A Distributed Privacy-Utility Tradeoff Method Using Distributed Lossy Source Coding with Side Information

Yonghao Gu[1*], Yongfei Wang[1], Zhen Yang[1] and Yimu Gao[2]
[1] Beijing Key Laboratory of Intelligent Telecommunication Software and Multimedia, School of Computer Science
Beijing University of Posts and Telecommunications
Beijing 100876 - P.R. China
[e-mail: guyonghao@bupt.edu.cn]
[2] Department of Computer and Information Sciences, University of Delaware
Newark DE, 19716 - USA
[e-mail: delphox0121@gmail.com]
*Corresponding author: Yonghao Gu

Received October 22, 2016; revised January 22, 2017; revised February 23, 2017; accepted March 16, 2017; published May 31, 2017

Abstract

In the age of big data, distributed data providers need to ensure the privacy, while data analysts need to mine the value of data. Therefore, how to find the privacy-utility tradeoff has become a research hotspot. Besides, the adversary may have the background knowledge of the data source. Therefore, it is significant to solve the privacy-utility tradeoff problem in the distributed environment with side information. This paper proposes a distributed privacy-utility tradeoff method using distributed lossy source coding with side information, and quantitatively gives the privacy-utility tradeoff region and Rate-Distortion-Leakage region. Four results are shown in the simulation analysis. The first result is that both the source rate and the privacy leakage decrease with the increase of source distortion. The second result is that the finer relevance between the public data and private data of source, the finer perturbation of source needed to get the same privacy protection. The third result is that the greater the variance of the data source, the slighter distortion is chosen to ensure more data utility. The fourth result is that under the same privacy restriction, the slighter the variance of the side information, the less distortion of data source is chosen to ensure more data utility. Finally, the provided method is compared with current ones from five aspects to show the advantage of our method.

Keywords: Privacy-utility tradeoff, rate distortion, distributed source coding, side information

This work was supported by the National Natural Science Foundation of China (No. 61173017, No. 61370195), Communication Soft Science Foundation of Ministry of Industry and Information (No. 2014-R-42, No. 2015-R-29), and Key Lab of Information Network Security Foundation of Ministry of Public Security (No. C14613).

https://doi.org/10.3837/tiis.2017.05.026 **ISSN : 1976-7277**

1. Introduction

In the era of Big Data, with the popularity of sensing devices and extensive use of data query and analysis services, the collection and mining of user's data has become a fast growing and common practice by a large number of institutions. We consider a user has two kinds of correlated data: sensitive data (or private data) that he/she would like to keep private and non-sensitive data (or public data) which he/she is willing to release to the third party for data analysis and query services. Data query and data analysis mean the utility of data, while avoiding the disclosure of sensitive attributes could protect user's privacy [1].

Unfortunately, Most of the work is concerned with privacy protection, ignoring usability. The question is how to find the balance between data utility and privacy to provide the data utility in the same time to preserve the data privacy [2]. More generally, it is needed to achieve the utility-privacy tradeoff region including all optimal tradeoff points. Besides, a data demander often requires the aggregate data from all the distributed users to conduct data analysis and mining [3]. Another challenge is that side information (referred to as auxiliary information in the database literature) from other databases (in general from sources external to the database of interest) in conjunction with one or more queries can also result in privacy breaches. Therefore, it is significant to solve the utility-privacy tradeoff problem and find the tradeoff region in the distributed environment with side information, which is the goal of this paper. To address the above problems, Ali [4] characterizes the privacy-accuracy tradeoff in terms of an optimization problem. The paper gives a lower bound on the probability of error in inferring the private data from the released data. Andreas [5] provides a utility-privacy tradeoff optimization method to find a provably near-optimal result. This paper evaluates the method using data got from the search activity log of many paticipants and assess their preferences about privacy and utility by a large-scale survey activity, aiming at geting users' willingness to trade data sharing privacy in returns for the efficiency of search result. Ping Xiong [6] defines two metrics, Utility Loss and Privacy Gain to evaluate the quality of anonymized results and to find the optimal anonymization solution to resist probabilistic inference attacks. Grigorios [7] provides a distance-based quality criterion that handles both QIDs(quasi-identifiers) and SAs(sensitive attributes) for k-anonymization data. Besides, they design an efficient heuristic algorithm for anonymization data with utility privacy tradeoff, which optimizes the weighted sum of the amount of generalization of QIDs and the amount of protection of SAs. Li [8] provides a privacy-utility tradeoff method for data publishing using the risk-return tradeoff in financial investment, whose concepts are borrowed from the Modern Portfolio Theory. Rashid [9] develops an analytical cost model to find the optimal tradeoff between privacy and data utility in terms of monetary value. Based on defining the utility $U(A)$ and privacy cost $C(A)$, and releasing any given set of attributes A without privacy disclosure, the goal of the paper [10] is to find a set A, that maximizes $U(A)$ while minimizing $C(A)$. In order to solve this tradeoff, Andreas [10] defines the objective function $F_\lambda(A) = U(A) - \lambda C(A)$, in which λ is a privacy-to-utility conversion factor, so that the tradeoff goal is transformed to solve the optimization problem of $A_\lambda^* = \arg\max_A F_\lambda(A)$. Different solutions A_λ^* can be found by varying λ that smaller λ leads to higher utility and higher cost, while larger values of λ leads to lower utility and privacy cost. Because solving the objective function $F_\lambda(A)$ is an NP-hard problem, which means it is very hard to find an optimal solution and also it cannot

give the utility and privacy cost region quantitatively. Grigorios [11] uses the R-U confidentiality map to assess the balance between disclosure risk and data utility. In addition, the effectiveness of three transaction algorithms (Aprior [12], COAT [13] and PCTA [14]) with privacy constraint is compared using R-U confidentiality maps, which is produced by applying the same method using different parameters. Aprior anonymization algorithm is designed to enforce km-anonymity using the full-subtree, global generalization model. COAT(Constrained-based Anonymization of Transactions) algorithm employs global generalization and suppression. PCTA(Privacy-constrained Clustering-based Transaction Anonymization) is a heuristic algorithm, which iteratively selects the privacy constraint that requires a small amount of generalization in order to be satisfied. Loukides [15] proposes the R-U map which constructing and demonstrating how these concepts can be used in assessing the disclosure risk and a tradeoff method offered by different anonymization solutions. In additions, several utility indexes are used, such as Normalized Certainty Penalty(NCP) [16], Utility Loss(UL), and Average Relative Error(ARE) [17]. NCP is expressed as the weighted average of the information loss of all generalized items, which are penalized based on the number of leaf-level descendants they have in the generalization hierarchy. UL quantifies information loss based on the size, weight and support of generalized items. ARE is a criterion that captures data utility, based on the accuracy of performing query answering on anonymized data. Salamatian [18] creates a practical framework allowing the design of privacy-preserving mechanisms that also maintain a certain level of data utility. Under the framework, data is distorted before it is published, based on a probabilistic privacy mapping, which is obtained by solving a convex optimization problem. The problem of privacy-utility tradeoff is how to minimize information leakage under a distortion constraint. The work in [19] also investigates the tradeoff between privacy and accuracy for the problem of differential privacy. Differential privacy requires that the answer to any query be “probabilistically indistinguishable” with or without a particular row in the database. Differential privacy is strong, which is not needed sometimes. In additions, differential privacy works under the assumptions that individuals are independent of each other. If the independence assumption is violated, differential privacy does not adequately limit inference about individual participation in the dataset. Ali [20] uses the results on maximal correlation and hypercontractivity of Markov processes, and provides utility-aware privacy mechanisms against inference attacks using partial statistical knowledge of the raw data prior distribution. Besides, Ali provides an upper-bound on the information leakage. Chakraborty [21] establishes the objective function with constraints, shows the quantitative relationship between risk and utility, and finds the optimal solution of the objective function by adjusting the parameters of the objective function. Reza [22] describes the problem of maximizing the privacy-utility tradeoff into a non-zero-sum Stackelberg game, then tries to find the optimal solution using linear programming and quadratic programming. Sankar [23] provides a theoretical analysis model to show relationship of source coding rate and privacy leakage versus distortion in the centralized environment. Nevertheless, this paper does not give the quantitative functions and detailed derivation of the relationship between the source rate R, privacy (equivocation) E and utility (distortion) D. Furthermore, the proposed method cannot be used directly in the distributed environment.

From the above analysis, existing solutions have the following disadvantages. Firstly, there are no formal metrics for utility and privacy and no formal privacy-utility tradeoff model in some works [7,9,11,12,15]. Secondly, some methods do not give the quantitative relationship between privacy and utility in the real sense [5,15,18,19,21]. Then, most of the existing solutions do not achieve the utility-privacy tradeoff region including all optimal

points [4-22], or the solution does not provide the quantitative tradeoff region [23]. Finally, no methods solve the privacy-utility tradeoff problem in the distributed environment with side information [4-23]. Motivated by the above analysis and disadvantages, our goal in the paper is to provide a distributed privacy-utility tradeoff method and find the tradeoff region using distributed lossy source coding theory with side information. In this paper we consider two application scenarios. The first is that the adversary has the background knowledge (or side information) and then combined with the released data to infer the privacy data. The second is that multiple data source stored separately need to be merged and analyzed with privacy constraints (means "distributed environment"). Our contributions are fourfold:

- We extend the model [23] to propose a distributed method using distributed lossy source coding theory with side information, which gives the quantitative relationship of each data source rate and privacy leakage versus each data source distortion and corresponding coding scheme for each source. In the distributed environment with side information, to achieve the utility-privacy tradeoff region and the RDL region, the coding scheme for each source must be $(n,2^{nR_i}, D_i)$ code with restricted source rate R_i and privacy leakage L_i satisfying Corollary 1.
- We provide four influence factors of source coding rate and privacy leakage degree, and show the quantitative analysis of the effect.
- We quantitatively give the privacy-utility tradeoff region and Rate-Distortion-Leakage (RDL) region, which includes all feasible results (coordinate points) of the distributed privacy-utility tradeoff problem.
- Using Public Health Care datasets, we compare the proposed method with the existing ones, and illustrate the advantage and usefulness of our method in five aspects. And also, this method could be applied to privacy issues in other application scenarios.

The remainder of this paper is organized as follows: Section 2 provides a distributed privacy-utility tradeoff method using distributed lossy source coding with side information, which contains theoretical analysis and formal derivation of this method. Simulation results and analysis are provided in Section 3 to justify the advantage of our method. Finally, Section 4 concludes this paper and outlines the future research.

2. Distributed privacy-utility tradeoff method with side information

In the distributed environment, each data participant (each data source) owns part of the dataset. All data participants send data to the data center. Each data participant must take efforts to keep its source data private from both other participants. In addition, each participant's dataset should be distorted separately. At the data center, the whole dataset will be reconstructed by the user, and also the user may have side information about the dataset. In order to solve the problem in the distributed environment with side information, this paper extends the centralized utility-privacy tradeoff model using distributed lossy source coding theory with side information.

2.1 Centralized utility-privacy tradeoff model without side information

The model provided by Sankar [23] encodes the public attributes to hide the private ones, which are correlated with public attributes. All of these attributes are stored in the centralized environment. Then, Sankar provided the concepts of data utility, data privacy and privacy leakage and quantified them. In addition, the relationships of the rate and privacy leakage versus distortion are shown to quantify the privacy-utility tradeoff.

In this model, a database d is a table with n entries (rows) and K public attributes (columns, denoted by X), in which private attributes denoted by Y are deleted before coding. The privacy-utility problem modeled by source coding theory with single source is shown in **Fig. 1**, in which X^n denotes public attributes(non-sensitive) to be distorted by source encoding, $\hat{X}^n$ denotes the revealed attributes, and R denotes the source coding rate.

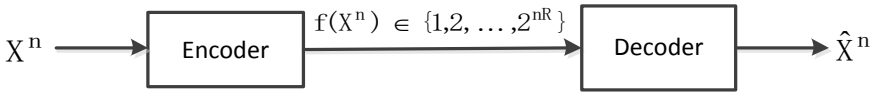

Fig. 1. Single source coding theme without side information

By quantitatively defining two metrics named utility u, privacy e (or privacy leakage l) and corresponding bounds D, E (or L), utility is mapped to distortion and privacy to information uncertainty by entropy (or privacy leakage by mutual information).

So, the problem of finding the privacy-utility tradeoff region is to obtain the Rate-Distortion-Equivocation (RDE) region or Rate-Distortion-Leakage (RDL) region. RDE region (or RDL region) is the set of all feasible tuples (R, D, E) (or (R, D, L)) for which there exists encoder *f* and decoder *g* with parameters (n, M, u, e) or (n, M, u, l) satisfying the constraints of each bound(D, E, L) and rate constraint ($M \le 2^{nR}$).

If a desired utility bound D is given, we need to obtain the set of all rate-equivocation tradeoff points (R, E) or the set of all rate-leakage tradeoff points (R, L). The set of all rate-equivocation tradeoff points (R, E) satisfies

$$\begin{cases} R \ge R(D) = I(X;\hat{X}) \\ E \le E(D) = H(Y \mid \hat{X}) \end{cases} \tag{1}$$

and the set of all rate-leakage tradeoff points (R, L) satisfies

$$\begin{cases} R \ge R(D) = I(X;\hat{X}) \\ L \ge L(D) = H(Y) - H(Y \mid \hat{X}) = I(Y;\hat{X}) \end{cases} \tag{2}$$

However, this method is not used directly in the distributed environment. So, this paper provides a distributed method using distributed lossy source coding with side information.

2.2 Distributed lossy source coding with side information

Distributed privacy-utility tradeoff is illustrated quantitatively by the relationship functions of the rate and privacy leakage versus distortion, which is based on the rate-distortion theory and distributed lossy source coding theory with side information(from other information sources).

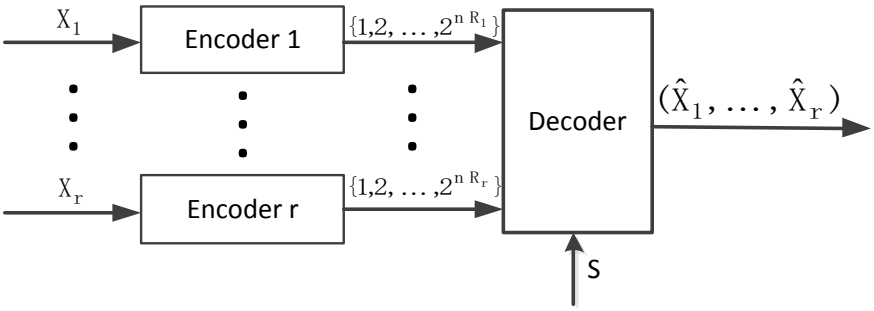

Fig. 2. Distributed source coding system with side information at the decoder

Distributed source coding (DSC) is an important problem in information theory and communication. DSC problems regard the compression of multiple independent information sources that do not communicate with each other. The distributed source coding system with side information is shown in **Fig. 2**. It is assumed that all links between encoders and decoders are noiseless and the data sources are noiseless ones.

In order to facilitate the description of distributed privacy-utility tradeoff model, a coding scheme with two sources and side information is shown in **Fig. 3**.

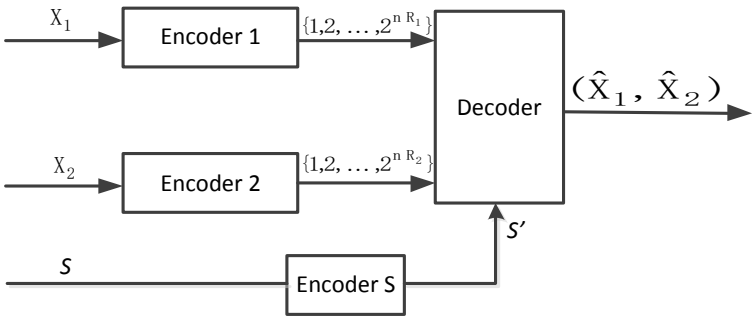

Fig. 3. Two source coding system with side information at the decoder

Let $\{X_1, X_2, S\}$ be real-valued independent and identically distributed (i.i.d.) sources and side information. Let $d(X_i, \hat{X}_i)$ be the difference distortion measure between X_i and $\hat{X}_i$ (i=1,2) .

Definition 1: A (n, R_1, R_2, D_1, D_2) coding scheme for the joint source (X_1, X_2) with side information S at the decoder consists of two encoding functions

$$\begin{cases} f_1 : X_1 \to X_1' = \{1,2,\ldots,M_1\} \\ f_2 : X_2 \to X_2' = \{1,2,\ldots,M_2\} \end{cases} \tag{3}$$

, satisfying $M_i \le 2^{nR_i}$ (i=1, 2) and two decoding functions

$$\begin{cases} g_1 : X_1' \times S' \to \hat{X}_1 \\ g_2 : X_2' \times S' \to \hat{X}_2 \end{cases} \tag{4}$$

, satisfying $Ed(X_1, g_1(f_1(X_1), S')) \le D_1$, $Ed(X_2, g_2(f_2(X_2), S')) \le D_2$ and the average distortion $\Delta_i \le D_i + \delta$ for i=1, 2.

2.3 Distributed privacy-utility tradeoff method

Based on the Definition 1, the goal of this paper is to find the RDE region or RDL region in distributed data sources with side information. So, we need to find the quantified relationship between source coding rate (R), privacy leakage (L) and distortion (D), which means $R(D_i)$ and $L(D_i)$ are quantified, and further to obtain the set of (R, D, E) or (R, D, L).

Corollary 1: The two source coding system with side information S is shown as Fig. 3 and the two sources and side information satisfy Gaussian distribution $X_i \sim N(0, \sigma_i^2)$ (i=1,2) and $S \sim N(0, \sigma_S^2)$ separately. Each source is coded with the data rate

$$R_i(D_i) \ge \frac{1}{2}\log\left[\frac{\sigma_i^2}{D_i}(1 - \rho_{iS}^2 + \rho_{iS}^2 \cdot 2^{-2R_S})\right] \tag{5}$$

, and the privacy leakage of each source at the decoder with the side information S is

$$L_i(D_i) \geq \frac{1}{2}\log\left(\frac{1}{1-\rho^2_{X_iY_i}+\rho^2_{X_iY_i}D_i/\sigma_i^2}\right)+\frac{1}{2}\log\left(\frac{1}{1-\rho^2_{SY_i}+\rho^2_{SY_i}D_S/\sigma_S^2}\right) \tag{6}$$

, in which R_S is the rate of side information S, D_S is the rate distortion of S, $\rho_{iS}=\frac{E(X_iS)}{\sigma_i\sigma_S}$, $\rho_{X_iY_i}=\frac{E(X_iY_i)}{\sigma_{X_i}\sigma_{Y_i}}$ and $\rho_{SY_i}=\frac{E(SY_i)}{\sigma_S\sigma_{Y_i}}$ (for i=1, 2) are correlation coefficients.

Proof of Corollary 1:

Because the side information is coded by the encoder S with the function $f_S: S\to S'=\{1,2,\ldots,2^{nR_S}\}$, we get $n(R_i+\delta)\geq \log M_i \geq h(X_i')\overset{(1)}{\geq} h(X_i'\mid S')\overset{(2)}{=} I(X_i;X_i'\mid S')\overset{(3)}{=} I(X_i;X_i',S')-I(X_i;S')\geq I(X_i;\hat{X}_i)-I(X_i;S')$.

(1) is obtained because conditioning S' reduces the entropy of X_i', (2) holds because of the fact that X_i' is the function of X_i, and (3) obeys the chain rule of mutual information.

Observe that $S'\to S\to X_i$ are Markov chains. Define

$$F_n(D_i)=\inf_{\hat{X}_i:\Delta_i\leq D_i}\frac{1}{n}I(X_i;\hat{X}_i) \text{ and } G_n(R_S)=\sup_{\substack{S':\frac{1}{n}I(S;S')\leq R_S\\ S'\to S\to X_i}}\frac{1}{n}I(X_i;S').$$

Then, we obtain $R_i+\delta\geq F_n(D_i+\delta)-G_n(R_S+\delta)$.So, the lower bound on $F_n(D_i)$ and the upper bound on $G_n(R_S)$ can be derived separately as in [24]:

$$F_n(D_i)\geq\frac{1}{2}\log\frac{\sigma_i^2}{D_i}, \text{ and } G_n(R_S)\leq\frac{1}{2}\log\left(\frac{1}{1-\rho^2_{iS}+\rho^2_{iS}\cdot 2^{-2R_S}}\right). \tag{7}$$

Finally, we get

$$R_i+\delta\geq\frac{1}{2}\log\frac{\sigma_i^2}{D_i+\delta}+\frac{1}{2}\log\left(1-\rho^2_{iS}+\rho^2_{iS}\cdot 2^{-2(R_S+\delta)}\right). \tag{8}$$

Letting $\delta\to 0$ on the above inequality, we have

$$R_i(D_i)\geq\frac{1}{2}\log\left[\frac{\sigma_i^2}{D_i}\left(1-\rho^2_{iS}+\rho^2_{iS}\cdot 2^{-2R_S}\right)\right], \text{ i=1, 2} \tag{9}$$

, in which ρ_{iS} is the correlation coefficient between X_i and side information S satisfying $\rho_{iS}=\frac{E(X_iX_S)}{\sigma_i\sigma_S}$. So (5) is proved.

As we know, $L\geq L(D_i)=H(Y_i)-H(Y_i\mid\hat{X}_iS')=I(Y_i;\hat{X}_iS')=I(Y_i;\hat{X}_i)+I(Y_i;S'\mid\hat{X}_i)$.

Because S and $\hat{X}_i$ are statistically independent, and also $S'\to S\to Y_i$ are Markov chains, we get $L(D_i)=I(Y_i;\hat{X}_i)+I(Y_i;S)$.

With the result in [25], we obtain $I(Y_i;\hat{X}_i)=\frac{1}{2}\log\left(\frac{1}{1-\rho^2_{X_iY_i}+\rho^2_{X_iY_i}D_i/\sigma_i^2}\right)$, and $I(Y_i;S)=\frac{1}{2}\log\left(\frac{1}{1-\rho^2_{SY_i}+\rho^2_{SY_i}D_S/\sigma_S^2}\right)$. So

$$L_i(D_i) \geq \frac{1}{2}\log\left(\frac{1}{1-\rho^2_{X_iY_i}+\rho^2_{X_iY_i}D_i/\sigma_i^2}\right)+\frac{1}{2}\log\left(\frac{1}{1-\rho^2_{SY_i}+\rho^2_{SY_i}D_S/\sigma_S^2}\right). \quad (10)$$

Corollary 1 is proved.

With the result in [23], we get $R_S=\frac{1}{2}\log\left(\frac{\sigma_S^2}{D_S}\right)$. (5) is changed into

$$R_i(D_i) \geq \frac{1}{2}\log\left[\frac{\sigma_i^2}{D_i}\left(1-\rho^2_{iS}+\rho^2_{iS}\cdot\frac{D_S}{\sigma_S^2}\right)\right]. \quad (11)$$

So, the minimal distortion rate of each source is the right part of (11). The set of all RDL tuples in (5) and (6) forms the RDL region, (5) and (6) specify the boundaries of this region. $R_i(D_i)$ given by (5) is the minimal rate and $L_i(D_i)$ given by (6) is the minimal privacy leakage for any choice of the distortion D of each source.

Therefore, in the distributed environment with side information, to achieve the utility-privacy tradeoff region and the RDL region, the coding scheme for each source must be $(n,2^{nR_i},D_i)$ code with restricted source rate R_i and privacy leakage L_i satisfying Corollary 1.

3. Result analysis

There are two main types of attributes in a database: categorical and numerical. In general, a database has these both two types of attributes. In order to simplify simulation process and analysis, this paper only consider the numerical attributes. In all databases, a medical database is a typical one containing lots of numerical data, which is used for privacy analysis.

The National Association of Health Data Organizations (NAHDO) reported that 37 states in the USA have legislative mandates to collect hospital level data and that 17 states have started collecting ambulatory care data from hospitals, physicians offices, clinics, and so forth. In Massachusetts, the Group Insurance Commission (GIC) is responsible for purchasing health insurance for state employees. GIC collected patientspecific data with nearly one hundred attributes for approximately 135,000 state employees and their families, which is named as Massachusetts Public Health Care Database. Among these attributes, there are Zip Code, Gender, Birth Date, Ethnicity, Weight, Blood Pressure, Diagnosis, Medication, Total Charge and etc. We only select the numberical attributes to form the new data set, such as Weight and Blood Pressure. Such numerical attributes are often assumed to be normally distributed or Gaussian distributed. We consider the numerical database with public attributes(Weight, Blood Pressure) expressed as X and private ones(Total Charge) expressed as Y, satisfying X and Y are jointly normally distributed with means μ_X, μ_Y and variances σ_X^2, σ_Y^2 respectively, and a correlation coefficient ρ_{XY}.

If we have 16 other states' Public Health Care Database, a distributed data sets with 16 data sources is formed. This paper considers two data sources and we suppose that each source is Gaussian distribution satisfying $X_i \sim N(0,\sigma_i^2)$ (i=1,2). If a user not only knows the revealed information $\hat{X}_i$, but has the side information X_S satisfying $X_S \sim N(0,\sigma_S^2)$, we need to know the quantitative results of R(D), L(D), privacy-utility region, RDL region, and by what factors these results are affected.

In this section, we use MATLAB software to do simulation tests and analyze the results from the following aspects.

Firstly, the following influencing factors of source rate and privacy leakage are analyzed.

ρ_{iS} : correlation coefficient between each data source i and side information source S,

$\rho_{X_iY_i}$: correlation coefficient between public attributes and private attributes of data source i,

σ_i^2 : statistical variance of data source i,

σ_S^2 : statistical variance of side information S.

Then, a quantitative representation of the privacy utility tradeoff region in a distributed scenario with side information is given.

Finally, the quantitative Rate-Distortion-Leakage (RDL) region in a distributed scenario with side information is given under the given parameters.

3.1 Influence of data distortion degree on the source rate with different ρ_{iS}

Using (11), determined values $\sigma_i^2 = \sigma_S^2 = 1$ and $D_S = 0.4$, the quantitative relationship of rate versus distortion with different correlation coefficient ρ_{iS} is shown in **Fig. 4**. **Fig. 4** shows that with the increase of distortion degree, the source rate decreases significantly and then slows down. In addition, the greater the correlation coefficient between data source X_i and side information S, the more obvious the influence of the source distortion increase on the source rate reduction. Under the same distortion (utility), the more relevant with the side information, the lower coding rate of the source to ensure a certain degree of privacy, which means that this source is coded with few symbols (less represented information source).

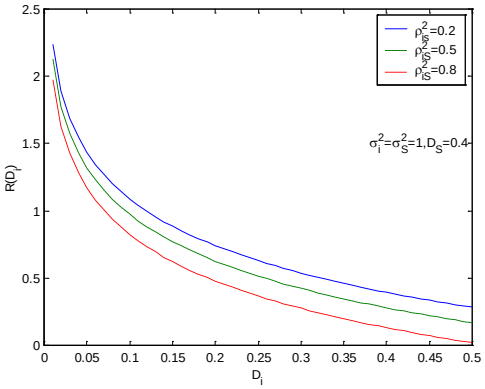

Fig. 4. Plot of R versus D with different ρ_{iS}

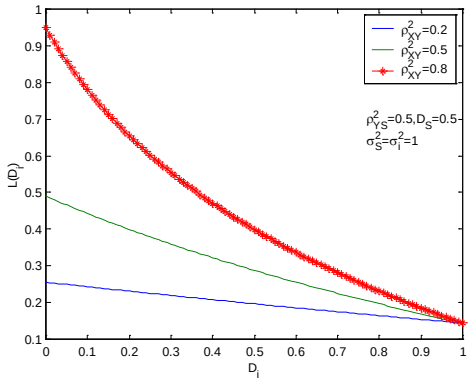

Fig. 5. Plot of L versus D with different $\rho_{X_iY_i}$

3.2 Influence of data distortion degree on the privacy leakage with different $\rho_{X_iY_i}$

Using (10), and determined values $\rho_{Y_iS}^2 = 0.5, D_S = 0.5, \sigma_S^2 = \sigma_i^2 = 1$, the quantitative relationship of privacy leakage versus distortion with different correlation coefficients between public attributes and private attributes is shown in **Fig. 5**. **Fig. 5** shows that with the increase of distortion degree, the privacy leakage decreases significantly and then slows down. In addition, the larger the correlation coefficient between public data X and private data Y, the more obvious the effect of the source distortion increase on the privacy leakage reduction.

Since privacy information Y is not released, but release information X. Therefore, if there is a strong correlation between these two information, it is possible to infer the privacy information Y using released information $\hat{X}$ through appropriate association analysis. That is to say, under the same privacy restrictions, the finer relevance between public data and private data of source, the finer perturbation of data source X is needed.

3.3 Influence of data source statistical variance σ_i^2 on the rate and privacy leakage

Using (10), (11), and determined values $\sigma_S^2 = 1, \rho_{iS}^2 = 0.6, D_S = 0.5, \rho_{X_iY_i}^2 = \rho_{Y_iS}^2 = 0.5$, the quantitative relationship of the source rate and privacy leakage versus distortion with different σ_i^2 is shown in **Fig. 6**. **Fig. 6** show that with the same distortion, the greater the variance of the data source, the higher the source rate and the privacy leakage. **Fig. 6** illustrates two points. The first is that the greater the variance of the data source's distribution, the smaller distortion is chosen to ensure more data utility. The second is that under the same distortion (utility), the greater the variance of the data source's distribution, the higher the source rate, which means that this source is coded with more symbols (more represented information source) while the privacy is less disclosed.

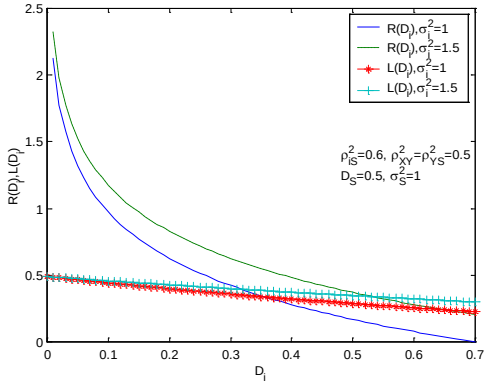

Fig. 6. Plot of R and L versus D with different σ_i^2

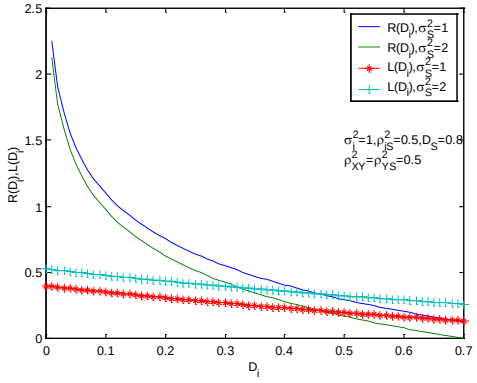

Fig. 7. Plot of R and L versus D with different σ_S^2

3.4 Influence of side information statistical variance σ_S^2 on the rate and privacy leakage

Using (10), (11), and determined values $\sigma_i^2 = 1, \rho_{iS}^2 = 0.5, D_S = 0.8, \rho_{X_iY_i}^2 = \rho_{Y_iS}^2 = 0.5$, the quantitative relationship of the source rate and privacy leakage versus distortion with different σ_S^2 is shown in **Fig. 7**. **Fig. 7** shows that the greater the variance of the side information, the greater effect of the source distortion increase on the source rate reduction and the greater privacy leakage level. **Fig. 7** illustrates two points. The first one is that under the same privacy restriction, the smaller the variance of the side information's distribution, the less distortion of data source is chosen to ensure more data utility. The second one is that under the same distortion (utility), the greater the variance of the side information's distribution, the lower the source rate, which means that this source is coded with less symbols(less represented information source) to ensure certain privacy.

3.5 Analysis of the privacy-utility region

Fig. 8 shows three regions, which are strong privacy region, privacy-utility tradeoff region, and strong utility region. Strong privacy region has very poor utility, and strong utility region has bad privacy. So, this paper quantitatively gives the privacy-utility tradeoff region using distributed source coding theory with side information.

3.6 Analysis of the Rate-Distortion-Leakage (RDL) region

Using (10), (11), and determined values $\rho_{iS}^2 = \rho_{X_iY_i}^2 = \rho_{Y_iS}^2 = 0.64$, $D_S = 0, \sigma_i^2 = \sigma_S^2 = 1$, the Rate-Distortion-Leakage (RDL) region described in section 3.3 is shown as **Fig. 9**. In the **Fig. 9**, the RDL region is the set of all RDL tuples satisfying (5) and (6). In this region, the rate boundary is minimal rate and the privacy leakage boundary is the minimal privacy leakage defining by (5) and (6) seperately.

Seperately given the statistical distributions of two sources and side information, we could get the distortion boundary Dmax, rate boundary, privacy leakage boundary and RDL region. All points that meet the privacy-utility tradeoff need are in the RDL region.

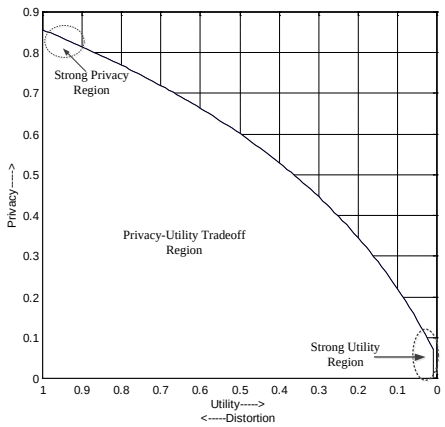

Fig. 8. Region of privacy-utility tradeoff

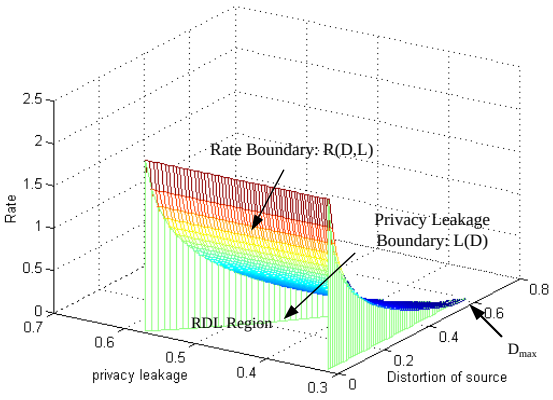

Fig. 9. Rate-Distortion-Leakage (RDL) region

3.7 Comparisons

The provided method in this paper is compared with the existing literatures from the following aspects:

(1) whether or not to propose a formal privacy-utility tradeoff model,

(2) whether or not to consider the impact of side information on the privacy disclosure and utility of revealed data,

(3) whether or not to qualitatively give the privacy-utility tradeoff region,

(4) whether or not to consider the privacy-utility tradeoff problem in distributed application scenarios with side information,

(5) whether or not to quantitatively give all feasible results(coordinate points) of the distributed privacy-utility tradeoff problem.

The comparision results are listed in **Table 1**. **Table 1** illustrates the advantages of our method in the above five aspects, compared with existing methods.

Table 1. Comparisons of different methods

methods / aspects	[4-6]	[7、9]	[8、10]	[11-12]	[13-14]	[15]	[16-22]	[23]	Our method
(1)	Y	N	Y	N	Y	N	Y	Y	Y
(2)	N	N	N	N	N	N	N	Y	Y
(3)	N	N	N	N	N	N	N	Y	Y
(4)	N	N	N	N	N	N	N	N	Y
(5)	N	N	N	N	N	N	N	N	Y

Remarks: Yes(Y), No(N)

4. Conclusion

In order to balance the privacy and utility of data in distributed environment and adversary having background knowledge, this paper proposes a privacy-utility tradeoff model using distributed source coding theory with side information. In this model, each variable representing the public data of different source is coded independently, and the quantitative relationship of each source rate and privacy leakage versus distortion is presented. Besides, the result gives the quantified privacy-utility tradeoff region and RDL region. In this paper, a theoretical analysis is made only on the distributed scene with two sources, which meet the assumption of independent and identically distributed (i.i.d.). In the future, this model is extended to multiple sources with non-i.i.d. dataset, and the privacy-utility tradeoff region and RDL region will be given. More real datasets will be used in other application scenarios (such as Location Based Services and Location Information Verification [26]) to valid the provided method in this paper.

References

[1] Huang X, Liu J, Han Z, et al, "A new anonymity model for privacy-preserving data publishing," *China Communications*, vol. 11, no. 9, pp. 47-59, 2014. Article (CrossRef Link)

[2] Alvim M S, Andrés M E, Chatzikokolakis K, et al, "Differential Privacy: on the trade-off between Utility and Information Leakage," in *Proc. of International Conference on Formal Aspects of Security and Trust*, vol. 7140, pp. 39-54, 2011. Article (CrossRef Link)

[3] V. Rastogi and S. Nath, "Differentially private aggregation of distributed time-series with transformation and encryption," in *Proc. of ACM SIGMOD International Conference on Management of Data*, Indianapolis, Indiana, USA, pp. 735-746, 2010. Article (CrossRef Link)

[4] Makhdoumi A, Fawaz N, "Privacy-utility tradeoff under statistical uncertainty," in *Proc. of Allerton Conference on Communication, Control, and Computing*, pp.1627-1634, 2013. Article (CrossRef Link)

[5] Krause A, Horvitz E, "A utility-theoretic approach to privacy in online services," *Journal of Artificial Intelligence Research*, vol. 39, no. 1, pp. 633-662, 2010. Article (CrossRef Link)

[6] Xiong P, Zhu T, "An Anonymization Method Based on Tradeoff between Utility and Privacy for Data Publishing," in *Proc. of International Conference on Management of E-Commerce and E-Government*, pp.72-78, 2012. Article (CrossRef Link)

[7] Loukides G, Shao J, "Data utility and privacy protection trade-off in k-anonymisation," in *Proc. of International Workshop on Privacy and Anonymity in Information Society*, Nantes, France, pp.36-45, March, 2008. Article (CrossRef Link)

[8] Li T, Li N, "On the tradeoff between privacy and utility in data publishing," in *Proc. of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.517-526, 2009. Article (CrossRef Link)

[9] Khokhar R H, Chen R, Fung B C M, et al, "Quantifying the costs and benefits of privacy-preserving health data publishing," *Journal of Biomedical Informatics*, vol. 50, no.8, pp. 107-121, 2014. Article (CrossRef Link)
[10] Andreas Krause, Eric Horvitz, "A Utility-Theoretic Approach to Privacy and Personalization," in *Proc. of Twenty-Third Conference on Artificial Intelligence*, pp. 1181-1188, July 2008. Article (CrossRef Link)
[11] Loukides G, Gkoulalas-Divanis A, Shao J, "On balancing disclosure risk and data utility in transaction data sharing using R-U confidentiality map," *Joint UNECE/Eurostat work session on statistical data confidentiality*, pp. 19, 2011. Article (CrossRef Link)
[12] M. Terrovitis, N. Mamoulis, and P. Kalnis, "Privacy-preserving anonymization of set-valued data," *PVLDB*, vol. 1, no.1, pp. 115-125, 2008. Article (CrossRef Link)
[13] G. Loukides, A. Gkoulalas-Divanis, and B. Malin, "COAT: Constraint-based anonymization of transactions," *KAIS*, vol. 28, no. 2, pp. 251-282, 2011. Article (CrossRef Link)
[14] Gkoulalas-Divanis A, Loukides G, "PCTA: privacy-constrained clustering-based transaction data anonymization," in *Proc. of International Workshop on Privacy and Anonymity in Information Society*, pp.1-10, March, 2011. Article (CrossRef Link)
[15] G. Loukides, A. Gkoulalas-Divanis, and J. Shao, "Assessing disclosure risk and data utility trade-off in transaction data anonymization," *International Journal of Software and Information*, vol. 6, no. 3, pp.399-417, 2012. Article (CrossRef Link)
[16] Terrovitis M, Mamoulis N, Kalnis P, "Privacy-Preserving Anonymization Of Set-Valued Data," in *Proc. of the Vldb Endowment*, 115-125, 2008. Article (CrossRef Link)
[17] G. Loukides, A. Gkoulalas-Divanis, and B. Malin, "COAT: COnstraint-based anonymization of transactions," *Knowledge and Information Systems*, vol. 28, no. 2, pp. 251-282, 2011. Article (CrossRef Link)
[18] Salamatian S., Zhang A., du Pin Calmon F, et al, "Managing your Private and Public Data: Bringing down Inference Attacks against your Privacy," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 7, pp. 1240-1255, 2015. Article (CrossRef Link)
[19] Ghosh A, Roughgarden T, Sundararajan M, "Universally Utility-Maximizing Privacy Mechanisms," *Siam Journal on Computing*, vol. 41, no. 6, pp. 351-360, 2009. Article (CrossRef Link)
[20] Ali Makhdoumi, Nadia Fawaz, "Privacy-Utility Tradeoff under Statistical Uncertainty," in *Proc. of Fifty-first Annual Allerton Conference*, pp. 1627-1634, October 2-3, 2013. Article (CrossRef Link)
[21] Chakraborty S, Charbiwala Z, Choi H, et al, "Balancing behavioral privacy and information utility in sensory data flows," *Pervasive and Mobile Computing*, vol. 8, no. 3, pp. 331-345, 2012. Article (CrossRef Link)
[22] Reza Shokri, "Privacy Games: Optimal Protection Mechanism Design for Bayesian and Differential Privacy," Submitted on 14 Feb 2014. Article (CrossRef Link)
[23] Sankar L, Rajagopalan S R, Poor H V, "A theory of utility and privacy of data sources," in *Proc. of 2010 IEEE International Symposium on Information Theory(ISIT)*, pp. 2642-2646, June 13-18, 2010. Article (CrossRef Link)
[24] Y Oohama, "Gaussian multiterminal source coding," *IEEE Transactions on Information Theory*, vol. 43, no. 6, pp. 1912-1923, November, 1997. Article (CrossRef Link)
[25] Wyner A D, "The rate-distortion function for source coding with side information at the decoder-II: General sources," *Probability Theory & Related Fields*, vol. 38, no. 1, pp. 60-80, 1978. Article (CrossRef Link)
[26] Sheet D, Kaiwartya O, Abdullah A, et al. "Location Information Verification using Transferable Belief Model for Geographic Routing in VANETs," *IET Intelligent Transport Systems*, 2016. Article (CrossRef Link)

Yonghao Gu received his Ph.D.(2007) from Beijing University of Posts and Telecommunications, China. Currently, he is a lecturer in the Beijing Key Laboratory of Intelligent Telecommunications Software and Multimedia, School of Computer, Beijing University of Posts and Telecommunications, China. His main research interests are network security and privacy preservation.

Yongfei Wang received the bachelor degree from Hebei North University, China, in 2012. Currently, he is a graduate student at Beijing University of Posts and Telecommunications, China. His main research interests are network security and privacy preservation.

Zhen Yang received the Ph.D. degree from Beijing University of Posts and Telecommunications, China, in 2007. He is an associate professor in the Beijing Key Laboratory Intelligent Telecommunications Software and Multimedia, School of Computer, Beijing University of Posts and Telecommunications, China. His current research focuses on cloud computing, and fault tolerant computing.

Yimu Gao received the bachelor degree from Beijing University of Posts and Telecommunications, China, in 2015. Currently, he is a graduate student at University of Delaware, USA. His main research interest is network security.