

大数据及其隐私保护

方贤进, 肖亚飞, 杨高明

安徽理工大学计算机科学与工程学院, 安徽 淮南 232001

摘要

在对大数据进行发布或数据挖掘的过程中, 隐私泄露是人们最关心的问题, 但目前关于大数据隐私保护的研究还处在初级阶段。介绍了有关隐私保护系统的基础知识, 包括数据参与角色与数据操作的定义, 给出了隐私保护系统的数学描述与隐私度量方法, 分析了隐私保护的数学模型, 包括 k -匿名模型与差分隐私模型。回顾了基于位置服务的隐私保护及其应用, 总结了大数据时代隐私保护的挑战与机遇, 指出了用于改进现有隐私保护方法的研究方向, 以满足大数据前所未有的各种计算需求。

关键词

大数据; k -匿名; 差分隐私; 隐私模型

中图分类号: TP311

文献标识码: A

doi: 10.11959/j.issn.2096-0271.2017051

Privacy preserving in the age of big data

FANG Xianjin, XIAO Yafei, YANG Gaoming

School of Computer Science and Engineering, Anhui University of Science & Technology,
Huainan 232001, China

Abstract

One of biggest concerns of big data is privacy, especially, in the processing of big data publishing or data mining. However, the study on big data privacy is still at a very early stage. The preliminary knowledge about the definition of roles and operations of privacy system were introduced. The mathematical description and measurement metrics of privacy study was given. The models of privacy preserving were analyzed, including k -anonymity and differential privacy. The current situation of privacy preserving in big data age was reviewed, especially, the privacy preserving based location-based services and its applications were summarized. The challenges and opportunities in the age of big data were summarized. The directions to improve the existing privacy protection methods satisfying the unprecedented computational requirements of big data were pointed out.

Key words

big data, k -anonymity, differential privacy, privacy model

1 引言

根据维基百科的定义,大数据(big data)是指一个特殊的数据集,其足够大而复杂,并且传统的数据处理应用软件不能对它进行有效处理。大数据面临的挑战包括大数据的获取、存储、管理、搜索、共享、传输、可视化、查询、更新与信息隐私等。一般来说,大数据分析是指预测分析、用户行为分析或者其他高级数据分析方法,并且能够从数据中获取价值。大数据很少是指数据集特别大^①,毫无疑问大数据的容量确实很大,但这并不是大数据生态系统的最具实质性的特性^②,正如麦肯锡全球研究所认为“大数据是一种规模大到在获取、存储、管理、分析方面大大超出了传统数据库软件工具能力范围的数据集合,具有海量的数据规模(volume)、快速的数据流转(velocity)、多样的数据类型(variety)和价值密度低(value)四大特征”。

数据集的快速增长在某种程度上是因为廉价、大量的信息传感移动设备、遥感设备、软件日志、摄像机、麦克风、射频识别(radio frequency identification, RFID)读取器、无线传感网络设备等采集数据的日益增长。从20世纪80年代开始,人均存储信息容量每40个月大约增加2倍,但截至2012年,每天产生的数据量为2.5 EB,包括人们在社交网络上的聊天记录、发出的照片与视频剪辑、电子文档、电子邮件、Web冲浪的足迹、在电子商务网站上的各种购物行为(包括用户的ID、IP地址、搜索的关键词、页面停留时间、点击链接记录、对商品的打分、购买商品的记录等)、电子支付记录等。另外,在公共场所的摄像头获取的视频、手机定位系统(包括基站

和全球定位系统(GPS)定位)留下的路线图、在各种情况下被录下的语音、驾车时的GPS信号、电子病历档案、公交刷卡记录等被动信息也是大数据的组成部分。还有各种传感器设备自动采集的有关温度、湿度、速度等万物信息,仍然是大数据的组成部分。总之,每个人、每种通信和控制类设备,无论它是软件还是硬件,都是大数据之源。

大数据挖掘是指在大数据集中发现知识、模式的一种计算过程,涉及人工智能、机器学习、统计分析方法、数据库系统等学科交叉。数据挖掘的目标就是从数据集中抽取信息,并且将其转换成可以理解的结构以供未来使用。大数据是信息时代的一个重要里程碑,它对人类社会产生了深远的影响。大数据挖掘能够极大地推进知识、服务以及社会各个领域生产力的发展。但是,针对大数据的数据挖掘和机器学习也对个人隐私造成了巨大的威胁。

随着社交网络和移动互联技术的迅猛发展,各类数据的采集、存储、分析、发布变得方便快捷。众多机构(如医疗机构、保险公司、电子商务公司、社交网站、电信运营商)发布行业数据以供研究,各种大数据分析公司、大数据分析竞赛等更是层出不穷。张华平等人^[1]根据1 700万条新浪微博对微博生态系统进行深度分析,从而掌握了新浪微博用户的各种宏观信息,构建了用户影响力模型,并深入研究了用户意图,反映了“大数据环境下无隐私”的状况;2006年,美国在线租赁公司网飞(Netflix)公司投资100万美元举办了一个为期3年的推荐系统算法竞赛,并发布一些经过匿名化处理的用户影评数据供参赛者测试。来自美国德克萨斯大学奥斯汀分校的两位研究人员利用该数据与公开的互联网电影数据库(internet movie database,

①
https://en.wikipedia.org/wiki/Big_data

②
<https://www.mendeley.com/research-papers/six-provocations-big-data/>

IMDb)用户影评数据之间的相关性,将网飞公司的一部分匿名用户与公开的IMDb用户进行了对应,并获得了IMDb用户在Netflix网站上观看的电影信息(包括涉及敏感题材的影片)^[2]。2013年8月,美国通用电气公司公布了由美国FlightStats公司提供的航空数据,希望参赛者开发出能够实时控制航班飞行路线、速度、高度和管制空域的模型,进而优化航班的总体运行效率;美国好事达保险公司提供2005—2007年的保险数据(包括具体的汽车情况及每辆车相关的赔偿支出次数和数量),悬赏1万美元寻求解决方案,希望更准确地预测汽车伤害索赔,以便优化保险定价方案。随着移动定位服务的流行,阿里巴巴和蚂蚁金服逐渐积累了来自用户和商家的海量线上线下交易数据。蚂蚁金服的线上到线下(online to offline, O2O)平台“口碑商家客流量预测系统”^③利用用户浏览行为数据(用户ID、商家ID、浏览时间)、用户支付行为数据(用户ID、商家ID、支付时间)、商家特征数据(商家ID、城市名、所在位置编号、人均消费、评分、评论数、门店等级)等为商家提供包括交易统计、销售分析和销售建议等在内的定制后端商业智能服务,为每个商家提供销售预测。基于预测结果,商家可以优化运营、降低成本,并改善用户体验,从而使该系统成为更加智能的商业平台,更好地服务社会。

由于数据集之间存在相关性,即使将数据进行匿名化处理,仍可导致各种敏感或隐私信息的泄露^[3]。而且,用于数据分析与数据挖掘而发布的数据本身就包含许多个人隐私信息,如健康医疗数据、个人消费习惯以及其他能够体现个人特征的数据,这些信息会随着数据集的发布、共享、挖掘而被泄露^[4,5]。对社交网络数据进行链接挖掘可以获得实体更丰富、更准确的信

息,发布与共享社交网络数据会导致隐私泄露,且社交网络中的隐私信息类型广泛,潜在的隐私泄露方式更加多样^[6]。删除数据集的标识符属性(如姓名、ID号等)能在一定程度上保护个人隐私,但一些攻击案例(如一致性攻击、合成式攻击、前景知识攻击、deFinetti攻击)表明^[7],这种简单操作远不足以保证隐私信息的安全。

总之,在数据分布(如计数查询、直方图发布、列联表发布、数据立方体、图数据发布等)、数据挖掘(如频繁项挖掘、回归、分类等)、机器学习(如电子商务中的推荐系统、预测分析)^[8]等领域都可能导致隐私或敏感信息的泄漏,隐私保护、防止敏感信息泄露成为当前面临的重大挑战。隐私保护研究的目标就是用修改隐私数据的技术,使修改后的数据在保护隐私的前提下,最大限度地保留原数据的整体信息,即隐私保护后的数据必须具有可用性,可以安全发布或供第三方研究,否则,被发布的数据将毫无价值。同时,隐私保护后的数据不会遭受抗匿名化等各种隐私攻击,即隐私保护后的数据具有隐私性。

③

<https://tianchi.shuju.aliyun.com/competition/information.htm?spm=5176.100068.5678.2.3TvCJe&raceld=231591>

2 隐私保护的基础知识

一个隐私保护系统包括各种参与者角色(participation role)、匿名化操作(anonymization operation)与数据状态(data status),它们之间的关系如图1所示。在隐私保护的研究中,有4个数据参与者角色^[9]。

数据生成者(data generator):指那些生成原始数据的个体或组织,例如病人的医疗记录、客户的银行交易业务。他们以某种方式主动提供数据(如发布照片到社交平台)或被动提供数据给别人(如

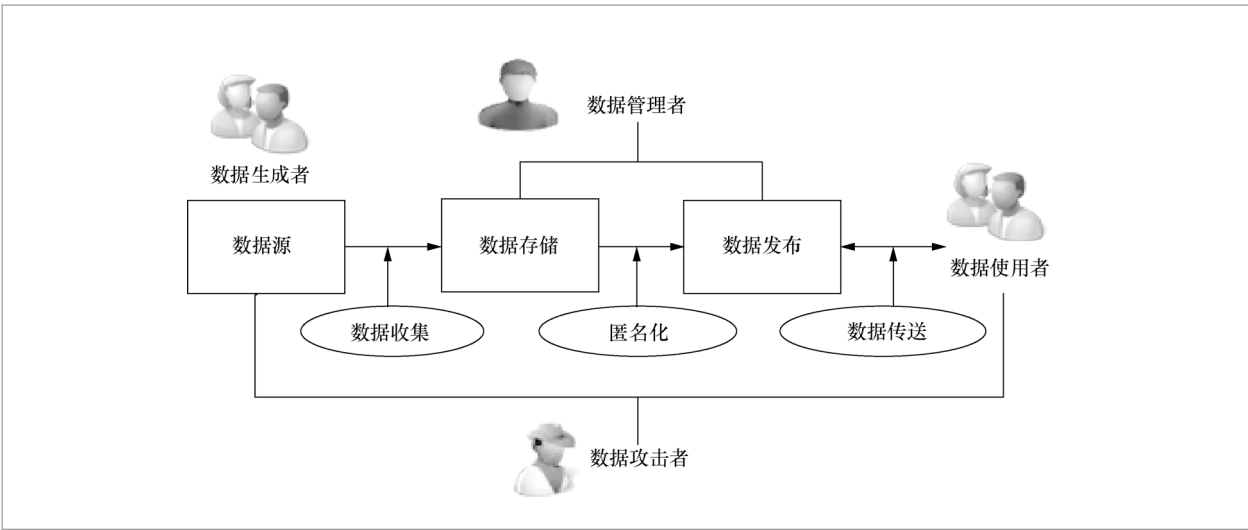


图1 一个隐私保护系统中的数据参与角色及其操作

在电子商务、电子支付系统中留下个人的信用卡交易记录等)。

- 数据管理者 (data curator): 指那些收集、存储、掌握、发布数据的个人或组织。
- 数据使用者 (data user): 指为了各种目的对发布的数据集进行访问的用户。
- 数据攻击者 (data attacker): 指那些为了善意或恶意的目的从发布的数据集中企图获取更多信息的人。数据攻击者是一种特殊类型的数据使用者。

隐私保护系统中存在3个主要的数据操作,包括数据收集、匿名化、数据传送。

- 数据收集 (collecting): 指数据管理者从不同的数据源收集数据。
 - 匿名化 (anonymizing): 指数据管理者为了发布数据而对收集的数据进行匿名化处理。
 - 数据传送 (communicating): 指数据使用者对发布的数据集执行信息检索。
- 隐私保护系统的数据集还存在以下3个不同的状态。
- 原始状态 (raw status): 指数据的原始格式。
 - 收集状态 (collected status): 指数

据已经被接收和处理 (如去除噪声、数据变换),并存储在数据管理者的存储空间中。

- 匿名化状态 (anonymized status): 指数据已经通过匿名化操作处理后的状态。

由图1可知,数据攻击者的目标可通过攻击任何数据角色与数据操作实现。

一般情况下,将数据集中的记录按照属性分成4类。

- 明确标识符 (explicit identifier): 指能清楚地标识个体的唯一属性,例如身份证号、驾驶证号等。
 - 准标识符 (quasi-identifier): 指当人们将这些标识符与其他辅助信息收集在一起时,可能会重新识别一个个体的属性的标识符,如年龄、职业、邮政编码等。
 - 敏感信息 (sensitive information): 指对手感兴趣、期望获得的信息,例如医疗记录中病人所患的“疾病”属性。
 - 其他信息 (other information)。
- 在表1中,驾照ID就属于明确标识符,而工作、性别与年龄都属于准标识符,疾病属于敏感信息。一条记录的所有准标识符的集合也称为一个等价类。

3 隐私研究的数学描述与数学模型

本节重点介绍隐私研究的数学描述与隐私保护的数学模型。

3.1 隐私研究的数学描述

为了更好地理解各种隐私保护方法，Yu S^[9]提出了隐私研究的数学模型，该模型如图2所示。

其中， $X = \{X_1, X_2, \dots, X_n\}$ 是原始数据，隐私保护系统或匿名化系统是一个映射函数 $F: X \rightarrow Y$ ，它将 X 转换为 $Y = \{Y_1, Y_2, \dots, Y_n\}$ ，作为期望的输出。函数 F 的目标是尽量保持 Y 的可用性，并且在一定程度上保护隐私。作为攻击者，其目标是获得 X' ，它是基于发布数据 Y 计算出来的，并且逼近于 X 。

隐私保护系统的两个目标是可用性 (utility) 与隐私性 (privacy)，也就是说可用性和隐私性是函数 F 的两个重要约束，既要考虑隐私性，也要考虑可用性，其目标是最小化隐私性泄露、最大化数据可用性。可用性通常用失真度 (distortion) D 进行度量，而隐私性通常用隐私泄露量 (leakage) L 标识。设 λ 表示勒贝格测度 (Lebesgue measure)，则数据失真度 D 可以表示为 $D = \lambda(X, Y)$ ，信息泄露量 L 可以表示为 $L = \lambda(X, Y')$ 。

有多种方法可以计算 D 和 L ，假如用平均均方差 (variance) 计算失真度 D ，可以利用式 (1)。

$$D = \frac{1}{n} \sum_{k=1}^n E[(x_k - y_k)^2] \quad (1)$$

其中， $E[\cdot]$ 表示 X 和 Y 联合分布的期望值。

如果用信息论中的互信息 (mutual

表1 数据表中不同类型标识符示例

驾照ID	工作 (job)	性别	年龄 (age)	疾病
3453543656	舞蹈演员	男	33	癌症
...	...	...	...	...

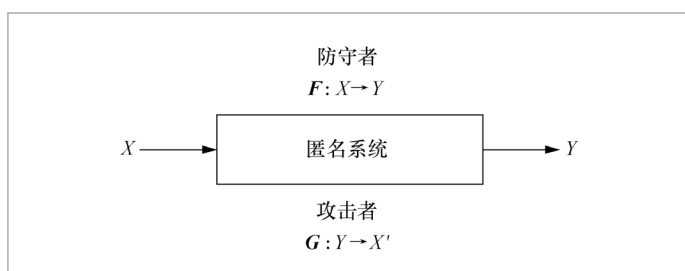


图2 隐私研究的数学模型

information) $I[\cdot]$ 度量信息泄露量 L ，计算式如式 (2) 所示。

$$L = I(X, X') = \sum_{x \in \text{dom}(X)} \sum_{y \in \text{dom}(X')} \Pr[X = x, X' = y] \log \frac{\Pr[X = x, X' = y]}{\Pr[X = x] \Pr[X' = y]} \quad (2)$$

其中， $\text{dom}(\cdot)$ 表示某一随机变量的值域。

彭长根等人^[10]将信息论中的信息熵 (entropy)、平均互信息 (mutual information)、条件熵、条件互信息等用于描述隐私保护系统信息源的隐私度量、隐私泄露量、含背景知识的隐私度量及隐私泄露量，据此提出了隐私保护方法的强度和对手攻击能力的量化测评方法，为隐私泄露的量化风险评估提供了支撑。

给定失真度阈值 D_0 和信息泄漏阈值 L_0 ，一个隐私保护系统的研究目标可以描述为一个优化问题，如式 (3) 所示。

$$\begin{aligned} &\text{optimize } F \\ &\text{s.t. } D \leq D_0 \\ &\quad L \leq L_0 \end{aligned} \quad (3)$$

从攻击者的角度来看，其只是对有关 $X' = \{X'_1, X'_2, \dots, X'_k\}$ 的知识感兴趣。一般而言，假设 $X' \subseteq X$ ，攻击者的目标是学习 X 中与

X' 相对应的知识。假设攻击者已拥有 X 中期望知识 X' 的背景知识,那么其目标是根据背景知识学习映射函数 $G: Y \rightarrow X'$,与防御操作相对应,攻击者的目标也可描述为类似于式(3)的优化问题。

3.2 隐私保护的数学模型

3.2.1 k -匿名模型

Machanavajjhala等人对 k -匿名模型的定义如下: 设 $T=\{t_1, t_2, \dots, t_n\}$ 为数据集 D 的一个数据表,其属性集为 $A=\{A_1, A_2, \dots, A_m\}$, $t_i[A_j]$ 表示记录 t_i 的属性 A_j 的值。如果 $C=\{C_1, C_2, \dots, C_k\} \subseteq A$,那么用 $T[C]=\{t[C_1], t[C_2], \dots, t[C_k]\}$ 表示记录 t 在属性集 C 上的投影, QI 表示一个数据表中所有准标识符集合。对一个数据表 T 进行 k -匿名化定义,对 T 中的每个记录 $t \in T$ 至少存在 $k-1$ 个其他记录 $t_{i1}, t_{i2}, \dots, t_{ik-1} \in T$,并且对所有的 $C \in QI$,满足 $t[C]=t_{i1}[C]=t_{i2}[C], \dots, t_{ik-1}[C]$ 。例如,在表1中,准标识符的集合为 $QI=\{job, age\}$,表2是进行 k -匿名化之后的数据($k=2$),以确保在数据表中的每个准标识符至少有 k 个记录与之对应,从而降低重新识别某个特定记录的概率。

3.2.2 差分隐私模型

差分隐私(differential privacy)^[11]是在2006年提出的有关统计数据库中数据记录的隐私标准与度量,它是一种高强度

的隐私保护框架,相关定义如下。

定义1 ϵ -差分隐私(ϵ -differential privacy): 所有的 $D1$ 、 $D2$ 为相邻数据集(即它们至多只有一条记录不同),一个给定的随机化函数 G ,对所有的 $S \in \text{Ranger}(G)$ (其表示随机化算法 G 的输出范围),满足:

$$\frac{\Pr[G(D1) \in S]}{\Pr[G(D2) \in S]} \leq e^\epsilon \quad (4)$$

也就是说,对于两个具有最小差别的数据集,经过匿名化之后的差异不超过一个给定值 e^ϵ ,参数 ϵ 称为隐私保护预算(privacy budget), ϵ 越小,隐私保护强度越高。

在差分隐私的框架中,全局敏感度(global sensitivity)是一个非常重要的度量参数,其定义如下。

定义2 全局敏感度: $\Delta f = \max_{D1 \sim D2} \|f(D1) - f(D2)\|_1$,其中 f 为函数 $f: D \rightarrow R^d$, $D1 \sim D2$ 为相邻数据集, $\|\cdot\|$ 为 $L1$ 范式距离。函数 f 敏感度是决定注入Laplace噪声量的因素之一,它是指函数 f 对两个具有最小差别的数据集 $D1$ 、 $D2$ 进行操作造成的结果的最大改变。函数 f 的全局敏感度由函数本身决定,不同的函数会有不同的全局敏感度。例如,若函数 f 是对数据集的计数查询,其全局敏感度显然为 $\Delta f=1$;若函数 f 是查询某一属性的最大值,其敏感度 Δf 可能远远大于1;若函数 f 是计算数据集中二个属性字段的联合分布(joint distribution),则 Δf 的计算将更为复杂。

差分隐私的实现机制有两种,如果函数 f 的输出是数值,则采用拉普拉斯机制(Laplace mechanism);如果函数 f 的输出是离散值,则采用指数机制(exponential mechanism)。

定义3 差分隐私的Laplace机制^[12]: 对数据集 D 上的操作函数 $f: D \rightarrow R^d$,称随机算法 $G(D)=f(D)+\eta$ 提供了 ϵ -差分隐私保护,

表2 经过 k - 匿名化处理后的数据表 ($k=2$)

工作	性别	年龄/岁	疾病
画家	女	[35, 40)	流感
画家	女	[35, 40)	癌症
专家	男	[30, 35)	流感
专家	男	[30, 35)	肝炎
...	...	...	...

其中, 注入的噪声 η 要求服从Laplace分布, 如式(5)所示。

$$\text{Lap}(\lambda): \Pr[\eta = x] = \frac{1}{2\lambda} e^{-\frac{|x|}{\lambda}} \quad (5)$$

将式(5)写成另一种形式为式(6)。

$$f(x) = \Pr[\eta = x] = \frac{1}{2\lambda} e^{-\frac{|x|}{\lambda}} = \begin{cases} \frac{1}{2\lambda} e^{-\frac{x}{\lambda}}, x \geq 0 \\ \frac{1}{2\lambda} e^{\frac{x}{\lambda}}, x < 0 \end{cases} \quad (6)$$

对式(6)中的Laplace分布进行积分, 得到累积分布函数(cumulative distribution function, CDF), 如式(7)所示。

$$F(x) = \int_{-\infty}^x f(u) du = \begin{cases} 1 - \frac{1}{2} e^{-\frac{x}{\lambda}}, x \geq 0 \\ \frac{1}{2} e^{\frac{x}{\lambda}}, x < 0 \end{cases} \\ = \frac{1}{2} \times \left[1 + \text{sgn}(x) \left(1 - e^{-\frac{|x|}{\lambda}} \right) \right] \quad (7)$$

根据式(7)计算逆累积分布函数得到式(8)。

$$x = F^{-1}(p) \\ = -\lambda \times \text{sgn}(p - 0.5) \times \ln(1 - 2 \times |p - 0.5|) \quad (8)$$

其中, p 是一个0~1.0均匀分布的随机数, λ 称为噪声规模(noise scale)参数, 在Laplace机制中, 其值满足 $\lambda \geq \Delta f / \epsilon$ 就能满足 ϵ -差分隐私, Δf 是定义2中函数 f 的全局敏感度。噪音规模 λ 与 Δf 成正比, 与 ϵ 成反比, 全局敏感度 Δf 越大或隐私保护预算越小, 则注入的噪音越大, 使得隐私保护强度增强, 但结果的可用性下降。

以下为一个差分隐私Laplace机制的例子。假设原始数据表为 D , 是存储某一地域人群艾滋病(HIV)阳性检测结果, 见表3。

假如函数 $f(D)$ 是定义在数据集 D 上的汇总查询(aggregate query)函数, 即 $f(D)$ =统计每个省份HIV检测结果为阳性的人群数量。显然, 如果直接发布数据表 D

表3 存储 HIV 阳性检测结果

...	省市	HIV阳性检测标志	...
...	安徽	0	...
...	浙江	1	...
...	安徽	1	...
...	上海	0	...
...	浙江	1	...
...	...	...	...

或直接发布函数 $f(D)$ 的统计结果(见表4)会造成敏感信息的泄漏(即HIV阳性检测标志), 可以利用差分隐私Laplace机制进行隐私保护数据发布。

显然, 函数 f 的敏感度为 $\Delta f=1$, 取隐私保护预算参数 $\epsilon=\ln 3$, 则注入噪声规模 $\lambda=\Delta f/\epsilon=0.910\ 239\ 227$, 利用式(9)计算注入的噪声, 其中, p 为0~1的随机数。

$$x = F^{-1}(p) \\ = -\lambda \times \text{sgn}(p - 0.5) \times \ln(1 - 2 \times |p - 0.5|) \\ = -\ln 3 \times \text{sgn}(p - 0.5) \times \ln(1 - 2 \times |p - 0.5|) \quad (9)$$

一次汇总查询经过差分隐私Laplace机制保护之后的结果见表5。

使用式(1), 采用方差计算失真度(Distortion)=0.693 457 201 943。

再一次汇总查询经过差分隐私Laplace机制保护之后的结果见表6。

使用式(1), 采用方差计算失真度(Distortion)=3.368 143 780 91。可见, 隐私保护预算 ϵ 越小, 噪音规模 λ 越大, 隐私保护强度增强了, 但输出结果的效用性却降低了。

表4 对表3进行统计查询结果

...	省市	计数
...	安徽	16
...	上海	8
...	浙江	10
...	...	...

表 5 利用差分隐私 Laplace 机制进行一次计数查询的结果

省市	原始计数	随机数 p	噪音规模, $\lambda=\Delta f/\epsilon=1/\ln 3$	噪音 x	带噪的计数
安徽	16	0.754 807 93	0.910 239 227	0.648 605 733 491	16.648 605 733 5
上海	8	0.876 941 22		1.276 106 244 03	9.276 106 244 03
浙江	10	0.411 761 98		-0.176 734 439 922	9.823 265 560 08

表 6 利用差分隐私 Laplace 机制进行再一次计数查询的结果

省市	原始计数	随机数 p	噪音规模, $\lambda=\Delta f/\epsilon=1/\ln 2$	噪音 x	带噪的计数
安徽	16	0.770 981 6	1.442 695 040 8	1.126 464 555 12	17.126 464 555 1
上海	8	0.872 567 94		1.972 199 839 02	9.972 199 839 02
浙江	10	0.892 972 47		2.223 946 209 72	12.223 946 209 7

定义4 差分隐私指数机制^[13]: 设随机算法 G 的输入为数据集 D , 输出为离散值 $\omega \in \Omega$, Ω 是 ω 的值域。 $f_s(D, \omega)$ 为得分函数 (score function), Δf_s 称为调节因子 (scaling factor), 用于控制隐私保护强度, 其值与 $f_s(D, \omega)$ 的敏感度有关, ϵ 为隐私保护预算。若算法 G 以正比于 $\exp(\frac{\epsilon \cdot f_s(D, \omega)}{2\Delta f_s})$ 的概率从 Ω 中选择输出 ω , 则称算法 G 实现了 ϵ -差分隐私保护, 如式 (10) 所示。

$$\Pr[\omega \text{ is selected}] = n \times \exp(\frac{\epsilon \cdot f_s(\omega)}{2\Delta f_s}) \quad (10)$$

得分函数的敏感度 Δf_s 可利用式 (11) 计算, 其中 $D1$ 、 $D2$ 为相邻数据集。

$$\Delta f_s = \max_{D1, D2, \omega' \in \Omega} \left\| f_s(\omega') - f_s(\omega'') \right\| \quad (11)$$

以下为差分隐私指数机制的一个例子。假如要举办一场体育项目比赛, 供选

择的体育项目为 $\Omega = \{Soccer, Volleyball, Basketball, Badminton\}$, 决策参与者拟通过投票决定哪个体育项目为比赛项目, 但要保证决策过程满足 ϵ -差分隐私保护要求, 即每个体育项目的得票数为敏感信息, 不便公开。在此用体育项目的得票数作为得分函数 f_s 的值, 显然其敏感度为 $\Delta f_s = 1$ 。按照差分隐私指数机制, 在给定参数 ϵ 值的情况下, 计算各体育比赛项目被选中的概率, 同时每个体育项目的得票数得到了保护, 见表7。

4 基于位置服务的隐私保护及应用

移动互联网、社交网络、云计算等新兴技术的发展和广泛使用, 使得基于位置服务 (locate-based service, LBS) 的隐私保护问题越来越严重, 成为工业界

表 7 利用差分隐私指数机制示例

可选的比赛项目	得票数 f_s (要保护的敏感信息)	被选中的概率 $\Pr[\cdot]$ (差分隐私保护下的输出结果)		
		$\epsilon=0$ $n=0.25$	$\epsilon=0.1$ $n=0.094\ 6$	$\epsilon=1$ $n=2.83E-8$
Soccer	30	0.25	0.424 0	0.092 513
Volleyball	25	0.25	0.330 2	0.007 594
Basketball	8	0.25	0.141 1	0.000 002
Badminton	2	0.25	0.104 5	0

和学术界广泛关注的热点问题。2003年, Beresford A R等人^[14]提出位置隐私的概念, 开始了对LBS隐私保护的研究; Ghinita G^[15]从私有查询和轨迹匿名两个方面对位置隐私进行了综述, 但没有涉及隐私度量和查询隐私; Krumm J^[16]着重评述了匿名、模糊化隐私保护技术和一些利用位置数据几何性质的隐私侵犯算法, 但没有涉及系统结构和查询隐私; 霍峥等人^[17]从传统关系数据隐私保护向时空方向拓展的角度对数据发布中的轨迹隐私保护和LBS中的轨迹隐私保护关键技术进行了分析和比较, 但没有涉及查询隐私; Shin X等人^[18]总结了LBS通用隐私威胁模型, 分析比较了各种LBS隐私保护技术及度量指标。张学军等人^[19]总结了LBS隐私保护关键问题、LBS隐私保护系统结构, 并探讨未来的研究方向。参考文献[20]针对位置隐私保护研究多基于欧氏空间的不足, 提出了路网环境下连续KNN查询方法, 该方法以路段交叉点为连续查询锚点, 将路段上所有用户的连续兴趣点查询转化为等价的必达路段顶点查询, 在保证路段上所有用户位置隐私的同时, 实现用户K最近邻兴趣点的精确查询。针对面向路网的隐私保护K近邻查询中, 保护位置隐私需求引起服务器端处理代价激增, 导致保护位置隐私前提下查询效率与查询准确性绝对对立问题, 参考文献[21, 22]提出支持查询者对查询准确性与查询效率进行偏好调控思想, 实现位置查询服务的安全化和个性化。王璐等人^[23]对时空数据发布中的隐式隐私保护进行研究, 首次提出并定义时空数据中的隐式隐私, 并提出保护该隐私的框架, 针对框架中的消除步骤提出了基于频繁移动对象的假数据添加算法, 并针对其数据效用差的缺陷设计了基于图的假数据添加方法, 大幅提高了数据效用性。参考文献[19]虽然系统地介绍了LBS隐私保护关

键问题、LBS隐私保护系统结构、LBS隐私保护度量指标, 并详细阐述和分析各种LBS隐私保护技术、主要LBS隐私保护技术的性能评估, 但没有涉及基于位置服务的隐私保护的数学描述与数学模型。

5 大数据隐私保护的挑战与机遇

5.1 隐私研究的挑战

大数据是当今计算机科学面临的新的计算环境, 而隐私保护又是其中极其重要的问题。在大数据隐私保护研究中存在诸多问题与挑战。

- 隐私度量问题。隐私是一个主观概念, 它根据不同的人、不同的时间变化而变化, 因此难以对其定义和度量, 它是基础性和挑战性问题, 不仅需要技术方面的努力, 也需要在社会学和心理学方面的研究。

- 隐私保护的理论框架问题。目前有数据聚类方法和差分隐私保护理论框架, 但由于数据聚类隐私保护方法(如k-匿名等)存在局限性, 目前在实际应用中采用差分隐私保护。在大数据时代能否研究出新的具有开创性的隐私研究理论基础, 将是另一个挑战。

- 隐私保护算法的可扩展性。目前存在的一些机制和策略处理大的数据库时主要是采用分治法, 但是大数据的规模远比较大的数据库要大, 设计可扩展性算法实现隐私保护也是一个挑战。

- 数据源的异构性。当前可用的隐私保护算法几乎都是面向同构数据(homogeneous data)的, 类似于数据库中的记录, 但实际上大数据的数据源都是异构数据(heterogeneous data)。以高效的方式处理异构大数据显然也是未来的挑战。

● 隐私保护算法的效率。对于体积庞大的大数据,算法的效率将是大数据计算过程中一个重要的因素。

5.2 大数据时代的机遇与研究方向

可以预见,未来人们需要改进当前的隐私保护方法,以满足大数据计算中存在的前所未有的需求,再者,期望新的隐私保护框架与机制的出现。本文认为下面的研究方向具有一定的前景,值得隐私保护的研究者投入时间与精力进行研究。

(1) 隐私保护的量子计算

量子计算机越来越接近于实际应用,量子计算在信息安全与隐私保护方面能够提供超乎寻常的功能。当前加密方法的主要优点是其时间复杂度,但它对于隐私保护与信息安全的计算是有条件安全。最近提出的基于量子计算模型的测量法为盲计算的实现提供了非常有前景的范例^[24],它的主要思想是客户端授权计算给量子服务器,量子服务器能在无需获取输入、输出与客户端等任何知识的情况下执行计算任务,Braz S等人^[25]已实现了这个模型的概念框架,并且展示了盲量子计算的可行性。因此,不仅应该在隐私与安全领域探索量子计算,也应该探索其他领域的量子计算。

(2) 将计算机技术与社会科学领域的研究进行集成

有必要将计算技术与社会科学的研究成果与需求进行紧密结合,而且这些工作必须得到相关领域领军人物的支持,例如那些从事新兴的计算心理生理学交叉学科的研究者等^④。

(3) 开发新的隐私保护框架

基于数据聚类的隐私保护方法(如k-匿名)具有一定的实用性,差分隐私保护框架具有较强的严谨性。但是前者易受各种不同

类型的攻击(如“合成式”攻击等),后者在实际应用中由于计算时间复杂度等问题缺乏灵活性和可行性。因此,有必要结合一系列不同学科理论知识解决目前隐私保护研究中的问题,如用于处理模糊概念的模糊逻辑(fuzzy logic)、用于解决不同方争论的博弈论(game theory)^[26]等。

参考文献:

[1] 张华平, 孙梦姝, 张瑞琦, 等. 微博博主的特征与行为大数据挖掘[J]. 中国计算机学会通讯, 2014, 10(6): 36-43.
ZHANG H P, SUN M S, ZHANG R Q, et al. Micro-blog blogger's characteristics and behavior big data mining[J]. Communications of the CCF, 2014, 10(6): 36-43.

[2] 张俊, 萧小奎. 数据分享中的差分隐私保护[J]. 中国计算机学会通讯, 2014, 10(6): 44-51.
ZHANG J, XIAO X K. Differential privacy protection based on data share[J]. Communications of the CCF, 2014, 10(6): 44-51.

[3] 刘雅辉, 张铁赢, 靳小龙, 等. 大数据时代的个人隐私保护[J]. 计算机研究与发展, 2015, 52(1): 229-247.
LIU Y H, ZHANG T Y, JIN X L, et al. Personal privacy protection in the era of big data[J]. Journal of Computer Research and Development, 2015, 52(1): 229-247.

[4] 董诚, 林立, 金海, 等. 医疗健康大数据: 应用实例与系统分析[J]. 大数据, 2015(2): 2015021.
DONG C, LIN L, JIN H, et al. Big data in healthcare: applications and system analytics[J]. Big Data Research, 2015(2): 2015021.

[5] 俞国培, 包小源, 黄新霆, 等. 医疗健康大数据的种类、性质及有关问题[J]. 医学信息学杂志, 2014, 35(6): 9-12.
YU G P, BAO X Y, HUANG X T, et al. Medical and health big data: types, characteristics and relevant issues[J].

④
http://www.
sciencemag.org/
custom-publishing/
collections/advan-
cescomputational-
psychophysiology

- Journal of Medical Intelligence, 2014, 35(6): 9-12.
- [6] 刘向宇, 王斌, 杨晓春. 社会网络数据发布隐私保护技术综述[J]. 软件学报, 2014, 25(3): 576-590.
- LIU X Y, WANG B, YANG X C. Survey on privacy preserving techniques for publishing social network data[J]. Journal of Software, 2014, 25(3): 576-590.
- [7] 熊平, 朱天清, 王晓峰. 差分隐私保护及其应用[J]. 计算机学报, 2014, 37(1): 101-122.
- XIONG P, ZHU T Q, WANG X F. A survey on differential privacy and applications[J]. Chinese Journal of Computers, 2014, 37(1): 101-122.
- [8] 张啸剑, 孟小峰. 面向数据发布和分析的差分隐私保护[J]. 计算机学报, 2014, 37(4): 927-949.
- ZHANG X J, MENG X F. Differential privacy in data publication and analysis[J]. Chinese Journal of Computers, 2014, 37(4): 927-949.
- [9] YU S. Big privacy: challenges and opportunities of privacy study in the age of big data[J]. IEEE Access, 2016(4): 2752-2763.
- [10] 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.
- [11] DWORK C. Differential privacy[C]// The 33rd International Colloquium on Automata, Languages and Programming, July 10-14, 2006, Venice, Italy. Venice: Springer-Verlag, 2006: 1-12.
- [12] DWORK C, MCSHERRY F, NISSIM K. Calibrating noise to sensitivity in private data analysis[C]// The 3rd Conference on Theory of Cryptography, March 4-7, 2006, New York, USA. New York: Springer, 2006: 265-284.
- [13] MCSHERRY F, TALWAR K. Mechanism design via differential privacy[C]// The 48th Annual IEEE Symposium on Foundations of Computer Science, October 21-23, 2007, Providence, USA. New Jersey: IEEE Press, 2007: 94-103.
- [14] BERESFORD A R, STAJANO F. Location privacy in pervasive computing[J]. IEEE Pervasive Computing, 2003, 2(1): 46-55.
- [15] GHINITA G. Private queries and trajectory anonymization: a dual perspective on location privacy[J]. IEEE Transaction on Data Privacy, 2009, 2(1): 3-19.
- [16] KRUMM J. A survey of computational location privacy[J]. Personal and Ubiquitous Computing, 2009, 13(6): 391-399.
- [17] HUO Z. A survey of trajectory privacy preserving techniques[J]. Chinese Journal of Computers, 2011, 34(10): 1820-1830.
- [18] SHIN K G, JU X, CHEN Z, et al. Privacy protection for users of location-based services[J]. IEEE Wireless Communications, 2012, 19(1): 30-39.
- [19] 张学军, 桂小林, 伍忠东. 位置服务隐私保护研究综述[J]. 软件学报, 2015, 26(9): 2373-2395.
- ZHANG X J, GUI X L, WU Z D. Privacy preservation for location-based services: a survey[J]. Journal of Software, 2015, 26(9): 2373-2395.
- [20] 周长利, 马春光, 杨松涛. 路网环境下保护LBS位置隐私的连续KNN查询方法[J]. 计算机研究与发展, 2015, 52(11): 2628-2644.
- ZHOU C L, MA C G, YANG S T. Location privacy-preserving method for lbs continuous knn query in road networks[J]. Journal of Computer Research and Development, 2015, 52(11): 2628-2644.
- [21] 倪巍伟, 陈萧, 马中希. 支持偏好调控的路网隐私保护k近邻查询方法[J]. 计算机学报, 2015, 38(4): 884-896.
- NI W W, CHEN X, MA Z X. Location privacy preserving k nearest neighbor query method on road network in presence of user's preference[J]. Chinese Journal of Computers, 2015, 38(4): 884-896.
- [22] 倪巍伟, 陈萧. 保护位置隐私近邻查询中隐私偏好问题研究[J]. 软件学报, 2016, 27(7): 1805-1821.
- NI W W, CHEN X. User privacy preference

support in location privacy-preserving nearest neighbor query[J]. Journal of Software, 2016, 27(7): 1805-1821.

[23] 王璐, 孟小峰. 位置大数据隐私保护研究综述[J]. 软件学报, 2014, 25(4): 693-712.

WANG L, MENG X F. Location privacy preservation in big data era: a survey[J]. Journal of Software, 2014, 25(4): 693-712.

[24] GOTTESMAN D, CHUANG I L. Demonstrating the viability of universal quantum computation using teleportation and single-qubit operations[J]. Nature, 1999, 402(6760): 390-393.

[25] BARZ S, KASHEFI E, BROADBENT A, et al. Demonstration of blind quantum computing[J]. Science, 2012, 335(6066): 303-308.

[26] 张伊璇, 何泾沙, 赵斌, 等. 一个基于博弈理论的隐私保护模型[J]. 计算机学报, 2016, 39(3): 615-627.

ZHANG Y X, HE J S, ZHAO B, et al. A privacy protection model base on game theory[J]. Chinese Journal of Computers, 2016, 39(3): 615-627.

作者简介



方贤进 (1970-), 男, 博士, 安徽理工大学计算机科学与工程学院教授, 主要研究方向为网络与信息安全、智能计算。



肖亚飞 (1994-), 男, 安徽理工大学计算机科学与工程学院硕士生, 主要研究方向为隐私保护。



杨高明 (1974-), 男, 博士, 安徽理工大学计算机科学与工程学院副教授, 主要研究方向为隐私保护。

收稿日期: 2017-07-30
基金项目: 国家自然科学基金资助项目 (No.61572034, No.61402012)
Foundation Items: The National Natural Science Foundation of China (No.61572034, No.61402012)