基于率失真的差分隐私效用优化模型

吴宁博[1] 彭长根[1),2)] 田有亮[1),2)] 牛 坤[1)] 丁红发[2),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)

摘 要 隐私保护与数据效用矛盾问题的解决方案是隐私保护领域中的一个研究热点.针对差分隐私离线数据发布场景中的隐私与效用平衡问题,利用率失真理论研究了平衡隐私与数据效用的最优化差分隐私机制.首先,基于Shannon通信理论抽象差分隐私的噪声信道模型,以互信息量与失真函数度量数据发布的隐私与效用,构建基于率失真理论的最优化模型.其次,考虑关联辅助背景知识对互信息隐私泄露的影响,提出基于联合事件的互信息隐私度量,并进一步修改率失真函数提出最小化隐私泄露模型.最后,针对 Lagrange 求解过程中计算困难性问题,基于Blahut-Arimoto 交替最小化算法提出了互信息隐私最优化信道机制的近似求解算法.通过实验仿真,验证了所提出的迭代近似计算方法的有效性.同时,实验结果表明所提出的方法比对称离散信道机制在限失真条件下互信息隐私泄露量平均降低了 21.7%,在相同的隐私容忍度条件下,数据效用提升了 38.3%.

关键词 率失真函数;隐私与效用平衡;差分隐私;互信息隐私泄露;数据效用优化

中图法分类号 TP309 **DOI 号** 10.11897/SP.J.1016.2020.01463

Utility Optimal Model for Differential Privacy Based on the Rate Distortion

WU Ning-Bo[1)] PENG Chang-Gen[1),2)] TIAN You-Liang[1),2)] NIU Kun[1)] DING Hong-Fa[2),3)]

[1)](*State Key Laboratory of Public Big Data, College of Computer Science and Technology, Guizhou University, Guiyang* 550025)

[2)](*Institute of Cryptography and Data Security, Guizhou University, Guiyang* 550025)

[3)](*College of Information, Guizhou University of Finance and Economics, Guiyang* 550025)

Abstract Privacy leakage has become a widely concerned issue and a major restricting factor for data releasing and sharing in the era of big data. This issue urgently needs effective privacy preserving mechanism to protect individual's private information while preserving the accuracy of released data. Commonly, the privacy-preserving data release is achieved by data distortion disturbance. Indeed, this will raise a problem of tradeoff between privacy protection intensity and the degree of data distortion, that is known as the problem of privacy-utility tradeoff. Recently years, the issue of privacy-utility tradeoff has become a hot topic in the field of privacy. Additionally, differential privacy (DP) as a strict privacy protection technology has been widely studied by researchers. The basic idea of DP is randomized perturbation for the purpose to release noisy data. Thus, the privacy-utility tradeoff is still the key research of differential privacy. In this paper, we mainly focus on the scenario of differential privacy offline data release. Based on Shannon's communication theory, we adopt privacy spread communication model to simulate data

收稿日期:2019-06-15;在线发布日期:2019-12-08.本课题得到国家自然科学基金(U1836205,61662009,61772008)、"十三五"国家密码发展基金(MMJJ20170129)、贵州省科技计划项目(黔科合重大专项字([2018]3001、[2018]3007、[2017]3002)、黔科合平台人才([2020]5017)、黔科合基础([2017]1045))、贵州省高等学校创新人才团队(黔教合人才团队([2013]09))及贵州省研究生科研基金立项课题(KYJJ2017005)资助.**吴宁博**,博士研究生,主要研究方向为数据安全、隐私保护等. E-mail: hn_dragon@163.com. **彭长根**(通信作者),博士,教授,博士生导师,中国计算机学会(CCF)会员,主要研究领域为密码学与信息安全、大数据隐私保护等. E-mail: peng_stud@163.com. **田有亮**,博士,教授,博士生导师,中国计算机学会(CCF)会员,主要研究领域为密码学与数据安全、理性密码协议等. **牛 坤**,博士研究生,讲师,主要研究方向为隐私保护、信息安全. **丁红发**,博士研究生,讲师,中国计算机学会(CCF)会员,主要研究方向为隐私保护、数据安全、访问控制.

releasing and model the randomized perturbation of differential privacy as a noisy channel model, i.e., a randomized probability mapping. Further, we adopt mutual information and expected distortion measure function to quantify privacy leakage and data usability. Thus, the problem of tradeoff between privacy and utility can be formalized as mutual information privacy optimization problem, which can be solved by celebrated rate-distortion theory. In this regard, Wang et al presented the mutual information privacy (PD-MIP), and utilized Karush-Kuhn-Tucker (KKT) condition to illustrate optimality criterion of mutual information privacy. And then, they proved that the optimal mutual information privacy also is close to the optimal differential privacy level when given a distortion constraint. Besides, Kalantari also pointed out that differential privacy leakage is the upper bound of mutual information privacy leakage. These researches establish a fundamental relation between mutual information privacy and differential privacy. However, they do not consider that how to obtain the optimal mutual information privacy mechanism, and the impact of related data on mutual information privacy leakage. That is, the data association of auxiliary background knowledge can help to obtain more private information, thus leading to the inference of individual's privacy with higher probability. Therefore, it is meaningful to study the impact of related data on mutual information privacy leakage. To address the problem mentioned above, we consider the influence of auxiliary background knowledge on the mutual information privacy leakage, and propose mutual information privacy metric based on joint events. Furthermore, we modify the current rate-distortion function, and propose a minimized mutual information privacy disclosure model. Finally, for the difficult problem of computing rate-distortion by using the Lagrange approach, we propose an approximate algorithm of differential privacy optimization channel mechanism based on the principle of Blahut-Arimoto algorithm. We verify the effectiveness of the proposed iterative approximate computation. The experimental results show that the amount of mutual information privacy disclosure is 21.7% which is lower than that in symmetric discrete channel mechanism under the condition of limited distortion. Meanwhile, the data utility improves 38.3% under the same privacy tolerance.

Keywords rate-distortion function; privacy-utility tradeoff; differential privacy; mutual information privacy leakage; data utility optimization

1 引 言

针对数据发布、共享及应用中存在的隐私泄露问题,Dwork[1]提出了差分隐私(Differential Privacy,DP)保护模型方法,利用随机化噪声扰动,力图解决差分攻击导致的隐私泄露问题.因差分隐私具有严格的数学理论支撑和忽略隐私攻击者的计算能力等优势,迅速成为隐私保护领域研究的热点[2-3].目前,差分隐私的研究主要集中于利用拉普拉斯(Laplace)机制[4]、高斯(Gaussian)机制[5]、指数(Exponential)机制[6]设计更有效的隐私保护算法,达到隐私度(Privacy)与查询响应精确度、可用性(Utility)之间的平衡折中,本文将其定义为经典差分隐私的研究范畴.与之对应的信息论差分隐私[7-8]逐渐成为差分隐私另一个重要的研究分支.围绕信息熵的内涵,以 Shannon 信息论基本通信模型[9]为出发点,基于量化信息流[10](Quantitative Information Flow,QIF)的思想,考虑隐私保护模型的隐私泄露风险与数据效用的量化以及平衡问题,不失为一种行之有效的解决方案.

自 Shannon 为解决信息度量问题提出熵的概念以来,基于熵的信息量化已广为科学研究者所接受[11-13].尤其近年来,利用信息熵的方法研究数据发布场景中隐私度与数据效用的平衡问题已成为新的研究方向,在隐私保护领域受到了国内外学者的高度关注.由于信息熵是一种有效可行的信息度量方法,自然地被引入到差分隐私框架中,成为量化隐私信息的基础.在此方面,已有不少学者借鉴条件熵、联合熵、互信息量、条件互信息量以及失真函数等[7-8,14-17]

表达差分隐私保护系统的隐私泄露风险和数据效用，并试图探讨最优化的信道机制问题[17-18]. 尽管如此，信息论差分隐私依然存在诸多有待深入研究和思考的问题：如差分隐私数据发布中隐私失真函数[14]获得最小信息传输率的下界以及达到最小信息率（最小化隐私泄露风险）时的最优化信道机制问题等.

现有差分隐私主要是针对数据独立情景，没有考虑敌手的先验知识和数据关联对隐私泄露的影响[19-20]. 通常情况下，数据之间存在关联且这些相关联的数据信息可以辅助识别特定个体的隐私信息[19]. 例如，临床医疗数据集中相同的属性信息可以辅助识别个体的疾病敏感信息. 这些关联信息被定义为差分隐私机制的背景知识. 也就是，如果敌手拥有一定的关联辅助背景知识，则有较高的概率获得更多的隐私信息量. 信息论方法应用于差分隐私研究中，条件互信息量能够度量敌手拥有先验背景知识条件下的隐私泄露量. 因此，从信息论视角下考虑背景知识的影响，研究最小化隐私泄露风险时的最优化信道机制问题凸显其重要性.

针对差分隐私离线数据发布场景中最小化隐私泄露风险以及可达的最优化机制问题. 本文抽象差分隐私噪声扰动机制为离散无记忆信道（Discrete Memoryless Channel，DMC）模型[21]，并以该通信模型为基础，分别引入平均互信息量和期望失真对隐私与效用（Privacy-Utility，P-U）进行度量. 在此基础上考虑关联背景知识对互信息隐私泄露的影响，从优化理论的视角将维持数据可用性约束条件下的隐私与效用平衡问题建模为多约束的一个凸优化问题. 进一步，借鉴信息率失真函数迭代计算的 Blahut-Arimoto 算法[22-23]设计本文中最优化机制的迭代收敛求解算法.

鉴于上述分析，本文的研究有别于 Dwork 经典差分隐私，是一种信息论视角的差分隐私研究，同时亦可视为经典差分隐私研究的拓广与延伸. 为此，本文立意于探讨限失真理论下平衡隐私与效用的差分隐私最优化机制问题，力图为数据发布场景中差分隐私算法设计提供一种基于信息论的理论支撑. 此外，本文的研究对差分隐私的最优化机制设计（最佳信道转移条件概率）、最佳信息传输率（率失真）以及差分隐私泄露风险评估（最小信息率）提供重要的理论支撑. 结合理论分析与实践应用，本文基于信息率失真理论对差分隐私模型中隐私与效用平衡问题的研究具有重要的理论和实践意义.

1.1　研究框架

本文利用信息论的基础方法，从量化信息流的角度考虑实现隐私与效用平衡的互信息隐私最优化机制问题. 为更好地阐述本文的研究思路，首先给出主要的研究框架示意图，如图 1 所示.

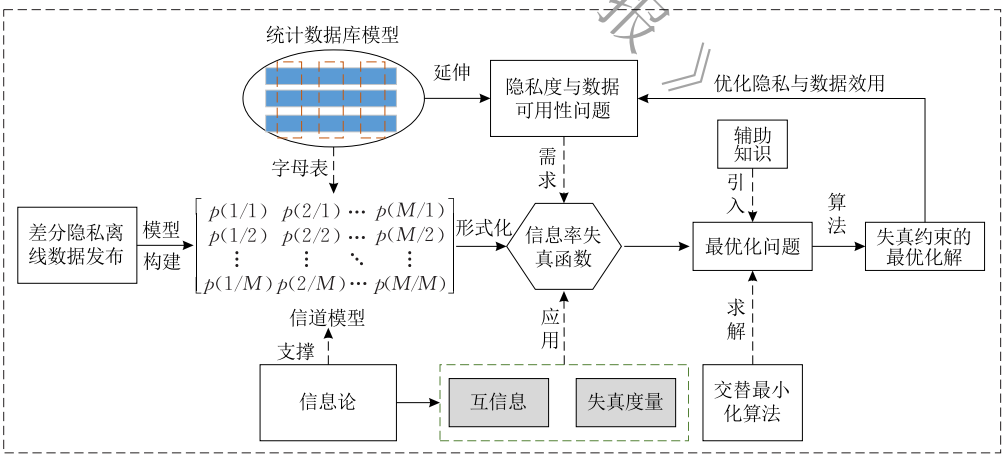

图 1　隐私与效用平衡的差分隐私最优化机制研究框架

在图 1 中，基于差分隐私离线数据发布场景为立意出发点，以信息论的方法解决数据发布中的隐私效用平衡问题为最终目标. 以下详细地阐述本文的研究思路：首先，考虑差分隐私离线数据发布，原始数据到近似数据副本的流程. 抽象原始统计数据库模型为关系模型，构建离散随机变量. 进而，从通信的角度，以信息论为基础建立差分隐私离散无记忆信道（DMC）模型. 在此基础上，针对统计数据库延伸出的隐私度与数据可用性之间的矛盾冲突问题，分别采用平均互信息量和期望失真量化隐私度与数据可用性，并以此为隐私与效用的度量基础，将隐私与效用平衡问题形式化为限失真约束条件下的最小化隐私信息泄露问题，即是信息率失真函数. 其次，基于差分隐私信道模型，引入敌手攻击的辅助背

景知识,将带有辅助背景知识条件下的隐私与效用的平衡问题形式化表述为多约束条件的一个凸优化问题.最后,结合理论分析结果,基于交替最小化算法设计本文中求解限失真约束条件下最优化问题的近似迭代求解算法,寻求最优化的噪声信道机制,实现隐私度与数据可用性之间的优化.

1.2 相关工作

国外学者针对信息论差分隐私的研究起步较早,在此本文综述了较具有代表性的研究成果.Mir[24]利用条件熵、互信息研究了隐私信息的度量问题,奠定了差分隐私的信息论基础.更重要的,Alvim 等人[8,11]几乎是最早提出基于量化信息流(QIF)的思想,将信息熵应用到差分隐私中量化隐私信息的不确定度,并从信息论的角度考虑了平衡隐私度与数据效用的方法,同时提出信息论对称信道机制能够达到理论最优性.Barthe 等人[25]利用信息熵研究了差分隐私的隐私边界问题.Cuff 等人[26]利用互信息给出了等价的隐私定义,研究了互信息差分隐私的性质及其隐私保护强度.此外,学者 Sankar 等人[13]构建了统计数据库概率模型,从最佳信源编码、译码方案的角度考虑了信息熵方法在数据库隐私与效用平衡中的应用.此后,Sankar 研究团队的成员 Kalantari 等人[7]以数据库记录为随机变量 X,且假设信源先验概率分布 $p(x)$ 未知,将信源概率分布划分为三种类别:均匀分布、单调递减分布以及其它分布情况,并以凸分析理论为基础理论方法,分别研究了汉明失真下平衡隐私度与数据效用的最佳差分隐私信道机制问题,理论分析证明了针对不同信源分布集合,信息论对称信道机制和非对称信道机制的最优性.Wang 等人[14]针对差分隐私非交互式数据发布场景,构建数据库实例为随机变量,并基于期望汉明失真从发布数据库全局层面,考虑了可辨识识别(Identifiability)、差分隐私(Differential Privacy)和互信息隐私(Mutual Information Privacy)三个不同隐私概念之间的基本联系,并基于失真测量建立了隐私失真最优化目标函数,给出了互信息隐私最优性的 KKT[27]条件,最终从理论上证明了互信息最佳机制满足给定隐私等级的 ε-差分隐私.更多的,隐私与失真问题在基于位置的隐私保护领域也得到了研究与应用.2019 年,Zhang 等人[28]在基于位置服务(Location Based Service,LBS)的隐私保护方面,基于平均互信息量和 Euclidean 距离失真度量考虑了一种在线位置轨迹发布的信息论方法.该方法考虑了隐私攻击的辅助背景知识,并基于条件互信息量解决隐私与效用平衡的率失真函数最优化问题,利用一阶马尔可夫模型设计了一个在线发布的位置隐私保护机制(LPPM),获得了隐私与数据可用性的最佳权衡.此外的研究者 Calmon 和 Fawaz[29]以及Makhdoumi 等人[30]针对隐私统计推断,利用互信息隐私度量将最佳的隐私保护机制形式化为率失真问题.

近几年来,国内学者也逐步开始关注信息论与隐私保护结合的研究点.彭长根等人[31]借鉴 Shannon 基本通信框架,结合信息论抽象隐私信息传播的基本通信模型,提出了隐私保护的基本信息熵模型、含敌手攻击的隐私信息熵模型和主观感受的加权隐私信息熵模型,并进一步提出了利用信息熵的方法度量隐私保护机制的信息泄露风险,为隐私保护效果评价提供了一种基于信息论的理论参考.此外,吴振强等人[32]提出了利用网络结构熵的方法度量社交网络的差分隐私保护效果.由此来看,信息论与差分隐私的结合已逐渐成为新兴的交叉研究方向,但目前国内学者针对该领域的研究尚处于起步状态.

已有的研究工作中,基于信息论的方法考虑了隐私与效用平衡的问题,并从理论上分析了保障差分隐私最优化机制存在的数学优化条件.但是,现有研究工作尚未给出获得最优化机制条件的理论实践方法,且未考虑相关联数据信息,即拥有背景知识对互信息隐私泄露风险的影响.

1.3 本文工作

鉴于上述分析可知,信息论的方法应用于解决差分隐私数据发布场景中隐私与数据效用的平衡问题具有理论可行性.本文考虑互信息与失真度量的特性,立足于 Shannon 经典信息论框架,借鉴限失真编码定理[33]的信息率失真函数,立意研究解决差分隐私离线数据发布场景中满足限失真约束的最小信息传输率,探讨最小化互信息隐私泄露风险的最优化机制问题.

由于差分隐私随机化噪声机制具有不确定性,且存在隐私与效用的极大极小对立关系,因此,基于信息论的方法如何从噪声信道的角度表述差分隐私机制,并将隐私与效用的矛盾冲突问题刻画为多约束条件的最优化求解问题是一个关键;其次,基于信息率失真函数的最优化问题,利用 Lagrange 乘子法求解最优分布是困难的,近似求解的 Blahut-Arimoto 算法是一种解决方案,并且理论上已经证明了算法计算的收敛性[21],但其关键是如何将带敌手背景知识攻击下隐私与效用平衡问题转变为信息论的双

重最小化可求解问题；最后，基于迭代近似求解方法，求解最优化信道条件转移概率也是一个关键问题.

围绕本文的预期目标，需解决上述几个关键问题，支撑信息论差分隐私最优化机制的研究. 本文的研究工作解决了差分隐私离线数据发布中平衡隐私与效用的最优化信道机制问题. 对比现有的相关研究工作，本文工作的主要研究贡献可归纳为：

(1) 考虑差分隐私离线数据发布场景，结合差分隐私的离散无记忆信道(DMC)模型，基于互信息的隐私风险度量和数据失真的效用度量提出了消息联合互信息隐私度量，建立了差分隐私数据发布与信息论通信模型之间的基本联系.

(2) 围绕信息率失真函数，将差分隐私数据发布场景中隐私泄露风险与数据效用平衡建模为带失真约束的最优化问题. 在此基础上考虑了辅助背景知识攻击下的隐私与效用平衡问题，建立拥有辅助背景知识的差分隐私数据发布隐私与效用优化模型，并进一步考虑了求解差分隐私最小信息率的计算方法.

(3) 基于 Blahut-Arimoto 迭代最小化算法设计了计算最小化互信息隐私泄露的迭代算法，解决如何利用迭代求解算法计算差分隐私数据发布的最优信道条件转移概率，为差分隐私最优化机制设计提供了一种基于信息论的支撑.

本文第 2 节介绍与本文主要研究内容密切相关的基础知识；第 3 节介绍本文内容的模型与符号表示，包含统计数据库模型与差分隐私噪声信道模型；第 4 节分别给出差分隐私数据发布场景中隐私与效用的信息论度量方法，提出本文的研究问题，并将其形式化表述为一个多约束条件的数学优化问题；第 5 节针对本文的研究问题，利用 Lagrange 乘子法、交替最小化算法设计出近似求解最小化互信息隐私泄露算法，计算最优化信道机制；第 6 节对提出的模型方法以及算法进行实验验证与分析；第 7 节总结本文的主要研究工作，并指出下一步的研究工作安排.

2　基础知识

本节分别介绍与本文研究内容相关的基础知识，主要包含有差分隐私的基本概念、最优化问题及信息率失真理论.

2.1　差分隐私

首先，回顾一下经典差分隐私的严格数学定义. 如果一个随机化噪声响应机制满足 ε-差分隐私，则其作用于至多相差一条数据记录的相邻兄弟数据集 $\mathcal{D}$ 和 $\mathcal{D}'$，满足查询输出空间 $\mathcal{S}\subseteq\mathcal{R}$ 概率的 ε 不可区分性. 经典差分隐私的定义忽视了敌手的计算能力和辅助知识，提供强隐私保护保障. 以下正式地将其定义为：

定义 1[1]. 一个随机化机制 $\phi:\mathbb{R}^n\to\mathbb{R}$ 称之为 ε-差分隐私，当且仅当对于任意的相邻输入数据集 $\mathcal{D}$ 和 $\mathcal{D}'$，满足 $\|\mathcal{D}-\mathcal{D}'\|_1\leq 1$ 和 $\mathcal{S}\subseteq\mathcal{R}$，

$$\Pr[\phi(\mathcal{D})\in S]\leq e^{\varepsilon}\Pr[\phi(\mathcal{D}')\in S] \tag{1}$$

其中，隐私参数 ε 决定差分隐私保护强度.

2.2　优化问题

考虑一个标准形式的优化问题[27]：

$$\begin{aligned}&\min\ f_0(x)\\&\text{s.t.}\ \ f_i(x)\leq 0,\ i\in\{1,2,\cdots,m\}\\&\qquad h_j(x)=0,\ j\in\{1,2,\cdots,p\}\end{aligned} \tag{2}$$

其中，$x\in\mathbb{R}^n$，优化问题(2)的最优解 p^* 满足

$$\begin{aligned}p^*=\inf\{f_0(x)\mid f_i(x)\leq 0,\ h_j(x)=0\},\\ i\in\{1,2,\cdots,m\},j\in\{1,2,\cdots,p\}\end{aligned} \tag{3}$$

构造 Lagrange 对偶函数对上述最优化问题进行求解，

$$L(x,\lambda,\mu)=f_0(x)+\sum_{i=1}^{m}\lambda_i f_i(x)+\sum_{j=1}^{p}\mu_j h_j(x) \tag{4}$$

其中 λ_i 为第 i 个不等式约束 $f_i(x)\leq 0$ 对应的 Lagrange 乘子；相似的，μ_j 称为第 j 个等式约束 $h_j(x)=0$ 对应的 Lagrange 乘子. 求解最优性条件得到目标函数的最小值.

2.3　率失真理论

由于通信系统中固有的噪声和干扰，Shannon 离散无记忆信道(DMC)是信源字母表 $\mathcal{X}$ 到再生字母表 $\hat{\mathcal{X}}$ 的条件概率映射. 大多数情况下，再生字母表 $\hat{\mathcal{X}}$ 与信源字母表 $\mathcal{X}$ 相同. 互信息量作为信息论中一个重要的概念，定量地描述了信宿获得有关信源的信息量，可用于测度整体上流经信道的平均信息量. 同时，互信息可以理解为信道的信息率，即码率.

定义 2. 给定两个随机变量 X 和 $\hat{X}$，对于所有的 $x\in\mathcal{X}$ 和 $\hat{x}\in\hat{\mathcal{X}}$，条件概率分布 $q(\hat{x}\mid x)$ 与边缘概率分布 $p(x)$ 和 $q(\hat{x})$，定义

$$I(X;\hat{X})=\sum_{x}\sum_{\hat{x}}p(x)q(\hat{x}\mid x)\log\frac{q(\hat{x}\mid x)}{q(\hat{x})} \tag{5}$$

度量随机变量 $\hat{X}$ 包含 X 的平均信息量.

定义 3[21]. 失真度量是一个非负函数 $d:\mathcal{X}\times\hat{\mathcal{X}}\to\mathbb{R}^+$ 映射信源字母表和再生字母表乘积空间到

一个非负实数.

对所有的 $x\in\mathcal{X}$ 和 $\hat{x}\in\hat{\mathcal{X}}$,失真函数 $d(x,\hat{x})$ 刻画了符号 $\hat{x}$ 表示 x 时的失真距离,定量地描述信道的统计特性.给定期望失真门限 δ,在率失真区域 (R,δ) 中所有码率 R 的下确界称为率失真函数 $R(\delta)$.Shannon 保真度准则(限失真编码定理)刻画了通信系统中满足信源失真约束条件下的最小信息传输率,反映数据压缩的临界极值[21].形式化地将 $R(\delta)$ 定义为:

定义 4[21,33]. 对于给定的失真门限 δ 和失真度量函数 $d(x,\hat{x})$,在满足 (R,δ) 的失真区域,所有码率 R 的下确界 $R(\delta)$

$$R(\delta)=\min_{q(\hat{x}|x):\sum_x\sum_{\hat{x}}p(x)q(\hat{x}|x)d(x,\hat{x})\le\delta} I(X;\hat{X}) \quad (6)$$

其中,$p(x)$ 和 $q(\hat{x}|x)$ 分别为信源概率分布和信道条件概率分布.

3 模型与符号

本节首先给出本文的一些基本模型假设,具体描述了统计数据库模型、差分隐私信道模型.表 1 列举出了本文中所使用的符号并给出具体的符号说明.

表 1 符号说明

符号	说明
$\mathcal{D},\mathcal{D}'$	相邻兄弟数据库
δ	期望失真门限
$\bar{\delta}$	平均失真度
R	码率,信息率
$\mathcal{X},\hat{\mathcal{X}},x,\hat{x}$	$\mathcal{X},\hat{\mathcal{X}}$ 符号字母表,$x,\hat{x}$ 表示初始数据库及合成数据库实例
$p(\cdot),r(\cdot,\cdot)$	分别表示先验概率分布,联合概率分布
$q(\cdot,\cdot\|\cdot),p(\cdot,\cdot)$	条件概率分布以及联合概率分布
$\|\cdot\|,\|\cdot\|_2$	集合基数,欧几里得距离 2 范数
$\varepsilon,\varepsilon^*$	差分隐私预算参数,信息论差分隐私最小的隐私参数
$d(\cdot,\cdot)$	距离度量函数
$D_{\mathrm{KL}}(\cdot\|\cdot)$	相对熵距离度量

3.1 统计数据库模型

数据库 $\mathcal{D}$ 是由个体记录组成的集合,每个记录是有关特定个体的多维属性描述信息.例如,一个临床医疗数据库通常包含有姓名、性别、社会保障号以及一些敏感属性集合.在这些属性集合中,通常有些属性值是敏感的隐私信息,应当属于受保护的私有隐私信息.

从关系代数的角度,本文中考虑将数据库建模为关系集合.正式的,一个数据库 $\mathcal{D}$ 含有 n 个元组,元组包含 k 个属性,记为 $\mathcal{D}=\{d_1,\cdots,d_n\}$,其中任意记录 $d_i(1\le i\le n)$ 都是数据取值域全集 $\mathcal{X}$ 的一个实例,$\mathcal{X}$ 表示所有 k 个属性集合元素的笛卡儿积,基数 $|\mathcal{X}|=M$,其中 $M\ge2$,表示数据库的离散 M 个原子符号.离散随机变量 X 表示数据库,取值于数据库域 $\mathcal{X}$.对于任意两个数据库 $x,x'\in\mathcal{X}$,汉明距离 $d(x,x')$ 度量两个数据库实例之间不同记录数目,如果满足 $d(x,x')\le1$,则称 x,x' 为相邻的兄弟数据库,记作 $x\sim x'$.

差分隐私的离线数据发布场景中,随机化噪声机制接收原始数据输入并输出原始数据库的近似副本.从 Shannon 信息论的视角,可以将其抽象为输入字母表 $\mathcal{X}$ 到输出再生字母表 $\hat{\mathcal{X}}$ 的随机化函数映射.基于这样的一个基本假设,可以确保针对统计数据库的差分隐私保护模型被抽象为信息论的信道模型.

3.2 差分隐私信道模型

差分隐私的数据发布框架中,随机化噪声机制是一个算法 ϕ,其接受输入原始数据库 $\mathcal{D}$,并产生输出响应噪声结果 $\phi(\mathcal{D})\to\mathbb{R}^n$[5].差分隐私保护机制的噪声扰动是一个随机化映射,类似于信息论中的噪声信道[34].例如二进制数据集 $\mathcal{D}=\{0,1\}^n$,具有先验概率分布 $p_x(0)$ 和 $p_x(1)$,基于差分隐私的随机化响应机制,可以将输入与输出关系表示为二进制交织信道模型.此外对于类别型数据(如城市地址、疾病类别等)均属于离散类型,诸如此类的有限离散集合之间的随机化映射可使用条件概率分布形式描述.基于噪声不确定性,可以借鉴离散的噪声信道模型对差分隐私进行研究.

为了建立差分隐私与信息论之间的一些基本联系,文中考虑将差分隐私的离线数据发布框架建模为经典的 Shannon 点到点的通信模型.也即是,差分隐私输入原始数据作为信源,输出噪声合成数据作为信宿,差分隐私保护机制构成噪声信道.正式地将其表述为,$\mathcal{X}$ 和 $\hat{\mathcal{X}}$ 为离散有限字符元素集合,考虑输出仅依赖于所对应的输入,与先前输入或输出字符条件独立.因此,抽象差分隐私为离散无记忆噪声信道,用条件概率 $q(\hat{x}|x)$ 的矩阵形式表示随机噪声扰动

$$q_{\hat{X}|X}(\hat{x}|x)=\begin{bmatrix} q(1|1) & q(2|1) & \cdots & q(M|1)\\ q(1|2) & q(2|2) & \cdots & q(M|2)\\ \vdots & \vdots & \ddots & \vdots\\ q(1|M) & q(2|M) & \cdots & q(M|M)\end{bmatrix} \quad (7)$$

噪声信道条件概率 $q(\hat{x}|x)$ 的统计特性反映了差分隐私的随机化特征.据此,从噪声信道条件概率的视

角将差分隐私定义重新表述.

定义 5[8,14]. 一个离散无记忆信道机制$\mathcal{C}:\mathcal{X}\times\hat{\mathcal{X}}\to\mathbb{R}^{+}$满足 ε-差分隐私(ε-DP)，当且仅当对所有的 $x,x'\in\mathcal{X}$ 和 $\hat{x}\in\hat{\mathcal{X}}$ 且 $x\sim x'$ 满足

$$q(\hat{x}|x)\leq e^{\varepsilon}q(\hat{x}|x') \tag{8}$$

信息论差分隐私定义独立于信源概率分布，反映了信道条件概率的统计特性. 差分隐私的隐私预算参数 ε 由信道条件概率 $q(\hat{x}|x)$ 的比值决定，刻画了信道统计特性. 由此定义差分隐私参数，

$$\varepsilon^{*}\triangleq\inf\left\{\varepsilon:\ln\frac{q(\hat{x}|x)}{q(\hat{x}|x')}\right\},x,x'\in\mathcal{X}\text{ 和 }\hat{x}\in\hat{\mathcal{X}} \tag{9}$$

噪声信道条件概率与信息传输率和失真紧密联系，从满足失真约束的角度，可以定义(ε,δ)可达信道集合.

定义 6. 固定信源概率分布 $p(x)$，期望失真满足给定失真门限 δ 的所有信道机制集合称为(ε,δ)可达试验信道，即是

$$(\varepsilon,\delta)\triangleq\left\{q(\hat{x}|x):\sum_{x,\hat{x}}p(x)q(\hat{x}|x)d(x,\hat{x})\leq\delta\right\} \tag{10}$$

基于定义 6 可知本文的(ε,δ)可达信道是满足差分隐私与限失真的试验信道交集. 由期望失真约束(式(10))可知可行集是超平面的闭半空间，因半空间是凸的[27]，且信道条件概率满足

$$\sum_{\hat{x}}q(\hat{x}|x)=1,q(\hat{x}|x)\geq0,\forall x\in\mathcal{X},\hat{x}\in\hat{\mathcal{X}} \tag{11}$$

依据凸集合交集的保凸运算规则，(ε,δ)可达信道集合依然是凸集. 此外，差分隐私预算参数计算(式(9))的函数由下水平集的角度已经证明是拟凸函数[34]. 因此，本文从噪声信道的角度，旨在利用凸可行集中寻找最佳条件概率分布解决研究问题.

此外，由于信道条件概率的不确定性，差分隐私的信道模型与隐私泄露和数据效用密切相关. 直观地说，信息论的方法可以用来量化差分隐私信道模型中的隐私信息流. 基于此，本文中第 4 节给出了隐私泄露与数据效用的信息论量化方法.

4　问题形式化

本节中首先给出隐私泄露与数据效用的信息论量化方法，进一步考虑辅助背景知识对隐私泄露的影响，将差分隐私离线数据发布中的隐私与效用的平衡问题形式化为多约束条件的最优化问题，以下详细地给出研究问题的描述.

4.1　隐私度量

差分隐私定义了输出概率比值的 ε-不可区分性，提供了不依赖于原始数据先验分布的强隐私保障. 然而，一些应用中敌手可能拥有关于原始数据的先验分布知识. 在此情景下，与差分隐私等价定义的语义安全差分隐私考虑了敌手先验分布与后验信念的概率分布距离，依此测量隐私泄露. 通常情况下，Kullback-Leibler(KL)散度可以用来度量两个概率分布之间的距离[21]. 在此具体场景下，定义敌手后验信念概率分布 $p_{X|\hat{X}}(x|\hat{x})$ 与先验分布 $p_X(x)$ 的 KL 距离度量平均意义上的隐私泄露量. 事实上，KL 距离量化后验贝叶斯概率分布和先验概率分布距离，其等效于互信息量 $I(X;\hat{X})$[29]，

$$I(X;\hat{X})\triangleq\mathbb{E}_{\hat{X}}D_{\mathrm{KL}}[p_{X|\hat{X}}\|p_X] \tag{12}$$

其中，$D_{\mathrm{KL}}[\cdot\|\cdot]$是 KL 距离函数.

互信息度量近似合成数据 $\hat{X}$ 包含原始数据 X 的信息量. 因此，可以定义互信息量度量差分隐私离线数据发布框架中的隐私泄露风险. 此外，互信息 $I(X;\hat{X})=H(X)-H(X|\hat{X})$，依据熵与互信息的关系，互信息 $I(X;\hat{X})$ 刻画观测到 $\hat{X}$ 后有关随机变量 X 不确定度的减少量，条件熵 $H(X|\hat{X})$ 表述随机变量仍具有的不确定度. 对于固定信源概率分布，$H(X)$ 固定. 由此，隐私泄露风险度量的互信息与条件熵可视为等价可替换的量化方法[28].

在本文中我们采用上述式(12)的形式度量隐私泄露量. 平均互信息量是先验分布与条件概率分布的二元函数，即是关于原始数据概率分布 $p(x)$ 的凸函数和差分隐私噪声信道转移条件概率 $q(\hat{x}|x)$ 的凹函数.

4.2　效用度量

差分隐私离线数据发布框架中依赖距离度量近似副本与原始数据之间的失真程度，采用非负的失真函数 $d(x,\hat{x})$ 度量发布数据集的整体数据可用性，即数据效用.

常用的失真函数有汉明距离和欧氏距离. 欧氏距离常用于数值型坐标距离计算. 然而，汉明距离度量函数可视为一种示性函数，具有测度敏感度高的特性，能够准确地表达出近似合成输出数据集与原始数据集之间记录变化数. 因此，本文中采用汉明距离 $d(x,\hat{x})$ 度量两个数据库 $x\in\mathcal{X}$ 和 $\hat{x}\in\hat{\mathcal{X}}$ 之间不同的数据记录数. 同时，为表达平均意义上的失真程度，定义期望汉明失真

$$\mathbb{E}[d(X,\hat{X})]=\sum_{x}\sum_{\hat{x}}p(x)q(\hat{x}|x)d(x,\hat{x}) \quad (13)$$

度量平均意义的失真程度. 由式(13)可知,期望汉明失真是有关原始数据先验分布、差分隐私信道条件概率和失真度量的函数. 本文中,我们基于期望汉明失真度量差分隐私离线数据发布近似数据 $\hat{x}$ 与原始数据 x 的效用,就是同时考虑了数据集先验分布与差分隐私条件概率信道的影响因素.

4.3 隐私效用平衡问题

采用互信息量度量差分隐私数据发布的隐私泄露,期望汉明失真量化发布数据与原始数据的失真程度,也即是数据的可用性. 直观上来说,差分隐私数据发布的隐私效用是一对极大极小问题. 根据隐私保护中的隐私效用原则,隐私效用的平衡问题是一个在满足数据可用性(限失真约束)条件下的一个最小值优化问题. 采用隐私与失真函数将隐私与效用的平衡问题形式化为问题 1 的表述形式.

问题 1. 差分隐私信道机制 $\mathcal{C}:\mathcal{X}\times\hat{\mathcal{X}}\to\mathbb{R}^{+}$ 获得最小的互信息泄露量,当且仅当对于给定的原始数据 X、发布近似合成数据 $\hat{X}$ 以及数据效用约束满足期望失真门限 $\bar{\delta}\leq\delta$ 的条件下,信道条件概率 $q(\hat{x}|x)$ 是以下凸优化问题的最优解:

$$\begin{aligned}&R(\delta)=\min_{q(\hat{x}|x)}I(X;\hat{X})\\&\text{s.t.}\sum_{x}\sum_{\hat{x}}p(x)q(\hat{x}|x)d(x,\hat{x})\leq\delta\\&\quad\sum_{\hat{x}}q(\hat{x}|x)=1\end{aligned} \quad (14)$$

其中,$I(X;\hat{X})$ 和 $\sum_{x}\sum_{\hat{x}}p(x)q(\hat{x}|x)d(x,\hat{x})\leq\delta$ 分别是上述优化问题的优化目标函数和数据效用约束条件.

上述问题 1 提供了隐私与效用平衡的一般性框架,上式(14)中最优化问题等价于信息率失真函数 $R(\delta)$,是关于信道条件概率的最小值问题. 信息率失真函数约束在满足失真门限条件下,获得最小互信息隐私泄露的差分隐私噪声扰动机制,等同于差分隐私数据发布维持数据效用前提下,最小化互信息隐私泄露量.

但是,上述隐私与失真函数没有考虑差分隐私离线数据发布场景中辅助背景知识对隐私泄露的影响. 实际中,由于相关联数据导致较高概率的隐私泄露,存在相关联差分隐私问题. 我们将关联辅助识别的数据表示为辅助背景知识 Z,并在本文中考虑辅助的背景知识 Z 对隐私泄露程度的影响. 更具体地说,根据辅助背景知识 Z 是否为隐私保护数据发布者和隐私攻击者先验已知的共同知识,将该问题划分为两种情形考虑:

(1) 假设辅助背景知识 Z 为隐私保护数据发布者和隐私攻击者(敌手)共同已知的背景知识. 在此情形下,基于互信息的隐私泄露度量演变为给定 Z 的条件下 X 与 $\hat{X}$ 的条件互信息量,也即是 $I(X;\hat{X}|Z)$. 与此同时,问题 1 中的目标函数演变为求解变量 $q(\hat{x}|x,z)$ 使得条件互信息 $I(X;\hat{X}|Z)$ 达到条件最小值.

(2) 假设 Z 为隐私攻击者(敌手)可通过外部数据源获得的辅助背景知识,数据发布者并不能准确地知道辅助背景知识,但是有关辅助背景知识具有一定的数据统计信息. 采用平均互信息量化该情形下的隐私泄露度量演变为 X 与 $\hat{X},Z$ 联合的互信息量,即是 $I(X;\hat{X},Z)$. 由此,问题 1 中的最优化问题演变为求解条件概率 $q(\hat{x},z|x)$ 使得目标函数 $I(X;\hat{X},Z)$ 达到最小值.

差分隐私离线数据发布框架公开发布原始数据集的合成副本,抽象差分隐私数据处理为原始数据到噪声扰动发布合成副本的过程. 假设隐私攻击的恶意敌手可公开获得合成数据副本,并关联可用的背景知识对原始数据进行推断攻击. 本文中,我们考虑隐私攻击者可获得的辅助背景知识 Z 为隐私攻击者的单边背景知识. 从隐私通信传播的角度,我们将数据发布的场景抽象为图 2 所示的系统模型架构.

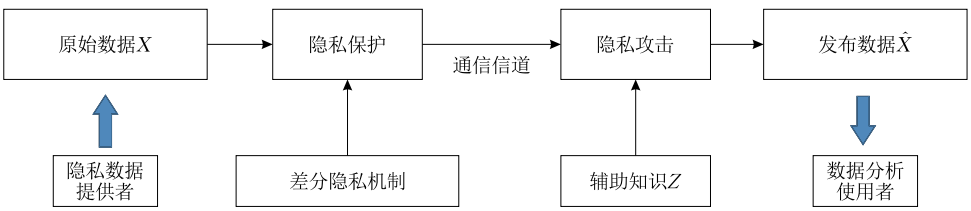

图 2 含辅助背景知识攻击的差分隐私数据发布模型

定理 1.　随机变量 X 与 $\hat{X}$，Z 的联合互信息量 $I(X;\hat{X},Z)$ 不小于互信息 $I(X;\hat{X})$.

证明.　由随机变量 X 与联合事件 X,Z 的互信息量定义，有

$$\begin{aligned}
&I(X;\hat{X},Z)\\
&=\sum_{x}\sum_{\hat{x}}\sum_{z}p(x,\hat{x},z)\log\frac{q(x|\hat{x},z)}{p(x)}\\
&=\sum_{x}\sum_{\hat{x}}\sum_{z}p(x,\hat{x},z)\log\left[\frac{q(x|\hat{x},z)}{q(x|\hat{x})}\cdot\frac{q(x|\hat{x})}{p(x)}\right]\\
&=\sum_{x}\sum_{\hat{x}}p(x,\hat{x})\log\frac{q(x|\hat{x})}{p(x)}+\\
&\quad\sum_{x}\sum_{\hat{x}}\sum_{z}p(x,\hat{x},z)\log\frac{q(x|\hat{x},z)}{q(x|\hat{x})}\\
&=I(X;\hat{X})+I(X;Z|\hat{X})\\
&\geq I(X;\hat{X})
\end{aligned}\tag{15}$$

即等于互信息量 $I(X;X)$ 与条件互信息量之和，此外，由于平均互信息的非负性质，故有上述结论成立.

证毕.

由此可知，从互信息隐私泄露量的角度，关联辅助背景知识 Z 可导致更多的隐私泄露量. 基于此，我们考虑将含关联辅助背景知识 Z 的差分隐私数据发布中隐私效用平衡问题表示为如下的定义形式.

定义 7.　对于给定的原始数据 X，差分隐私信道机制 $\mathcal{C}:\mathcal{X}\times\hat{\mathcal{X}}\to\mathbb{R}^{+}$ 以及发布近似合成数据 $\hat{X}$. 在给定的失真度量函数 $d(x,\hat{x})$ 满足期望失真约束 $\bar{\delta}\leq\delta$ 条件下，差分隐私可达的隐私泄露最小信息率为 $R_z(\delta)$. 其中，$R_z(\delta)$ 为下述问题 2 的最优值.

问题 2.　考虑图 2 所示的差分隐私离线数据发布模型，关联辅助背景知识 Z 为隐私攻击者完全知识，数据发布者拥有统计知识. 在给定失真度量函数 $d(x,\hat{x})$，满足期望失真效用约束 $\bar{\delta}\leq\delta$，条件概率 $q(\hat{x},z|x)$ 是可达互信息隐私泄露 $R_z(\delta)$ 的最优点，即是下述凸优化问题的最优解，

$$\begin{aligned}
&R_z(\delta)=\min_{q(\hat{x},z|x)}I(X;\hat{X},Z)\\
&\text{s.t.}\sum_{x,\hat{x},z}p(x)q(\hat{x},z|x)d(x,\hat{x})\leq\delta\\
&\quad\sum_{\hat{x},z}q(\hat{x},z|x)=1
\end{aligned}\tag{16}$$

其中，$I(X;\hat{X},Z)=\sum_{x,\hat{x},z}p(x,\hat{x},z)\log\frac{q(\hat{x},z|x)}{p(\hat{x},z)}$.

上述问题 2 即是互信息隐私泄露 $I(X;\hat{X},Z)$ 达到最小值的 $q(\hat{x},z|x)$ 求解问题. 换言之，上述优化问题(16)的最优解就是满足期望失真的效用约束条件下，使得互信息隐私泄露达到最小值.

5　差分隐私最优化机制

本节中围绕上述问题，基于 Lagrange 对偶函数介绍隐私与效用平衡的最优化问题求解方法. 在分析计算最优条件概率分布困难性基础上，提出基于 Blahut-Arimoto 算法设计差分隐私数据发布场景中最优化机制的迭代近似求解算法.

5.1　隐私效用平衡的最优解

最优化问题(16)是满足期望失真约束和条件概率分布约束条件下，关于凸函数的标准最小化问题，可利用 Lagrange 乘子法求解. 首先构造以下泛函 $L(q)$，

$$\begin{aligned}
L(q)=&\sum_{x}\sum_{\hat{x},z}p(x)q(\hat{x},z|x)\log\frac{q(\hat{x},z|x)}{q(\hat{x},z)}+\\
&\lambda\sum_{x}\sum_{\hat{x},z}p(x)q(\hat{x},z|x)d(x,\hat{x})+\\
&\sum_{x}\upsilon(x)\sum_{\hat{x},z}q(\hat{x},z|x)
\end{aligned}\tag{17}$$

$L(q)$ 关于 $q(\hat{x},z|x)$ 求偏导，得到如下表达式，

$$\begin{aligned}
\frac{\partial L(q)}{\partial q(\hat{x},z|x)}=&p(x)\log\frac{q(\hat{x},z|x)}{q(\hat{x},z)}+p(x)-\\
&\sum_{x'}p(x')q(\hat{x},z|x')\frac{1}{q(\hat{x},z)}p(x)+\\
&\lambda p(x)d(x,\hat{x})+\upsilon(x)\\
=&0
\end{aligned}\tag{18}$$

由于先验概率分布 $p(x)\geq 0$，可以计算得到条件概率 $q(\hat{x},z|x)$ 的参量表达式，

$$q(\hat{x},z|x)=\frac{q(\hat{x},z)e^{-\lambda d(x,\hat{x})}}{\sum_{\hat{x},z}q(\hat{x},z)e^{-\lambda d(x,\hat{x})}}\tag{19}$$

通常情况下，直接联合方程组解出最优条件概率分布比较困难. Blahut 和 Arimoto[22-23] 提出了针对上述问题的迭代求解算法，该算法是计算两个概率分布凸集之间最小相对熵距离的一种特殊情形. 算法在两个凸集之间交替最小化距离，迭代计算的最小化过程存在一个极限，且极限为率失真 $R(\delta)$. 为描述优化问题 2 的求解算法，给出以下预处理流程.

5.2　交替最小化算法

给定两个凸集 A 和 B 及其 Euclid 范数距离，目标是计算它们之间的最小距离. 首先，对于任意的 $a\in A$，计算 $b\in B$ 使得 $\min\|a-b\|_2$. 进一步，固定 b 计算集合 A 中与它最近的点，重复该过程，随着重复次数增加，距离收敛于两个集合的最小值[21].

特别地，如果 A 和 B 是概率分布集合，在相对熵(Kullback-Leibler，KL)距离度量中，上述求解最小距离的算法收敛到两个概率分布集合之间的最小相对熵[35].

基于上述算法的迭代原理求解本文中的凸优化问题. 首先，需要将其改写为两个概率分布集合之间相对熵最小化的形式，为此给出以下引理 1.

引理 1. 设 $p(x)q(\hat{x},z|x)$ 是给定的联合分布，使得最小化相对熵 $D(p(x)q(\hat{x},z|x)\|p(x)r(\hat{x},z))$ 的分布 $r(\hat{x},z)$ 是对应于条件概率 $q(\hat{x},z|x)$ 的边缘分布 $r^*(\hat{x},z)$，也即是

$$D_{\mathrm{KL}}(p(x)q(\hat{x},z|x)\|p(x)r^*(\hat{x},z)) = \min_{r(\hat{x},z)} D_{\mathrm{KL}}(p(x)q(\hat{x},z|x)\|p(x)r(\hat{x},z)) \quad (20)$$

其中，$r^*(\hat{x},z)=\sum_x p(x)q(\hat{x},z|x)$.

证明. 依据相对熵的定义，构造等式有

$$\begin{aligned}
&D_{\mathrm{KL}}(p(x)q(\hat{x},z|x)\|p(x)r(\hat{x},z))-\\
&D_{\mathrm{KL}}(p(x)q(\hat{x},z|x)\|p(x)r^*(\hat{x},z))\\
&=\sum_{x,\hat{x},z} p(x)q(\hat{x},z|x)\log\frac{p(x)q(\hat{x},z|x)}{p(x)r(\hat{x},z)}-\\
&\quad\sum_{x,\hat{x},z} p(x)q(\hat{x},z|x)\log\frac{p(x)q(\hat{x},z|x)}{p(x)r^*(\hat{x},z)}\\
&=\sum_{x,\hat{x},z} p(x)q(\hat{x},z|x)\log\frac{r^*(\hat{x},z)}{r(x,z)}\\
&=\sum_{\hat{x},z} r^*(\hat{x},z)\log\frac{r^*(\hat{x},z)}{r(x,z)}\\
&=D_{\mathrm{KL}}(r^*(\hat{x},z)\|r(\hat{x},z)) \qquad (21)
\end{aligned}$$

相应的，依据相对熵的非负性，故有上式≥ 0 的结论. 进一步，当满足条件时，上述等式$=0$ 成立.

证毕.

基于引理 1 和平均互信息 $I(X;\hat{X},Z)$ 的计算公式，可以将问题 2 中的最优化问题表述为如下形式的一个基于相对熵距离的双重最小化问题，即

$$\begin{aligned}
&\min_{q(\hat{x},z|x);\bar{\delta}\leq\delta} I(X;\hat{X},Z)\\
&=\min_{r(\hat{x},z)}\min_{q(\hat{x},z|x);\bar{\delta}\leq\delta}\sum_{x,\hat{x},z} p(x)q(\hat{x},z|x)\log\frac{q(\hat{x},z|x)}{r(\hat{x},z)} \qquad (22)
\end{aligned}$$

针对上述(20)中的双重最小化问题，利用交替最小化算法近似求得最优解. 首先，如果集合 A 为边际分布 $p(x)$ 满足失真限制的所有联合概率分布 $p(x,\hat{x},z)$ 的集合，B 为乘积分布 $p(x)r(\hat{x},z)$ 构成的集合，则对于任意的分布 $r(\hat{x},z)$，则式(20)中的双重最小化可以表述为以下形式，

$$\min_{q(\hat{x},z|x);\bar{\delta}\leq\delta} I(X;\hat{X},Z)=\min_{q\in B}\min_{p\in A} D_{\mathrm{KL}}(p\|q) \qquad (23)$$

将问题 2 的凸优化问题，转化为了在两个概率分布集合中寻找最小相对熵距离的双重最小化问题. 借鉴计算率失真函数的方法[35-36]，设计迭代交替最小化算法近似求解本文中问题 2 中的最优化问题. 以下部分详细阐述本文中的最优化机制的迭代近似计算方法.

5.3 最优化机制的迭代计算方法

针对问题 2 中描述的具有多约束条件的凸优化问题，我们考虑利用两个概率分布集合之间相对熵距离双重最小化的迭代计算方法求解. 本文中，基于 Blahut-Arimoto 算法思想设计计算最优输出分布的交替最小化算法. 算法接受预设参数输入 Lagrange 乘子 λ、数据库先验概率分布 $p(x)$、汉明失真矩阵 $\boldsymbol{d}(x,\hat{x})$ 以及算法收敛门限阈值 T. 根据交替最小化算法计算流程，首先，初始化输出联合概率分布 $r_0(\hat{x},z)$ 为均匀分布. 其次，利用 Lagrange 乘子法得到的条件概率表达式(式(19))计算条件概率分布 $q_0(\hat{x},z|x)$. 进一步，利用引理 1 计算输出联合分布 $r(\hat{x},z)$，并以该分布为基础进行下一轮迭代计算，直到互信息隐私泄露量收敛于门限阈值 T. 最后，算法输出可达最小互信息隐私泄露量、条件概率分布 $q(\hat{x},z|x)$ 以及对应的期望汉明失真. 算法 1 详细描述了最小化互信息隐私泄露量的计算过程.

算法 1. 最小化互信息隐私泄露量.

输入：λ：Lagrange 乘子，$p(x)$：数据库先验分布，$\boldsymbol{d}(x,\hat{x})$：汉明失真矩阵，$T$：收敛门限阈值参数

输出：$q(\hat{x},z|x)$：条件概率分布，MI^*：最小的互信息泄露量 $I(X;\hat{X},Z)$，$\bar{\delta}$：达到最小互信息时的期望失真度

1. 初始化 $r_0(\hat{x},z)$ 为均匀分布
2. $q_0(\hat{x},z|x)=\dfrac{r_0(\hat{x},z)e^{-\lambda d(x,\hat{x})}}{\sum_{\hat{x},z} r_0(\hat{x},z)e^{-\lambda d(x,\hat{x})}}$
3. $I_0\leftarrow$算法 2 利用 $p(x),q_0(\hat{x},z|x),r_0(\hat{x},z)$
4. $r(\hat{x},z)=\sum_x p(x)q_0(\hat{x},z|x)$
5. WHILE TRUE DO
6. $\quad q(\hat{x},z|x)=\dfrac{r(\hat{x},z)e^{-\lambda d(x,\hat{x})}}{\sum_{\hat{x},z} r(\hat{x},z)e^{-\lambda d(x,\hat{x})}}$
7. $\quad I\leftarrow$算法 2 利用 $p(x),q(\hat{x},z|x),r(\hat{x},z)$
8. $\quad$IF ($I_0-I\leq T$) THEN
9. $\quad\quad MI^*\leftarrow I$
10. $\quad\quad$期望失真度 $\bar{\delta}=\sum_{x,\hat{x},z} p(x)q(\hat{x},z|x)d(x,\hat{x})$
11. $\quad\quad$RETURN $MI^*,q(\hat{x},z|x),\bar{\delta}$
12. $\quad$ELSE

13. $I_0 \leftarrow I$

14. $r(\hat{x},z)=\sum_x p(x)q(\hat{x},z|x)$

15. END IF

16. END WHILE

其中，算法 1 中步骤 3 的计算利用算法 2，依据 X 与 $X,\hat{Z}$ 联合的互信息计算公式

$$I(X;\hat{X},Z)=\sum_{x,\hat{x},z}p(x)q(\hat{x},z|x)\log\frac{q(\hat{x},z|x)}{q(\hat{x},z)} \tag{24}$$

计算互信息隐私泄露量，具体的计算过程如算法 2 所描述.

算法 2.　计算互信息量隐私泄露量.

输入：$p(x)$：数据库先验概率分布，$q(\hat{x},z|x)$：条件概率分布，$r(\hat{x},z)$：联合概率分布

输出：MI：平均互信息泄露量

1. 初始化 $MI=0$

2. FOR 循环遍历 $X,\hat{X}$，以及 Z 取值空间，$i\in|\mathcal{X}|,j\in|\hat{\mathcal{X}}|,k\in|\mathcal{Z}|$

3. $MI\leftarrow\sum p(x_i)q(\hat{x}_j,z_k|x_i)\log\frac{q(\hat{x}_j,z_k|x_i)}{r(\hat{x}_j,z_k)}$

4. END FOR

5. RETURN MI

上述迭代算法 1 可以计算可达最小互信息隐私泄露量的最佳条件概率分布 $q(\hat{x},z|x)$. 然而，差分隐私噪声信道机制与条件概率 $q(\hat{x}|x)$ 相关. 因此，我们考虑利用条件概率公式计算联合概率分布 $p(x,\hat{x},z)$. 进一步，基于联合概率分布，关于辅助背景知识 Z 计算边缘分布 $q(x,\hat{x})$. 以先验概率分布 $p(x)$ 为基础，类似的计算得到差分隐私的最佳信道转移条件概率分布 $q(\hat{x}|x)$. 在此，文献[14,24]提出了最佳的率失真同时保证一个确定等级的差分隐私. 基于此，针对条件概率分布 $q(\hat{x}|x)$ 利用式(9)依次计算差分隐私参数，根据定义 5 计算隐私预算参数 ε^*. 具体的计算过程如算法 3 所描述.

算法 3.　求解最佳信道转移条件概率及差分隐私参数 ε^*.

输入：$q(\hat{x},z|x)$：条件概率分布，$p(x)$ 先验概率分布

输出：$q(\hat{x}|x)$ 信道条件概率，ε^*：差分隐私预算参数

1. 计算边缘分布 $q(x,\hat{x})=\sum_z p(x)q(\hat{x},z|x)$

2. 计算条件概率分布 $q(\hat{x}|x)=\frac{q(x,\hat{x})}{p(x)}$

3. FOR 循环遍历 $X,\hat{X}$，$i\in|\mathcal{X}|,j\in|\hat{\mathcal{X}}|$

4. 求解 $\varepsilon^*=\min\left(\log\max\left\{\frac{p(\hat{x}_j|x_i)}{p(\hat{x}_j|x_i')}\right\}\right)$

5. END FOR

6. RETURN $q(\hat{x}|x),\varepsilon^*$

上述迭代算法的基本操作是计算 $q(\hat{x},z|x)$、$r(\hat{x},z)$ 及互信息量 $I(X;\hat{X},Z)$. 每一轮的计算过程中，步骤 6 计算 $(\hat{x},z)$ 联合需要进行 $O(|\hat{X}||Z|)$ 基本运算，进而对所有的 x 计算条件概率需要的基本运算次数为 $O(|X||\hat{X}||Z|)$. 其次，步骤 7 计算互信息隐私量需要 $O(|X||\hat{X}||Z|)$ 次基本运算操作. 最后，步骤 14 联合概率分布 $r(\hat{x},z)$ 的计算中，对特定的每一个变量 x 基本运算复杂度为 $O(|\hat{X}||Z|)$. 因此，$r(\hat{x},z)$ 的计算复杂度为 $O(|X||\hat{X}||Z|)$. 鉴于上述分析，算法 1 的总体计算时间复杂度为问题规模源字母表空间、再生字母表空间及辅助背景知识空间大小的函数，即 $O(|X||\hat{X}||Z|)$.

6　实验分析

本节对文中隐私与效用的平衡优化理论方法和迭代双重最小化求解算法进行实验仿真. 基于 Java 语言及第三方 Math 数学工具包编程实现文中算法，在 Windows10 平台仿真测试. 实验中，我们根据有或无辅助背景知识考虑了两种差分隐私离线数据发布场景.

(1) 针对差分隐私数据发布中无辅助背景知识的场景. 选择 MovieLens 用户电影评分数据集①，包含有 943 个用户对 1682 部电影的 100000 个评分数据，评分等级 rating 是范围(1～5)的整数. 选择电影(编号 785)计算评分 1～5 的各项概率分布 $p(x)=\{0.0513,0.1538,0.4872,0.2051,0.1026\}$，信源分布的信息熵 $H(X)=1.9464$. 其次，基于汉明失真距离构建汉明失真矩阵. 选定 Lagrange 乘子区间 $\lambda\in[0.6,10]$ 并设置门限阈值 $T=10^{-8}$，计算不同 λ 取值对应的互信息隐私泄露量、期望失真以及满足的差分隐私预算参数.

图 3 给出了 MovieLens 数据集评分的期望失真与互信息泄露的率失真函数曲线. 从图中可以看出随着期望失真度趋于 0 时，平均互信息泄露量逐渐趋近于信源熵. 其次，随着期望失真度大于 0.5，平均互信息隐私泄露量逐步趋于 0. 图中展示了数据失真与互信息隐私泄露量的变化曲线.

此外，考虑 Lagrange 乘子 λ 对期望失真以及互信息隐私泄露的影响. 图 4 所示收敛门限 $T=10^{-8}$ 时，λ、互信息隐私泄露以及期望失真的三维效果图.

① MovieLens Dataset [Online]. Available: https://grouplens.org/datasets/movielens/

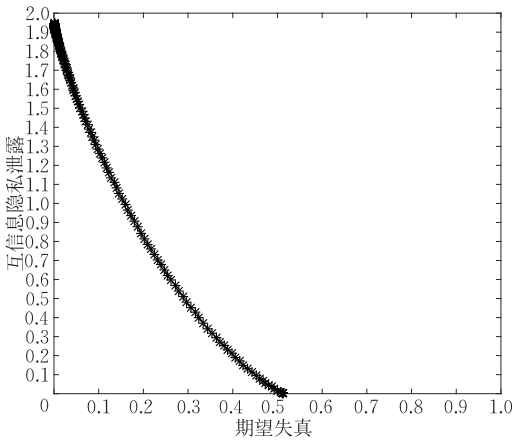

图 3 MovieLens 数据集评分的率失真曲线

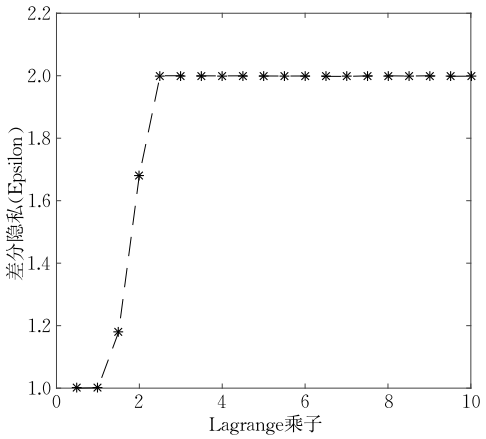

图 5 Lagrange 乘子与差分隐私参数

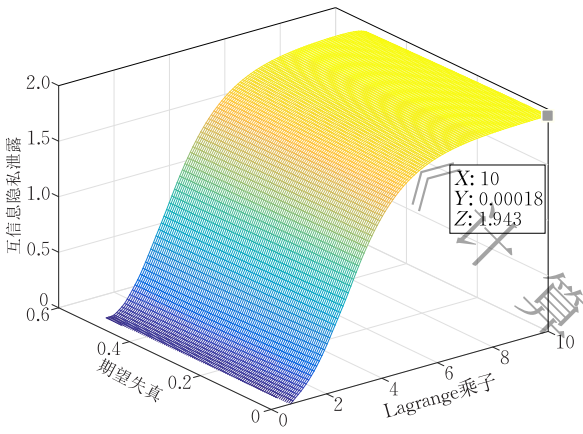

图 4 期望失真、互信息隐私与 Lagrange 乘子关系图

从图中可知，当 $\lambda=10.0$ 时，互信息隐私泄露量趋于信源熵，同时期望失真度近似于 0，与图 3 中互信息隐私泄露与期望失真曲线图显示结果具有一致性.

Lagrange 参数 λ 与收敛门限 T 影响算法求解最佳信道转移概率和差分隐私预算参数. 图 5 中给出了信道机制满足差分隐私预算参数曲线（其中，以 ln 为单位）. 从图中可知，隐私参数随着 λ 增加而趋于稳定. 对比图 4 与图 5 可知，一方面由于互信息隐私泄露随着 λ 增加而增大，隐私保护强度减弱；另一方面，λ 对互信息隐私泄露的影响逐渐减弱，致使隐私保护强度趋于平稳.

(2) 考虑具有辅助背景知识的差分隐私离线数据发布场景. 具体实验验证分析使用机器学习的 Adult 数据集①，选择其中的婚姻状态(marital-status)作为需要发布的原始数据 X，婚姻状态拥有 7 个不同的类别型取值. 同时，考虑职业(occupation)作为辅助背景知识 Z，其拥有 14 个不同的类别型取值. 针对类别型变量总是假设原始数据的取值域空间与发布失真数据取值域空间相同. 首先，统计原始数据 X 的概率分布 $p(x)=\{0.1386,0.0007,0.4668,0.0127,0.322,0.0312,0.0273\}$，依据先验概率分布计算信息熵 $H(X)=1.82$. 其次，基于汉明失真构建汉明失真矩阵. 最后，Lagrange 乘子区间 $\lambda\in[0.5,10]$，设置收敛门限阈值 $T=10^{-8}$ 进行实验分析.

从期望失真与互信息隐私泄露量的角度，对比本文中基于迭代最小化的最优化信道机制与强对称信道机制. 如图 6 给出了两种不同信道机制的期望失真与互信息隐私泄露之间的变化曲线. 从图中可以明显地看出，在相同的失真度(限失真效用)门限约束下，本文中的迭代最小化算法所计算的最优化信道机制具有更小的互信息隐私泄露量. 与此同时，表 2 给出了相同失真度条件下，两种不同机制互信息隐私泄露的对比，定量地表述迭代优化机制具有更小的互信息隐私泄露量. 此外，在相同的互信息隐私泄露容忍度条件下，表 3 对比了期望失真度. 由图 6 及表 3 可知，迭代优化机制具有更小的期望失真度. 由此可以得出以下结论，在满足失真约束条件

表 2 限失真的互信息隐私泄露量对比

期望失真度	互信息隐私泄露量	
	对称信道机制	迭代优化机制
0.509	0.3397	0.0154
0.423	0.4963	0.1073
0.355	0.6379	0.2293
0.270	0.8387	0.4437
0.203	1.0185	0.6593
0.156	1.1587	0.8440
0.120	1.2761	1.0143
0.081	1.4166	1.2242
0.050	1.5428	1.3923
0.033	1.6205	1.5290
0.021	1.6813	1.6237
0.013	1.7261	1.6880

① Adult dataset [Online]. Available: http://archive.ics.uci.edu/ml/

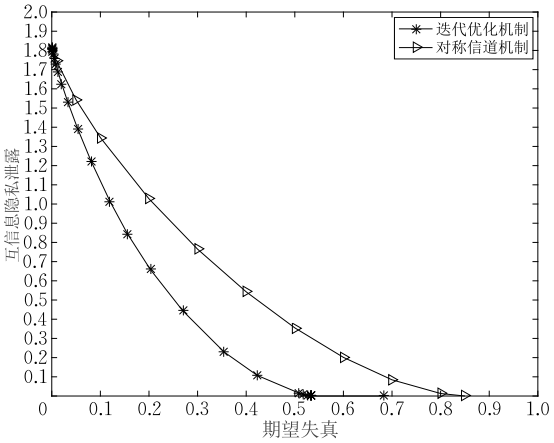

图 6　两种差分隐私信道机制的对比

表 3　相同互信息隐私泄露的期望失真度对比

互信息隐私泄露	期望失真度	
	对称信道机制	迭代优化机制
0.02	0.780	0.509
0.11	0.675	0.423
0.23	0.580	0.355
0.44	0.450	0.270
0.66	0.345	0.203
0.84	0.270	0.156
1.01	0.205	0.120
1.54	0.050	0.033
1.69	0.020	0.013

下，本文中的最优化机制方法较对称信道机制具有更好的隐私保护的效果.

选取 Lagrange 乘子区间 $\lambda\in[0.5,10]$，收敛门限阈值 $T=0.01$ 研究互信息隐私泄露、差分隐私预算参数与 λ 之间的影响关系. 图 7 所示为互信息隐私泄露量随 λ 增加而增大，表明 λ 增加则会导致隐私保护强度减弱. 另外，可以得知，随 λ 增加互信息隐私泄露量的增长率减少. 当 $\lambda>5$ 时，呈现出平稳的趋势. 与之对应的，图 8 所示为相应的隐私预算参

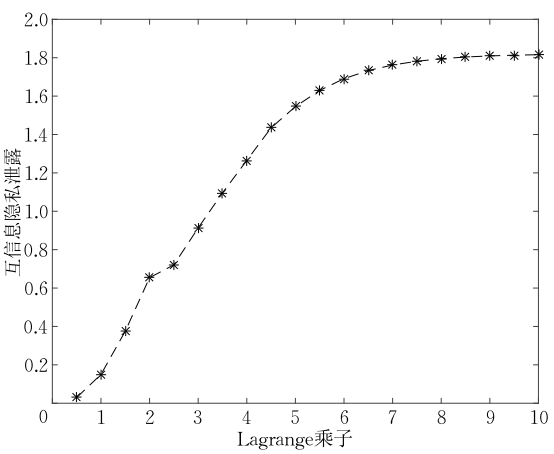

图 7　Lagrange 乘子与互信息隐私泄露

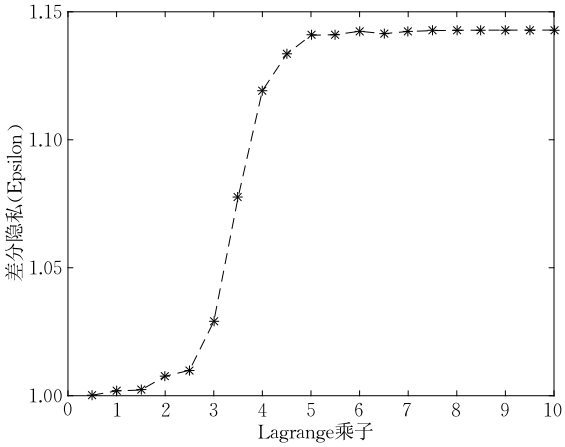

图 8　具有辅助知识的 Lagrange 乘子与差分隐私参数

数的变化曲线(图中以 ln 为单位). 从图中可以看出，隐私预算参数呈现增长趋势，反映了隐私保护强度减弱. 当 $\lambda>5$ 时，呈现出平稳的趋势，与图 7 结果具有一致性.

更多地，为详细阐述 Lagrange 乘子 λ、收敛门限阈值 T、互信息隐私泄露以及期望失真度之间的关系，考虑 5 种不同组合，$\lambda\in\{0.5,1.0,1.5,2.0,2.5\}$ 和 $T\in\{10^{-4},10^{-5},10^{-6},10^{-7},10^{-8}\}$ 进行实验. 如图 9 所示的实验结果表明，固定收敛门限阈值，互信息隐私泄露随 λ 增加而增大. 对应的，固定 Lagrange 参数 λ 取值时，但互信息隐私随 T 逐步增加 10 的倍数，但变化微弱.

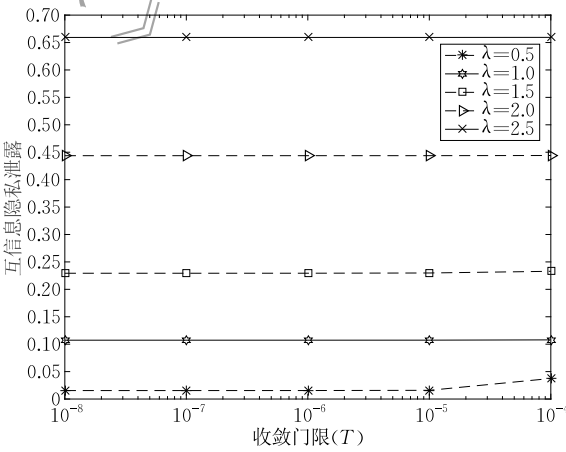

图 9　收敛门限与互信息隐私泄露量

图 10 所示期望失真与算法收敛门限阈值及 λ 之间的关系. 从图中容易看出，固定算法收敛门限 T，期望失真度随 λ 增加而增大. 与之对应的，固定 λ 参数取值时，期望汉明失真随 T 逐步增加 10 的倍数，但变化趋势并不明显.

由此说明，适当地选择参数 λ 和 T 可以在获得

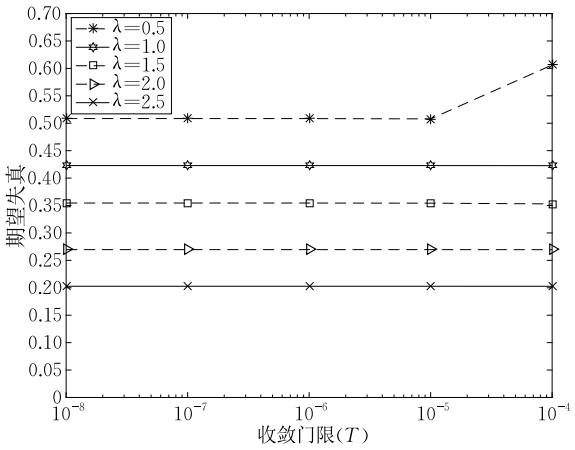

图 10 收敛门限与期望失真度

最佳互信息隐私的同时提供一个确定等级的差分隐私保护,达到理想的隐私保护效果.

本节通过两组不同的数据集进行实验,分别针对差分隐私离线数据发布中有无辅助背景知识,给出了不同场景下的差分隐私最优化信道机制以及互信息隐私泄露与期望失真的变化曲线.在互信息隐私泄露和期望失真度方面,与差分隐私的一种对称信道进行了对比分析,结果表明本文中考虑的有辅助背景知识的最优化信道机制具有更小的互信息隐私泄露量.其等同于,在相同的互信息隐私泄露容忍度条件下,本文中所提出的最优化信道机制具有更高的发布数据效用.

7 结 论

本文以 Shannon 信息论为基础,针对差分隐私离线数据发布场景中隐私与效用的平衡问题,考虑了隐私攻击者可能具有的关联辅助背景知识对互信息隐私泄露的影响,研究了满足期望失真约束条件的互信息隐私最优化机制问题.在此基础上,基于 Blahut-Arimoto 算法思想设计了最优化信道条件概率的求解算法.通过实验仿真,验证分析了所提出方法的有效性.基于实验结果从互信息隐私泄露量、期望失真变化的角度,阐述了本文理论研究成果.与此同时,对比差分隐私的对称信道机制,表明本文中的方法在满足数据效用约束的条件下,具有更小的互信息隐私泄露量.

在下一步的研究工作中,拟基于信息论视角考虑交互式差分隐私计算模型,构建查询与响应的隐私相关度模型,并基于迭代的思想研究差分隐私每一轮查询响应的最优化机制,总结分析整体的隐私泄露与数据失真问题.力图为权衡交互式差分隐私计算模型中隐私度与数据可用性的研究提供一种信息论的方法.

参 考 文 献

[1] Dwork C. Differential privacy//Proceedings of the 33rd International Colloquium on Automata, Languages, and Programming- Volume Part II (ICALP). Venice, Italy, 2006: 1-12

[2] Zhu T, Li G, Zhou W, et al. Differentially private data publishing and analysis: A survey. IEEE Transactions on Knowledge and Data Engineering, 2017, 29(8): 1619-1638

[3] Zhang Xiao-Jian, Meng Xiao-Feng. Differential privacy in data publication and analysis. Chinese Journal of Computers, 2014, 37(4): 927-949(in Chinese)
(张啸剑, 孟小峰. 面向数据发布和分析的差分隐私保护. 计算机学报, 2014, 37(4): 927-949)

[4] Dwork C, Mcsherry F, Nissim K, Smith A D. Calibrating noise to sensitivity in private data analysis//Proceedings of the 3rd Theory of Cryptography Conference (TCC). New York, USA, 2006: 265-284

[5] Dwork C, Roth A. The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science, 2014, 9(3-4): 211-407

[6] Dwork C. Differential privacy: A survey of results//Proceedings of the 5th International Conference on Theory and Applications of Models of Computation(TAMC). Xi'an, China, 2008: 1-19

[7] Kalantari K, Sankar L, Sarwate A D. Robust privacy-utility tradeoffs under differential privacy and hamming distortion. IEEE Transactions on Information Forensics and Security, 2018, 13(11): 2816-2830

[8] Alvim M S, Andrés M E, Chatzikokolakis K, Degano P, Palamidessi C. Differential privacy: On the trade-off between utility and information leakage. Formal Aspects in Security and Trust, 2011, 7140: 39-54

[9] Shannon C E. A mathematical theory of communication. Bell System Technical Journal, 1948, 27(3): 379-423

[10] Smith G. On the foundations of quantitative information flow//Proceedings of the 12th International Conference on Foundations of Software Science and Computational Structures (FoSSaCS). York, UK, 2009: 288-302

[11] Alvim M S, Andrés M E. On the relation between differential privacy and quantitative information flow. International Colloquium on Automata Languages and Programming, 2011, 6756: 60-76

[12] Alvim M S, Andrés M E, Chatzikokolakis K. On the information leakage of differentially-private mechanisms. Journal of Computer Security, 2015, 23(4): 427-469

[13] Sankar L, Rajagopalan S R, Poor H V. Utility-privacy tradeoffs in databases: An information-theoretic approach. IEEE Transactions on Information Forensics and Security, 2013, 8(6): 838-852

[14] Wang W, Ying L, Zhang J. On the relation between identifiability, differential privacy, and mutual-Information privacy. IEEE Transactions on Information Theory, 2016, 62(9): 5018-5029

[15] Sarwate, A D, Sankar L. A rate-disortion perspective on local differential privacy//Proceedings of the 52nd Annual Allerton Conference on Communication, Control, and Computing. Monticello, Illinois, USA, 2014: 903-908

[16] Wang W, Ying L, Zhang J. A minimax distortion view of differentially private query release//Proceedings of the 49th Asilomar Conference on Signals, Systems and Computers. Pacific Grove, USA, 2015: 1046-1050

[17] Kalantari K, Sankar L, Sarwate A D, Optimal differential privacy mechanisms under hamming distortion for structured source classes//Proceedings of the IEEE International Symposium on Information Theory (ISIT). Barcelona, Spain, 2016: 2069-2073

[18] Kairouz P, Oh S, Viswanath P. Extremal mechanisms for local differential privacy. Journal of Machine Learning Research, 2016, 17(1): 492-542

[19] Zhu T, Xiong P, Li G, Zhou W. Correlated differential privacy: Hiding information in non-IID data set. IEEE Transactions on Information Forensics and Security, 2015, 10(2): 229-242

[20] Li Y, Ren X, Yang S, Yang X. Impact of prior knowledge and data correlation on privacy leakage: A unified analysis. IEEE Transactions on Information Forensics and Security, 2019, 14(9): 2342-2357

[21] Cover T M, Thomas J A. Elements of Information Theory. 2nd Edition, Hoboken, USA: Wiley, 2006

[22] Arimoto S. An algorithm for calculating the capacity of an arbitrary discrete memoryless channel. IEEE Transactions on Information Theory, 1972, 18(1): 14-20

[23] Blahut R E. Computation of channel capacity and rate-distortion functions. IEEE Transactions on Information Theory, 1972, 18(4): 460-473

[24] Mir D J. Information theoretic foundations of differential privacy//Proceedings of the 5th International Conference on Foundations and Practice of Security. Montreal, Canada, 2012: 374-381

[25] Barthe G, Köpf B. Information-theoretic bounds for differentially private mechanisms//Proceedings of the IEEE 24th Computer Security Foundations Symposium (CSF). Cernay, France, 2011: 191-204

[26] Cuff P, Yu L. Differential privacy as a mutual information constraint//Proceedings of the ACM SIGSAC Conference on Computer and Communications Security. New York, USA, 2016: 43-54

[27] Boyd S, Vandenberghe L, Convex Optimization. New York, USA: Cambridge University Press, 2004

[28] Zhang W, Li M, Tandon R, et al. Online location trace privacy: An information theoretic approach. IEEE Transactions on Information Forensics and Security, 2019, 14(1): 235-250

[29] du Pin Calmon F, Fawaz N. Privacy against statistical inference //Proceedings of the 50th Annual Allerton Conference on Communication, Control, and Computing. Monticello, Illinois, USA, 2012: 1401-1408

[30] Makhdoumi A, Fawaz N. Privacy-utility tradeoff under statistical uncertainty//Proceedings of the 51st Annual Allerton Conference on Communication, Control, and Computing. Allerton, Illinois, USA, 2013: 1627-1634

[31] Peng Chang-Gen, Ding Hong-Fa, Zhu Yi-Jie, et al. Information entropy models and privacy metrics methods for privacy protection. Journal of Software, 2016, 27(8): 1891-1903(in Chinese)
(彭长根，丁红发，朱义杰等. 隐私保护的信息熵模型及其度量方法. 软件学报，2016，27(8)：1891-1903)

[32] Wu Zhen-Qiang, Hu Jing, Tain Yu-Pan, et al. Privacy preserving algorithms of uncertain graphs in social networks. Journal of Software, 2019, 30(4): 1106-1120(in Chinese)
(吴振强，胡静，田垿攀等. 社交网络下的不确定图隐私保护算法. 软件学报，2019，30(4)：1106-1120)

[33] Shannon C E. Coding theorems for a discrete source with a fidelity criteria. IRE National Convention Record, 1959, (4): 93-126

[34] Xiong S, Sarwate A D, Mandayam N B. Randomized requantization with local differential privacy//Proceedings of the 41st IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Shanghai, China, 2016: 2189-2193

[35] Csiszár I, Tusnády G. Information geometry and alternating minimization procedures, Statistics and Decisions, 1984, Supplement(1): 205-237

[36] Csiszár I. On the computation of rate distortion functions. IEEE Transactions on Information Theory, 1974, 20(1): 122-124

WU Ning-Bo, Ph. D. candidate. His research interests include data security and privacy protection, etc.

PENG Chang-Gen, Ph. D., professor, Ph. D. supervisor. His research interests include cryptography, information security and privacy protection in big data, etc.

TIAN You-Liang, Ph. D., professor, Ph. D. supervisor. His research interests include cryptography and data security, cryptographic protocols, etc.

NIU Kun, Ph. D. candidate, lecturer. Her research interests include data security and privacy protection.

DING Hong-Fa, Ph. D. candidate, lecturer. His research interests include privacy protection, data security and access control.

Background

Recently years, the problem of tradeoff between privacy and utility attracts many attentions from researchers. Under differential privacy (DP), some researchers investigate the problem of privacy-utility tradeoffs by using the information theoretic approach. They propose some metric methods for privacy leakage and data utility from the perspective of Quantitative Information Flow (QIF), i. e., privacy and utility quantified via mutual information (MI) and distortion measure, respectively. In addition, some researchers begin to study the optimal DP mechanism by using MI privacy. Although many achievements have been achieved, there are still some problems to be solved. Especially, how to obtain the optimal MI privacy mechanism.

To this end, this paper focused on the scenario of the DP offline data publishing and utilize rate-distortion theory to study the problem of optimal MI privacy mechanism which can achieve minimized MI privacy leakage for given distortion constraint. Moreover, this research can provide important information theoretic reference for risk assessment of privacy.

This work is supported by the National Natural Science Foundation of China (U1836205, 61662009, 61772008), the 13th Five-Year National Cryptography Development Foundation (MMJJ20170129), the Science and Technology Program of Guizhou Province (Guizhou Science Contract Major Program ([2018]3001, [2018]3007, [2017]3002), Guizhou Science Contract Platform Talent ([2020]5017), Guizhou Science Contract Foundation ([2017]1045)), the Project of Innovative Group in Guizhou Education Department ([2013]09), and Research Fund Project for Graduate Students of Guizhou Province (KYJJ2017005).