

A Density-based Clustering Method for K-anonymity Privacy Protection

Jie Liu^{1,2*}, Shou-Lin Yin¹, Hang Li¹ and Lin Teng¹

¹Software College
Shenyang Normal University
No.253, HuangHe Bei Street, HuangGu District, Shenyang, P.C 110034 - China
352720214@qq.com;nan127@sohu.com;1451541@qq.com;1532554069@qq.com

²Department of Information Engineering
Harbin Institute of Technology
92 West Straight Street In Nangang District of Harbin Harbin,150001-China
*Corresponding author:nan127@sohu.com

Received July, 2016; revised August, 2016

ABSTRACT. *In order to prevent sensitive information leakage in the cloud storage, we propose a density-based clustering method for K-anonymity privacy protection. We also analyze the effect of identifier on sensitive properties. Density clustering method is used to make sensitive properties cluster for data, which makes the data more similarity in same class. Through this new clustering method, it can make differentiation for query function sensitive properties and improve data availability. The experimental results show that the new algorithm can effectively realize K-anonymity privacy protection. Compared to traditional K-anonymous method, new algorithm can reduce the loss of data information. And the execution time is shorter.*

Keywords: Density, Clustering, K-anonymity privacy protection, Sensitive properties

1. Introduction. With the rapid development of computer network and database technology, medical, bank accounts, email, etc, systems are widely used in people's life. Individual data in the database is excessively used in data mining and data distribution resulting in the personal information privacy leaking[1]. Therefore, personal privacy protection become a ticklish problem, many privacy protection methods are put forward. Generally, it needs data preprocessing before releasing data by deleting identity property (such as name, ID number, etc.). Although this method has a certain effect, the privacy information leakage still exists. Aiming to this question, we propose a density-based clustering method for K-anonymity privacy protection taking background knowledge attack and consistency attack into consideration to reduce the information loss of data anonymization.

1.1. Related work. In the data analysis field, the privacy protection technology can be roughly divided into numerical disturbance[2], query limit[3], anonymous group technology[4] and data distribution[5], etc,. Numerical disturbance technology makes a disturbance for the original data value by adding random noise to hide the real data. Query limit technology limits data inquiry from two aspects: 1) strictly limiting the query number; 2) limiting continuity query. K-anonymous mechanism is actually a group strategy in anonymous group technology. It combines the data with same quasi-identifiers to achieve

the overall data record group. In each group, there are at least k data, so one record can be hidden in k data. Data distribution technology conducts vertical or horizontal partitioning for data to achieve the goal of data hiding. Above privacy protection technologies often do not define background knowledge of data obtained by attacker. Therefore, when dealing with complex attacking model, once attackers acquire more background knowledge, the attacking may become joint attacking and consistency attacking.

So Dwork[6] proposed differential privacy protection model provided privacy proof. But it had poor data availability. LATANYA[7] provided a formal presentation of combining generalization and suppression to achieve k-anonymity. Bhattarai[8] drawn a parallel between face de-identification and oracle attacks in digital watermarking for preventing automatic matching with public face collections. Tang[9] proposed a novel delay-aware privacy-preserving transmission scheme based on a combination of two-phase forwarding and secret sharing. Xie[10] combined distributed randomization with the K-anonymity algorithm to reduce the information loss rate in order to increase the data availability and avoid the leakage of privacy data information. But these methods still exists security risk. Therefore, we propose density-based clustering method for K-anonymity privacy protection. The following are the structures of this paper. In section2, we give some definitions. Section3 detailed introduces the K-anonymity privacy protection with density-based clustering method. Section4 is the experiments part, which is used for demonstrating our method. There is a conclusion in section5.

2. Preliminaries. In this section, we give some definitions to pave below.

Definition 1. Identifier attribute. A specific record directly connects identifier called identifier attribute (such as name, ID, Bank number).

Definition 2. Quasi-identifier. A attribute set connects external attribute called quasi-identifier.

Definition 3. K-anonymity. $T(A_1, A_2, \dots, A_n)$ is a table. In corresponding equivalence class of Quasi-identifier (QI), if the attribute number is at least $K \geq 2$, then T meets K-anonymity.

Definition 4. Equivalence class. After K-anonymization, each group in QI at least has k same elements. The set is one equivalence class.

Definition 5. Reference matrix. Reference matrix of impact probability of QI on sensitive attribute is J . J is $m \times n$ matrix. Where m is the sensitive attribute number. n is QI number. $q_i(1 \leq i \leq n)$ is QI attribute. $p_i(1 \leq i \leq m)$ is sensitive attribute.

$$J = J_{ij}(m \times n) = \begin{bmatrix} 2 & 4 & \dots & 5 \\ 6 & 1 & \dots & 3 \\ \vdots & \vdots & \dots & \vdots \\ 3 & 0 & \dots & 2 \end{bmatrix} \quad (1)$$

Where J_{ij} denotes impact probability of i -th QI on j -th sensitive attribute. Dis_{ab} is the impact distance between a -th sensitive attribute and b -th sensitive attribute.

$$Dis_{ab} = \sum_{i=1}^n |J_{ai} - J_{bi}|. \quad (2)$$

Definition 6. Differential privacy[11-13]. Differential privacy model defines the privacy quantization of probability distribution. Assuming algorithm M meets ϵ -differential privacy model, when M satisfies probability constraint of equation(3).

$$Pr[M(D) = S] \leq e^\varepsilon Pr[M(D') = S]. \quad (3)$$

Where ε denotes privacy protection budget, D is original data set. D' is the adjacent data set of D . It can delete or add any record in D . S is output results set. Differential privacy protection requires that M is applied into adjacent data set getting the same probability ratio of output results (upper bound of probability ratio is e^ε).

3. K-anonymity privacy protection with density-based clustering method. When K-anonymity is attacked by link, it can effectively prevent privacy information leakage. However, if the attack is homogeneity attack and background knowledge attack. K-anonymity cannot prevent it. In order to better protect sensitive information, decrease the information loss and process data diversity. We adopt density-based clustering method in this paper. It is divided into two steps:

1. Calculating the reference matrix of data set background knowledge, so the effect of each QI on sensitive attribute can be obviously reflected. We use density distribution to make clustering for sensitive attribute. The data is more similar in one class and more different outside class. Clustering results are more higher than traditional k-means method.
2. Using density clustering method to make clustering, data will be divided into different equivalence groups according to the density distribution. Computing minimum information loss and merging equivalence class.

3.1. Density-based clustering for sensitive attribute. Density-based clustering method is different from hierarchical method and portioning method. It defines the cluster as points maximizing set of density connecting. Also it can divide the area of enough high density as cluster. This method can find clustering with any shape, which has great advantage on data grouping.

This paper proposes a density-based clustering method. Using density clustering method to make cluster, data will be divided into different equivalence groups according to the density distribution. Data record number in each group is at least k . When it realizes data anonymous, meanwhile, sensibility of inquiry function will be divided into k records in each group to reduce sensibility of inquiry function. In the clustering division, supposing all the attributes are QI, so the data set obtained by clustering can satisfy K-anonymity.

Algorithm 1 $DCM(D, r, k)$

Input. Original data set D , non-sensitive clustering neighbourhood radius r , size of minimum clustering k .

Output. Data set \overline{D} after clustering partition.

- Using density clustering algorithm for D , getting clustering result D_c .
 - Dividing D_c into different clusters.
 - For clusters, cluster centroid replaces each data in clusters.
 - Then return the new data set \overline{D} .
-

Theorem 3.1. *Original data set D , query function f_i , return the i – th record in data set. So $\Delta(f_i \cdot DCM) \leq \frac{\Delta(f_i)}{k}$ is true.*

Proof. We apply DCM algorithm into numeric data set and get \overline{D} . Minimum cluster size is k . Obviously, when the query function f_i acts on data set, because the difference of adjacent data set is divided into k data record, then when $f_i \cdot DCM$ acts on data set,

it will return cluster's centroid of $i - th$ data record. The sensitiveness is $\frac{\Delta(f_i)}{k}$ at most. Therefore, the theorem is true.

3.2. Density-based clustering for K-anonymity privacy protection. To realize K-anonymity, it needs to make a conversion for data. That can result in information loss. For numeric attribute and classification attribute, we use different methods to calculate information loss respectively. Assuming $QI = (N_1, N_2, \dots, N_n, M_1, M_2, \dots, M_m)$. Where $N_i (1 \leq i \leq n)$ is the $i - th$ numeric attribute. $M_j (1 \leq j \leq m)$ is $j - th$ classification attribute. When processing K-anonymity, one tuple $t = (x_{N_1}, x_{N_2}, \dots, x_{N_n}, x_{C_1}, x_{C_2}, \dots, x_{C_m})$ is transformed into $t' = ([y_{N_1}, z_{N_1}], [y_{N_2}, z_{N_2}], \dots, [y_{N_n}, z_{N_n}], D_{C_1}, D_{C_2}, \dots, D_{C_m})$. And its information loss can be defined as following (setting $R = (r_1, r_2, \dots, r_k)$):

Definition 7. Tuple information loss.

$$I_L = \frac{J_{bi}}{\sum_{i=1}^n J_{bi}} \left(\sum_{j=1}^m \sum_{i=1}^n \frac{z_{N_i} - y_{N_i}}{|N_i|} + \sum_{j=1}^m \sum_{i=1}^n \frac{H(\wedge(x_{C_j}, D_{C_j}))}{H(T_{C_j})} \right). \quad (4)$$

Where $\frac{J_{bi}}{\sum_{i=1}^n J_{bi}}$ is the influence weight of QI N_i on sensitive attribute P_b . $|N_i|$ is the range of numeric attribute. $H(\wedge(x_{C_j}, D_{C_j}))$ is the subtree height and its root is the minimum common ancestor of x_{C_j} and D_{C_j} . $H(T_{C_j})$ is the classification tree height with C_j .

Algorithm 2 K-anonymity Privacy Protection with Density-based Clustering Method

Input. Table T with N data records, parameter K .

Output. Table T' after K-anonymization.

- Initializing each record in data set as equivalence class.
 - From G clusters, each cluster selects a record as centroid of equivalence class.
 - Repeat.
 - According to formula (4), computing the information loss when all the records in same cluster in **Algorithm1** are transformed equivalence class.
 - Allocating the record with minimum information loss to corresponding equivalence class.
 - If record number reaches to K in each equivalence class, then stop allocating.
 - Updating the centroid of G equivalence classes.
 - Repeat.
 - Until centroid does not change.
-

4. Experiments and analysis. In this section, we adopt the "Adult" data set to make experiment. Data preparation method is used as reference[14]. Incomplete data or information with name, ID i.e., will be deleted. And we get the experiment data set including 45222 records. (age, work class, education, martial status, race, gender, native country) are as QI attributes. Where "age" and "education" are numeric attributes. "work class" is sensitive attribute. The rest are classification attributes. In order to further demonstrate the effectiveness of our method, we use our scheme to make a comparison with traditional K-anonymity privacy protection method[15], N-K-anonymity[16] and LBS-K-anonymity[17], which are abbreviated to A, B, C, D respectively. Following is the experiment from data quality, running time and information distortion.

4.1. Data quality. Information loss with different algorithms is as figure1. From figure1, we can know that the information loss with new algorithm is smaller than that of other three methods with the increase of K . When K is bigger, the advantage is more obvious, data quality is higher.

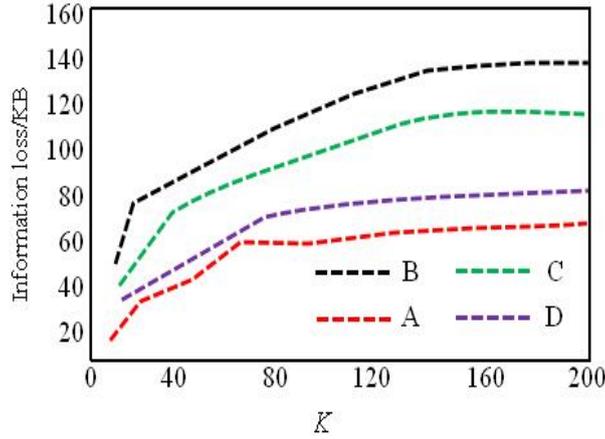


FIGURE 1. Information loss

4.2. **Running time.** The comparison of running time with different algorithms is as shown figure2. We can see that the running time with new method is shorter than original method. This can demonstrate the efficiency of our method.

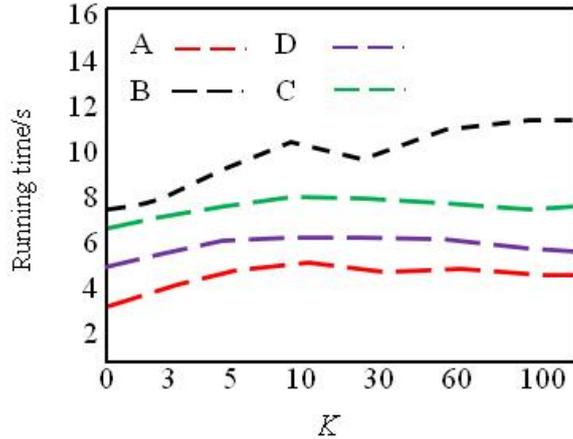


FIGURE 2. Running time

4.3. **Information distortion.** We define that Information distortion denotes data ratio of query results in an uncertain range including "Possible-Sometimes-Inside, PSI and Definitely-Always-Inside, DAI". Information distortion(IDs) can be described as:

$$IDs = [N(\Omega) - N(\Omega')]/N(\Omega). \quad (5)$$

Where $N(\Omega)$ denotes the number of query results in Ω . Ω is area range. $N(\Omega')$ denotes the number of query results in Ω' . Ω' is a new area range after K -anonymity with the different methods. With the increase of K , Information distortion of PSI and DAI will get a higher value as figure3 and figure4. From the two figures, we can know that density-based clustering method for K -anonymity privacy protection is a better choice.

5. **Conclusions.** In this paper, we adopt density-based clustering method for K -anonymity privacy protection. The data will be aggregated as K clusters. Data are similar to each other in one cluster, however, they are greatly different from each other in different clusters. The data with similar sensitive attributes are aggregated together. Then it makes

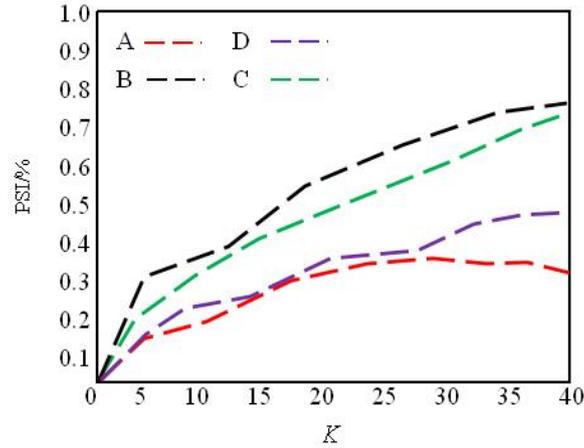


FIGURE 3. Infirmation distortion of PSI

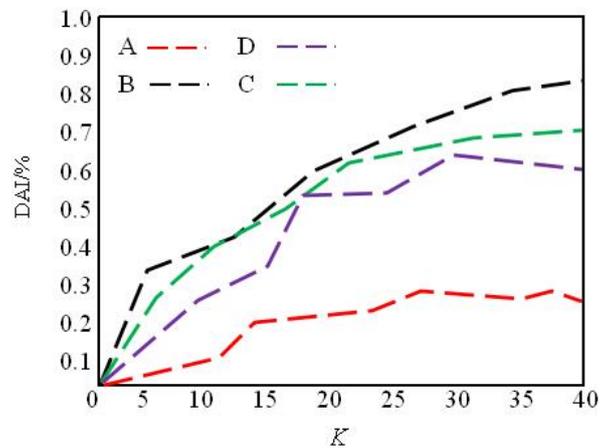


FIGURE 4. Infirmation distortion of DAI

a clustering for the whole data according to aggregation results to achieve K-anonymous. Finally, we make a comparison from data quality and running time. The results show that when the K value increases, this new algorithm can obtain high quality data. In the future, we will study how to realize the efficient protection for huge amounts of data.

Acknowledgment. The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

- [1] J. Ge, J. Peng, Z. Chen, Your privacy information are leaking when you surfing on the social networks: A survey of the degree of online self-disclosure (DOSD) *Proc. in IEEE, International Conference on Cognitive Informatics & Cognitive Computing. IEEE Computer Society*, pp. 329-336, 2014.
- [2] Z. Radojevic, V. Terzija, Numerical algorithm for overhead lines protection and disturbance records analysis, *[J]. Iet Generation Transmission & Distribution*, vol. 1, no. 2, pp. 357-363, 2007.
- [3] M. Herrmann, C. Grothoff, Privacy-Implications of Performance-Based Peer Selection by Onion-Routers: A Real-World Case Study Using, *I2P[C]// International Conference on Privacy Enhancing Technologies. Springer-Verlag*, pp. 155-174, 2011.
- [4] T. Gao, Q. Miao, N. Guo, Anonymous Authentication Scheme Based on Proxy Group Signature for Wireless MESH Network, *Proc., on Eighth International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing*, pp. 533-537, 2014.

- [5] C. Guo, Q. Shen, Y. Yang, et al. User Rank: A User Influence-Based Data Distribution Optimization Method for Privacy Protection in Cloud Storage System, *Proc., on Computer Software and Applications Conference. IEEE*, pp. 104-109, 2015.
- [6] C. Dwork, Differential privacy, [M]. *Encyclopedia of Cryptography and Security. Springer US*, pp. 338-340, 2011.
- [7] L. Sweeney, Achieving k-anonymity privacy protection using generalization, and suppression, *International Journal of Uncertainty Fuzziness and Knowledge-Based Systems*, , vol. 10, no. 5, pp. 571-588, 2012.
- [8] B. Bhattarai, A. Mignon, F. Jurie, et al., Puzzling face verification algorithms for privacy protection, *Proc., on IEEE International Workshop on Information Forensics and Security. IEEE*, pp. 66-71, 2014.
- [9] D. Tang, J. Ren, A Novel Delay-Aware and Privacy Preserving (DAPP) Data Forwarding Scheme for Urban Sensing Network, [J]. *IEEE Transactions on Vehicular Technology*, pp. 1-10, 2015.
- [10] Y. Xie ,Q. He, D. Zhang, et al., Medical ethics privacy protection based on combining distributed randomization with K-anonymity, *Proc., on International Congress on Image and Signal Processing. IEEE*, 2015.
- [11] A. Nikolov, K. Talwar, Z. Li, The Geometry of Differential Privacy: The Small Database and Approximate Cases, [J]. *Siam Journal on Computing*, vol. 45, no. 2, pp. 575-616, 2016.
- [12] K. Nissim, U. Stemmer, On the Generalization Properties of Differential Privacy, [J]. *Computer Science*, 2015.
- [13] T. Zhu, P. Xiong, G. Li, et al., Correlated Differential Privacy: Hiding Information in Non-IID Data Set, [J]. *IEEE Transactions on Information Forensics & Security*, vol. 10, no. 2, pp. 229-242, 2015.
- [14] N. Dalal, and B. Triggs, Triggs, B.: Histograms of Oriented Gradients for Human Detection, *In: CVPR*. vol. 1, pp. 886-893, 2005.
- [15] L. Sweeney, k-anonymity: A model for protecting privacy, [J]. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 5, pp. 557-570, 2002.
- [16] J. Jia, F. Zhang, Non-deterministic K-anonymity Algorithm Based Untrusted Third Party for Location Privacy Protection in LBS, [J]. *International Journal of Security and Its Applications*, vol. 9, no.9, pp. 387-400, 2015.
- [17] Y. M. Ye , C. C. Pan , G. K. Yang An Improved Location-Based Service Authentication Algorithm with Personalized K-Anonymity, *Proc., on China Satellite Navigation Conference (CSNC) 2016 Proceedings: Volume I*. Springer Singapore, pp. 257-266, 2016.