AQ-DP: A New Differential Privacy Scheme based on Quasi-Identifier Classifying in Big Data

Haifeng Ke[1], Anmin Fu[1], Shui Yu[2,3], and Si Chen[1]
[1] School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China;
[2] School of Computer Science, Guangzhou University, 510006, China.
[3] School of Software, University of Technology Sydney, 2007, Australia.
fuam@njust.edu.cn

***Abstract*—The rapid development of big data has brought great convenience to human's lives. The circulation and sharing of information are two main characteristics of the big data era. However, the risk of privacy leakage is also greatly increased when we enjoy the various services of big data. Therefore, how to protect the data privacy in the complex context of big data has become a research hotspot in academic circles. Most of the current researches on privacy protection are divided into two research fields: *k-anonymity* and *differential privacy*. Some existing research shows that traditional methods of privacy protection, such as k-anonymity and its extension, cannot achieve absolutely security. The emergence of differential privacy provides a new solution for privacy protection. We draw the lessons from exiting work and propose a new privacy method based on differential privacy: AQ-DP. We propose the first method for classifying quasi-identifiers based on sensitive attributes, which divide quasi-identifiers into associated quasi-identifiers (AQI) and non-associated quasi-identifiers (NAQI). The purpose is not to lose the correlation between quasi-identifiers and sensitive attributes. Our model AQ-DP carries out random shuffling of NAQIs, generalizes the AQIs, and adds random noise that satisfies the laplacian distribution to the statistics. We have conducted extensive experiments, confirming that our model can achieve a satisfying privacy level and data utility.**

Keywords—Big data; Privacy protection; Data utility

I. INTRODUCTION

In the information age, big data is an epoch-making product, which brings great convenience for society and provides a new direction for scientific research. Meanwhile, the application of information technology has been widely expanded, increasingly enjoying great popularity. As such, various information systems can store and accumulate abundant data, such as patient diagnostic data sets established by medical institutions and online transaction data sets collected by e-commerce enterprises, allowing people to gain more valuable knowledge about the real world.

Nonetheless, there are also some potential problems in processing big data. The most noteworthy by-product is the privacy threat, which may put users at risk. The data set usually contains a large amount of personal privacy information, such as medical diagnosis results, personal consumption habits and other data that can reflect individual characteristics. This information will be leaked along with the data set released and shared. Publishing unprocessed or improperly processed data may expose users' sensitive information to adversaries [1]. Therefore, privacy protection research has become a worldwide hot topic to scholars.

The concept of data privacy was first put forward by a statistician, Dalenius, in the late 1970s [2]. He asserted that to protect the privacy of information in the database, we must disenable users to obtain precise information involving any individual in the process of accessing the database. With the development of privacy protection data publishing, two main branches have gradually arisen. One is *k-anonymity* and its extensions, *l-diversity*, *t-closeness*, and so on. The other is *differential privacy* and its extensions, which provides strict privacy protection to statistics data sets.

The *k-anonymity* [3] and its extended models [4-6] are far-reaching, which are extensively studied in the field of privacy protection. The basic idea of these models is to generalize and compress the recorded quasi-identifier value. Specifically, all records are divided into several equivalence groups, and the records in each group have the same quasi-identifier value. it enables single record to be hidden in a set of records. Hence, such models are also known as group-based privacy protection models.

However, follow-up studies have shown that such models cannot achieve the full protection of privacy [7-10]. The essential reason is that the security of the group-based privacy protection model is related to the background knowledge possessed by adversaries, while all possible background knowledge is difficult to define adequately. Therefore, a privacy model independent of background knowledge can resist any new type of attacks.

Differential privacy, the other method presented by Dwork in 2006 [11-12], can solve the issues mentioned above. The basic background is that an attacker may obtain expected information by multiple queries to a statistical database on top of his background knowledge of victims. The defence strategy is: for two data sets with a minimum difference, the difference between the queries on the two data sets is very limited, therefore limiting the information gain for attackers. *Differential privacy* protection does not allow for any possible background knowledge possessed by attackers. Meanwhile, it is built on a solid mathematical foundation that defines privacy strictly and provides a quantitative assessment method. In this regard, the theory of

978-1-5386-4727-1/18/$31.00 ©2018 IEEE

differential privacy is quickly recognized by the industry, and has gradually become a hot topic in privacy protection research [15-20].

We design a new *differential privacy* scheme to address the issues in the existing schemes. More specifically, a method we propose aims to divide quasi-identifiers into associated quasi-identifiers (*AQIs*) and non-associated quasi-identifiers (*NAQIs*). *AQIs* refers to the quasi-identifier which is associated with the sensitive attributes. *NAQIs* is the opposite. In addition, we put forward a new privacy model by randomizing shuffling of *NAQIs* and generalization of *AQIs*, finally merging and adding Laplace noise to the statistics. This new model of privacy protection has not been studied before, named as AQ-DP model.

In this paper, we have three contributions as following.

1) Firstly, we propose a QI-classifying method. This is the first method to classify quasi-identifiers based on sensitive attributes. According to this classification, the correlation between the quasi-identifier and the sensitive information cannot be destroyed when we process the data.

2) Secondly, we design a novel privacy model which can solve the problem of the loss of correlation between the quasi-identifier and the sensitive information. The mutual information comparison indicates that our scheme has less privacy loss, hence greatly facilitating the utility of the data.

3) Thirdly, we introduce differential privacy protection mechanism, which is different from the existing differential privacy protection mechanism. For different groups of data, we can provide different differential privacy processing methods, which is more flexible and can provide stronger privacy protection level.

The remainder of this paper is organized as follows. Section II introduces the related work. Then we detailedly present AQ-DP and propose two main algorithms in Section III and Section IV. We analyze our scheme in section V. Section VI evaluates the performance of AQ-DP. Finally, we conclude our paper in Section VII.

II. Related Work

In this section, we review the mainstream approach of privacy preserving data publication.

Generally, there are two different types of privacy preserving data publication. Sweeney put forward the *k-anonymity* [3] model, which grouped the quasi-identifiers by tuple generalization and suppression. The quasi-identifiers in each group are the same and contain at least *k* tuples, so each tuple cannot be distinguished from the other *k-1* tuples. Since the *k-anonymity* model considers all the attribute sets, and specific attributes are not completely defined, which makes it difficult for an attribute to be processed anonymously. An attacker can effectively get the value of an attribute if the value of the sensitive attribute is equal to that of an equivalent class.

Then *l-diversity* [4] was proposed, which is characterized by that the diversity of sensitive data in each anonymous group is greater than or equal to *l*. *l-diversity* makes the average frequency of sensitive data possible. However, the attackers can guess the value when the data in the same equivalence class is not big enough. Similarly, the *t-closeness* [5-6] scheme requires the distribution of sensitive data in the equivalence class to be consistent with that of the data in the entire table.

On the other hand, Dwork [11-12] presented the *differential privacy*, which builds on a solid mathematical foundation and provides a strict quantitative assessment method to privacy protection. Massive statistical methods have been applied to this model. Therefore, it opens up a new direction of privacy protection and has been widely studied by scholars around the world.

Hardt et al. [13] proposed an efficient mechanism, i.e., PMW. The theoretical basis of the mechanism is weighted majority algorithm in machine learning, which is used to construct a composite algorithm through voting mechanism. The mechanism is able to answer more queries under a given privacy protection budget. Xiao et al. [14] proposed a histogram publishing algorithm based on *k-d* tree. The original algorithm first generates a histogram according to the given data set and *k* attributes, gets all the data grid and its frequency, and adds noise with the privacy budget ε/2 to frequency. Then the *k-d* tree algorithm is used to recursively divide the space, and the frequency after the interference is used as the data point in the *k*-dimension space.

Li et al. [15] studied the merits and demerits of *k-anonymity* and *differential privacy* protection models, and proposed "security k-anonymity" with *k-anonymity* and the *differential privacy* combined. This study provides a new idea for the realization of (ε, δ)-*differential privacy*, but only a theoretical framework is proposed. There is no specific solution to the implementation of a sampling function which satisfies the condition of ε-differential privacy.

III. Preliminary And System Modeling

Some concepts and theory of differential privacy are demonstrated in this section. Moreover, we show our system model, which can effectively protect the privacy of users.

A. Preliminary

Generally, based on privacy considerations, attributes can be classified into the following four categories in a published data table T. Table I shows the elements of big data sets.

1) Explicit Identifier: The unique attribute that can clearly identify an individual, such as name, passport number, and so on.

2) Quasi-Identifier: The attributes that can re-identify an individual with a high probability, combine other external information, like gender, age, code, etc.

3) Sensitive Attribute: The information that need to be protected, like medical records, income information. It cannot uniquely identify a user's record.

4) Other Attributes: All other attributes that have no effect on privacy preserving data publishing.

Table I. Elements of Datasets

Explicit Identifier	EI
Quasi-Identifier	QI
Sensitive Attribute	SA
Other Attributes	OA

Table II. Elements of Quasi-Identifiers

Associated quasi-identifier	AQI
Non-associated quasi-identifier	NAQI

Definition 1 (quasi-identifiers classification**).** We divide quasi-identifiers into associated quasi-identifiers (AQIs) and non-associated quasi-identifiers (NAQIs). AQIs is the QI which associated with the SA. NAQIs is the opposite.

As table II shows, we can process them separately by sorting the quasi-identifiers, thus ensuring the association between the quasi-identifier and the sensitive information.

Definition 2 (ε-differential privacy**).** A randomized function K meets ε-differential privacy if all data sets T_1 and T_2 differ at most one element, and for any possible output $O \in$ Range(K),

$$\Pr[K(T_1) \in O] \leq e^{\varepsilon} \times \Pr[K(T_2) \in O] \quad (1)$$

In the above formula, $K(T_1)$ and $K(T_2)$ represent the output obtained by using T_1 and T_2 as the input of algorithm K, where K is a differential privacy stochastic algorithm. $\Pr[K(T_1) \in O]$ and $\Pr[K(T_2) \in O]$ denote the probability that the outputs $K(T_1)$ and $K(T_2)$ belong to O. Therefore, under the circumstances of an appropriate ε, it is difficult for a specific output O to determine whether the original data table is T_1 or T_2.

Definition 3 (Global Sensitivity**).** For any query function *f*: $D \rightarrow R^d$, the global sensitivity Δf of *f* is

$$\Delta f = \max_{T_1,T_2} ||f(T_1) - f(T_2)||_1 \quad (2)$$

for all T_1 and T_2 differing in one record.

Therefore, functions with lower sensitivity mean the differentially private mechanism is more tolerant towards changes in the database.

Definition 4 (Laplace Mechanism**).** The Laplace mechanism achieves *differential privacy* by adding Laplace noise to real values. Specifically, the noise is generated according to a Laplace distribution Lap(λ) with the probability dense function $\Pr(\eta) = \frac{1}{2\lambda} e^{\frac{-|\eta|}{\lambda}}$, where the parameter λ is determined by the global sensitivity Δf and the desired *differential privacy* parameter ε.

B. System Model

In this section, we present the proposed differential privacy model based on the *AQIs*. As shown in Figure 1, companies and governments collect data from users, then process the data through our system AQ-DP and finally release for inquiries. As the data in the procedure of collecting and processing may be "internal disclosure", which we cannot decide, we only discuss the possibility of data disclosure during the release process.

A general representative of a record is shown in equation (3-5). We remove the identifier and other information and segment QIs into two parts: *AQIs* and *NAQIs*.

$$\text{data} = \{AQIs, NAQIs \mid SA\} \quad (3)$$

$$AQIs = \{AQI_1, AQI_2, \dots, AQI_i\} \quad (4)$$

$$NAQIs = \{NAQI_1, NAQI_2, \dots, NAQI_j\} \quad (5)$$

Our program AQ-DP mainly process user data in three steps. We list the changes of a record in the dataset as equation (6-8).

Steps 1 (Grouping and Shuffling)**.** We group the data by SA and take the strategy of random shuffling for each group. But it is important to note that we only carry out random shuffling of *NAQIs*.

Steps 2 (Generalization)**.** Next we need to generalize the remaining QIs for each group, which we call AQIs. Its purpose is to protect the privacy of the data without shuffling.

Steps 3 (Merging and Adding noise)**.** We merge the same items in each group and count the statistics for each item. Finally, we add random noise which satisfies the Laplacian distribution to the statistic. Of course, the negative statistics is meaningless, so our scheme ensures that the statistics are nonnegative after adding noise processing.

$$data_t\_S = \{AQ_{1t}, AQ_{2t}, \dots, AQ_{it}, NAQ_{1r_1}, NAQ_{2r_2}, \dots, NAQ_{jr_j} | SA\} \quad (6)$$

$$data_t\_G = \{GAQ_{1t}, GAQ_{2t}, \dots, GAQ_{it}, NAQ_{1r_1}, NAQ_{2r_2}, \dots, NAQ_{jr_j} | SA\} \quad (7)$$

$$\text{output\_data} = \{GAQ_{1t}, GAQ_{2t}, \dots, GAQ_{it}, NAQ_{1r_1}, NAQ_{2r_2}, \dots, NAQ_{jr_j} | SA | nc_t\} \quad (8)$$

Where $data_t\_S$ is one record of the data table after grouping and shuffling, and $data_t\_G$ is the record after the process of generalization. output_data represents the data table released for analysis.

After this three-step process, we can solve the trouble of information loss caused by random shuffle. In addition, our program can provide greater level of privacy protection and theoretical support.

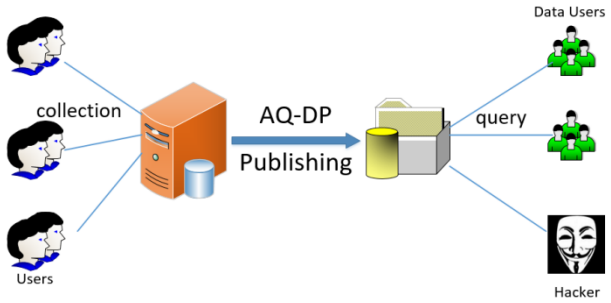

Fig.1. Privacy System

IV. ALGORITHMS

In this section, we present our data publishing algorithm, which includes random shuffle algorithm and differentially

private noise generation algorithm. We use the *Fisher-Yates Shuffle algorithm* in **Algorithm 1**. This is a classic algorithm, and we use it to process our *NAQIs*. Meanwhile, we design a differentially private noise generation algorithm with bounded noise.

The key notations are summarized in Table III. *Rand(i)* generate a random number belonging to [1, n].

Table III. Key Notations

Notation	Definition
tc	true count
nc	noisy count
ln	Laplace noise
T	Raw data table
TS	Data table after random shuffling
AQIs	The *s*th AQI
NAs	The *s*th NAQI
NAs[*i*]	The *i*th record of the column of the NAs
[α, β]	true count range
[γ, η]	Laplace noise range

When we shuffled the data table, we need to generalize *AQIs* of every *TS*. Because some medical researchers or government officials are the users of the final model, they have more intimate knowledge of the data than us. According to the experimental requirements of the generalization of the data operation, the specific scope of the generalization and operation steps need them to decide.

Algorithm 1 Random Shuffling Algorithm

Input: T
Output: TS

```
for each table T_k
  for each NAQI NA_s
    for(int i = T_k.max; i>=1; --i)
      int j=rand(i)
      exchange(NA_s[i], NA_s[j])
    end for
  end for
end for
return TS
```

Existing work has added Laplace noise to the statistics. However, the Laplace noise has its own attribute. The resulting value may be negative after adding noise to it. Direct use of Laplace noise without restricted will make the released datasets meaningless. Hence, we introduce a bounded noise generation mechanism to solve this trouble.

In practical application, the value of true count should belong to a certain range [α, β] and $\alpha = 0$ in normal situation. The Laplace noise should belong to another range [γ, η]. So it must be ensured that $\mathrm{tc} + \mathrm{nc} > 0$ when generating noise.

Next we will introduce our differentially private noise generation algorithm in **Algorithm 2**. We guarantee that our statistics will not be negative after adding random noise. At this point, we have completed the data processing.

Algorithm 2 Differentially Private Noise Generation Algorithm

Input: $\alpha = 0$, β , $\{tc_i \mid i = 1, \ldots, n_1\}$, γ, η
Output: $\{nc_i \mid i = 0, \ldots, \mathrm{N}\}$

```
01: if α < 0 || α > β || n_1 < 0, then
02: return ⊥
03: end if
04: for all p, q∈[1, n_1]
05:     μ = Average { tc_p }
06:     Δf = max_{p,q}{|tc_p − tc_q|}
07: end for
08: b = Δf / μ , β = 2 * μ
09: ln_i ← pdf(x)
10: if ln_i < η && ln_i > γ && i < n_1, then
11:     nc_i ← tc_i + ln_i,  i++
12: end if
13: if i ≠ n_1, goto line 09
14: end if
15: return {nc_i | i = 0, . . ., N}
```

V. Analysis

Suppose that a database T_1 need to be published for analysis. Our objective is to hide T_1 such that the generated database T_2 meets the following two requirements.

(1) **Satisfying ε-differential privacy:** Formally, for any database T_1 differing in at most one record with T_2, the anonymization function K satisfies $\Pr[\mathrm{K}(T_1) \in \mathrm{O}] \leq e^{\varepsilon} \times \Pr[\mathrm{K}(T_2) \in \mathrm{O}]$.

(2) **Achieving a high data utility:** Differential privacy usually need to add noise to the original dataset. Therefore, the data utility is inevitably influenced. The proposed scheme should perform fewer negative impacts on the data utility.

A. Data Utility

The data utility of the published datasets can be described as the degree of data distortion. Relative entropy, also known as the KL-divergence (Kullback-Leibler divergence), is a method of describing the difference between the two probability distributions P and Q [16]. In our model, the KL-divergence of the original data set P and the expected output data set Q should be

$$\mathrm{D}(P, Q) = \sum_{i}^{x} P(i) * log \frac{P(i)}{Q(i)} \tag{9}$$

where x is the sample space in the whole data set.

KL-divergence is a popular method to measure the distance between different distributions, and called KL-distance. However, it does not satisfy the concept of distance because of its asymmetry. It is not accurate enough to use it directly. Therefore, we make appropriate improvements to make it into a symmetric metric which is

$$D_{KL}(P, Q) = \frac{1}{2}\big(D(P, Q) + D(Q, P)\big)$$
$$= \frac{1}{2}\left\{\sum_{i}^{x} P(i) * log \frac{P(i)}{Q(i)} + \sum_{i}^{x} Q(i) * log \frac{Q(i)}{P(i)}\right\} \tag{10}$$

Then the unbalance of the original KL-divergence has been eliminated by $D_{KL}(P,Q)$. The $D_{KL}(P,Q)$ is regarded as zero as $lim_{i\to 0}\, ilogi = 0$ whenever $P(i)$ or $Q(i)$ equals to 0.

B. Privacy Level

The Mutual Information (MI) metric is applied to evaluate the total privacy lever in our scheme. MI measures the mutual dependence between two sets X and Y. In addition, it can be seen as a random variable containing information about the other, or the uncertainty that a random variable changes with another known random variable reduces. For example, if X and Y are two identical variables, if we get the value of X and then we can determine the value of Y and vice versa. In other words, if X and Y are independent, then knowing X cannot help us get any information about Y and vice versa, so the MI between them is zero. The MI we used is defined as follows. Given x and y, the MI between x and y is defined as:

$$\mathrm{MI}(\mathrm{x},\mathrm{y}) = \sum_{x}\sum_{y} \Pr(x,y)\, log\, \frac{\Pr(x,y)}{\Pr(x)\Pr(y)} \tag{11}$$

VI. Performance Evaluation

Our prototype of the data publishing scheme runs on Windows10 with an Intel i7-7700K processor and an 8 GB memory. The database we use in this paper is the "Adult" data set released by UC Irvine machine learning repository, which lean close to the real datasets. We have already remove the "name" as EI and some records of other attributes. The privacy budget we choose is 0.1 to compare with other scheme. In this section, we compare our scheme to SA-Group and k-anonymity, to evaluate the performances including data utility, privacy level, and efficiency.

A. Data Utility

In terms of comparison of data utility, KL-divergence is widely used to measure the similarity of two datasets and we can measure the utility of dataset D with respect to dataset D'. Fig.2. tells us that our model has significantly better performance than SA-Group and K-anonymity.

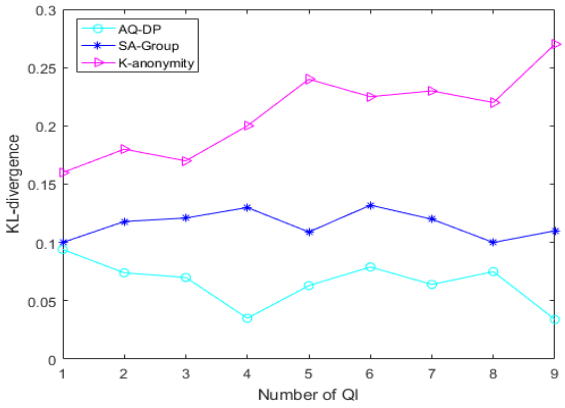

Fig.2. KL-divergence for different QI

B. Privacy Level

We evaluate the total privacy lever under the Mutual Information (MI) metric. The larger the MI value is, the higher the similarity between the processed data set and the original data set is. Therefore, the privacy information is easily leaked. We can conclude that our privacy lever is significantly higher than SA-Group model and *k-anonymity* as it shown in figure 3.

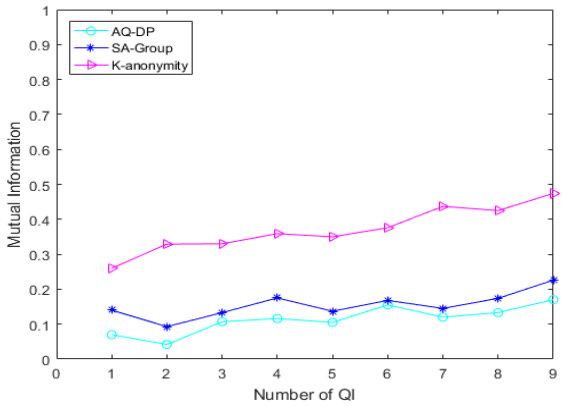

Fig.3. MI changes when QI increase

C. Efficiency

We measure the performance of AQ-DP, SA-Group and *k-anonymity* in this part. To facilitate our experiments, we set the number of AQI to the half of the number of QI in our model experiments. Figure 4 and figure 5 show that running time of the AQ-DP model does not increase significantly with QI's increase and the size increase of dataset. That means our model can meet different requirements in different scenarios. The robust of the proposed scheme is satisfying. At the same time, our scheme has similar performance as SA–Group and costs less running time than K-anonymity. Through our analysis, we find that the first two steps of our scheme have the same time complexity as SA-Group, we only have a little time consumption in the third step.

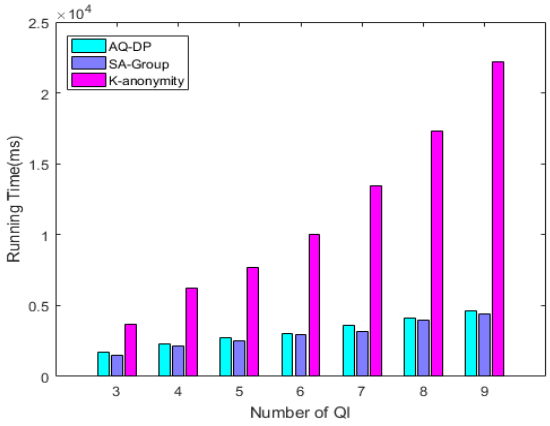

Fig.4. Efficiency changes when QI increase

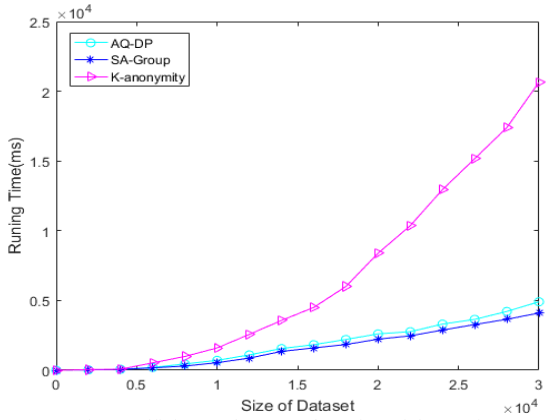

Fig.5. Efficiency changes when size of dataset increase

D. Model Analysis

We finally analyze the performance of our own scheme. There are 9 *QIs* in our experiment, and we take the quantity of *AQI* as independent variable, so we can analyze its performance change with the increase of *AQI*. We analyze their respective performance under different privacy budget ε (we choose 0.1, 0.2, 0.5 and 0.8). We can see the running time is in the slow reduction, as Fig.6. shown, with the increase of AQI amount. This illustrates that our scheme has excellent stability. The reason for the decrease is that the data elements of the database have changed, and the frequency of noise added has been reduced.

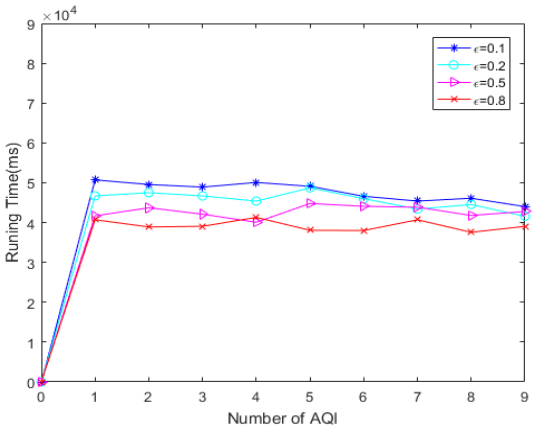

Fig.5. Efficiency changes when AQI increase

VII. Conclusion

Our privacy is being threatened while we are enjoying the conveniences of big data. So how to protect users' privacy becomes more and more important. In this paper, we propose the first method of classifying quasi-identifiers and a novel approach for differentially private publishing data. Our model can solve some shortcomings of existing schemes, and can greatly improve the availability of data under the premise of privacy protection. Our massive experiments on real-world data set prove that the proposed model is effective and efficient.

Acknowledgment

This work is supported by National Science Foundation of China (61572255, 61702266), Six talent peaks project of Jiangsu Province, China (XYDXXJS-032).

References

[1] S. Yu. "Big privacy: Challenges and opportunities of privacy study in the age of big data." *IEEE Access*, vol. 4, pp. 2751–2763, 2016.

[2] Dalenius T. "Towards a methodology for statistical disclosure control." *Statistic Tidskrift*, vol. 15, no. 2, pp. 429-444, 1977.

[3] P. Samarati, L. Sweendey. "Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression." in *Proc. IEEE S&P*, pp. 1–19, 1998.

[4] A. Machanavajjhala, J. Gehrke J, D. Kifer, et al. "L-diversity: privacy beyond k-anonymity." in *Proc. IEEE ICDE*, pp. 24-24. 2006.

[5] N. Li, T. Li, S. Venkatasubramanian. "Closeness: A New Privacy Measure for Data Publishing." *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 7, pp. 943-956, 2009.

[6] J. Soriacomas, J. Domingoferrer, D. Sanchez, et al. "t-closeness through microaggregation: Strict privacy with enhanced utility preservation." *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 11, pp. 3098–3110, 2015.

[7] A. Fu, S. Yu, Y. Zhang, et al. "NPP: A New Privacy-Aware Public Auditing Scheme for Cloud Data Sharing with Group Users." *IEEE Transactions on Big Data*, DOI: 10.1109/TBDATA.2017.2701347, 2017.

[8] A. Fu, Y. Li, S. Yu, et al. "DIPOR: An IDA-based Dynamic Proof of Retrievability Scheme for Cloud Storage Systems." Journal of Network and Computer Applications, vol.104, pp. 97-106, 2018.

[9] A. Fu, S. Li, S. Yu, Y. Zhang, and Y. Sun, "Privacy-Preserving Composite Modular Exponentiation Outsourcing with Optimal Checkability in Single Untrusted Cloud Server", *Journal of Network and Computer Applications*, 118（2018）：102-112.

[10] S. Yu, M. Liu, W. Dou, X. Liu, S. Zhou, "Networking for Big Data: A Survey," IEEE Communications Surveys and Tutorials, Vol 19, No 1, 2017, pp531-549

[11] C. Dwork. "Differential privacy." *Lecture Notes in Computer Science,* vol. 26, no. 2, pp. 1–12, 2006.

[12] C. Dwork, F. McSherry, K. Nissim, et al., "Calibrating noise to sensitivity in private data analysis." in *Proc. TCC*, pp. 265–284, 2006.

[13] M. Hardt, G. N. Rothblum. "A Multiplicative Weights Mechanism for Privacy-Preserving Data Analysis." in *Proc. IEEE FOCS*, pp. 61-70, 2010.

[14] Y. Xiao, L. Xiong, C. Yuan. "Differentially private data release through multidimensional partitioning." in *Proc. VLDB*, pp.150-168, 2010.

[15] N. Li, W. Qardaji, D. Su. "On sampling, anonymization, and differential privacy or, k -anonymization meets differential privacy." in *Proc. ACM CCS*, pp. 32-33, 2012.

[16] W. Qardaji, W. Yang, and N. Li. "PriView: practical differentially private release of marginal contingency tables." in *Proc. ACM SIGMOD*, pp. 1435–1446, 2014.

[17] H. H. Nguyen, J. Kim, and Y. Kim. "Differential privacy in practice." *Journal of Computing Science and Engineering*, vol.7, no. 3, pp. 177–186, 2013.

[18] Y. Qu, S. Yu, L. Gao, et al. "Big data set privacy preserving through sensitive attribute-based grouping." in *Proc. IEEE ICC*, pp.1-6, 2017.

[19] X. Yang, T. Wang, X. Ren, et al. "Survey on Improving Data Utility in Differentially Private Sequential Data Publishing." *IEEE Transactions on Big Data*, DOI: 10.1109/TBDATA.2017.2715334, 2017.

[20] J. Brickell, V. Shmatikov. "The cost of privacy: destruction of data-mining utility in anonymized data publishing." in *Proc. ACM SIGKDD*, pp. 70-78, 2008.