计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.190400098

基于图论与互信息量的差分隐私度量模型

王毛妮[1,2]　彭长根[1,2]　何文竹[1,2]　丁　兴[1,2]　丁红发[3]
1 贵州大学计算机科学与技术学院公共大数据国家重点实验室　贵阳 550025
2 贵州大学密码学与数据安全研究所　贵阳 550025
3 贵州财经大学信息学院　贵阳 550025
(1265071449@qq.com)

摘　要　差分隐私是数据发布、数据挖掘领域内隐私保护的重要工具，但其强度和效果仅能后验评估，且高度依赖于经验性选择的隐私预算。文中提出一种基于图论和互信息量的差分隐私量化模型和隐私泄露量计算方法。利用信息论通信模型重构了差分隐私保护框架，构造了差分隐私信息通信模型和隐私度量模型；基于图的距离正则和点传递提出隐私泄露互信息量化方法，证明并计算了差分隐私泄露量的信息量上界。分析和对比表明，该隐私泄露上界与原始数据集的属性数量、属性值数量以及隐私预算参数具有较好的函数关系，且计算限制条件较少。文中所提方法优于现有方法，能够为差分隐私算法的设计及评价、隐私泄露风险评估提供理论支撑。

关键词：差分隐私；隐私度量；互信息；汉明图；隐私泄露
中图法分类号　TP309

Privacy Metric Model of Differential Privacy via Graph Theory and Mutual Information

WANG Mao-ni[1,2], PENG Chang-gen[1,2], HE Wen-zhu[1,2], DING Xing[1,2] and DING Hong-fa[3]
1 Sate Key Laboratory of Pubic Big Data, College of Computer Science and Technology, Guizhou University, Guiyang 550025, China
2 Institute of Cryptography and Data Security, Guizhou University, Guiyang 550025, China
3 School of Information, Guizhou University of Finance and Economics, Guiyang 550025, China

Abstract　Differential privacy is an important tool for privacy preserving in many fields, such as data publishing and data mining. However, the strength and effectiveness of differential privacy cannot be evaluated previously, and highly rely on empirical selection of privacy budget. To this end, a privacy metric model and a privacy leakage method via graph theory and mutual information were proposed. This work models differential privacy as an information theoretic communication channel, and constructs an information channel and privacy metric model for differential privacy. Then, a mutual information based privacy metric method is proposed by employing the distance-regular and vertex-transitive of graphs, the upper bound of this metric is proofed, and an explicit formula is proposed for the bound. Delicate analysis and comparison show that the proposed upper bound has a function relationship limited by fewer computational constraints among the original dataset's attributes, attribute values and privacy budget. This work benefits more than related works, and provides theoretical foundation for algorithm design, algorithm evaluation, and privacy assessment.

Keywords　Differential privacy, Privacy metric, Mutual information, Hamming graph, Privacy leakage

1　引言

大数据时代的到来和移动互联网的普及，产生了巨大的商业价值和社会价值，同时也引发了人们对隐私的广泛关注和担忧。更加隐蔽、多样的数据收集存储以及数据挖掘，导致隐私泄露和隐私窃取更加频繁，从而产生更加巨大的危害和

到稿日期：2019-04-17　返修日期：2019-09-01　本文已加入开放科学计划(OSID)，请扫描上方二维码获取补充信息。
基金项目：国家自然科学基金(U1836205，61662009，61772008，11761020)；贵州省科技计划项目(黔科合重大专项字[2018]3001；黔科合重大专项字[2018]3007；黔科合重大专项字[2017]3002；黔科合支撑[2019]2004；黔科合支撑[2018]2159；贵州省教育厅青年科技人才成长项目(黔教合KY字[2016]171)
This work was supported by the National Natural Science Foundation of China(U1836205, 61662009, 61772008, 11761020), Science and Technology Program of Guizhou Province(Guizhou-Science-Contract-Major-Program [2018]3001, Guizhou-Science-Contract-Major-Program [2018]3007, Guizhou-Science-Contract-Major-Program [2017]3002, Guizhou-Science-Contract-Support [2019]2004, Guizhou-Science-Contract-Support [2018]2159, and Youth Science and Technology Talents Development Project of Guizhou Education Department(Guizhou-Education-Contract [2016]171).
通信作者：丁红发(hongfa.ding@foxmail.com)

影响。一方面，数据拥有者未进行任何保护处理直接发布含有隐私信息的数据，将会造成个人隐私信息的泄露；另一方面，恶意攻击者利用成熟的数据挖掘等技术窃取发布数据中的敏感信息。因此，隐私保护与隐私泄露量化是数据采集与应用过程中的重要研究主题。隐私度量可以作为隐私保护算法的隐私保护强度指标，是评价和优化隐私保护算法的重要依据；同时也是隐私泄露风险评估量化的重要方法，能够为隐私泄露量的评估、降低和隐私泄露风险控制提供量化指标。信息论作为重要的信息量化工具，在信息安全[1]、隐私保护[2-4]领域广泛应用，特别是在隐私保护算法的隐私保护强度评估方面，已被应用于匿名算法和差分隐私算法，但其理论有待完善。差分隐私的隐私预算参数 ε 代表隐私保护强度，该参数的选取高度依赖经验，仍然缺乏有效的信息量化方法对差分隐私强度和隐私泄露量进行预先量化。如何利用信息论对其隐私泄露量进行量化，并基于信息熵模型界定差分隐私保护程度的上界，已成为优化差分隐私算法和设计隐私风险评估方案的关键。

数据隐私保护问题最早由 Dalenuis[5] 针对关系数据库提出；随后，基于匿名模型的隐私保护技术如 k-anonymity[6]，l-diversity[7] 及其扩展的方法被相继提出。数据隐私保护的基本思想是通过对记录中的准标识符进行匿名化处理，使得所有记录被划分为若干个等价类，从而实现将一条记录隐藏在另一组记录中。匿名保护模型及其衍生算法尽管能一定程度地保护用户个人隐私，但均无法抵御背景知识攻击、同质攻击和相似性攻击，隐私保护程度无法得到严格证明。在此背景下，2006 年 Dwork[8] 提出差分隐私，即使攻击者具有无限背景知识，其也可保证相邻数据集上的查询具有概率不可区分性，同时能够提供严格有效的隐私保护水平证明。差分隐私坚实的数学基础，使其成为了近年来隐私保护研究领域的热点。文献[9]提出了差分隐私保护的 Laplace 实现机制，该方法通过向查询结果添加服从 Laplace 分布的噪声实现了差分隐私保护，但其仅适用于数值型查询结果。McSherry 等[10] 针对非数值型查询结果提出了指数机制，并将该机制应用于数据发布中。Roth 等[11] 针对交互式数据发布中查询数量的问题，提出了中位数机制，使得在相同的隐私预算下能够呈现更多的查询数量。Hardt 等[12] 基于机器学习中的加权多数算法提出了 PMW 机制来减少隐私预算的消耗。信息熵[13] 是信息量化的基础方法，可用以量化隐私信息量。2002 年，Dazi 等[14] 最早用信息熵对匿名保护模型中的匿名性进行了量化。2016 年，Peng 等[15] 将隐私保护系统描述为通信模型，并基于信息论的通信模型提出了几种隐私保护信息熵模型。2018 年，Wagner 等[16] 从不确定度、误差、信息增益等多个角度系统地介绍了 80 多种隐私度量标准及其适用领域。2009 年，针对差分隐私的隐私量化，Heusser 等[17] 提出了一种自动计算定量信息流的方法，该方法通过将数据库查询转换成相应的程序形式，对查询泄露的信息进行统计分析，推导出数据库存在的威胁；但该方法未涉及基于信息论的隐私量化。同年，Clarkson 等[18] 提出一种发布数据集的效用最大机制，通过几何机制在查询结果中添加噪声，研究差分隐私与隐私泄露之间的关系。Alivim 等[19-20] 将数据库查询系统看作一个有噪信道，并基于差分隐私的扩展定义提出一种求解互信息量泄露和 Rényi 熵泄露的方法；该方法所给的互信息量泄露的上界不严紧。Brathed 等[21] 提出了一种在考虑机制输入域值大小的情况下，基于信息论编码原理给出差分隐私泄露上界的方法，并用有理函数证明了差分隐私泄露的最优上界问题是可判定的。2016 年，Wang 等[22] 在同一隐私失真框架下研究了识别性、差分隐私和互信息隐私 3 种不同隐私概念之间的关系，并证明在最大失真函数范围条件下，存在一种同时优化可识别性的级别和互信息隐私的机制。Cuff 等[23] 利用互信息给出差分隐私的等效定义，将攻击者拥有的背景知识与从发布结果中获取的信息结合起来，用互信息量描述原始数据集中隐私信息的不确定度减少的量。信息论已成为量化差分隐私的有效工具，并被作为不断完善差分隐私算法优化、隐私量化的理论基础，但已有的研究主要针对非交互式差分隐私，如何利用互信息量对差分隐私保护系统中的隐私保护与隐私泄露之间的约束关系进行量化，仍需探索。

针对差分隐私保护的隐私泄露量化问题，本文提出一种基于图论与互信息的差分隐私度量模型。该模型利用信息论通信模型重构了差分隐私保护框架，构造了差分隐私的信息通信模型，将原始数据集表示为信源，发布数据集表示为信宿，查询机制和噪音机制表示为通信信道；进一步将信源和信宿视为图，以此将信道转移矩阵视为信源图和信宿图的复合图，并基于图的距离正则和点传递将信道转移矩阵转换为汉明图，提出差分隐私的隐私泄露互信息量化方法；利用图的自同构、邻接关系，通过放缩公式的方法证明差分隐私保护机制的隐私泄露量存在上界，并提出一个计算隐私泄露上界的公式。本文所提出的差分隐私度量模型以信息通信模型为基础，利用图的特性，结合信息熵，给出隐私泄露量的互信息化计算方法。隐私泄露量的界仅依赖于原始数据集的属性数量、属性值数量及差分隐私预算参数，对任意分布的原始数据集、任意攻击能力的敌手都成立。通过分析，差分隐私保护机制的隐私泄露上界与原始数据集的属性数量、属性值数量及差分隐私预算参数具有较好的函数关系；通过对比，该隐私量化模型可给出差分隐私保护的隐私泄露互信息上界，限制条件较少，适用于所有信道，且不依赖原始数据集的分布。

2 基础定义

差分隐私[24] 保护是一种基于数据失真的隐私保护技术，通过在原始数据集或统计结果中添加噪声扰动来实现隐私保护，同时保持某些数据属性或统计属性不变。差分隐私保护技术确保了数据集中单个记录的变化不会影响查询结果，根据实现环境可分为交互式差分隐私和非交互式差分隐私。

定义 1（邻近数据集） 如果数据集 D 和 D' 具有完全相同的数据属性结构，且两个数据集之间至多相差一条记录，则称数据集 D 和 D' 是邻近数据集（或相邻的），记为 $D\sim D'$。

定义 2（差分隐私） 设有随机算法 K，若算法 K 对于邻近数据集 D 和 D' 上的任意输出结果 $S(S\in Range(K))$ 满足

$$\Pr[K(D)\in S]\leqslant e^{\varepsilon}\Pr[K(D')\in S] \tag{1}$$

则称随机算法 K 满足 ε-差分隐私。其中，$\Pr[\cdot]$ 表示隐私被披露的概率；ε 是隐私预算参数，表示隐私保护程度，值越小

则隐私保护程度越高。

3 基于信息论的差分隐私度量模型

在交互式差分隐私保护框架下，用户通过查询接口向数据拥有者递交查询请求，数据拥有者根据查询请求在源数据集中进行查询，然后将查询结果添加噪声扰动之后反馈给用户，其基本框架如图 1 所示。在非交互式差分隐私保护框架下，数据管理者直接发布一个满足差分隐私保护的数据集，再根据用户的请求对发布数据集进行查询操作，其基本框架如图 2 所示。

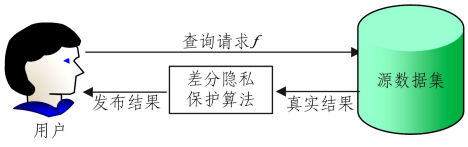


图 1 交互式差分隐私保护框架

Fig. 1 Interactive differential privacy protection framework

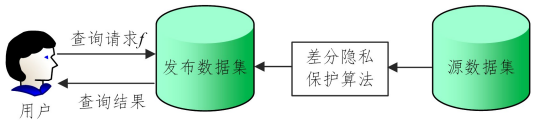


图 2 非交互式差分隐私保护框架

Fig. 2 No-Interactive differential privacy protection framework

为了更精确地量化差分隐私保护框架中的隐私泄露量，将差分隐私保护框架描述成一种通信信道$(X,Y,\boldsymbol{M})$，其中 X,Y 分别为信道的信源和信宿，$\boldsymbol{M}$ 是信道转移概率矩阵(简称信道矩阵)。$\boldsymbol{M}$ 中每个元素 $M_{x,y}$ 表示已知输入 X 的值为 x 后，输出 Y 的值为 y 的概率，即 $M_{x,y}=p(y|x)$。信道矩阵 $\boldsymbol{M}$ 满足 ε-差分隐私，即 $M_{x,y}\leqslant e^{\varepsilon}M_{x',y}(x'\in X,x\sim x'$ 且 $y\in Y)$。差分隐私的基本通信模型如图 3 所示。

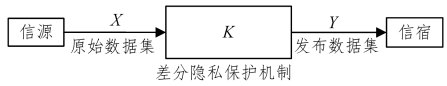


图 3 差分隐私的通信模型

Fig. 3 Differential privacy channel model

隐私泄露量是指攻击者从发布数据集中获取信息后，原始数据集 X 中信息不确定度的减少量，记为 $L(X,Y)$。

差分隐私度量模型中包含查询机制和噪音机制。查询机制指数据管理者根据查询请求在原始数据集(发布数据集)中查询后得到真实结果；噪音机制指将真实结果(原始数据)进行差分隐私机制扰动后得到发布结果(发布数据集)。该模型的目的是量化发布数据集对原始数据集的隐私泄露量，即 Y 对 X 的互信息量 $I(X;Y)$。差分隐私度量模型如图 4 所示。

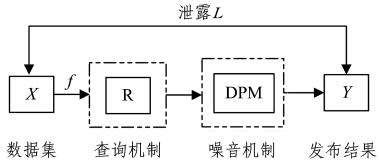


图 4 差分隐私度量模型

Fig. 4 Differential privacy metric model

4 面向差分隐私的隐私泄露上界

本节在差分隐私保护框架下，基于图 4 中差分隐私度量模型对原始数据集与发布数据集间的隐私泄露量进行量化，其原理与过程如图 5 所示。

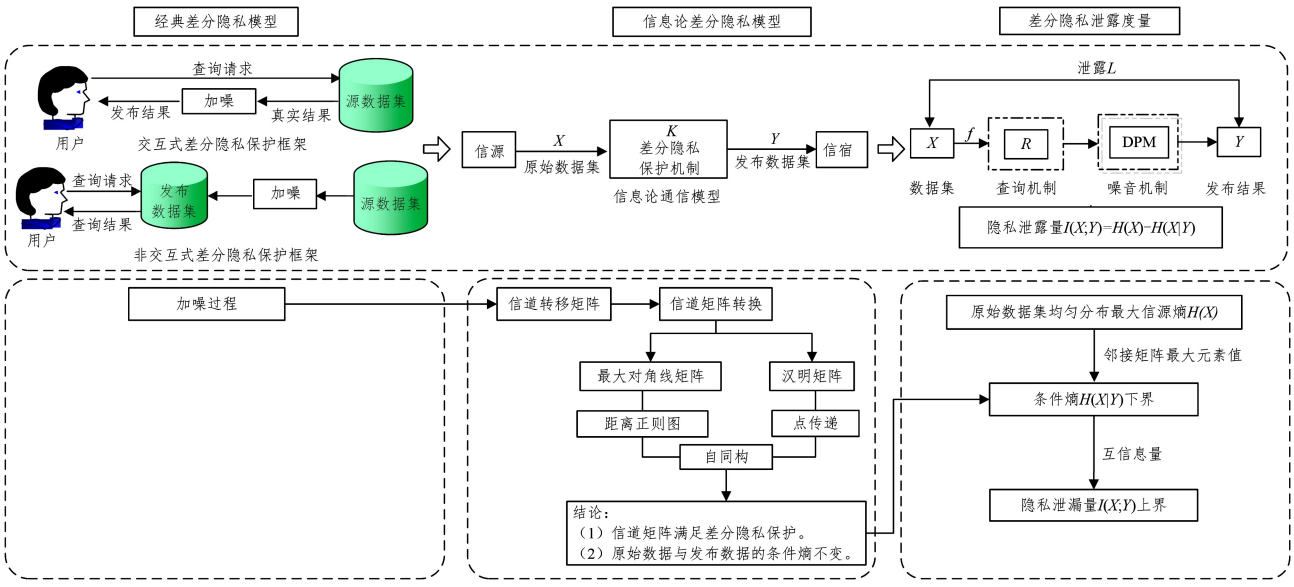


图 5 基于信息论的差分隐私量化计算模型

Fig. 5 Computing model of differential privacy metric based on information theory

首先，基于第 3 节所提的差分隐私量化模型，将经典差分隐私模型的加噪过程转换为信息论差分隐私模型的通信信道矩阵；接着，利用图论的相关性质对信源到信宿的信道矩阵进行转换，证明该信道矩阵满足差分隐私保护，且原始数据集与发布数据集间的条件熵不变(引理 1)；然后，利用转换后的信道矩阵的汉明矩阵对称性和自同构关系，证明并计算原始数据集与发布数据集间的条件熵下界(定理 1)；最后，在原始数据集均匀分布的条件下，利用互信息的计算方法计算隐私泄露量的信息量上界(定理 2)。

4.1 信道矩阵转换

给定一个满足 ε-差分隐私的信道矩阵 $\boldsymbol{M}$，将信道输入数据集中的属性和属性值视为图形结构中顶点与边，且将原始数据集中的相邻关系映射为图中的点与点之间的邻接关系。因此，原始数据集和发布数据集可分别表示为无向图$(X,\sim)$和$(Y,\sim)$，其中数据集中的属性与属性值对应于图的点和边。基于此，利用图的距离正则和点传递对通信模型中的信道矩阵进行转换处理，得到汉明矩阵，并将其用于求解通信信道中原始数据集与发布数据集间的条件熵，且保证转换后的信道矩阵仍满足 ε-差分隐私保护，信道中原始数据集与发布数据集间的条件熵不变。

定义 3（条件熵[13]） 条件熵是在联合符号集合 XY 上的条件自信息量的数学期望。在已知随机变量 Y 的条件下，随机变量 X 的条件熵 $H(X|Y)$ 定义为：

$$H(X|Y)=-\sum_{y\in Y}p(y)\sum_{x\in X}p(x|y)\log_2 p(x|y) \tag{2}$$

定义 4（距离正则图[26]） 如果存在整数 b_d 和 c_d（$d=0,1,\cdots,d_{\max}$）使得对于图 G 中的任意顶点 v 和 v'，在顶点 v' 的邻点中，到顶点 v 距离为 $i-1$ 的顶点数目为 b_d，到顶点 v' 距离为 $i-1$ 的顶点数目为 c_d，则称图 G 为距离正则图，其中参数 b_d 和 c_d 为图的交叉数。

定义 5（点传递[25]） 如果图 G 的全自同构群作用在顶点集 $V(G)$ 上传递，即对于图 G 中任意顶点 $v,v'\in V$，存在自同构使得 $\sigma(v)=v'$，则称图 G 为点传递。

定义 6（自同构[25]） 顶点集 $V(G)$ 上的置换 σ 称为图 G 的自同构，即对于任意顶点 $v,v'\in V$，均有：如果 $v\sim v'$，则 $\sigma(v)\sim\sigma(v')$。

假设信道的输入为随机变量 $X=\{x_0,x_1,\cdots,x_{n-1}\}$，输出为随机变量 $Y=\{y_0,y_1,\cdots,y_{m-1}\}$，其中 $n\leqslant m$，原始数据集的图形结构 X 上存在邻近关系$\sim$，且原始数据集的概率分布为均匀分布。下文中用下标来表示随机变量 X 和 Y 的元素，即 $i=x_i$，$j=y_j$，记 $i\sim h=x_i\sim x_h$，$d(i,h)=d(x_i,x_h)$。对于信道矩阵 $\boldsymbol{M}$，记行 i 中第 j 列的最大值为 $\max_j^{\boldsymbol{M}}$，矩阵中的最大值记为 $\max^{\boldsymbol{M}}$。

信道矩阵 $\boldsymbol{M}$ 转换为汉明矩阵 $\boldsymbol{M}''$ 的步骤如图 6 所示。

$$\boldsymbol{M}=\begin{bmatrix} M_{0,0} & M_{0,1} & \cdots & M_{0,m-1}\\ M_{1,0} & M_{1,1} & \cdots & M_{1,m-1}\\ \vdots & \vdots & & \vdots\\ M_{n-1,0} & M_{n-1,1} & \cdots & M_{n-1,m-1}\end{bmatrix}\rightarrow\boldsymbol{M}'=\begin{bmatrix} \max_0^{\boldsymbol{M}'} & - & \cdots & -\\ - & \max_1^{\boldsymbol{M}'} & & -\\ \vdots & \vdots & & \vdots\\ - & - & \cdots & \max_{n-1}^{\boldsymbol{M}'}\end{bmatrix}\rightarrow\boldsymbol{M}''=\begin{bmatrix} \max^{\boldsymbol{M}''} & - & \cdots & -\\ - & \max^{\boldsymbol{M}''} & \cdots & -\\ \vdots & \vdots & & \vdots\\ - & - & \cdots & \max^{\boldsymbol{M}''}\end{bmatrix}$$

图 6 信道矩阵的距离正则图和点传递图的矩阵转换

Fig. 6 Matrix transformation for distance-regular and vertex-transitive graphs

1）信道矩阵 $\boldsymbol{M}$ 转换为最大对角线矩阵 $\boldsymbol{M}'$。将信道矩阵 $\boldsymbol{M}$ 前 n 列中每一列元素的最大值移动到对角线上，即 $M'_{i,i}=\max_j^M(0\leqslant i\leqslant n-1)$，且令其余列均为 0，即 $M'_{i,j}=0(0\leqslant i\leqslant n-1,n\leqslant j\leqslant m-n)$。此时，最大对角线矩阵 $\boldsymbol{M}'$ 为 $n\times n$ 的方阵，矩阵 $\boldsymbol{M}'$ 仍满足 ε-差分隐私，且原始数据集与发布数据集间的条件熵 $H(X|Y)$ 不变。

2）最大对角线矩阵 $\boldsymbol{M}'$ 转换为汉明矩阵 $\boldsymbol{M}''$。假设图 X 中存在一个邻接关系$(M'_{i,i}\sim M'_{i,j})$，使得原始数据集的图形结构$(X,\sim)$是连通的且为距离正则图和点传递，则最大对角线矩阵 $\boldsymbol{M}'$ 转化为汉明矩阵 $\boldsymbol{M}''$ 后仍满足 ε-差分隐私，且原始数据集与发布数据集间的条件熵 $H(X|Y)$ 不变，其对角线上的元素都相等且均为矩阵中的最大元素，即 $M''_{i,i}=\max^{\boldsymbol{M}''}(0\leqslant i\leqslant n-1)$。

引理 1 信道矩阵 $\boldsymbol{M}$ 经过变换后得到汉明矩阵 $\boldsymbol{M}''$，则汉明矩阵 $\boldsymbol{M}''$ 仍满足 ε-差分隐私，且原始数据集与发布数据集间的条件熵不变。

$$H^{\boldsymbol{M}}(X|Y)=H^{\boldsymbol{M}'}(X|Y)=H^{\boldsymbol{M}''}(X|Y) \tag{3}$$

证明：当 $0\leqslant k\leqslant m-1$ 时，构造集合 C_k 为：

$$C_k=\{j\,|\,M_{k,j}=\max_j^{\boldsymbol{M}}\ \text{且}\ \forall\, i<k, M_{i,j}<\max_j^{\boldsymbol{M}}\} \tag{4}$$

因每个列 j 都属于集合 C_k，故：

$$\bigcup_k C_k=\{0,1,\cdots,m\} \tag{5}$$

且当 $h\neq k$ 时，$C_h\cap C_k=\varnothing$，故：

$$M'_{i,k}=\sum_{j\in C_k}M_{i,j}(0\leqslant i\leqslant n-1,0\leqslant k\leqslant m-1) \tag{6}$$

又因 $\boldsymbol{M}$ 满足 ε-差分隐私，故：

$$M'_{i,k}=\sum_{j\in C_k}M_{i,j}\leqslant\sum_{j\in C_k}\mathrm{e}^{\varepsilon}M_{h,j}=\mathrm{e}^{\varepsilon}\sum_{j\in C_k}M_{h,j}=\mathrm{e}^{\varepsilon}M'_{h,k} \tag{7}$$

因此，最大对角线矩阵 $\boldsymbol{M}'$ 满足 ε-差分隐私。

由定义 7 中的自同构知：

$$M''_{i,j}=\frac{1}{|\Gamma|}\sum_{\sigma\in\Gamma}M'_{\sigma(i),\sigma(j)} \tag{8}$$

因 $\boldsymbol{M}'$ 满足 ε-差分隐私，故

$$M''_{i,j}=\frac{1}{|\Gamma|}\sum_{\sigma\in\Gamma}M'_{\sigma(i),\sigma(j)}\leqslant\frac{1}{|\Gamma|}\mathrm{e}^{\varepsilon}\sum_{\sigma\in\Gamma}M'_{\sigma(i),\sigma(j)}=\mathrm{e}^{\varepsilon}M''_{h,j} \tag{9}$$

因此，汉明矩阵 $\boldsymbol{M}''$ 满足 ε-差分隐私。

由条件熵的定义可知：

$$\begin{aligned}H^{\boldsymbol{M}''}(X|Y)&=-\sum_{y\in Y}\sum_{x\in X}p''(y)p''(x|y)\log_2 p''(x|y)\\&=-\sum_{y\in Y}\sum_{x\in X}p''(y)p''(x|y)\log_2 p''(x|y)+\\&\quad\sum_{y\in Y}p''(y)\log_2 p''(x|y)\end{aligned} \tag{10}$$

因此，汉明矩阵 $\boldsymbol{M}''$ 中的对角线元素都相等，故：

$$H^{\boldsymbol{M}''}(X|Y)=-\sum_{i=0}^{n-1}M''_{i,i}\log_2\frac{M''_{i,i}}{n}+\sum_{y\in Y}p''(y)\log_2 p''(x|y) \tag{11}$$

因 σ 是自同构，故：

$$\begin{aligned}H^{\boldsymbol{M}''}(X|Y)&=-\sum_{i=0}^{n-1}\frac{1}{|\Gamma|}\sum_{\sigma\in\Gamma}M'_{\sigma(i),\sigma(i)}\log\frac{\frac{1}{|\Gamma|}\sum_{\sigma\in\Gamma}M'_{\sigma(i),\sigma(i)}}{n}+\\&\quad\sum_{y\in Y}p''(y)\log_2 p''(x|y)\\&=-\frac{1}{|\Gamma|}\sum_{\sigma\in\Gamma}\sum_{i=0}^{n-1}M'_{i,i}\log\frac{\frac{1}{|\Gamma|}\sum_{\sigma\in\Gamma}M'_{i,i}}{n}+H(Y)\\&=\sum_{i=0}^{n-1}M'_{i,i}\log_2\frac{M'_{i,i}}{n}+H(Y)\\&=-\sum_{y\in Y}\sum_{x\in X}p'(y)p'(x|y)\log_2 p'(y)p'(x|y)+\end{aligned}$$

$$\sum_{y\in Y} p'(y)\log_2 p'(y) = H^{\boldsymbol{M}'}(X|Y) \tag{12}$$

故：

$$H^{\boldsymbol{M}}(X|Y) = H^{\boldsymbol{M}''}(X|Y) \tag{13}$$

证毕。

同理可证 $H^{\boldsymbol{M}}(X|Y) = H^{\boldsymbol{M}'}(X|Y)$。

通过引理 1 可知，信道矩阵在经过两次转换后仍满足 ε-差分隐私，且原始数据集与发布数据集间的条件熵大小不变。因此，在信道$(X,Y,\boldsymbol{M}'')$中，基于汉明矩阵的图性质，对原始数据集与发布数据集间的条件熵下界的证明和推导对原始矩阵 $\boldsymbol{M}$ 同样有效。

4.2 隐私泄露的量化

本小节对差分隐私度量模型中原始数据集与发布数据集间的隐私泄露量进行量化，记隐私度量模型中发布数据集对原始数据集的最大隐私泄露量为 ML。

定义 7(信息熵[13]) 假设存在随机变量 $X=\{x_1,x_2,\cdots,x_n\}$，每个随机变量的概率分布 $P(X)=\{p(x_1),p(x_2),\cdots,p(x_n)\}$，定义信源各个离散消息的自信息量的数学期望为信源的平均信息量，称为信源的信息熵，记为 $H(X)$。

$$H(X) = -\sum_{x\in X} p(x)\log_2 p(x) \tag{14}$$

定义 8(互信息量[13]) 在通信模型中，用平均互信息量来刻画信道中的隐私泄露程度，Y 对 X 的平均互信息量是在 Y 一无所知的情况下，X 的先验不定度与收到 Y 后的后验不定度之差，即表示收到 X 前后关于 Y 的不确定度减少的量。

$$I(X;Y) = H(X) - H(X|Y) \tag{15}$$

定理 1 设信道矩阵 $\boldsymbol{M}$ 满足 ε-差分隐私，原始数据集 X 的概率为均匀分布，且$(X,\sim)$为距离正则图和点传递，则：

$$H(X|Y) \geqslant -\log_2 \max^{\boldsymbol{M}} \tag{16}$$

证明：由式(2)中的条件熵定义知：

$$\begin{aligned} H^{\boldsymbol{M}''}(X|Y) &= -\sum_{y\in Y} p''(y)\sum_{x\in X} p''(x|y)\log_2 p''(x|y) \\ &= -\sum_{y\in Y} p''(y)\sum_{x\in X} p''(x|y)\times[\log_2 p''(y)p''(x|y) - \log_2 p''(y)] \\ &= -\sum_{y\in Y}\sum_{x\in X} p''(y)p''(x|y)\log_2 p''(y)p''(x|y) + \sum_{y\in Y} p''(y)\log_2 p''(y) \end{aligned} \tag{17}$$

由式(14)中信息熵的定义及均匀分布最大熵原理得：

$$\begin{aligned} H^{\boldsymbol{M}''}(X|Y) &\geqslant -\sum_{y\in Y}\sum_{x\in X} p''(x)p''(y|x)\log_2 p''(x)p''(y|x) - \log_2 n \\ &= -\sum_{y\in Y}\sum_{x\in X} p''(x)p''(y|x)\log_2 \frac{1}{n}p''(y|x) - \log_2 n \end{aligned} \tag{18}$$

又因 $M''_{i,j} \leqslant \max^{\boldsymbol{M}''}$，故：

$$\begin{aligned} H^{\boldsymbol{M}''}(X|Y) &\geqslant -\sum_{y\in Y}\sum_{x\in X} p''(x)p''(y|x)\log_2 \frac{1}{n}\max^{\boldsymbol{M}''}\log_2 n - \log_2 n \\ &= -\log_2 \frac{\max^{\boldsymbol{M}''}}{n}\sum_{y\in Y}\sum_{x\in X} p''(x)p''(y|x) - \log_2 n \end{aligned} \tag{19}$$

又因 $p(y)=\sum_{x\in X} p(x)p(y|x)$，且 $\sum_{y\in Y} p(y)=1$。故：

$$H^{\boldsymbol{M}''}(X|Y) = -(\log_2 \max^{\boldsymbol{M}''} - \log_2 n)\sum_{y\in Y} p''(y) - \log_2 n = -\log_2 \max^{\boldsymbol{M}''} \tag{20}$$

由引理 1 知，$H^{\boldsymbol{M}}(X|Y) = H^{\boldsymbol{M}''}(X|Y)$，故：

$$H(X|Y) \geqslant -\log_2 \max^{\boldsymbol{M}} \tag{21}$$

证毕。

由定理 1 可知，差分隐私通信信道中信源与信宿的条件熵下界，可归结为求矩阵的最大元素的上界。

引理 2 如果信道矩阵 $\boldsymbol{M}$ 满足 ε-差分隐私，且对于每个 $0\leqslant i\leqslant n-1$ 有 $M_{i,i}=\max^{\boldsymbol{M}}$，则对任意顶点 $i\in X$，有：

$$\max^{\boldsymbol{M}} \leqslant \frac{1}{\sum_{d\in S_G} |X_{\langle d\rangle}(i)|/\mathrm{e}^{\varepsilon d}} \tag{22}$$

证明：由文献[24]中定义 2 的扩展知，假设信道矩阵 $\boldsymbol{M}$ 满足 ε-差分隐私，则对于任意列 j，以及任意一对行 i 和 h($i\sim h$)，有：

$$\frac{1}{\mathrm{e}^{\varepsilon d(i,h)}} \leqslant \frac{M''_{i,j}}{M''_{h,j}} \leqslant \mathrm{e}^{\varepsilon d(i,h)} \tag{23}$$

当 $h=j$ 时，矩阵 $\boldsymbol{M}''$对角线上的元素相等，且等于最大元素值，故对于每一个元素 $M''_{i,j}$有：

$$\max^{\boldsymbol{M}''} \leqslant \mathrm{e}^{\varepsilon d(i,j)} M''_{i,j} \tag{24}$$

又因为矩阵 $\boldsymbol{M}''$中任意行元素均为概率分布，则 $\sum_j M''_{i,j}=1$，故：

$$\sum_j \frac{\max^{\boldsymbol{M}''}}{\mathrm{e}^{\varepsilon d(i,j)}} \leqslant \sum_j M''_{i,j} = 1 \tag{25}$$

且根据图形结构元素的距离分组知：

$$\sum_{d\in S_G} (|X_{\langle d\rangle}(i)|\frac{\max^{\boldsymbol{M}''}}{\mathrm{e}^{\varepsilon d}}) \leqslant 1 \tag{26}$$

其中，$|X_{\langle d\rangle}(i)|$表示图形结构$(\sim,X)$中到顶点 i 的距离为 d 的顶点的个数，$S_G=\{0,1,\cdots,d_{\max}\}$。

通过不等式变换得到：

$$\max^{\boldsymbol{M}''} \leqslant \frac{1}{\sum_{d\in S_G} |X_{\langle d\rangle}(i)|/\mathrm{e}^{\varepsilon d}} \tag{27}$$

由引理 1 可知：

$$\max^{\boldsymbol{M}} \leqslant \frac{1}{\sum_{d\in S_G} |X_{\langle d\rangle}(i)|/\mathrm{e}^{\varepsilon d}} \tag{28}$$

证毕。

若通信模型的输入图形结构为距离正则图和点传递，则对于每一个 $d\in S_G$，$|X_{\langle d\rangle}(i)|$的值均相同且只取决于 d，将其值记为 N_d，即 $N_d=|X_{\langle d\rangle}(i)|$。故：

$$\max^{\boldsymbol{M}} \leqslant \frac{1}{\sum_{d\in S_G} N_d/\mathrm{e}^{\varepsilon d}} \tag{29}$$

X 与 Y 间的隐私泄露量 $L(X,Y)$度量的是攻击者通过观察发布数据集所获取的原始数据集中的隐私信息，利用互信息来量化度量模型中的隐私泄露量，即 $L(X,Y)=I(X;Y)$。在 4.1 节信道矩阵转换的基础上，结合汉明矩阵的距离正则和点传递，证明了原始数据集与发布数据集间的条件熵存在下界；再基于互信息量计算方法，求出发布数据集对原始数据集的隐私泄露上界。

定理 2 若随机算法 K 满足 ε-差分隐私，则对于原始数据集的任意概率分布，发布数据集对原始数据集的互信息量上界为：

$$I(X;Y)=u\log_2\frac{v\mathrm{e}^{\varepsilon}}{v-1+\mathrm{e}^{\varepsilon}} \tag{30}$$

证明：通过改变表示 i 的 u 元组中的个体的值，可以得到与 x 距离为 d 的每个元素 j。这些个体有 $\binom{u}{d}$ 种可能的选择，每一种选择有 $v-1$ 种可能情况，故：

$$N_d=|X_{\langle d\rangle}(i)|=\binom{u}{d}(v-1)^d \tag{31}$$

则：

$$\begin{aligned}\sum_{d\in S_G}\frac{N_d}{\mathrm{e}^{\varepsilon d}}&=\sum_{d\in S_G}\frac{\binom{u}{d}(v-1)^d}{\mathrm{e}^{\varepsilon d}}\\&=\frac{1}{\mathrm{e}^{\varepsilon u}}\sum_{d\in S_G}\frac{\binom{u}{d}(v-1)^d\mathrm{e}^{\varepsilon u}}{\mathrm{e}^{\varepsilon d}}\\&=\frac{1}{\mathrm{e}^{\varepsilon u}}\sum_{d\in S_G}\binom{u}{d}(v-1)^d\mathrm{e}^{\varepsilon(u-d)}\\&=\frac{(v-1+\mathrm{e}^{\varepsilon})^u}{\mathrm{e}^{\varepsilon u}}\end{aligned} \tag{32}$$

由式(29)知：

$$\max^{\mathbf{M}}\leqslant\frac{1}{\sum\limits_{d\in S_G}N_d/\mathrm{e}^{\varepsilon d}}=\frac{\mathrm{e}^{\varepsilon u}}{(v-1+\mathrm{e}^{\varepsilon})^u} \tag{33}$$

由式(21)知：

$$H^{\mathbf{M}'}(X|Y)\geqslant-\log_2\max^{\mathbf{M}'}=-\log_2\frac{\mathrm{e}^{\varepsilon u}}{(v-1+\mathrm{e}^{\varepsilon})^u} \tag{34}$$

故：

$$H(X|Y)\geqslant-\log_2\frac{\mathrm{e}^{\varepsilon u}}{(v-1+\mathrm{e}^{\varepsilon})^u} \tag{35}$$

当原始数据集的概率为均匀分布时，信息熵有最大值，即 $H(X)=\log_2 n=\log_2 v^u$。根据式(15)中互信息量的定义知：

$$\begin{aligned}I(X;Y)&=H(X)-H(X|Y)\\&\leqslant\log_2 n+\log_2\max^{\mathbf{M}}\\&=u\log_2\frac{v\mathrm{e}^{\varepsilon}}{v-1+\mathrm{e}^{\varepsilon}}\end{aligned} \tag{36}$$

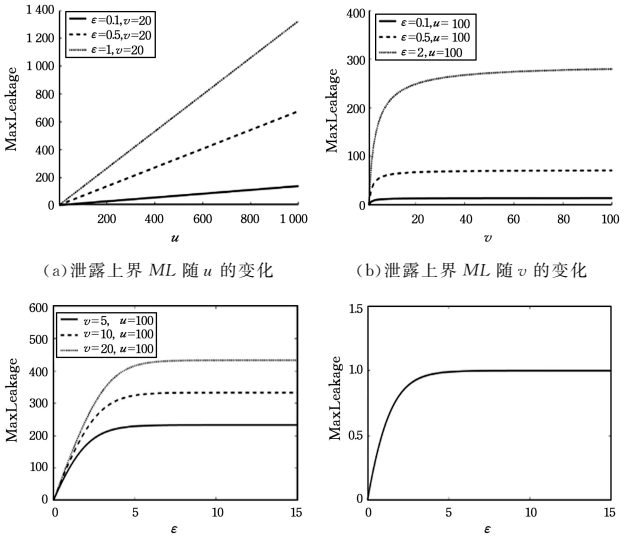
由定理 1 至引理 2 可知，当原始数据集的概率为均匀分布时，原始数据集有最大信息熵，此时互信息量泄露最大，故当原始数据集为任意概率分布时，式(36)中的结果仍然成立。因此，互信息量上界对原始数据集上的任意分布都是有效的。此外，由于所提出的模型仍满足差分隐私机制，因此互信息上界对对手可能具有的任何背景知识都是有效的。证毕。

由定理 2 证明可知，若随机算法 K 满足 ε-差分隐私，则差分隐私保护框架中原始数据集与发布数据集间的隐私泄露有界，即 $I(X;Y)\leqslant ML(u,v,\varepsilon)$，其中 $ML(u,v,\varepsilon)=u\log_2\frac{v\mathrm{e}^{\varepsilon}}{v-1+\mathrm{e}^{\varepsilon}}$。同时，差分隐私保护系统中的最大隐私泄露上界对任意信道、任意原始数据集分布都成立。

5 隐私度量模型分析及对比

5.1 隐私度量模型分析

在差分隐私保护框架中，隐私泄露与差分隐私间存在约束关系，对任意原始数据集和拥有任意攻击能力的敌手，隐私泄露量的上界为 $ML(u,v,\varepsilon)=u\log_2\frac{v\mathrm{e}^{\varepsilon}}{v-1+\mathrm{e}^{\varepsilon}}$，其中 $ML(u,v,\varepsilon)$ 是 ε 上的连续函数。当 $\varepsilon=0$ 时，信道矩阵中的所有行都是相同的，因此，差分隐私查询系统没有隐私泄露，隐私泄露量为 0；当 ε 值趋于无穷时，矩阵同一列中两个元素的比值可能变得无界，原始数据集与发布数据集间的条件熵可能为零，则通信模型中的隐私泄露量等于原始数据集的信息熵，即 $ML(u,v,\varepsilon)$ 收敛于 $\log_2 v^u$。差分隐私保护中，最大泄露量 $ML(u,v,\varepsilon)$ 随参数 u,v,ε 取值的变化趋势如图 7 所示。



(a)泄露上界 ML 随 u 的变化　(b)泄露上界 ML 随 v 的变化

(c)泄露上界 ML 随 ε 的变化　(d)二元信道泄露上界 ML 随 ε 的变化

图 7　差分隐私最大泄露量随参数 u,v,ε 的变化趋势

Fig. 7　Maximum leakage amount varies with the parameter u,v,ε of differential privacy

从图 7(a)可以看出，差分隐私的最大隐私泄露量 ML 与 u 成正比；且随差分隐私预算 ε 的增加，最大泄露上界的增长速率不断增加。这表明当属性值的取值数量相同时，属性数量越多，差分隐私的最大隐私泄露量越大；隐私保护强度越低，隐私泄露量的增长速度越快。

从图 7(b)可以看出，差分隐私的最大隐私泄露量随属性值取值数量的增加而增加。这表明当属性数量相同时，属性值的取值数量越多，则差分隐私的最大隐私泄露量越大，且随着隐私预算 ε 的增加，最大隐私泄露量最终收敛于 $\log_2 v^u$。

从图 7(c)可以看出，差分隐私的最大隐私泄露量随隐私预算参数的增大而增加。这表明在属性数量相同的情况下，属性值的取值数量越多，则最大隐私泄露量越大，并且当隐私保护强度越高时，差分隐私的最大隐私泄露量越大，最终收敛于 $\log_2 v^u$。

从图 7(d)可以看出，二元信道下差分隐私的最大隐私泄露量 ML 随隐私预算 ε 的增大而增加，并趋于稳定。

综上所述，基于互信息与图论的差分隐私度量模型能够给出差分隐私的互信息上界，与第 3－4 节的模型和隐私量化方法的证明结果一致。

5.2 对比分析

为了说明本文所提模型的特点和有效性，将其与差分隐私量化相关模型进行对比，如表 1 所列。

表 1 差分隐私度量模型对比

Table 1 Comparison of differential privacy metric model

模型	场景	信道类型	原始数据集分布	目的	量化工具	方法	结果
Alvim[19]	差分隐私	任意信道	均匀分布	隐私泄露上界	香农熵与 Rényi 熵	输出扰动 (Laplace 机制)	$I(X;Y)\leqslant(e^{2\delta}+e^{-2\delta})\log e^{\delta}+(e^{2\delta}-e^{-2\delta})\sum_{y}p(y\|x^{*})\log p((y\|x^{*}))$ $I_{\infty}(X;Y)\leqslant 2\delta\log e$
Alvim[20]	差分隐私	任意信道	任意分布	隐私与效用平衡	Rényi 熵	通信模型	$I_{\infty}(X;Y)\leqslant u\log_2\frac{ve^{\varepsilon}}{v-1+e^{\varepsilon}}$
Barthe[21]	差分隐私	二元信道	均匀分布	组合性下的隐私泄露上界	Rényi 熵	差分隐私等价定义与信息论编码	$ML(P_{Y\|X})\leqslant d\log_2 e^{\varepsilon}+\log_2 m$
Wang[22]	差分隐私	—	均匀分布	隐私失真框架下的可识别性差分隐私、互信息间的关系	互信息量	汉明距离	$\varepsilon_i^{*}-\varepsilon_X\leqslant\varepsilon_d^{*}(D)\leqslant\varepsilon_i^{*}(D)$
Cuff[23]	差分隐私	—	—	差分隐私与互信息量的关系	互信息量	互信息量隐私	$\varepsilon\text{-}DP\geqslant MI\text{-}DP\geqslant(\varepsilon,\delta)\text{-}DP$
本文模型	差分隐私	任意信道	任意分布	隐私泄露上界	互信息量	图论与互信息量	$I(X;Y)\leqslant-\log_2\frac{e^{\varepsilon u}}{(v-1+e^{\varepsilon})^{u}}$

由表 1 可知，文献[19-21]和本文所提出的差分隐私度量模型都是为了量化差分隐私的隐私泄露量，且给出了隐私泄露的上界。文献[19]以香农熵和 Rényi 熵为度量工具，并基于输出扰动(Laplace 机制)进行差分隐私机制的信息理论分析，利用差分隐私的扩展定义给出差分隐私保护系统中发布数据集对原始数据集的隐私泄露上界，但所提出的隐私泄露量的计算公式复杂，本文所提出的隐私泄露上界仅与原始数据集的属性数量、属性值数量及差分隐私预算参数有关。文献[20]在文献[19]的基础上仅考虑 Rényi 熵，利用信息论通信模型重构了差分隐私保护系统，并基于贝叶斯公式给出了隐私泄露与效用的上界，本文是基于图论并利用互信息量提出的一种隐私泄露上界计算方法。文献[21]中的通信信道为二元信道，在考虑差分隐私组合特性的情况下，利用信息论编码提出了差分隐私泄露的量化方法，并证明其存在隐私泄露上界。本文提出了一种基于图论和互信息的差分隐私泄露量化模型并证明了隐私泄露上界，将差分隐私查询系统建模为含有查询机制和噪音机制的信息论信道，将信道的输入和输出视为一种图形结构，并通过将图的距离正则和点传递与互信息公式相结合提出一种差分隐私泄露上界的计算方法。该模型对任意信道、任意原始数据集分布的差分隐私泄露上界都成立。

文献[22-23]和本文都研究了差分隐私与互信息间的关系。其中，文献[22]在隐私失真框架下，基于输入与输出间的汉明距离提出了可识别性、差分隐私与互信息隐私三者之间的偏序关系；文献[23]研究了互信息对差分隐私的约束，并基于互信息量的定义更清晰地定义了差分隐私；本文模型利用信息论中的通信机制将差分隐私查询系统建模为信息论通信信道，并利用图形结构的自同构等性质将差分隐私的加噪过程转化为信道转换过程，最后再基于互信息提出差分隐私泄露量的上界及其量化方法，但是本文模型未考虑发布数据的效用度量问题。

结束语 针对差分隐私的算法评价和隐私泄露量化问题，本文基于信息论和图论提出了一种差分隐私量化模型和基于互信息的隐私泄露上界量化方法，并证明了差分隐私在信息论通信框架下，隐私泄露量上界的存在性和可计算性。首先，本文对差分查询系统进行信息通信建模，构造原始数据、差分噪声机制和发布数据到信源、信道和信宿的映射；其次，利用图论将差分隐私通信模型的信源、信宿和信道进行图实例化，并基于图的距离正则与点传递将信道矩阵转换为汉明矩阵；最后，利用数据集的邻接关系、图的自同构及放缩证明了差分隐私保护的隐私泄露量存在互信息上界，并提出了基于信源和信宿互信息的隐私泄露上界计算公式。本文所提出的差分隐私量化模型和方法对任意分布的原始数据、任意攻击能力的敌手、任意通信信道都成立，且限制条件少。此外，分析表明，所提出的隐私泄露上界的计算公式仅与原始数据集的属性数量、属性值数量及差分隐私预算参数有关，具有较好的函数关系，能够为差分隐私的算法隐私保护能力的评价、差分隐私泄露的风险评估提供基础。

下一步，将探究交互式差分隐私中隐私泄露与效用之间的平衡问题，并探索连续查询下的交互式差分隐私量化问题。

参 考 文 献

[1] LIANG Y, POOR H V, SHAMAI S. Information theoretic security[J]. Foundations and Trends in Communications and Information Theory, 2009, 5(4/5): 355-580.

[2] CHAKRABORTY B, SADHYA D, VERMA S, et al. Information Theoretic Analysis of Privacy in a Multiple Query-Response Based Differentially Private Framework[C]// International Conference on Communication, Networks and Computing. Singapore: Springer, 2018: 262-272.

[3] PADAKANDLA A, KUMAR P R, SZPANKOWSKI W. The Trade-off between Privacy and Fidelity via Ehrhart Theory[J]. arXiv: 1803.03611, 2018.

[4] MIRONOV I. Rényi differential privacy[C]// 2017 IEEE 30th Computer Security Foundations Symposium (CSF). IEEE, 2017: 263-275.

[5] DALENIUS T. Towards a methodology for statistical disclosure control[J]. Statistic Tidskrift, 1977, 15(2): 429-444.

[6] SWEENEY L. k-anonymity: A model for protecting privacy[J]. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002, 10(5): 557-570.

[7] MACHANAVAJJHALA A, GEHRKE J, KIFER D, et al. l-diversity: Privacy beyond k-anonymity[C]// 22nd International Conference on Data Engineering (ICDE'06). IEEE, 2006: 24-24.

[8] DWORK C. Differential privacy[M]// Encyclopedia of Cryptography and Security. Boston, MA: Springer, 2011: 338-340.

[9] DWORK C, MCSHERRY F, NISSIM K, et al. Calibrating noise to sensitivity in private data analysis[C]// Theory of cryptography conference. Berlin: Springer, 2006: 265-284.

[10] MCSHERRY F, TALWAR K. Mechanism Design via Differential Privacy[C]// IEEE 48th Annual IEEE Symposium on Foundations of Computer Science (FOCS 2007). ACM, 2007: 94-103.

[11] ROTH A, ROUGHGARDEN T. Interactive privacy via the median mechanism[C]// Proceedings of the Forty-second ACM Symposium on Theory of Computing. ACM, 2010: 765-774.

[12] HARDT M, ROTHBLUM G N. A multiplicative weights mechanism for privacy-preserving data analysis[C]// 2010 IEEE 51st Annual Symposium on Foundations of Computer Science. IEEE, 2010: 61-70.

[13] COVER T M, THOMAS J A. Information Theory and Statics [M]// Elements of Theory. Hoboken, NJ: Wiley-Blockwell, 2006: 279-355.

[14] DIAZ C, SEYS S, CLAESSENS J, et al. Towards measuring anonymity[C]// International Workshop on Privacy Enhancing Technologies. Berlin: Springer, 2002: 54-68.

[15] PENG C G, DING H F, ZHU Y J, et al. Information entropy model of privacy protection and its measurement method[J]. Journal of Software, 2016, 27(8): 1891-1903.

[16] WAGNER I, ECKHOFF D. Technical privacy metrics: a systematic survey[J]. ACM Computing Surveys (CSUR), 2018, 51(3): 1-45.

[17] HEUSSER J, MALACARIA P. Applied quantitative information flow and statistical databases[C]// International Workshop on Formal Aspects in Security and Trust. Berlin: Springer, 2009: 96-110.

[18] CLARKSON M R, SCHNEIDER F B. Quantification of integrity[J]. Mathematical Structures in Computer Science, 2015, 25(2): 207-258.

[19] ALVIM M S, CHATZIKOKOLAKIS K, DEGANO P, et al. Differential Privacy versus Quantitative Information Flow [J]. arXiv: 1012.4250, 2010.

[20] ALVIM M S, ANDRÉS M E, CHATZIKOKOLAKIS K, et al. Differential privacy: on the trade-off between utility and information leakage[C]// International Workshop on Formal Aspects in Security and Trust. Berlin: Springer, 2011: 39-54.

[21] BARTHE G, KOPF B. Information-theoretic bounds for differentially private mechanisms[C]// 2011 IEEE 24th Computer Security Foundations Symposium. IEEE, 2011: 191-204.

[22] WANG W, LEI Y, ZHANG J. On the Relation Between Identifiability, Differential Privacy, and Mutual-Information Privacy [J]. IEEE Transactions on Information Theory, 2016, 62(9): 5018-5029.

[23] CUFF P, YU L. Differential privacy as a mutual information constraint[C]// Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. ACM, 2016: 43-54.

[24] DWORK C. A firm foundation for private data analysis[J]. Communications of the ACM, 2011, 54(1): 86-95.

[25] BONDY J A, MURTY U S R. Graph theory with applications [M]. London: Macmillan, 1976.

[26] BROUWER A E, HAEMERS W H. Distance-regular graphs [M]// Spectra of Graphs. New York: Springer, 2012: 177-185.

WANG Mao-ni, born in 1994, graduate student. Her main research interests include privacy and data security.

DING Hong-fa, born in 1988, doctor, is a member of China Computer Federation. His main research interests include privacy and data security.