大数据环境下差分隐私保护技术及应用

付钰，俞艺涵，吴晓平
（海军工程大学信息安全系，湖北 武汉 430033）

摘 要：大数据中的隐私保护问题是当前网络空间安全领域的一个研究热点，差分隐私保护作为严格且可证明的隐私保护定义，研究其在大数据环境下的应用现状能够为其后续的系统性应用等提供参考与指导。在系统分析差分隐私保护的相关概念与技术特性的基础上，通过对差分隐私保护技术在数据发布与分析、云计算与大数据计算、位置与轨迹服务及社交网络中的应用等进行综述，阐述了当前具有代表性的研究成果并分析了其存在的问题。研究表明，现有成果从差分隐私保护机理、噪声添加机制与位置、数据处理方式等方面对差分隐私保护应用进行了卓有成效的创新与探究，且相关成果在不同场景下实现了交叉应用。最后提出了差分隐私保护在大数据环境下进一步系统性应用还需要注意的四大问题。

关键词：差分隐私；隐私保护；大数据；数据发布；云计算；位置服务；社交网络
中图分类号：TP393
文献标识码：A
doi: 10.11959/j.issn.1000−436x.2019209

Differential privacy protection technology and its application in big data environment

FU Yu, YU Yihan, WU Xiaoping
Department of Information Security, Naval University of Engineering, Wuhan 430033, China

Abstract: The privacy protection in big data is a research hotspot in the field of cyberspace security. As a strict and provable definition of privacy protection, studying application status of differential privacy protection in big data environment can provide reference and guidance for its subsequent system applications. Based on the analysis of the related concepts and technical characteristics of differential privacy protection, the application of differential privacy protection technology was reviewed in data distribution and analysis, cloud computing and big data computing, location and trajectory services and social networks, which expounded the current representative research results and analyzed its existing problems. The research shows that the existing results have made effective innovation and exploration of differential privacy protection applications from the aspects of differential privacy protection mechanism, noise addition mechanism and location, and data processing methods, and the related results have been cross-applied in different scenarios. Finally, four major problems that need to be studied in the further systematic application of differential privacy protection in the big data environment are proposed.

Key words: differential privacy, privacy protection, big data, data publishing, cloud computing, location service, social network

1 引言

互联网创新已经进入一个前所未有的大数据时代，为利用其中海量有价值的数据信息，不可避免地会对数据进行收集与分析，而过度的数据收集使隐私泄露问题日益凸显[1]。以近年来兴起的互联

收稿日期：2019−08−19；**修回日期**：2019−09−02
通信作者：俞艺涵，chenyike1992@163.com
基金项目：国家重点研发计划基金资助项目（No.SQ2018YFGX210002）；国家自然科学基金资助项目（No.2015CFC867）
Foundation Items: The National Key Research and Development Program of China (No.SQ2018YFGX210002), The National Natural Science Foundation of China (No.2015CFC867)

网金融点对点（P2P, point to point）借贷[2-3]模式为例，目前国内 P2P 理财平台数量在 5 000 家左右，这些平台在为用户提供服务时，收集用户相关隐私信息来分析用户的还贷能力，其中最基本的身份验证要求是用户的身份证信息或者经实名制验证的手机号码信息，除此之外，用户的个人征信报告、信用卡账单邮箱、网购记录和通讯录等个人隐私信息也普遍被 P2P 平台获权查阅并记录。一方面，绝大多数 P2P 平台对于用户隐私数据的隐私保护级别不高，甚至存在为拉拢用户而进行信息共享的行为；另一方面，在众多 P2P 平台中，有信息安全隐患的问题平台约占总数的$\frac{1}{3}$，这些问题平台自身并不可信，且会频繁遭受网络黑客的攻击，其用户个人隐私信息极不安全。

随着数据挖掘及信息安全技术的不断发展，针对特定数据的隐私保护问题，国内外学者已经开展了卓有成效的研究工作，主要通过匿名化技术[4]和数据加密技术[5]等来实现隐私保护。近年来，有学者提出隐私计算的概念[6]，探讨了隐私计算的应用[7-8]，这为隐私识别与量化评估提供了条件。考虑到互联网下的隐私信息具有数据量大、类别多、层次关系复杂等特点，而基于匿名化和数据加密的隐私保护技术又需要紧密依赖背景知识假设，只能保证单一数据集上隐私不被泄露的局限性往往难以满足互联网下针对大数据隐私保护的要求。所以，人们开始关注差分隐私（DP, differential privacy）保护技术及其应用，差分隐私是 Dwork[9]在 2006 年针对统计数据库的隐私泄露问题所提出的一种隐私的概念。差分隐私保护模型就是一种建立在严格的数学基础之上，通过对隐私泄露风险做定量的形式化证明，并保有数据极大可用性的数据安全模型，该模型假设隐私信息的攻击者在获取目标之外所有信息的情况下，也不能判断出目标信息是否在被攻击的数据中，即差分隐私保护能够抵御攻击方的最大背景知识攻击。由于差分隐私保护模式可以提供可度量的隐私保护等优势，故被广泛应用于网络空间安全等领域。特别是在大数据环境下，从理论研究的角度看，差分隐私保护表现出极高的兼容性[10]。一方面，大数据环境下的数据集体积大，数据集中存在大量的记录，这一条件有利于区分隐私，因此可以用较小的噪声来实现差分隐私保护；另一方面，差分隐私保护在为大数据提供隐私保护时，不需要改变原始数据，一般只在输出中加入随机噪声，不会对原始数据处理的速度造成影响；同时，差分隐私保护独立于底层数据结构并兼容多种数据类型，能够兼容所有类型的数据集，适用于大数据中存在结构化、非结构化以及半结构化等多种数据形式的现实情况。但从实际应用来看，差分隐私保护在大数据中的应用也面临诸多问题。

由此，本文首先对差分隐私保护的基本概念与相关技术进行了系统介绍，随后全面综述了大数据环境下差分隐私保护技术在数据发布与分析、云计算与大数据计算、位置与轨迹服务和社交网络中的应用与存在的问题，最后提出了大数据环境下差分隐私保护的系统性应用所面临的挑战并展望其发展方向。

2 差分隐私保护的相关概念

差分隐私保护可以克服传统隐私保护技术应用时其安全性依赖攻击者的相关背景知识、保护效果，难以用有效严格的数学方法定量化描述等缺陷，从而可在大大降低保护对象数据集隐私泄露风险的同时，尽可能保证数据集数据的可用性，其过程就是通过对真实数据添加随机扰动，并保证数据在被干扰后仍具有一定的可用性来实现的，即要使保护对象数据失真且同时保持数据集中特定数据或数据属性（如统计特性等）不变。

2.1 基本概念

为方便理解与讨论，首先给出差分隐私保护的概念。

定义 1 对于一个有限域 Z，数据集 $D=\{z|\ z\in Z\}$，其样本量为 n，属性的个数为维度 d，若 $F=\{f_1,f_2,\cdots\}$ 表示一组操作（如查询等），而 M 是对系列操作的某种处理，且使之满足某种隐私保护的条件，则称此过程为针对数据集 D 的隐私保护机制。

定义 2 设有限域 Z 上的 2 个完全相同或至多相差一条记录的数据集 D 和 D'，则称 D 和 D' 为邻接数据集（adjacent dataset），即邻接数据集 D 和 D' 具有相同的属性结构，且二者的对称差 $D\Delta D'$ 中记录的数量为 1。

定义 3 对于任意 2 个邻接数据集 D 和 D'，设随机算法 A 的值域为 $R(A)$，事件 X 发生的可能性为 $\Pr[X]$，若对任意 $S,S'\in R(A)$，都满足

$$\Pr[A(D)=S]\leqslant \mathrm{e}^{\varepsilon}\Pr[A(D')=S'] \tag{1}$$

则随机算法 A 提供 ε-差分隐私保护，称 ε 为差分隐私保护预算[11]。

由此可以看出，若用户由查询函数 F 对数据集 D 进行查询操作时，随机算法 A 通过对查询函数 F 进行扰动，使之能满足差分隐私保护的条件[12-13]。

2.2 相关概念

1) 差分隐私保护预算

差分隐私保护预算 ε 是差分隐私保护所能提供的隐私保护级别的度量，ε 的值越低，差分隐私保护所提供的隐私保护级别就越高。一般 ε 取很小的值，以保证差分隐私保护的效果，例如取 ε=0.1。被保护数据的可用性也与 ε 密切相关，一般来说，ε 越小，差分隐私保护所提供的随机扰动越大，被保护数据的可用性则越差。

2) 敏感度

差分隐私保护通常对查询函数的返回值添加随机扰动来达到隐私保护的目的，随机扰动的大小与查询函数的敏感度密切相关。查询函数的敏感度是指当数据中仅发生一条记录的改变时查询结果的最大改变量。通常，差分隐私保护利用查询函数的全局敏感度（GS, global sensitivity）[8]来度量随机扰动的大小。全局敏感度定义如下。

定义 4　设 D 和 D' 是邻接数据集，则称

$$\Delta F_{\mathrm{GS}} = \max_{D,D'} \| F(D) - F(D') \|_1 \tag{2}$$

为查询函数 F 的全局敏感度。这里 $\|\cdot\|_1$ 表示向量元素绝对值之和。

查询函数的全局敏感度由函数本身决定，全局敏感度越大，差分隐私保护标准下所需的随机扰动也越大。对于一些全局敏感度小的查询函数（如计数函数），用其全局敏感度来度量随机扰动的大小较为合适；对于平均值函数、中位数函数等全局敏度较大的函数，用其全局敏感度来度量随机扰动的大小容易造成随机扰动量过大，造成不必要的数据可用性方面的损失。由此，局部敏感度（LS, local sensitivity）[14]的概念被提出，其定义如下。

定义 5　设 D 和 D' 是邻接数据集，则称

$$\Delta F_{\mathrm{LS}} = \max_{D'} \| F(D) - F(D') \|_1 \tag{3}$$

为查询函数 F 的局部敏感度。

查询函数的局部敏感度与函数自身以及具体数据共同决定，一般比全局敏感度要小，两者存在如下关系。

$$\Delta F_{\mathrm{GS}} = \max_{D}(F_{LS}(D)) \tag{4}$$

3) 扰动机制

差分隐私保护中存在多种扰动机制，其中拉普拉斯机制[15]在对数值型结果的保护中应用最为广泛。

定义 6　对于数据集 D，查询函数 F 及其全局敏感度 ΔF_{GS}，如果随机噪声 Y 服从尺度为 $\frac{\Delta F_{\mathrm{GS}}}{\varepsilon}$ 的拉普拉斯分布，则称随机算法 $A(D) = F(D) + Y$ 可以提供 ε-差分隐私保护[11]。

拉普拉斯机制只能对数值型查询结果进行保护，而在实际应用中，存在许多查询结果不是数值型的情况。由此，指数机制[16]被提出。

定义 7　对于数据集 D，其输出为一实体对象 $r \in \mathrm{Range}$，$q(D,r)$ 为 r 的可用性函数，其敏感度记作 ΔF_q。若随机算法 M 以正比于 $\exp\left(\frac{\varepsilon q(D,r)}{2\Delta F_q}\right)$ 的概率从 Range 中选择并输出 r，则 M 提供 ε-差分隐私保护[16]。

4) 组合性质

差分隐私保护存在以下两方面的组合性质[17]，它们是将差分隐私保护运用到反复迭代过程中，证明算法满足差分隐私保护以及合理分配差分隐私预算的基础。

性质 1　若存在 n 个随机算法序列 $A_i (1 \leqslant i \leqslant n)$ 提供 ε_i 差分隐私保护，则对于同一数据集 D，$\{A_1, \cdots, A_n\}$ 在 D 上的序列组合算法也提供 ε-差分隐私保护，其中，$\varepsilon = \sum_{i=1}^{n} \varepsilon_i$。

性质 2　若存在随机算法 A 提供 ε-差分隐私保护，数据集 D 可分为不相交的子集 $D_1, \cdots, D_m$，则随机算法 A 在 $\{D_1, \cdots, D_m\}$ 上的组合运算所构成的算法也提供 ε-差分隐私保护。

因为每一种隐私保护方法其保护效果都是基于某种攻击模型而度量，如 K-匿名就是基于攻击者对对象数据集信息全不知晓的假设，否则 K-匿名算法也无法对数据隐私实施保护。从差分隐私的定义及相关特性可以看出，其基于的攻击模型是最坏的可能攻击者已知除一条记录以外的对象数据集所有的敏感属性，但这条记录的敏感属性信息也可得到有效保护。所以，差分隐私在数据发布与分析、云计算与大数据计算、面向位置与轨迹服务、社交

网络等领域有着越来越广泛的应用。

3 面向数据发布与分析的差分隐私保护

差分隐私保护作为当前数据隐私保护技术应用较为广泛的一种，其应用最早出现在数据库领域。如何在大数据时代在保证对数据隐私保护的前提下对海量数据进行发布与分析，已经成为近年来数据库应用，尤其是数据发布领域的研究热点。

3.1 基于差分隐私的数据发布

Han 等[18]指出差分隐私保护在保证数据效用时，由于可能存在的非独立推理（NIR, non-independent reasoning），敏感数据将面临泄露的风险。由于差分隐私保护的固有机制，必然造成基于差分隐私保护的数据发布存在数据隐私性和数据可用性之间的矛盾。目前的研究大多集中在选择最优噪声机制、优化噪声添加策略和优化数据发布策略等方面，目的是寻求数据发布中的隐私性与可用性之间的平衡。

差分隐私保护中存在多种噪声机制，众多学者对于基于差分隐私保护数据发布中的普遍最优机制展开了研究。Hai 等[19]对差分隐私保护下的普遍最优噪声机制做了定义，并且证明了不可能存在一种普遍最优机制能够保证诸如直方图等一般查询函数达到隐私性和可用性的最优。进一步的研究表明，目前不存在一种噪声机制能够在没有侧面信息或额外条件限制的情况下使差分隐私保护下数据发布的隐私性和可用性达到普遍最优。在查询函数为输出为整数且有界的各类查询函数的条件下，Ghosh 等[20]和 Gupte 等[21]分别采用贝叶斯和非贝叶斯风险规避模型来对普遍最优噪声机制展开了研究，他们得出的结论是在贝叶斯或非贝叶斯风险规避模型下，对于各类计数查询函数，几何噪声机制是普遍最优的。Geng 等[22]则在一般风险规避模型下，针对实值查询函数提出一种阶梯噪声机制，当满足以下 3 个条件时，所提的阶梯噪声机制是隐私性与可用性最优的。1) 查询函数的值域为实值，范围为 $(-\infty,+\infty)$；2) 查询生成器（QG, query generator）没有可用的侧面信息；3) 查询函数的局部敏感度等于全局敏感度或保证敏感度在所有可能的输出上保持不变。然而，在实际应用中，大部分查询函数难以满足条件 1)和条件 3)，且其所得结论默认了查询函数的局部敏感度和全局敏感度相等，这是不符合实际的。Chen 等[23]对差分隐私保护中的最佳扰动机制进行了分析与实验，在对单实数（标量）进行查询时，提出以侧面信息为基础设计噪声扰动机制的设想，使差分隐私保护的效用尽可能最大化。如何在未知数据侧面信息的情况下，合理设计噪声扰动机制是下一步的研究方向。

在优化噪声添加策略方面，Nissim 等[14]指出用局部敏感度进行计算有泄露隐私数据分布特征的风险，因此为局部敏感度定义了平滑上界与平滑敏感度的概念，并证明对于某些非线性查询函数可以利用查询函数的平滑敏感度对所添加的噪声大小进行校准，从而提高数据可用性。Lin 等[24]提出了差分隐私保护中动态噪声阈值的概念，证明了差分隐私保护方案中的噪声添加与受保护数据集大小之间的关系为

$$H(x)=\left|\frac{\sum_{i=1}^{n-1}x_{n-1}}{n-1}-x_n\right|\ell \tag{5}$$

其中，$H(x)$ 为第 n 个数据所需的噪声扰动，x 为数据，ℓ 为数据的扰动校正值。这样做保证了所添加的噪声不会因为过小而不能满足差分隐私要求，也不会因为过大而影响数据可用性，并减小了计算开销。但是，该方案所提的差分隐私方法需充分了解所保护数据的特征，其应用具有一定的局限性。Ji 等[25]在非交互模式下对多维数据发布中的差分隐私保护方法进行了研究，通过 Haar 小波变换对原始多维数据进行处理以构建一般关系表的紧凑概要，并在概要的每一个记录中添加多对数噪声之后，得到一个扰动后的概要。随后在扰动后的概要中对查询进行评估，输出一般关系表中受扰动小波系数的紧凑集合，最终将其扩展回受扰动的关系元组直到查询结束，实现了多维数据发布中的快速查询处理与可证明的隐私保证。De[26]则寻求以近似的差分隐私要求来完成对数据的保护，主要策略是放宽差分隐私保护的隐私性要求，减少因噪声添加带来的误差，降低算法复杂度。他们证明了纯 ε-差分隐私与近似 (ε,δ)-差分隐私之间的界限，并证明当 $\delta>0$ 时，存在一种查询可以添加方差为 $O\sqrt{n\log\frac{1}{\delta}}$ 的噪声达到差分隐私保护的要求；当 $\delta=0$ 时，则需要方差为 $\Omega(n)$ 的噪声。

在优化数据发布策略方面，Zhang 等[27]提出一种差分隐私框架下应答计数范围查询的算法 DPAV，该算法首先通过频率矩阵生成平均树，树

中每个节点的值等于其叶子节点值的平均值，随后通过权重函数合理分配拉普拉斯噪声大小并添加到平均树中不同层级的节点中，最后将平均树转化为新的频率矩阵，以此来应答计数范围查询。Li 等[28]提出一种矩阵机制来提供谓词计数查询服务，该机制以矩阵的形式表示线性查询集合，生成一个策略查询集，以策略查询集中的最小误差估计值作为查询的应答，在保留数据差异性的同时提高查询的准确性，但其实际应用过程中存在多维度的矩阵计算，计算成本过高。Koufogiannis 等[29]对于身份查询函数，提出了一种复合差分隐私保护机制，采用马尔可夫随机过程的延迟抽样来作为数据的逐步输出，可以在保证差分隐私固有属性的情况下，逐步提高数据释放的准确性，并推测其方案可以适用于更一般的应用条件。Hay 等[30]对被噪声扰乱后的数据在一定约束条件下进行转化，使差异化的输出保有了一致性，以此提高了数据发布的准确性，但该方法只对一维数据集的查询起效。

目前对于差分隐私保护下单一时间点静态数据的发布已有较多的成果，但在大数据环境下，实际数据的发布往往还要求能够对数据流中动态数据提供差分隐私保护。

Kellaris 等[31]定义了在动态数据流中的两类隐私保护级别——事件级别和用户级别。事件级别表示对于保护某一特定时间戳下的某一数据的隐私保护级别；用户级别则表示对于保护数据流中全数据的隐私保护级别。Kellaris 等提出一种 ω-事件 ε-差分隐私保护模型，该模型合并了事件级别和用户级别，来保证在 ω 个时间戳内对于数据流的保护达到隐私性与可用性平衡的效果。但是其所提的模型并不能适用于所有类型的数据流，且对于隐私预算 ε 采用简单平均分配的方法也没有考虑到不同时间戳内的数据特性差异，造成了 ε 没有被充分利用，引起数据可用性降低的不利结果。

Fan 等[32]在有限数据流的情况下提出一种满足用户级别的实时聚合统计数据发布框架 FAST。该方案基于过滤和自适应抽样，根据检测到的数据动态地对长时间序列进行采样，并预测非采样点的数据值，然后以此纠正经噪声扰动后的采样点的数值，来最小化整体隐私成本并最大程度地提高每个时间戳的数据准确性。Chan 等[33]与 Dwork 等[34]针对动态数据中的统计数据发布分别做出了相似的研究。Dwork 等建立完整的二叉树更新流，在二叉树的节点中存储对固定长度的数据流所添加的成比例对的噪声值，对实时更新数据流识别其所属的节点并在数据流中添加节点中的噪声值后给计数器计数。Chan 等在对二叉树更新时，其节点包含所有子树中经添加成比例对的随机噪声后的更新总和，经识别后对最大子树进行更新并将其根节点中存储的数值总和上报给计数器。

Wang 等[35]提出一种 RescueDP 的实时动态空间众包数据发布方案，该方案利用数据变化相似性的特征动态对数据进行区域分组，通过自适应抽样和自适应预算分配，在每个分组中加入拉普拉斯噪声扰动，最终使用卡尔曼滤波片提高了实时数据发布的准确性。

Chen 等[36-37]最先将差分隐私保护应用到序列数据中，利用嘈杂的前缀树结构将具有相同前缀的序列分组到相同的分支中，从而缩小输出域，随后还提出利用一组可变长度的 n-gram 模型顺序提取数据库的基本信息，来简化一般时序数据，缓解了因序列数据的固有顺序性和高维度所造成的应用差分隐私到序列数据中困难的问题。

Kang 等[38]针对物流领域中时间序列位置信息的隐私泄露问题，提出了一种基于差异隐私的时间序列位置数据发布方法。首先，利用聚类优化算法构造与时间相关的公共感兴趣区域，用质心点代表公共利益区的位置，由此建立位置搜索树，保证了位置数据之间的内在联系；然后将拉普拉斯噪声添加到位置搜索树节点，减少了添加噪声的次数，以此确保了所发布数据的可用性。

基于差分隐私的数据发布技术对比如表 1 所示。

3.2　基于差分隐私的大数据分析

基于机器学习、聚类分析、神经网络等传统的数据分析技术直接应用在经过差分隐私保护后的数据上时，往往会出现效率差与准确性低等问题，这就需要提出相应适用于差分隐私数据中的数据分析算法。优化基于差分隐私的数据分析算法，与差分隐私自身理论的发展联系紧密。目前的研究主要集中在如何解决在当前数据分析技术中通过添加随机噪声来满足差分隐私保护要求的同时，提高算法的准确性并降低计算复杂度的问题上。

在随机梯度下降算法中加入随机噪声是实现机器学习中差分隐私保护的一般方法，由于随机梯度下降算法需通过反复迭代来达到学习目的，随机噪声的添加方式将直接影响算法的效用。Abadi 等[39]

表 1 基于差分隐私的数据发布技术对比

技术	相关文献	具体方式	主要优点	主要缺点
选择最优噪声机制	[19-23]	针对特定数据类型和查询函数选择最优噪声机制	对查询函数为计数函数等情况优化效果明显	普适性差
优化噪声添加策略	[14,24-25]	为噪声、敏感度定界	算法计算复杂度降低，可适用于多维数据	数据可用性降低
	[26]	采取近似差分隐私策略	数据可用性提升	数据隐私性降低
优化数据发布策略	[27-38]	通过各类转换技术、划分技术扩大查询范围和提高查询精度；合理分配隐私预算，提高效用	能够适用于大数据环境下动态数据流中，数据发布准确性高，优化了查询精度与查询范围	存在计算复杂度高、通信开销大的问题，一般只适用于特定的数据类型

通过随机排列构建不同的批量和批次，将计算分为适当大小的组来进行，并对算法中的隐私预算、学习步长等参数进行了有目的性的优化，提高了算法的效率并使数据保有了较好的可用性。Cai 等[40]则研究了平均估计和线性回归中统计准确性与隐私之间的权衡问题，主要通过改进最小极大值下界、迭代阈值等参数的设定策略来保证满足差分隐私的前提下提升统计准确性。Mcsherry 等[41]在 Netflix Prize 数据集上对其提出的端对端差分隐私保护系统进行评估，并与文献[39]采取了类似的高维数据输入方式，不同的是其非凸目标函数的构造是在确定学习任务的核心后进行的，这样做能够充分利用统计数据，并以高斯机制进行差分隐私计算。Xu 等[42]则针对生成式对抗网络（GAN, generative adversarial network）应用于隐私数据时可能因记忆样本而产生隐私泄露威胁的问题，提出了一种满足差分隐私的生成式对抗网络——GANobfuscator，其通过在学习过程中为梯度添加精心设计的噪声来实现 GAN 下的差异隐私，并设计了梯度修剪策略，在保证差分隐私的同时提高了数据训练的可扩展性和稳定性。

Li 等[43]提出了一种支持差分隐私保护的分布式在线学习算法，该算法通过一个时变的双重随机矩阵来控制各分布式节点之间的通信，并采用平均权重的策略共享算法迭代中的更新参数，保证每个节点可以充分利用全局数据的通信，大大缩减了通信开销。同时，由每轮迭代的更新步长来控制随机噪声的大小，使随机噪声随着迭代轮次的进行而越来越小，在满足差分隐私保护要求的同时，进一步增加了数据的可用性。Beimel 等[44]在机器学习中提出了一种宽松要求的差分隐私保护方法，通过保护学习样本中样本标签的隐私来达到隐私保护的目的，经证明该算法的复杂度与不提供差分隐私保护的学习算法复杂度相同。

Kasiviswanathan 等[45]指出了噪声条件下的机器学习与保证隐私条件下机器学习的区别，对 2 种典型的机器学习模型 PAC（probabilistically approximately correct）和 SQ（statistical query）进行了研究，提出了与无隐私条件机器学习时复杂度相同的 PAC 算法，并讨论了交互式与非交互式（自适应与非自适应）的 SQ 算法，表明可以在与 $\log|C|$ 成比例的样本复杂度下对每个概念类别 C 进行隐私学习。Beimel 等[46]发现对于概念类别 C 的有隐私机器学习中，不同于无隐私的一般机器学习，输出结果在 C 中的学习模型与输出结果不在 C 中的学习模型的样本复杂度有很大的区别。他们证明了每一个输出在 C 类的 $\varepsilon-\mathrm{POINT}_d$ 隐私学习模型的样本复杂度必须达到 $\Omega(d)=\Omega(\log|\mathrm{POINT}_d|)$，并在文献[47]提出一种代替隐私学习模型。

可以看出，将差分隐私保护技术应用在数据分析中同样也面临着差分隐私在数据发布应用中所面临的挑战，当前存在更突出的挑战介绍如下。

1) 无论通过哪种数据分析技术，在对大数据进行分析的过程中，对原始数据和计算结果往往需要反复调取并多次迭代，用现有的差分隐私保护机制来实现对数据分析过程及结果的保护所需要的噪声量以及计算复杂度仍旧过大，不适合应用在大规模数据集中。如何创新选择随机算法进行差分隐私保护并有效降低算法的计算复杂度是未来的研究方向。

2) 当前在数据分析中应用差分隐私保护的算法适用性较差，在不同的场景假设下或应用不同的数据分析技术时都需有特定的算法，且针对的目标函数也较为简单，在大数据环境中的实际应用性不佳。如何设计通用于数据分析中多种目标函数的差分隐私保护算法是接下来的研究方向。

4　云计算与大数据计算中的差分隐私保护

隐私数据在云计算与大数据计算的过程中往往要经历多个环节，每一个环节的数据泄露都会造成隐私泄露的风险。当前，差分隐私保护技术已被广泛地运用在云计算与大数据计算中，如部署在苹果手机上的用户数据收集模块[48]。基于云计算与大数据计算的自身特征，可以总结出要将差分隐私保护运用到云计算与大数据计算中主要存在以下两方面的问题：1) 计算资源提供方不可信所造成的隐私信息在数据计算过程中难以保证不被泄露的问题；2) 海量、高频数据流所造成的差分隐私保护机制难以实现的问题。

针对问题 1)的现有研究较少，通过访问控制与差分隐私保护相结合是一种可行的研究思路。Roy 等[49]在 MapReduce 计算架构下，研究了云计算中的信息安全问题。他们认为云计算中隐私保护所面临的一个巨大问题是用户和云服务商都不想耗费太多的计算资源来进行隐私保护，而在云计算的过程中，数据的输入输出、数据的分布式运算以及数据在用户和云服务商直接传输的过程中，任意一条数据信息的泄露都有可能造成不可挽回的隐私泄露。Roy 等将差分隐私保护与强访问控制相结合，提出了 Airavat 系统。在差分隐私保护部分中，Airavat 系统充分考虑用户的信息安全知识水平、云服务商是否可信等因素，通过一系列的参数限制与组合策略，在 Reduce 环节添加满足差分隐私条件的随机噪声，达到差分隐私保护的要求。Airavat 系统主要通过提前声明 Mapper 的输出范围，估计 Reducer 中的数据敏感度，从而限定一部分超出阈值的数据的输出，来优化所添加的噪声量。Airavat 系统的优势在于其不需要额外审计不受信任的代码，但过多的参数和策略限制也影响了其在应用上的可扩展性。

针对问题 2)，当前研究主要利用抽样、分组等方法来缓解数据样本的海量与高频特性，以便于差分隐私保护机制的实现。Mir 等[50]在模拟 CDR（通话记录）模型 WHERE 的基础上，提出一种 DP-WHERE 模型，该模型在 WHERE 模型的各个计算环节所得的概率分布中添加受控随机噪声实现差分隐私保护，并在关键步骤中采用文献[30]中提的后处理技术和文献[51]中所提的 grouping 技术优化噪声量，产生相应的概率分布，通过系统地抽样这些分布，合成含有位置和相关时间的模拟数据信息并发布，实现了在海量数据计算中的差分隐私保护。该模型不需要进一步访问原始 CDR，并且用来合成数据的参数有很强的通用性，在保持隐私性的同时，在应用方面极具扩展性。Cormode 等[52]针对可以由树结构索引的多维数据的差分隐私保护进行了研究，提出了通过差分隐私空间分解来对数据分布进行差分隐私保护下的发布。分割点的选择以及分布区域的描述策略在其所提方法中作为基本步骤来保证高维数据差分隐私保护中的隐私性，基于约束推理的后处理技术以及样本均匀采样技术则作为噪声优化方法来保证海量、高频数据流中高维数据差分隐私保护中的可用性。Wang 等[53]也做了类似的研究，不同的是其采用了斐波那契混合数列来估计并分配树中各层节点的隐私预算量，以较大的隐私预算量来提高数据查询的准确性。

与此同时，云计算与大数据计算中的差分隐私保护应用还需解决因分布式计算架构自身特点所引起的开销过大的挑战。不同于差分隐私保护应用于集中式数据库，在分布式计算架构下，各个节点的数据相互独立。严格来说，需使各个节点数据都满足差分隐私保护要求才能保证整个分布式计算架构下的数据处理是差分隐私的，这就要求在各个节点都需进行随机扰动，那么随后的数据通信与协同计算都将需要更大的开销。为此，如何设计分布式计算架构下的差分隐私机制构成下一步研究方向。

5　面向位置与轨迹服务的差分隐私保护

当前，随着公共交通、网购快递、网络订餐等行业的兴起，使以<用户名, 位置, 轨迹>为格式的数据集呈爆炸式增长，形成了位置与轨迹大数据。其隐私保护数据发布机制主要有两类：一是发布轨迹数据集，每一条轨迹作为一个记录，目的是保护轨迹信息；二是发布一条轨迹信息，轨迹中的每个位置作为一个记录，目的是保护每个点的位置信息。而将差分隐私保护应用于位置轨迹大数据中，面临以下 3 个方面的挑战。

一是在满足差分隐私保护要求的基础上，数据集的稀疏性将引起随机化机制产生大量的噪声。二是敏感度的计算。对于位置与轨迹数据集，随机化机制与距离测量相关联。然而，如果人们只是通过传统的方式测量敏感度，当涉及位置与轨迹之间的

最大距离时，敏感度将会非常大。为了达到严格的隐私保证，必须增加大量的噪声，这将大大降低位置与轨迹数据集的效用。三是位置与轨迹数据集的语义保留。当对位置与轨迹数据集进行随机化时，传统的差分隐私机制不考虑位置与轨迹数据集的语义，仅基于距离的测量来辨别位置，这样做难以辨别其属于哪个具体的区域，如不能判断某一位置点属于哪个城区。

Lin 等[54]针对人体传感器网络（BSN, body sensor network）中的隐私大数据提出了一种差分隐私保护方案。该方案基于非交互式数据发布框架，通过 Haar 小波变换将直方图转化为二叉树，由二叉树的高度来决定全局敏感度，在一定程度上缓解了因全局敏感度过大造成提供差分隐私保护随机噪声过大的问题。

Xiong 等[55]提出一种 PriLocation 算法，首先通过聚类来对相邻位置信息进行分组，并限制随机域，以此来收缩随机区域减少所需噪声；随后利用位置信息的分层结构，通过对聚类权重扰动隐藏位置信息权重，提出了基于层次结构的敏感度概念的位置数据集，并对隐私位置选择来隐藏每一个用户的真实位置。

He 等[56]提出一种基于个人原始 GPS 轨迹合成支持差分隐私保护的移动数据系统——DPT。该系统利用参考系统的层次结构，以多个分辨率来离散空间域，并为每个分辨率维护一个前缀树。不同的参考系统捕获不同速度的运动。在每个参考系统中，个体仅能从一个点移动到限定个数的相邻点。因此，尽管有较多数量的前缀树，但是每个树具有小得多的分支因子，从而使维持模型的计数数量呈指数减少趋势，选择差分隐私的方式设置相关参数，并通过自适应机制以及方向加权来提高效用。其存在的不足与文献[36-37]的问题相同，他们的结论都是建立在发布的原始轨迹包含许多公共前缀这一无效假设的基础上的。

Hua 等[57]旨在消除上述假设，提出一种提供差分隐私保护的位置泛化算法，该算法基于轨迹距离并利用指数机制来概率地合并位置，然后以差分隐私化的方式发布泛化后的位置轨迹信息。

Li 等[58]首先使用 k-means 聚类来分割原始位置，由此获得 $n-1$ 个新的广义轨迹，并随机选择 $n-n_1$ 个原始轨迹来近似代替原始轨迹数据库，随后产生有界的拉普拉斯噪声，将它们添加到不同轨迹的计数值中，并用发布这些轨迹与其计数值。Chatzikokolakis 等[59]提出了一种基于预测机制的位置轨迹差分保护方法，预测机制通过利用数据的关联性和历史数据记录来对用户现有位置进行预测，由测试函数对预测进行测试，仅对不满足测试要求的预测结果进行新的噪声添加，与独立噪声添加机制相比，大大降低了噪声量。

Asada 等[60]将差分隐私用于位置偏好推荐系统，在利用矩阵分解提高推荐的准确性的同时，有选择性地在局部实施差分隐私以严格保护用户隐私，并实现了位置偏好推荐与隐私保护的最佳平衡，在保证推荐准确有效的情况下也保证了位置数据的隐私性。

虽然当前研究针对降低因位置轨迹数据稀疏性所引起的噪声量以及降低数据敏感度这两方面挑战有了较好的成果，但是针对第三个挑战，当前并没有专门针对位置与轨迹数据集保留语义方面的工作。在具体应用中，虽然差分隐私保护已广泛应用于各类位置与轨迹服务系统中，目前还是缺乏成熟完备的应用系统，且应用系统在隐私保护的完整性、适用数据的可扩展性等方面仍需进行进一步的研究。以上方面可构成面向位置与轨迹服务的差分隐私保护下一步的研究方向。

6 社交网络中的差分隐私保护

近年来，社交网络的迅猛发展在给人们生活带来极大便利的同时，由于各个网络社交平台一般都需要用户在注册时提供一定程度的个人身份信息，也给人们网络生活中的个人隐私保护带来了巨大挑战。在对社交网络进行数据分析时，通常利用图结构来描述社交网络活动，其中图中的节点代表用户，边代表用户之间的关系或社交活动。将差分隐私保护应用到社交网络中等同于将其应用到图结构中。

对图结构进行差分隐私保护一般分为基于节点的差分隐私保护和基于边的差分隐私保护，基于出度的差分隐私保护和基于分区的差分隐私保护则作为 2 种新颖的差分隐私保护概念在近年来被广泛应用。Task 等[61]认为基于节点的差分隐私保护方法存在在数据中添加的噪声量过大、代价昂贵的问题，且存在噪声多余添加的问题；基于边的差分隐私的隐私性并不严密。基于出度的差分隐私与基于分区的差分隐私作为 2 个新颖的社交网络隐私保护

机制，能够在引入极小噪声的情况下提供强有力的隐私保护。基于出度的差分隐私通过在隐私数据中增加或删除任意节点的外部链接来保护数据参与者所提供数据的隐私，在此机制下，Task 等[62]提出一种基于自组网样式的分析算法，对以前标准执行过于敏感的查询提供近似结果。而基于分区的差分隐私则针对社交网络中属性繁多的特点，在差分数据中任意添加或删除一个子图来满足差分隐私要求，为社交网络多属性研究提供比基于节点的隐私保护更高的隐私级别，并且能够进行在基于节点隐私保护机制下难以进行的分析研究。

另一方面，要将差分隐私保护应用到社交网络的数据挖掘中需研究差分隐私保护下的社交网络分析技术，即差分隐私保护下的图分析技术。针对子图计数查询问题，Karwa 等[63]在利用文献[14]所提的查询函数局部敏感度与平滑上界来计算随机噪声量的基础上，将对查询结果添加随机噪声的方法[64]应用到了 *K*-星计数查询中，缓解了子图计数查询函数一般全局敏感度较大的问题，并提出了 *K*-三角形数的查询算法；针对聚类系数计算问题，Wang 等[65]提出一种 D&C 算法，首先将聚类系数计算函数分解为多个分计算函数，然后通过预先设定的噪声添加与隐私预算分配策略对每个分计算函数结果添加随机噪声，最后通过满足第 2 节中的差分隐私保护性质 1 和性质 2 的数学运算合成输出结果，以达到差分隐私保护的要求；针对边隐私保护问题，Costea 等[66]在不考虑目标节点和边中信息是否为敏感信息的情况下，利用拉普拉斯机制在边权重中添加随机噪声并生成新的图，并利用迪杰斯特拉算法计算原图与生成图中的最短路径来评估差分隐私保护效果；针对度分布问题，Hay 等[35,67]通过一致性约束进行图中度分布的估计，并给出了图中度分布计算的最小二乘解算法，在满足差分隐私保护的条件下极大程度地提高了度分布查询结果的准确性；针对目的节点隐私保护问题，Javidbakh 等[68]在路由开销约束下，研究了目的节点在差分隐私保护下的最优路由开销方案，采取了使数据传递经过多个目的节点的策略，并将目的节点信息差分化的方式来实现网络路由中的数据保护。虽然，Javidbakh 等通过概率优化意图将通信开销降低到最低值，但其所提方案不可避免地仍将造成更多的通信负担。Li 等[69]研究了图合成的工作流程，提出了一种基于加权隐私综合查询方法，针对不准确的协同系数以最佳种子图替换一般种子图，以此作为马尔可夫链中的初始状态，并且通过交叉三角形的信息一步一步地进行以增加合成图中的三角形的数量，以此发布满足差分隐私的社交图。

在社交网络大数据环境下，以图结构来重构社交网络展开差分隐私保护应用研究，不可避免地将面临与差分隐私保护在分布式计算应用中类似的通信开销过大问题；同时，社交网络中各个节点的数据具有关联性，这也将不可避免地要求更多的噪声。下一步研究可在充分考虑社交网络中数据相关性的情况下，寻求新的差分隐私保护策略来降低因数据反复传递引起的通信开销。

7　结束语

差分隐私保护作为一种严格可证明的数学模型，在大数据环境下已经被广泛地应用于各个领域。随着近年来相关研究的不断深入，差分隐私保护的理论及相关概念日益完善，其在数据发布、数据分析、云计算等领域的应用也越来越成熟。

本文首先对差分隐私保护的基本理论进行了介绍，随后综述了差分隐私保护在数据发布与分析、云计算与大数据计算、面向位置与轨迹服务和社交网络中的应用。从本文所做综述中不难看出，针对差分隐私保护应用的研究主要集中在保证数据满足差分隐私保护要求的同时，如何提高数据可用性与降低算法复杂度上。针对这一核心问题，相关研究从差分隐私保护原理、噪声添加机制与位置以及数据处理方式等方面已对差分隐私保护的应用取得了卓有成效的优化，且相关应用成果可以在多个领域的不同场景下交叉应用。但是，在大数据环境下，差分隐私保护想要得到系统性、普遍性的应用，还需从以下几个方面开展下一步的研究。

1) 隐私识别与量化评估技术

数据隐私主动防护体系的基础是隐私分级与量化评估。大数据环境下，数据来源复杂、数据操作繁多等因素都将造成隐私识别与量化评估的困难[70]。对于差分隐私保护而言，隐私识别能够区分数据集中的隐私数据与一般数据，明确差分隐私保护的对象，减少因保护范围的无故扩大造成的开销；量化评估则可以区分隐私数据的重要程度，明确差分隐私保护的隐私级别，利于解决数据隐私性与可用性的平衡问题，即隐私预算量 ε 的设置与分配问题。合理的隐私预算分配，将使差分隐私保护

的效用最大化。特别是在高速动态数据流的情况下，只有在对隐私预算量ε进行合理设置与分配的情况下，差分隐私保护才能持续高效地对隐私数据提供保护，在这个过程中，隐私识别与量化评估技术则显得尤为重要。今后隐私识别与量化评估可通过隐私概念公理化表述、隐私及隐私集合测度概念提出、隐私计算体系构建、隐私计算方法探究、隐私保护效果评估及系统化保护应用研究等展开。

2) 主动差分隐私保护框架确立

从当前的研究情况来看，差分隐私保护在大数据环境下的应用往往是被动进行的，存在场景依赖缺陷。针对不同隐私保护对象与不同隐私保护需求的差分隐私保护过程中，需重新设计相应算法，没有形成智能化的主动差分隐私保护框架。

在大数据环境下，建立主动差分隐私保护框架主要需解决两方面的问题：① 由于查询函数种类多造成的查询函数敏感度主动计算困难问题；② 由于数据类型复杂引起的噪声机制自动优化困难问题。针对问题①，可进一步研究查询函数敏感度边界设定与差分隐私保护效用之间的博弈关系，寻求以近似敏感度代替精确敏感度的同时使额外噪声量降到最低；针对问题②，可将噪声机制的优化过程拆分成可合并的多个方面，研究噪声机制的选择性优化方法。

3) 多元融合的差分隐私保护体系构建

差分隐私保护通过引入足够的噪声量来满足其严格的隐私定义。但在大数据环境下，往往会出现由于添加的噪声量过大造成大数据不可用的情况。针对这一问题，一些学者试图通过降低差分隐私保护的隐私级别来提高数据的可用性[26,71-72]。从该角度出发，可研究建立多元融合的差分隐私保护体系，寻求将现有的多种隐私保护技术与差分隐私保护融合到同一体系中，当出现因需满足数据可用性而造成差分隐私保护隐私性降低的情况时，研究应用其他隐私保护技术弥补隐私性的相关方法。

4) 新信息技术框架下差分隐私应用方法

随着信息技术的不断发展，大数据环境下的数据隐私特征也在不断发展和变化，具体表现在数据使用权由封闭转向开放、数据计算由单极转向多级、数据所有权由固化转向流通、数据隐私边界由粗放转向精细。当前的隐私保护技术、数据管理策略、运营保障制度在新信息技术框架下存在许多短板。因此，对于差分隐私保护，需研究新信息技术框架下其具体应用方法，当前而言，具体可研究在5G 通信网络下差分隐私如何在工业互联网中有效实施、在基于人工智能所衍生的隐私威胁下如何保证差分隐私不失效、依托人工智能成果辅助优化差分隐私保护效用等多个方面。

参考文献：

[1] BERTINO E, FERRARI E. Big data security and privacy[M]// A Comprehensive Guide Through the Italian Database Research Over the Last 25 Years. Springer International Publishing, 2018:757-761.

[2] YANG Z, ZHANG Y, JIA H. Influencing factors of online P2P lending success rate in China[J]. Annals of Data Science, 2017, 4(2): 1-17.

[3] HUANG, HUI R. Online P2P lending and regulatory responses in China: opportunities and challenges[J]. European Business Organization Law Review, 2018, 19(1): 63-92.

[4] SREEVANI P, NIRANJAN D P, SHIREESHA P. A novel data anonymization technique for privacy preservation of data publishing[J]. International Journal of Engineering Sciences & Research Technology, 2014, 3(11): 201-205.

[5] ZENG L, POLYTECHNIC L. Research on new data encryption algorithm in big data environment[J]. Bulletin of Science & Technology, 2017, 33(6): 205-208.

[6] 李凤华, 李晖, 贾焰, 等. 隐私计算研究范畴及发展趋势[J]. 通信学报, 2016, 37(4): 1-11.
LI F H, LI H, JIA Y, et al. Privacy computing: concept, connotation and its research trend[J]. Journal on Communications, 2016, 37(4): 1-11.

[7] 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[J]. 软件学报, 2016, 27(8): 1891-1903.
PENG C G, DING H F, ZHU Y J, et al. Information entropy models and privacy metrics methods for privacy protection[J]. Journal of Software, 2016, 27(8): 1891-1903.

[8] 熊金波, 王敏燊, 田有亮, 等. 面向云数据的隐私度量研究进展[J]. 软件学报, 2018, 29(7): 1963-1980.
XIONG J B, WANG M S, TIAN Y L, et al. Research progress on privacy measurement for cloud data[J]. Journal of Software, 2018, 29(7): 1963-1980.

[9] DWORK C. Differential privacy[M]// Automata, Languages and Programming. Springer Berlin Heidelberg, 2006: 1-12.

[10] SHRIVASTVA K M P, RIZVI M A, SINGH S. Big data privacy based on differential privacy a hope for big data[C]// International Conference on Computational Intelligence and Communication Networks. IEEE, 2015: 776-781.

[11] HAEBERLEN A, PIERCE B C, NARAYAN A. Differential privacy under fire[C]// Usenix Conference on Security. USENIX Association, 2011: 33.

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

[13] DWORK C, MCSHERRY F, NISSIM K. Calibrating noise to sensitivity in private data analysis[J]. Proceedings of the VLDB Endowment, 2006, 7(8): 637-648.

[14] NISSIM K, RASKHODNIKOVA S. Smooth sensitivity and sampling in private data analysis[C]// Thirty-Ninth ACM Symposium on Theory

of Computing. ACM, 2007: 75-84.

[15] DWORK C, ROTH A. The algorithmic foundations of differential privacy[M]. Now Publishers Inc. 2014.

[16] MCSHERRY F, TALWAR K. Mechanism design via differential privacy[C]//48th Annual IEEE Symposium on Foundations of Computer Science. IEEE, 2007: 94-103.

[17] CHAUDHURI K, MONTELEONI C, SARWATE A D. Differentially private empirical risk minimization[J]. Journal of Machine Learning Research, 2009, 12(2): 1069-1109.

[18] HAN C, WANG K. Sensitive Disclosures under differential privacy guarantees[C]//IEEE International Congress on Big Data. IEEE Computer Society, 2015: 110-117.

[19] HAI B, NISSIM K. Impossibility of differentially private universally optimal mechanisms[J]. Foundations of Computer Science Annual Symposium on, 2010, 43(5): 71-80.

[20] GHOSH A, ROUGHGARDEN T, SUNDARARAJAN M. Universally utility-maximizing privacy mechanisms[C]// ACM Symposium on Theory of Computing. ACM, 2009: 351-360.

[21] GUPTE M, SUNDARARAJAN M. Universally optimal privacy mechanisms for minimax agents[C]//Twenty-Ninth ACM Sigmod-Sigact-Sigart Symposium on Principles of Database Systems. ACM, 2010: 135-146.

[22] GENG Q, VISWANATH P. The optimal mechanism in differential privacy[C]// IEEE International Symposium on Information Theory. IEEE, 2013: 2371-2375.

[23] CHEN C L, PAL R, GOLUBCHIK L. Oblivious mechanisms in differential privacy: experiments, conjectures, and open questions[C]// Security and Privacy Workshops. IEEE, 2016: 41-48.

[24] LIN C, SONG Z, SONG H, et al. Differential privacy preserving in big data analytics for connected health[J]. Journal of Medical Systems, 2016, 40(4): 1-9.

[25] JI Z, XIN D, YU J, et al. Differentially private multidimensional data publication[J]. China Communications, 2014, 11(s1): 79-85.

[26] DE A. Lower bounds in differential privacy[J]. Lecture Notes in Computer Science, 2013, 7194: 321-338.

[27] ZHANG X, WU Y, WANG X. Differential privacy data release through adding noise on average value[M]// Network and System Security. Springer Berlin Heidelberg, 2012: 417-429.

[28] LI C, HAY M, RASTOGI V, et al. Optimizing linear counting queries under differential privacy[C]// Twenty-Ninth ACM Sigmod-Sigact-Sigart Symposium on Principles of Database Systems. DBLP, 2010: 123-134.

[29] KOUFOGIANNIS F, HAN S, PAPPAS G J. Gradual release of sensitive data under differential privacy[J]. Journal of Privacy and Confidentiality, 2015(12): 1-25.

[30] HAY M, RASTOGI V, MIKLAU G, et al. Boosting the accuracy of differentially private histograms through consistency[J]. Proceedings of the VLDB Endowment, 2010, 3(1-2): 1021-1032.

[31] KELLARIS G, PAPADOPOULOS S, XIAO X, et al. Differentially private event sequences over infinite streams[J]. Proceedings of the VLDB Endowment, 2014, 7(12): 1155-1166.

[32] FAN L, XIONG L. An adaptive approach to real-time aggregate monitoring with differential privacy[J]. IEEE Transactions on Knowledge & Data Engineering, 2014, 26(9): 2094-2106.

[33] CHAN T H H, SHI E, SONG D. Private and continual release of statistics[J]. ACM Transactions on Information & System Security, 2011, 14(3): 1-24.

[34] DWORK C, NAOR M, PITASSI T, et al. Differential privacy under continual observation[C]//STOC'10—Proceedings of the 2010 ACM International Symposium on Theory of Computing. ACM, 2010: 715-724.

[35] WANG Q, ZHANG Y, LU X, et al. RescueDP: real-time spatio-temporal crowd-sourced data publishing with differential privacy[C]//International Conference on Computer Communications. IEEE, 2016: 1-9.

[36] CHEN R, FUNG B C M, DESAI B C. Differentially private trajectory data publication[J]. arXiv Preprint, arXiv: 1112.2020, 2011.

[37] CHEN R, ACS G, CASTELLUCCIA C. Differentially private sequential data publication via variable-length n-grams[C]// ACM Conference on Computer and Communications Security. ACM, 2012: 638-649.

[38] KANG H Y, ZHANG S X, JIA Q Q. A method for time-series location data publication based on differential privacy[J]. Wuhan University Journal of Natural Sciences, 2019(2): 107-115.

[39] ABADI M, GOODFELLOW I. Deep learning with differential privacy[C]//ACM Sigsac Conference on Computer and Communications Security. ACM, 2016: 308-318.

[40] CAI T T, WANG Y, ZHANG L. The cost of privacy: optimal rates of convergence for parameter estimation with differential privacy[J]. Statistics, 2019.

[41] MCSHERRY F, MIRONOV I. Differentially private recommender systems: building privacy into the net[M]// Differentially Private Recommender Systems. 2009: 627-636.

[42] XU C, REN J, ZHANG D, et al. GANobfuscator: mitigating information leakage under GAN via differential privacy[J]. IEEE Transactions on Information Forensics and Security, 2019, 14(9): 2358-2371.

[43] LI C, ZHOU P, JIANG T. Differential privacy and distributed online learning for wireless big data[C]// International Conference on Wireless Communications & Signal Processing. IEEE, 2015: 1-5.

[44] BEIMEL A, NISSIM K, STEMMER U. Private learning and sanitization: pure vs. approximate differential privacy[M]// Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques. Springer Berlin Heidelberg, 2013: 363-378.

[45] KASIVISWANATHAN S P, LEE H K, NISSIM K, et al. What can we learn privately?[J]. Siam Journal on Computing, 2008, 40(3): 793-826.

[46] BEIMEL A, KASIVISWANATHAN S P, NISSIM K. Bounds on the sample complexity for private learning and private data release[C]// International Conference on Theory of Cryptography. Springer-Verlag, 2010: 437-454.

[47] BEIMEL A, NISSIM K, STEMMER U. Characterizing the sample complexity of private learners[J]. Computer Science, 2014: 97-110.

[48] TANG J, KOROLOVA A, BAI X, et al. Privacy loss in Apple's implementation of differential privacy on MacOS 10.12[J]. arXiv Preprint, arXiv:1709.02753, 2017.

[49] ROY I, SETTY S T V, KILZER A, et al. Airavat: security and privacy for MapReduce.[C]// Usenix Symposium on Networked Systems Design and Implementation. DBLP, 2010: 297-312.

[50] MIR D J, ISAACMAN S, CACERES R, et al. DP-WHERE: differentially private modeling of human mobility[C]// IEEE International Conference on Big Data. IEEE, 2013: 580-588.

[51] KELLARIS G, PAPADOPOULOS S. Practical differential privacy via

grouping and smoothing[J]. Proceedings of the VLDB Endowment, 2013, 6(5): 301-312.

[52] CORMODE G, PROCOPIUC C, SRIVASTAVA D, et al. Differentially private spatial decompositions[C]// International Conference on Data Engineering. IEEE, 2012: 20-31.

[53] WANG J, LIU S, LI Y K, et al. Differentially private spatial decompositions for geospatial point data[J]. China Communications, 2016, 13(4): 97-107.

[54] LIN C, WANG P, SONG H, et al. A differential privacy protection scheme for sensitive big data in body sensor networks[J]. Annals of Telecommunications, 2016, 71(9-10): 465-475.

[55] XIONG P, ZHU T, NIU W, et al. A differentially private algorithm for location data release[J]. Knowledge & Information Systems, 2016, 47(3): 647-669.

[56] HE X, CORMODE G, SRIVASTAVA D, et al. DPT: differentially private trajectory synthesis using hierarchical reference systems[J]. Proceedings of the VLDB Endowment, 2015, 8(11): 1154-1165.

[57] HUA J, GAO Y, ZHONG S. Differentially private publication of general time-serial trajectory data[C]// Computer Communications. IEEE, 2015: 549-557.

[58] LI M, ZHU L, ZHANG Z, et al. Achieving differential privacy of trajectory data publishing in participatory sensing[J]. Information Sciences, 2017, 400-401: 1-13.

[59] CHATZIKOKOLAKIS K, PALAMIDESSI C, STRONATI M. A predictive differentially-private mechanism for mobility traces[J] Privacy Enhancing Technologies, 2014,8555: 21-41.

[60] ASADA M, YOSHIKAWA M, CAO Y. When and where do you want to hide? Recommendation of location privacy preferences with local differential privacy[C]// IFIP Annual Conference on Data and Applications Security and Privacy. Springer, 2019: 1-20.

[61] TASK C, CLIFTON C. A guide to differential privacy theory in social network analysis[C]// International Conference on Advances in Social Networks Analysis and Mining. IEEE Computer Society, 2012: 411-417.

[62] TASK C, CLIFTON C. What should we protect? defining differential privacy for social network analysis[M]// State of the Art Applications of Social Network Analysis. Springer International Publishing, 2014: 139-161.

[63] KARWA V, RASKHODNIKOVA S, SMITH A, et al. Private analysis of graph structure[J]. ACM Transactions on Database Systems, 2011, 39(3): 1146-1157.

[64] DWORK C, MCSHERRY F, NISSIM K. Calibrating noise to sensitivity in private data analysis[C]// Theory of Cryptography Conference. Springer, 2006: 265-284.

[65] WANG Y, WU X, ZHU J, et al. On learning cluster coefficient of private networks[C]// International Conference on Advances in Social Networks Analysis and Mining. IEEE Computer Society, 2012: 395-402.

[66] COSTEA S, BARBU M, RUGHINIS R. Qualitative analysis of differential privacy applied over graph structures[C]// Roedunet International Conference. IEEE, 2013: 1-4.

[67] HAY M, LI C, MIKLAU G, et al. Accurate estimation of the degree distribution of private networks[C]// Ninth IEEE International Conference on Data Mining. IEEE Computer Society, 2009: 169-178.

[68] JAVIDBAKHT O, VENKITASUBRAMANIAM P. Differential privacy in networked data collection[C]// Conference on Information Science and Systems. IEEE, 2016: 117-122.

[69] LI X Y, YANG J, SUN Z J, et al. Publishing social graphs with differential privacy guarantees based on wPINQ[J]. Chinese Journal of Electronics, 2019, 28(2): 273-279.

[70] KIFER D, MACHANAVAJJHALA A. No free lunch in data privacy[C]// ACM SIGMOD International Conference on Management of Data. DBLP, 2011:193-204.

[71] LI N, QARDAJI W, DONG S. On sampling, anonymization, and differential privacy or, *k*-anonymization meets differential privacy[C]// ACM Symposium on Information, Computer and Communications Security. ACM, 2012: 32-33.

[72] GEHRKE J, HAY M, LUI E, et al. Crowd-blending privacy[C]// Cryptology Conference on Advances in Cryptology. Springer-Verlag, 2012: 479-496.

[作者简介]

付钰（1982− ），女，湖北武汉人，博士，海军工程大学副教授、硕士生导师，主要研究方向为信息安全、风险评估。

俞艺涵（1992− ），男，浙江金华人，海军工程大学博士生，主要研究方向为信息安全、隐私保护。

吴晓平（1961− ），男，山西新绛人，博士，海军工程大学教授、博士生导师，主要研究方向为信息安全、密码学。