1 \le j \le b$)
    if ($Pleak$ ($t^*_j$)>$pleak_j$) then
        {Compute the leakage collection $LC$ ($t^*_j$) of tuple $t^*_j$;
        do{
            search the tuple $r$ in $LC(t^*_j)$, which has the largest information left degree;
if ($r[A^S]$= the root of the $GT_{A_i}$ ($A^S$))
$LC(t^*_j)$= $LC(t^*_j)$-{$r$};
else
$r[A^S]$= the parent of $r[A^S]$;
} while($Pleak(t^*_j)$> $pleak_j$);
}
```

7. Experiment

Our experimental environment is: Intel Pentium IV CPU, Memory 2G, Microsoft Visual C++ 6.0 and SQL Server. The data we selected is a company employees database, where attributes are {country, age, sex, zipcode, profession}, {country, age, sex, zipcode} is quasi-identifier and {profession} is sensitive attribute. In the experiment we sets the privacy requirement of every data owners is *pleak*=0.25 (that is equivalent to *k*=4). We compare the variation of privacy preservation of *Anonymity_Generalization* and other traditional *k*-anonymization algorithm by changing tuples from 0 to 50000, Figure 2 shows the influence on the privacy leakage probability (here refers to the average disclosure probability).

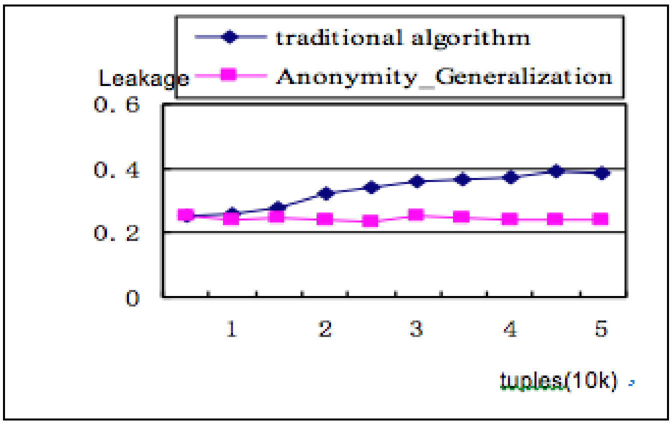


Figure 2: The experimental result

As can be seen from Figure 2, when the tuples in the publishing table is less, both the *Anonymity_Generalization* algorithm and traditional *k*-anonymity algorithm has a lower privacy leakage probability, which satisfies the data owner's requirements. But when the number of the tuples is larger enough, the privacy leakage probability of the anonymous table generated by traditional algorithm is obviously increased, while *Anonymity_Generalization* algorithm is almost invariable and less than the privacy requirement of data owners. It means that traditional *k*-anonymity algorithm cannot guarantee the privacy of the publishing data under the uneven distribution of the sensitive value, but *Anonymity_Generalization* algorithm can satisfy the privacy requirement of all data owners in deed.

8. Conclusions

In this paper, we analyses the disclosure caused by k-anonymization even under uniform privacy requirements because the slope of sensitive values. We present a k-anonymization method which aims at personalized privacy preservation. The experiments show that our method can make the publishing dataset satisfying the privacy request of each data owner and prevent the slope attack effectively.

Acknowledgements

This work is supported by The Natural Science youth Foundation of Hebei Province (No. F2015407039), Hebei Province Department of science and technology Fund (NO.13227427), Hebei Province Department of Education Fund (NO. sz132075, sz141097, 2012-112), Doctoral Foundation of Hebei Normal University of Science & Technology (2013YB007, 2013YB001).

References

Cao J.M., Carminati B., Ferrari E., Tan K., 2011, CASTLE: Continuously anonymizing data streams. IEEE Transactions on Dependable and Secure Computing, 8(3): 337□-352. [DOI: 10.1109/TDSC.2009.47]

Fung B.C.M., Wang K., Yu P.S., 2007, Anonymizing Classification Data for Privacy Preservation, IEEE Transactions on Knowledge and Data Engineering, 19(5): 711-725. [DOI: 10.1016/j.datak.2008.12.001]

Fung B.C.M., Wang K., Wang L., 2009, Privacy-Preserving Data Publishing for Cluster Analysis, Data & Knowledge Engineering, 68(6): 552-575. [DOI: 10.1145/1857947.1857950]

Hyunji L., Jae-Woo C., 2013, Density-based k-anonymization scheme for preserving users' privacy in location-based services. Lecture Notes in Computer Science, 7861: 536-545. [DOI: 10.1007/978-3-642-38027-3_57]

Lefvre K., DeWitt D., Ramakrishnan R., 2005, Incognito: Efficient full-domain k-anonymity, In Proceedings of the International Conference on Management of Data, 49-60. [DOI: 10.1145/1066157.1066164]

LeFevre K., DeWitt D.J., Ramakrishnan R., 2006, Mondrian Multidimensional K-Anonymity, In Proceedings of International Conference on Data Engineering, 25. [DOI: 10.1109/ICDE.2006.101]

Li J.Y., Wong R.C.W., Fu A.W.C., 2008, Anonymisation by Local Recoding in Data with Attribute Hierarchical Taxonomies, IEEE Transactions on Knowledge and Data Engineering, 20(9): 1181-1194. [DOI: 10.1109/TKDE.2008.52]

Song J.L., Huang L.M., Zhang C., Zhang G.B., 2014, Multi-objective Optimization of K in K-anonymity Model, Journal of Computational Information Systems, 10(22): 9759- 9770. [DOI: 10.12733/jcis13119]

Skarkala M.E., Maragoudakis M., Gritzalis S., 2012, Privacy Preservation by k-Anonymization of Weighted Social Networks. In Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM), IEEE Computer Society: 423-428. [DOI: 10.1109/ASONAM.2012.75]

Sweeney L., 2002, K-Anonymity: a model for protecting privacy, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(5): 557-570. [DOI: 10.1142/S0218488502001648]

Sweeney L., 1997, Guaranteeing anonymity when sharing medical data: the Datafly system, In Proceedings of the 1997 AMIA Annual Fall Symposium, 4 (suppl): 51-55. [DOI: 10.1109/ITAB.1997.649428]

Sweeney L., 2002, Achieving k-anonymity privacy protection using generalization and suppression, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(5): 571-588. [DOI: 10.1142/S021848850200165X]

Wang B., Yang J., 2012, Personalized (α, k)-Anonymity Algorithm Based on Entropy Classification, Journal of Computational Information Systems, 8(1): 259- 266. [DOI: 10.12733/jcis11309]

Yuan M.X., Chen L., Yu P.S., Yu T., 2013, Protecting sensitive labels in social network data anonymization. IEEE Transactions on Knowledge and Data Engineering, 25(3): 633-647. [DOI: 10.1109/TKDE.2011.259]

Zakerzadeh H., Osborn S.L., 2011, FAANST: Fast anonymizing algorithm for numerical streaming data. In Proceedings of the 5th International Workshop on Data Privacy Management and 3rd International Conference. on Autonomous Spontaneous Security, 36-50. [DOI: 10.1007/978-3-642-19348-4_4]