doi:10.11959/j.issn.1000-436x.2015326

基于多变量信源编码的隐私效用均衡方法

谷勇浩[1]，林九川[2]
（1. 北京邮电大学 计算机学院 智能通信软件与多媒体北京市重点实验室，北京 100876；2. 公安部第三研究所，上海 201204）

摘 要：在大数据时代，数据提供者需要保证自身隐私，数据分析者要挖掘数据潜在价值，寻找数据隐私性与数据可用性间的均衡关系成为研究热点。现有方法多数关注隐私保护方法本身，而忽略了隐私保护方法对数据可用性的影响。在对隐私效用均衡方法研究现状分析的基础上，针对数据集中不同公开信息对隐私保护需求不同的问题，提出基于多变量信源编码的隐私效用均衡方法，并给出隐私效用均衡区域。分析表明，隐私信息与公开信息的关联度越大，对公开信息扰动程度的增加会显著提高隐私保护效果。同时，方差较大的变量对应的公开信息，可选择较小的扰动，确保公开信息可用性较大。

关键词：隐私保护；隐私效用均衡；信源编码；率失真
中图分类号：TP309 **文献标识码**：A

Privacy-utility tradeoff method using multi-variable source coding

GU Yong-hao[1], LIN Jiu-chuan[2]
(1. Beijing Key Laboratory of Intelligent Telecommunication Software and Multimedia, School of Computer Science, Beijing University of Posts and Telecommunications, Beijing 100876, China;
2. The Third Research Institute of Ministry of Public Security, Shanghai 201204, China)

Abstract: In the age of big data, data providers need to ensure their privacy, while data analysts need to mine the value of data. So, how to find the privacy-utility tradeoff has become a research hotspot. Current works mostly focus on privacy preserving methods, ignoring the data utility. Based on the current research of privacy utility equilibrium methods, a privacy-utility tradeoff method using multi-variable source coding was proposed to solve the problem that different public datasets in the same database have different privacy requirements. Two results are obtained by simulations. The first result is that the greater the association degree between the private information and public information, the increase of the distortion degree of public information will significantly improve the effect of privacy preservation. The second result is that public information with larger variance should be less distorted to ensure more utility.

Key words: privacy preservation; privacy-utility tradeoff; source coding; rate distortion

1 引言

随着信息技术的快速发展，感知设备的普及，数据查询分析业务的广泛使用，个人数据信息收集的种类和数量呈指数增长。由于知识决策、信息共享、科学研究等工作的需要，数据拥有者、采集者（个人、企业、政府等）需要将数据发布共享，从而得到基于数据分析的服务。但是，被采集共享的信息中可能涉及到个体隐私，如果将收集到的原始数据或者经过分析后的统计数据直接发布，个人敏感信息可能泄露。隐私保护是数据分发、共享、分析、挖掘过程中必须要解决的问题[1~4]。

目前，多数研究者考虑的是信息泄露带来的隐私问题，而忽略了隐私保护导致可用性降低的问

收稿日期：2015-03-24；**修回日期**：2015-10-01
基金项目：国家自然科学基金资助项目（61173017）；工信部通信软科学基金资助项目（2014-R-42，2015-R-29）；信息网络安全公安部重点实验室开放课题基金资助项目（C14613）
Foundation Items: The National Natural Science Foundation of China(61173017)；Communication Soft Science Foundation of Ministry of Industry and Information(2014-R-42，2015-R-29)；Key Lab of Information Network Security Foundation of Ministry of Public Security (C14613)

题。如果把数据的采集、汇聚、融合、传输、决策与控制等过程称为数据的生命周期，隐私保护的目的是在数据生命周期中不泄露数据提供者（或所有者）的隐私信息。数据可用性的目的是在数据生命周期中数据使用者能正确获得数据并进行数据分析、挖掘、统计和决策。研究表明，保护隐私性和保证可用性是一对相互排斥的需求。不公开任何数据信息可以获得最佳的数据隐私，但牺牲了数据的可用性；相反，全部公开原始数据可使数据的使用效果最佳，但这样没有任何隐私可言[5]。因此，如何寻找数据隐私性与数据可用性之间的关系，并给出获得合理的隐私效用均衡方法成为研究热点。

本文在对隐私效用均衡方法研究现状分析的基础上，对数据集中不同数据对隐私保护需求不同的问题[6]，提出基于多变量信源编码的隐私效用均衡方法。

2　相关工作

差分隐私技术[7]在数据集个体隐私的定义中引入攻击者能力的概念，即攻击者能准确判断某条数据是否属于（或不属于）在该数据集。差分隐私保护方法的最大优点是，在数据失真技术中加入的噪声量与数据集大小无关。虽然差分隐私保护技术不受限于特定应用，可用于任何具有统计特征的数据集，但是，差分隐私保护技术对于隐私匿名的定义过于严格，而这种严格的隐私匿名方法在很多应用场景中并不需要。同时，差分隐私保护技术只适用于处理数值型数据，还存在局限性[8]，不适合处理非数值型数据。文献[9]将差分隐私保护技术与传统的统计假设检验建模方法相结合，应用于临床实验数据挖掘，并且在如何调节样本数量以获取统计效率和隐私保护等级的均衡问题上总结了一些规律。但是，多数基于差分隐私的隐私保护技术导致数据可用性损失显著[8]，尤其是在一些特定应用场景下，可用性损失更为严重[10]。

文献[10]指出，数据隐私性和数据可用性是分别针对数据集个体和数据集整体而言的。在此基础上，采用现代投资理论中风险收益率的概念来描述隐私性和可用性间的均衡关系，并且采用隐私损失和可用性损失指标分别量化隐私性和可用性。此外，文中采用的量化及均衡方法是基于特定数据分布特性以及所有敏感属性值同等重要的假设条件下提出的。这种假设会随着数据分布特性的变化以及不同应用场景对敏感属性需求的不同而变化，对分析结果产生影响。

文献[11]以特定应用场景下隐私及可用性需求为基础，寻找匿名数据发布方法。该文对隐私性和可用性需求以约束的形式建模，同时研究如何以形式化方法定义约束，提出基于约束的匿名方法，并通过实验验证方法的有效性和灵活性。由于寻找特定约束条件下的最优匿名方法是一个 NP 难题[12]，所以该文献所提方案在实际应用中不能直接使用，需要寻找替代的近似最优匿名方法，同时满足特定约束条件和信息损失最小的需求。

文献[12]采用 Risk-Utility 二维关系图描述隐私泄露风险和数据可用性之间的关系，然后通过不同的可用性量化指标（normalized certainty penalty[13], utility loss, average relative error[11]）比较多种匿名隐私保护方法在隐私效用均衡方面的差异。文献[14]采用隐私损失和可用性损失 2 个指标，通过 f 散度量化指标后对 k 匿名、l 多样以及 t 近似 3 种匿名保护方法在参数(k、l、t)分别取不同值时进行比较，得到 3 种方法的隐私损失程度及可用性损失程度对比关系的走势。但是，这些方法还存在不足，如未考虑潜在攻击者可能具有的背景知识、未充分考虑数据分析需求对数据可用性量化的影响等。

有些文献采用信息论中的概念（如信息熵、条件熵、期望等）对隐私性和可用性量化，从而寻找可用性和隐私性之间的量化关系。如文献[15, 16]分别采用 Renyi 熵和 Shannon 熵量化隐私，文献[16]将具有可用性约束的隐私保护最优化问题归结成凸规划问题。

文献[6]所提模型对公开信息(能通过关联分析推断出隐私信息的其他信息)进行编解码，隐藏用户隐私信息，给出编码率与隐私性、可用性间的量化关系，实现数据可用性和隐私性之间的均衡。该模型对敏感信息（表示为信源）采用统一编码，同时假设表示信源的变量具有独立同分布的特性。

3　基于多变量信源编码的隐私效用均衡方法

文献[6]采用信源统一编码方式，实现发布信息的隐私效用均衡，所提模型未考虑信源中不同变量表示的公开信息对隐私保护需求程度的差异。本文在对该模型分析的基础上，提出基于多变量信源编码的隐私效用均衡方法。

3.1　基于信源统一编码的隐私效用均衡方法

文献[6]在给定数据源中用变量 X、Y 和 $\hat{X}$ 分别

表示公开信息、隐私信息和经过编解码后还原的公开信息（经过隐私保护处理后的发布信息），编码器映射关系为

$$F_E : X \to X' \tag{1}$$

其中，X'为编码映射后序列，序列元素个数为 M（$M=2^{nR}$），n 为数据源中数据集个数，R 为编码率。解码器映射关系为

$$F_D : X' \to \hat{X} \tag{2}$$

文献[6]中分别给出可用性、隐私性和隐私泄露程度的定义和度量方法如下。

可用性为经过图 1 所示的编解码前后数据集间的失真程度。采用距离量化数据集的可用性 u，同时保证数据集的失真程度具有上界 D，即满足

$$u = E\left[\frac{1}{n}\sum_{i=1}^{n}\rho(X_i, \hat{X}_i)\right] \leqslant D \tag{3}$$

其中，$\rho()$ 为失真函数（如汉明失真函数、欧拉距离函数等）。

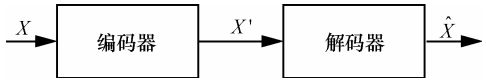

图 1 基于信源统一编解码映射关系

隐私性为在得到发布信息 $\hat{X}$ 的条件下，推断出隐私信息 Y 的不确定程度。隐私性 e 采用条件熵度量，同时保证隐私程度具有下界 E，即满足

$$e = H(Y \mid \hat{X}) \geqslant E \tag{4}$$

隐私泄露程度为在得到发布信息 $\hat{X}$ 的条件下，推断出隐私信息 Y 的不确定程度的缩减量。隐私泄露程度 l 采用互信息度量，同时保证隐私泄露程度具有上界 L，即满足

$$l = I(Y; \hat{X}) \leqslant L \tag{5}$$

在上述信源编解码映射关系（如图 1 所示）下，可用性和隐私性（或隐私泄露程度）分别采用失真度和条件熵（或互信息）度量。由此，给出隐私效用均衡区域的定义如下。

隐私效用均衡区域：所有（D, E）二元组构成的点集，且满足式（1）、式（2）对应的编解码方案，该方案中的参数（n, M, u, e）或（n, M, u, l）需满足式（3）和式（4）（或式（5））的约束。

从上述定义可以看出，寻找隐私效用均衡区域，就是要寻找具有特定参数的最优编解码方案，得到所有（D, E）点集。由于 M 满足 $M=2^{nR}$，因此对发布信源信息的隐私效用均衡问题的求解变为：在给定扰动程度 D 的条件下，求解率失真 $R(D)$ 和隐私度量 $E(D)$，从而获得隐私效用均衡区域为所有（R, D, E）三元组构成的点集，满足条件

$$R \geqslant R(D) = I(X; \hat{X}) \tag{6}$$

$$E \leqslant E(D) = H(Y \mid \hat{X}) \tag{7}$$

如果用隐私泄露程度 L 代替隐私度量 E，则信源隐私效用均衡问题的求解变为：在给定扰动程度 D 的条件下，求解率失真 $R(D)$和隐私泄露度量 $L(D)$，从而获得隐私效用均衡区域为所有（R, D, L）三元组构成的点集，满足条件

$$R \geqslant R(D) = I(X; \hat{X}) \tag{8}$$

$$L \geqslant L(D) = H(Y) - H(Y \mid \hat{X}) = I(Y; \hat{X}) \tag{9}$$

3.2 基于多变量信源编码的隐私效用均衡方法概述

3.1 节所述方法是针对信源采用统一编码及失真度量的情形。如果信源表示的公开信息集中不同信息具有不同的统计分布特性，而且这些公开信息对隐私保护的需求存在差异时，基于统一编码的隐私效用均衡方法将失效。为解决信源表示的公开信息集中不同信息统计特性及隐私保护需求差异的问题，本文提出基于多变量信源编码的隐私效用均衡方法。

假设信源是由 n 个统计特性不同的数据集构成，这些数据集分别用变量 X_i (i=1,2,⋯,n)表示，且满足 $X_i \sim \mathrm{N}(0, \sigma_i^2)$。由于这些变量代表的公开信息对隐私信息 Y 的泄露具有不同的影响程度，需要对各自变量独立编码，提出多变量信源编码方法。多变量信源编解码映射关系如图 2 所示。

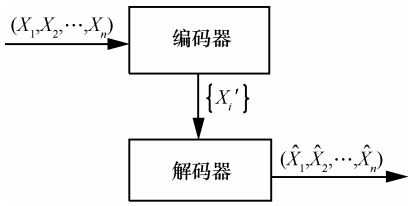

图 2 基于多变量信源编解码映射关系

由 3.1 节可知，信源隐私效用均衡问题的求解如下。在给定扰动程度 D 的条件下，求解率失真 $R(D)$和隐私泄露度量 $L(D)$，从而获得隐私效用均衡的最佳区域 (R, D, L)。那么，基于多变量信源编码的隐私效用均衡方法，即在给定信源编码速率 R 的条件下，如何给不同变量分配编码速率 R_i，使信源总体失真和隐私泄露程度最小。

由于

$$\begin{aligned} I(X^n;\hat{X}^n) &= h(X^n) - h(X^n \mid \hat{X}^n) \\ &= \sum_{i=1}^{n} h(X_i) - \sum_{i=1}^{n} h(X_i \mid X^{i-1}, \hat{X}^n) \\ &\geqslant \sum_{i=1}^{n} h(X_i) - \sum_{i=1}^{n} h(X_i \mid \hat{X}_i) \\ &= \sum_{i=1}^{n} I(X_i;\hat{X}_i) \\ &\geqslant \sum_{i=1}^{n} R(D_i) \\ &= \sum_{i=1}^{n} \left(\frac{1}{2}\log\frac{\sigma_i^2}{D_i} \right)^{+} \end{aligned} \tag{10}$$

因此，求解率失真函数 $R(D)$ 的问题则变为

$$\begin{aligned} R(D) &= \min I(X^n;\hat{X}^n) \\ &= \min_{\sum D_i = D} \left(\sum_{i=1}^{n} \max\left\{ \frac{1}{2}\log\frac{\sigma_i^2}{D_i}, 0 \right\} \right) \end{aligned} \tag{11}$$

然后，确定 n 个变量 X_i (i=1,2,⋯,n)的最优失真分配方案。利用拉格朗日乘子法建立函数

$$J(D) = \sum_{i=1}^{n} \frac{1}{2}\log\frac{\sigma_i^2}{D_i} + \lambda \sum_{i=1}^{n} D_i \tag{12}$$

对式（12）求 D_i 的偏导，并且令其等于 0，得到

$$\begin{aligned} \frac{\partial J}{\partial D_i} &= -\frac{1}{2}\frac{1}{D_i} + \lambda = 0 \\ \Rightarrow D_i &= \frac{1}{2\lambda} = \lambda' \end{aligned} \tag{13}$$

由式（13）可知，最优的码率分配方法是让各变量具有相等的失真 λ'。但是，当总体失真 D 增大时，各变量取相等的失真 λ' 无法满足，运用 Kuhn-Tucker 条件得到如下结论。

设 $X_i \sim \mathrm{N}(0,\sigma_i^2)$（$i$=1,2,⋯,$n$）是独立不同分布的高斯随机变量，其中 $D_i \in [0,\sigma_i^2]$，各变量对应的率失真值为

$$R(D_i) = \frac{1}{2}\log\left(\frac{\sigma_i^2}{D_i}\right) \tag{14}$$

该信源的率失真函数为

$$R(D) = \sum_{i=1}^{n} \frac{1}{2}\log\left(\frac{\sigma_i^2}{D_i}\right) \tag{15}$$

其中，$D_i = \begin{cases} \lambda, & \lambda < \sigma_i^2 \\ \sigma_i^2, & \lambda \geqslant \sigma_i^2 \end{cases}$，选择适当的 λ，满足 $\sum_{i=1}^{n} D_i = D$。

由于 X_i 和 Y 分别满足 $X_i \sim \mathrm{N}(0,\sigma_i^2)$ 和 $Y \sim \mathrm{N}(0,\sigma_Y^2)$，因此 X_i 与 Y 的相关系数为

$$\begin{aligned} \rho_{X_iY} &= \frac{Cov(X_i,Y)}{\sigma_i\sigma_Y} = \frac{E(X_iY) - E(X_i)E(Y)}{\sigma_i\sigma_Y} \\ &= \frac{E(X_iY)}{\sigma_i\sigma_Y} \end{aligned} \tag{16}$$

由文献[6]可知，X、Y 和 $\hat{X}$ 之间具有马氏链关系 $Y - X - \hat{X}$，因此概率分布函数满足

$$\begin{aligned} p(Y,X,\hat{X}) &= p(Y)p(X \mid Y)p(\hat{X} \mid X,Y) \\ &= p(Y)p(X \mid Y)p(\hat{X} \mid X) \\ &= p(Y,X)p(\hat{X} \mid X) \end{aligned} \tag{17}$$

对每个变量进行编解码并还原发布后，由式（9）、式（14）、式（16）和式（17）得到隐私泄露程度为

$$L(D_i) = \frac{1}{2}\log\left(\frac{1}{\left[(1-\rho_{X_iY}^2) + \frac{\rho_{X_iY}^2 D_i}{\sigma_i^2} \right]} \right) \tag{18}$$

信源的总体隐私泄露程度为

$$L(D) = \sum_{i=1}^{n} \frac{1}{2}\log\left(\frac{1}{\left[(1-\rho_{X_iY}^2) + \frac{\rho_{X_iY}^2 D_i}{\sigma_i^2} \right]} \right) \tag{19}$$

因此，由式（15）、式（19）可获得隐私效用均衡的最佳区域为 (R, D, L)。

4　结果与分析

4.1　扰动程度变化对码率及隐私泄露程度影响的分析

假设信源是均值为 0，方差为 σ_X^2 的高斯信源，由式（14）、式（18）得到

$$R(D) = \frac{1}{2}\log\left(\frac{\sigma_X^2}{D}\right) \tag{20}$$

$$L(D) = \frac{1}{2}\log\left(\frac{1}{\left[(1-\rho_{XY}^2) + \frac{\rho_{XY}^2 D}{\sigma_X^2} \right]} \right) \tag{21}$$

其中，$D \in [0,\sigma_X^2]$，ρ_{XY} 为 X 与 Y 的相关系数。在给定 $\sigma_X^2 = 1$ 的情况下，率失真 $R(D)$ 和隐私泄露度量 $L(D)$ 各自与扰动程度 D 之间的关系走势如图 3 所示。

从图 3 中看出，随着扰动程度 D 的增加，码率 R 和隐私泄露程度 L 明显减小，然后趋缓。同时，

如果 X 与 Y 的相关系数越大，扰动程度 D 的增加对隐私泄露程度 L 降低的作用越明显。说明隐私信息 Y 与公开信息 X 的关联度越大，对公开信息扰动程度的增加会显著提高隐私保护效果。

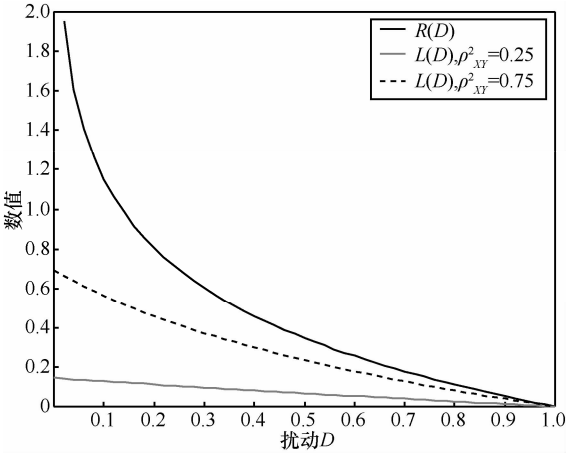

图 3 信源统一编码下的率失真和隐私泄露走势

4.2 多变量编码下不同统计特性的变量如何选择扰动程度的分析

为解决公开信息集中不同信息统计特性及隐私保护需求的差异，本文提出多变量信源编码。

假设信源是由 n 个统计特性不同的高斯随机变量构成，满足 $X_i \sim \mathrm{N}(0,\sigma_i^2)$ (i=1,2,⋯,n)，其中，$D_i \in [0,\sigma_i^2]$，由式（14）、式（18）得到信源各分量 X_i 的率失真值 $R(D_i)$ 和隐私泄露度量 $L(D_i)$ 各自与扰动程度 D_i 之间的关系走势如图 4 所示。

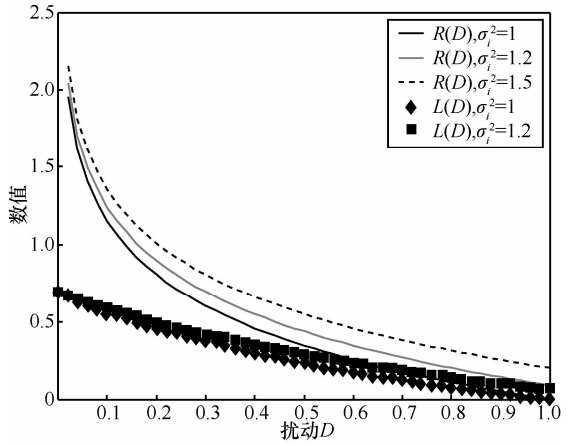

图 4 多变量信源编码下的率失真和隐私泄露走势

从图 4 中看出，数据源中分布方差越小的变量，$R(D)$和 $L(D)$随扰动 D 的变化越显著。在给定扰动 D 的条件下，分布方差越大的变量，$R(D)$和 $L(D)$越大。因此，对具有不同分布数据变量的数据源来说，统计方差越大的变量，可选择较小的扰动，确保信息可用性较大。

4.3 基于多变量编码的隐私效用均衡区域的分析

由式（20）得到曲线 $L(D)$，$L(D)$与 2 个坐标轴围成如图 5 所示的区域，该区域包含 D_1、D_2、D_3 3 个子区域。其中，D_1 为近似零扰动区域，该区域为可用性较强且无隐私保护的区域。D_2 为基于多变量信源编码生成的隐私效用均衡区域。D_3 为近似无隐私泄露区域，该区域的隐私保护程度较高但数据可用性很差。

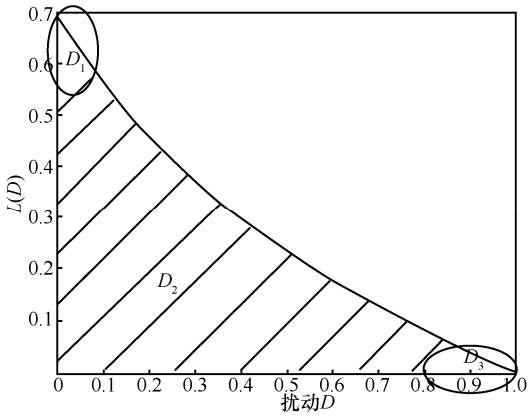

图 5 基于多变量信源编码的隐私效用均衡区域

5 结束语

为解决数据集中不同属性数据对隐私保护需求不同的问题，本文提出一种基于多变量信源编码的隐私效用均衡方法。该方法对信源中具有不同分布的变量进行独立编码，给出信源各分量的率失真值和隐私泄露度量各自与扰动程度 D 之间的关系。分析表明，对这种数据源来说，统计方差越大的变量，可选择较小的扰动，确保信息可用性较大。此外，本文对数据源的统计特征做了高斯信源的假设，便于结果的分析，后续工作需要拓展到其他类型信源（如随机信号信源）的建模工作上。

参考文献：

[1] 刘向宇, 王斌, 杨晓春.社会网络数据发布隐私保护技术综述[J]. 软件学报, 2014,25(3):576-590.
LIU X Y, WANG B, YANG X C. Survey on privacy preserving techniques for publishing social network data[J]. Journal of Software, 2014, 25(3):576-590.

[2] 王璐, 孟小峰. 位置大数据隐私保护研究综述[J] 软件学报, 2014, (4): 693-712.
WANG L, MENG X F. Location privacy preservation in big data era: a survey [J]. Journal of Software, 2014,(4):693-712.

[3] XU J, ZHANG Z J, XIAO X K. Differentially private histogram pub-

lication[A]. Proc of IEEE 28th International Conference on Data Engineering (ICDE)[C]. Washington DC, USA, 2012. 32-43.

[4] 李瑞轩, 董新华, 辜希武, 等. 移动云服务的数据安全与隐私保护综述[J].通信学报, 2013, 34(12):158-166.
LI R X, DONG X H, GU X W, et al. Overview of the data security and privacy-preserving of mobile cloud services[J]. Journal of Communications, 2013, 34(12):158-166.

[5] CHAWLA S, DWORK C, MCSHERRY F. Towards privacy in public databases[A]. Proc of the 2nd IACR Theory Crypto[C]. Cambridge, MA. 2005. 363-385.

[6] SANKAR L, RAJAGOPALAN S R, POOR H V. A theory of utility and privacy of data sources[A]. Proc of 2010 IEEE International Symposium on Information Theory(ISIT)[C]. Austin, TX, 2010. 2642-2646.

[7] DWORK C. Differential privacy[A]. Proc of the 33rd Int. Colloq Automata Lang, Prog[C]. Venice, Italy, 2006. 1-12.

[8] SARATHY R，MURALIDHAR K. Some additional insights on applying differential privacy for numeric data[A]. Proc of International Conference on Privacy in Statistical Databases[C]. 2010. 210-219.

[9] VU D，SLAVKOVIC A. Differential privacy for clinical trial data: preliminary evaluations[A]. Proc of the 9th IEEE International Conference on Data Mining, Miami[C]. FL, USA, 2009. 138-143.

[10] LI T, LI N. On the tradeoff between privacy and utility in data publishing[A]. 15th ACM SIGKDD Int Conf Knowledge Discovery and Data Mining[C]. Paris, France, 2009. 517-526.

[11] LOUKIDES G, GKOULALAS-DIVANIS A, MALIN B. COAT: onstraint-based anonymization of transactions[J]. Knowl Inf Syst (KAIS), 2011, 28(2):251-282.

[12] LOUKIDES G, GKOULALAS-DIVANIS A, SHAO J. Assessing disclosure risk and data utility trade-off in transaction data anonymization[J]. International Journal of Software and Information, 2012, 6(3): 399-417.

[13] TERROVITIS M, MAMOULIS N, KALNIS P. Privacy-preserving anonymization of set-valued data[J]. PVLDB, 2008, 1(1): 115-125.

[14] GU Y H. A quantifying method for trade-off between privacy and utility[A]. IET International Conference on Information and Communications Technologies (IETICT 2013)[C]. Beijing, China, 2013. 270-273.

[15] ALVIM M, ANDRÉS M. On the relation between differential privacy and quantitative information flow[A]. 38th Int Conf, Automata, Languages and Programming Volume Part II[C]. Zurich, Switzerland, 2011. 60-76.

[16] CALMON F, FAWAZN N. Privacy against statistical inference[A]. 50th Annual Allerton Conf on Commun, Control, and Computing, Monticello[C]. IL, USA, 2012. 1-8.

作者简介：

谷勇浩（1980-），男，山西太原人，北京邮电大学讲师，主要研究方向为网络安全、隐私保护等。

林九川（1980-），男，江苏盐城人，公安部第三研究所助理研究员，主要研究方向为信息安全。