# 第一章 大数据概述（耿玥，李映乐）

随着博客、社交网络、基于位置的服务LBS为代表的新型信息发布方式的不断涌现，以及云计算、物联网等技术的兴起，数据正以前所未有的速度在不断地增长和累积。目前，数据已经从原先的可以简单管理的对象逐渐转变为更为基础的资源。如何对大数据进行更为高效地合理应用也逐渐成为衡量各个国家科技水平的硬指标。

大数据的快速发展，使之成为信息时代的一大新兴产业，并引起了国内外政府、学术界和产业界的高度关注。大数据时代的到来，为数据的存储、管理及分析带来了前所未有的挑战。数据管理方式正随着社会的进步而潜移默化地发生转变，云计算因此应运而生。云计算为大数据在存储、管理及分析等方面带来了极佳的技术支撑，成为分布式计算、互联网技术、大规划资源管理等技术的整合和发展。

但是大数据在收集、存储与使用过程中往往也会伴随着非常多的安全问题，如何有效地保护数据安全与隐私也成为了各国必须要面对的技术挑战。

本章节以此出发，简要介绍近几年国内外有关大数据的发展历史及相关概念，以及随着大数据而蓬勃发展的云计算技术，分析了两者的联系，并说明了典型大数据的分析流程及对当今社会引发的变革，最后对大数据时代必定带来的安全隐私方面进行了简要的分析。

## 1.1 大数据基本概念

新华社于2016年3月17日全文播发《中华人民工和国国民经济和社会发展第十三个五年规划钢要》[1]，钢要提出实施国家大数据战略，把大数据作为基础性战略资源，全国实施促进大数据发展行动，加快推动数据资源共享开放和开发应用，助国产业转型升级和社会治理创新。自“十三五”计划以来，我国的大数据产业也在蒸蒸日上，全国各个地区都在落实并加强发展大数据的工作，各大行业应用都得到超速推广，市场的规模也随着显著提升。如产业界的诸多报告、纽约时报等都遍布了大数据的有关讯息，国内外政府部门也设立了重大项目，从而进一步地发展大数据；Nature、Science等期刊也发表了有关大数据的诸多论点。

近年来随着研究方法的不断成熟，人工智能的技术以及科学研究也正在经历着不断的蜕变与突破，而大数据分析不仅作为一种技术，更作为一种哲学手段，为人工智能的解题破解提供了独特的研究进路。大数据思维的哲学研究不仅有助于推进人工智能的认知难题破解，还能为人工智能的实践发展起到启发性及支撑性的作用。

本小节对大数据的基本概念进行了系统性的介绍。小节首先介绍了大数据本身的定义以及目前普遍被人们接受的大数据五大特性，并简要介绍了国内外近几年有关大数据的发展情况。

### 1.1.1 大数据的发展

进入2012年以来，大数据的关注度与日俱增。近几年我国也开始积极加入到这行动中来。2015年，我国发布了第一个大数据国家行动计划——《促进大数据发展行动钢要》[2]，此纲要提出在未来五至十年我国大数据发展和应用实现的目标，其中，2017年底前形成跨部门数据资源共享共用格局、2018年底前建成政府数据放平台等。2016年12月，工业和信息化部编制印发了《大数据产业发展规划（2016~2020年）》[3]，此规划提出“十三五”是实施国家大数据迅猛发展的重要窗口期，要紧抓发展，实现数据强国的转变，进一步地确实了推动我国大数据产业发展为当前的主要任务、重要工程以及保障国家经济发展的必要措施。2017年4月10日，工业和信息化部编制印发了《云计算发展三年行动计划（2017-2019年）》[4]，提出截止到2019年，我国云计算产业规模将达到4300亿元，云计算的相关服务能力将达到国际领先水平。国家自然科学基金委员会也于同年8月发布“大数据驱动的决策研究”重大研究计划2017年度项目指南[5]，针对大数据驱动的决策范式转变机理与理论、大数据资源治治机制设计与协同管理、领域导向的大数据价值发现理论与方法三个关键科学问题开展了研究。2018年1月，中商产品研究院发布了《2018-2023年中国大数据行业发展前景及投资机会研究报告》[6]数据显示，2017年我国的大数据行业市场规模已达到3615亿元，预计2018年中国大数据行业市场规模将达到近6000亿。2018年4月份在“2018大数据产业峰会”上，中国信息通信研究院重磅发布《大数据白皮书（2018年）》。主要提出了技术创新趋势、数据资产管理、应用渗透路径三点聚焦。同年5月份，大数据战略重点实验室研究发布了《大数据蓝皮书：中国大数据发展报告No.2》，该报告认为，中国已经进入大数据创新突破与应用落地的发展上升期，并对中国大数据发展的十大趋势进行了展望。

国外情况，2016年5月23日，美国发布“联邦大数据研发战略计划”[7]，该计划为针对数据科学、数据密集型应用、大规模数据管理及分析领域开展及主持各项研发工作的联邦各机构提供了相互关联的大数据研发战略，该战略主要涵盖以下几点：充分利用新兴的大数据基础与技术，创建新一代的能力； 探索与理解数据及知识的可信程度，进而实现突破性地科学发现，从而做出更好的决策；创建并改善科研网络基础设施，实现大数据的创新；通过促进数据的共享与管理政策来提升研究价值；针对隐私、安全及伦理进一步地对大数据的收集与共享进行管理；完善大数据教育与培训的国家布局，从而满足高级分析人才的需求；在国家大数据创新生态系统中建立并加强各种联系。瑞士国家重点科研计划（NFP）大数据专项（Big Data，NFP75）于2017年正式启动[8]，计划为期四年投入资金2.5亿瑞士法郎。此项目于2016年11月开始公开征集，最终确定了共36个项目，其中分为大数据信息技术、大数据相关社会及法律问题、大数据应用这三个版块。欧盟及其成员国也已经制定大数据发展战略，主要包括：数据价值链战略计划、资助“大数据”和“开放数据”领域的研究和创新活动、实施开放数据政策、促进公共资助科研实验成果和数据的使用及再利用。2018年4月，英国发布《工业战略：人工智能》报告，立足引领全球人工智能和大数据发展，从鼓励创新、培养和集聚人才、升级基础设施、优化营商环境以及促进区域均衡发展等五大维度提出一系列实实在在的举措。

世界经济正加速向以网络信息技术产业为重要内容的经济活动转变。各国都在紧紧把握住这一历史契机，不断以信息化培育新功能，从而失去新的发展。只有不断加大投入、加强信息基础设施建设，推动互联网和实体经济深度融合，加快传统产业数字化、智能化，才能做大做强数字经济，从而拓展经济发展新空间。

### 1.1.2 大数据的定义

随着大数据的流行，大数据的定义也开始呈现多样化的趋势，达成统一的共识非常困难。由于大数据本身是一个比较抽象的概念，单从字面来看，它表示数据规模的庞大。但是仅仅数量上的庞大显然无法看出大数据这一概念和以往的“海量数据”（massive data）、“超大规模数据”（very large data）等概念之间有何区别。对于大数据尚未有一个公认的定义，不同的定义基本是从大数据的特征出发，通过这些特征的阐述和归纳试图给出其定义。

国际数据中心IDC(International Data Corporation)在2011 年的报告中定义了大数据[9]：“大数据技术描述了一个技术和体系的新时代, 被设计于从大规模多样化的数据中通过高速捕获、发现和分析技术提取数据的价值”。这个定义刻画了大数据的4个显著的特点，即容量(volume)、多样性(variety)、速度(velocity) 和价值(value)。

IDC对大数据的定义法通常被称为属性定义法，此外比较著名的还有比较定义法与体系定义法。针对于比较定义法，2011 年, McKinsey 公司的研究报告中将大数据定义[10]为“超过了典型数据库软件工具捕获、存储、管理和分析数据能力的数据集”。这种定义是一种主观定义，没有描述与大数据相关的任何度量机制，但是在定义中包含了一种演化的观点(从时间和跨领域的角度，说明了什么样的数据集才可以被认为是大数据。而对于体系定义法，美国国家标准和技术研究院NIST[11]则认为 “大数据是指数据的容量、数据的获取速度或者数据的表示限制了使用传统关系方法对数据的分析处理能力，需要使用水平扩展的机制以提高处理效率”。

因此眼下在大数据定义问题上很难达成一个完全的共识，这点和云计算的概念刚提出时的情况是相似的，每种定义都反映了大数据的特定方面。故在面对实际问题时，不必过度地拘泥于具体的定义之中。

**1.1.3 大数据的5V特性**

关于大数据的特性，IBM提出的“5V”[12]理论一直广受业界肯定。5V，即Volume（大量）、Velocity（高速）、Value（价值）、Variety（多样）、Veracity（真实）这五大特性，如下图1-1所示。

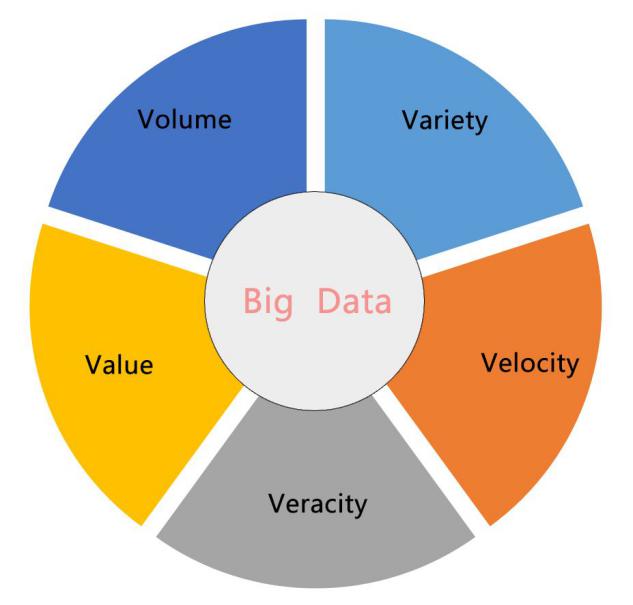


图1-1 大数据的5V特性

然而在近几年研究中，很多与大数据相关的都会减少一个“V”—Veracity（真实）。IBM之所以提出Veracity特性，原因是他们认为互联网上人们所留下的，都是用户在操作电脑过程中所留下来的真实的电子踪迹，这些踪迹都可以真实地反映出用户的想法。故本书采用IBM公司的“5V”理论。

下面简单介绍大数据的5V特性：

Volume：数据量极大，主要包括对数据的采集、存储及计算。大容量是大数据区分于传统数据的最显著的特征，因此大数据能更多地发现大量数据的潜在价值。随着大数据时代数据量和数据处理的提升，也极有可能从海量数据中挖掘到更多的数据价值。

目前，大数据的起始计量已经由TB增长到了PB，甚至是EB、ZB级别，其中，PB代表1000个TB，EB代表100万个TB。而截止目前，全人类自文明史后所有的话语数据量大约是5EB。

Velocity：处理速度极快。由于数据的增长速度快，故要求处理速度必须提升，时效性的要求非常高。比如搜索引擎，必须满足用户可以搜索到几分钟前的新闻，且个性化的推荐算法也要尽可能要求实时性的完成推荐。由于大数据对时效性的高要求，故很多数据的价值也会随着时间而不断流逝。据IDC的“数字宇宙”报告，预计在2020年为止，各个国家数据使用量之和将会达到35.2ZB。故在海量的数据面前，数据的处理效率就是企业的生命。

Value：价值高，通常价值密度的高低与数据的总量大小成反比例关系，因此，如何通过强大的机器算法可以快速地完成海量数据的价值提纯，也是当今大数据背景下的热门研究课题。

Variety：数据类型繁多，包括结构化、半结构化和非结构化数据。具体主要表现为网络日志、音频、视频、图片以及地理位置信息等等。不同类型的数据对数据的处理能力也提出了相对较高的要求。

Veracity：准确性能高，也就是说数据的质量要求较高。只有数据真实且准确，数据才能发挥其最大的作用。

## 1.2 大数据与云计算

云计算（Cloud Computing）是基于互联网的相关服务的增加、使用和交付模式。通常，云计算会涉及通过互联网来提供的动态易扩展且经常是虚拟化的资源。过去，云在图中经常用来表示电信网，后来也被用来表示互联网和底层基础设施的抽象，因此又把云作为网络、互联网的一种比喻说法。

同样，对于云计算的定义，也有非常多的说法。现阶段被广为接受的，是美国国家标准与技术研究院（NIST）的定义[13]：云计算是一种按照用户的使用量来进行付费的一种模式，这种模式提供可用的、便捷的、按需的网络访问，且云计算可以利用互联网进入可配置的计算资源共享池。其中，计算资源包括网络、服务器、存储、应用软件与服务等。这些资源能够被快速提供，且只需投入很少的管理工作，或与服务供应商进行很少的交互。故通过云计算，用户可以凭借相关的业务负载进行遍历与释放资源，在提升服务质量的同时降低运维成本。

云计算不同于网格计算等传统的分布式计算，他们之间有着明显的区别。首先，云计算是弹性的，也就是说，云计算可以凭借工作负载量的大小将资源进行动态分配，而原先部署在云平台上的相关应用，也需要及时地适应分配所造成的资源的变化，并且针对相关变化做出及时响应。再者，相较于强调异构资源共享的网格计算，云计算会更强调超大规模的资源池的共享。云计算通过资源池分享从而显著提高资源利用率，并凭借此规模有效地降低运行成本。最后，云计算需要考虑经济代价，故而相关的硬件设备与软件构造不可盲目追求高性能，更应该从可用性、成本代价等多个角度进行考量。

本小节简要介绍了云计算与云服务的联系与区别，阐明大数据与云计算彼此的关系，进而提出典型大数据的分析流程，并提出大数据时代对各行各业引发的变革。

### 1.2.1 云计算与云服务的关系

虽然云计算近些年来已广为人知，然而在谈论到云计算的过程当中，我们还会涉及到云服务的概念。那么，云计算与云服务有什么区别与联系呢？两者在商业运作以及平台管理上又有何不同？

中国云计算服务网将云服务进行了严格定义。云服务，即云计算服务，指可以拿来作为服务提供使用的云计算产品。它将大量用网络计算资源进行统一的管理与调度，从而构成了一个计算资源池，并向用户提供按需服务，用户通过网络以按需、易扩展的方式获得所需资源及相关服务。因此，云服务是一种商业模式，其通过大量创业公司提供丰富的个性化产品，以满足市场上日益膨胀的个性化需求。其繁殖方式是为创业公司提供资金、推广、支付、物流、客服一整套服务，把自己的运营能力像水和电一样让外部随需使用。

众所周知，云计算的表现形式是一系列服务的集合。针对于云计算的服务类型，一般可以将其分为三个层面：核心服务、服务管理和用户访问。其中，核心服务层又分为四个子层，分别为：基础设施即服务层（IaaS，infrastructure as a service）、平台即服务层（Paas，platform as a service）、软件即服务层（SaaS，software as a service）和数据即服务层（DaaS，Date as a service）

IaaS，基础设施即服务层。此层中，云端公司把IT环境的基础设施建设好，然后直接对外出租硬件服务器或虚拟机，而消费者得以利用所有计算基础设施。消费者不管理或控制任何云计算基础设施，但可以控制操作系统的选择、存储空间或部署的应用，也可以获得有限制的网络组件的控制等。

Paas，平台即服务层，它提供了企业进行定制化研发的中间件平台，同时也涵盖数据库和应用服务器等。换句话说，PaaS把运行用户所需的软件平台作为服务出租。此层主要包括操作系统、数据库、中间件和运行库。相对于IaaS来说，Paas租户的灵活性降低了，只能在云端提供的有限平台范围内做软件，但是此层能最大化的利用租用的资源。

SaaS，软件即服务层，它是一种通过Internet提供软件的模式，厂商将应用软件统一部署在自己的服务器上，客户可以根据自己实际需求，通过互联网向厂商定购所需的应用软件服务，按定购的服务多少和时间长短向厂商支付费用，并通过互联网获得厂商提供的服务。

DaaS，数据即服务层，是大数据时代的产物。它包含的主要技术有数据虚拟化、SOA、BPM等。在这一层，云端公司负责建立全部的IT环境，收集用户需要的基础数据并且做数据分析，最后对分析者算法提供编程接口，让数据成为服务。DaaS是大数据时代的象征，能做DaaS服务的云端公司需要从数据积累，数据分析和数据交付三个方面来积累自身的核心竞争力。



图1-2 大数据时代下的云架构

这四个层次共同组成了云计算技术层面的整体架构。其中，这些服务层里会包括一些虚拟化的应用技术及相关应用、自动化的部署及分布式计算等技术。而云服务则是在云计算的上述技术架构下，对外提供的按需分配，可以计量化的一种IT服务模式，这种服务以代替本地用户自己建立的IT服务。

总而言之，云服务其实是运行在云计算之上的，云计算作为整个云平台的技术基础架构，对于推动整个云计算业务快速发展具有非常重要的作用，巨头们可以通过自研的形式弥补在基础设施上的不足，创业者也可以通过租赁公有云或私有云的形式打破进入云计算行业的门槛。

### 1.2.2 大数据与云计算的关系

大数据的复杂需求也对其技术的实现提出了更高的标准，而云计算凭借其特有的弹性伸缩、动态调配、资源虚拟化、支持多租户、支持按量计费等特征更好地迎合了大数据所提出的需求，成为现如今解决大数据问题的重要方向。

借用大数据云计算关系一文中的简单介绍就是云计算是硬件资源的虚拟化，而大数据是海量数据的高效处理。两者的区别主要是以下几点：

1）从两者的目的来分析，大数据是要尽可能地挖掘数据中的信息，提取数据中的价值，而云计算则是通过互联网进而更好地扩展、管理、存储资源，以节省企业的成本。

2）从两者的对象角度，大数据的处理对象是数据，而云计算的处理对象是IT资源、处理能力及应用。

3）从推动力量上来看，大数据的推动力量是从事数据存储与处理的软件厂商，或是拥有海量数据的企业，而云计算的推动力量则是IT设备厂商，或是拥有计算或存储资源的企业。

4）从结果来分析，云计算注重资源分配，大数据注重的是资源处理。也可以说，大数据需要云计算支撑，云计算为大数据处理提供平台。

5）从带来的价值上来看，大数据能发现数据中的潜在价值，从而带来收益，而云计算则是节省了IT的部署成本。

但是两者却又存在着千丝万缕的联系。首先，两者都是为数据的存储和处理来服务的，故而都需要占用大量的存储资源和计算资源，同时，两者都要用到海量数据存储技术、海量数据管理技术以及MapReduce等并行处理技术。再者，随着数据处理复杂程度不断增加，其相应的性能和扩展需求也不断增大，而云计算所具备的弹性伸缩和动态调配、资源的虚拟化，按需使用，以及绿色节能等基本要素正好契合了新型大数据处理技术的需求。在数据量爆发增长以及对数据处理要求越来越高的先当下，实现大数据和云计算的结合，才能最大程度上发挥二者的优势，满足用户的需求。

以此看来，大数据与云计算之间，并非独立概念，而是关系非比寻常。无论在资源的需求上还是在资源的再处理上，都需要二者共同运用。这也使得不少地区在做出相关产业规划时，都会同时推进大数据与云计算建设。因此，与其计较大数据与云计算之间怎么区分，不如规划在一起，让云计算为大数据提供强大平台，以大数据分析出的结论完成云计算价值。

### 1.2.3 典型的大数据分析流程

随着数据工具自动化程度的不断提高，各行各业都在应用大数据来进行知识评判和预测分析。传统的数据分析流程，已经不能满足当今大数据的发展趋势。大数据带来的高成本、多变化性、海量数据集等问题，都为大数据的分析带来了严峻挑战。

从大数据的特征以及其产生的领域来看，其来源非常广泛，因此产生的数据类型及相关的处理方法也就迥然不同。但总体来说，大数据的分析流程可以分为以下四个阶段：数据采集、数据的处理与集成、数据分析和数据解释。如图1-3所示。

图1-3 大数据处理基本流程

1. 数据采集

当前，众多大型企业和政府机构，在信息化过程中，与自身业务充分结合，搭建了各式各类的软件系统。其中，这些系统里面累积了大量的用户数据，他们及需将这些数据整合起来，从而形成自己独有的大数据平台，来进一步地做数据挖掘、处理与分析，更好的为自己的客户服务。因此，针对于数据采集，这是大数据处理流程中最为基础，却也是最为重要的一步。

数据采集方法的选择不但要依赖于数据源的物理性质，更要明白数据分析的目标。目前主要有三种常用的数据采集方法：传感器、日志文件及web爬虫。

传感器主要用于测量物理环境的变量并将其转化为可读的数字信号。其中，有线传感器网络通过网线收集传感器的信息，而无线传感器则利用无线网络作为信息传输的载体，适用于没有能量或通信的基础设施的场合。日志文件则是以特殊的文件格式记录系统的活动，如web服务器通常要在访问日志文件中记录网站用户的点击、键盘的输入、访问行为及其它属性，日志文件常可以看作为“软件传感器”。Web爬虫则是在搜索引擎下载并存储网页的程序，其数据采集过程由选择策略、重访策略、礼貌策略以及并行策略[14]共同来决定。

除上述三种采集方法外，还有许多和领域相关的数据采集方法及系统。根据采集方式的不同，数据采集方法大致可分为以下两类：

基于拉（pull-based）的方法，数据由集中式或分布式的代理来主动收集；

基于推（push-based）的方法，数据由源或第三方推向数据汇聚点。

1. 数据的处理与集成

数据的处理与集成这一阶段，主要是完成对于已经采集到的数据进行适当的处理、清洗去噪以及进一步的集成存储。传统数据库领域在此已有了较为成熟的研究，但随着新型的数据源不断涌现，有关数据的处理与集成方法也在不断进行更新。

数据处理只用计算机收集、记录数据，经加工产生新的信息形式的技术。数据处理是系统工程和自动控制的基本环节，旨在从大量的原始数据抽取出有价值的信息，也可以说是完成数据转换成信息的过程。该过程主要对所输入的各种形式的数据进行加工整理。具体是要完成对已接收数据的抽取、清洗等操作。

1）抽取：因获取的数据可能具有多种结构和类型，数据抽取过程可以帮助我们将这些复杂的数据转化为单一的或者便于处理的构型，以达到快速分析处理的目的。

2）清洗：对于大数据，并不全是有价值的，有些数据并不是我们所关心的内容，而另一些数据则是完全错误的干扰项，因此要对数据通过过滤“去噪”从而提取出有效数据。

数据集成则是把不同来源、格式、特点性质的数据在逻辑上或物理上有机地集中，从而为企业提供全面的[数据共享](https://baike.baidu.com/item/%E6%95%B0%E6%8D%AE%E5%85%B1%E4%BA%AB" \t "_blank)。其核心任务，就是将互相关联的分布式异构数据源集中在一起，使用户能够以完全透明的方式来对这些资源进行访问。

例如现阶段研究较为广泛的知识图谱，其针对关联密集型的数据进行处理，基于图数据库的优点，对于不需要进行关系延伸计算的数据不放入图谱，这些数据可使用适应的存储并与知识图谱中实体作连接；而对于结构固定、实体属性信息丰富的实体类，使用其它数据库存储更能体现优势。

一般地，在对大数据进行“去噪”和清洗后，为使得数据拥有较高质量，通常也会设计一些数据过滤器，通过采用聚类或是关联分析的规则方法，来将没有用处的、错误的离群数据挑选出来从而进行过滤，然后再对这些整理好的数据进行集成和存储。如果只是单纯对放置数据，势必会丢将来的数据访问造成影响，现在针对不同种类的数据建立相应的数据库，将这些数据分别放置，既能有效减少数据查询及访问时间，同时增加了数据的提取速度。

目前，在企业数据集成领域，已经有了很多成熟的框架可以利用。通常采用的是联邦式、基于[中间件](https://baike.baidu.com/item/%E4%B8%AD%E9%97%B4%E4%BB%B6" \t "_blank)模型和数据仓库等方法来构造集成的系统，这些技术在不同的着重点和应用上解决数据共享和为企业提供决策支持。

1. 数据分析

数据分析是整个大数据处理流程里最为核心、是最为关键的部分。

数据分析是指用适当的[统计分析方法](https://baike.baidu.com/item/%E7%BB%9F%E8%AE%A1%E5%88%86%E6%9E%90%E6%96%B9%E6%B3%95" \t "_blank)对收集来的大量数据进行分析，提取有用信息和形成结论而对数据加以详细研究和概括总结的过程。这一过程也是[质量管理体系](https://baike.baidu.com/item/%E8%B4%A8%E9%87%8F%E7%AE%A1%E7%90%86%E4%BD%93%E7%B3%BB" \t "_blank)的支持过程。在实用中，数据分析可帮助人们做出判断，以便采取适当行动。

常用的大数据分析方法分为数据可视化法、统计分析法及数据挖掘。其中，数据可视化主要针对信息绘图学或信息可视化，其宗旨是以图形化的方式有效地展示信息，有效辅助算法设计和软件开发，如Tabusvis[15]，它是一个轻型的可视化系统，可以提供对多维数据的灵活、可定制的数据可视化。统计分析则基于统计学理论，也是应用数学的一个分支，此过程中与不确定性均由概率理论建模。而较火热的数据挖掘是发现大数据集中数据模式的计算过程。现如今，许多数据挖掘算法已在人工智能、机器学习、模式识别等领域得到了广泛地应用。

大数据的类别不同，其分析方法也存在着差异。如针对于结构化数据，可采用数据挖掘和统计分析技术，如统计机器学习，基于精准的数据模型以及强有力的算法，可以被应用于异常检验[16]及能量控制[17]上；针对于文本分析，即电子邮件、文档、社交媒体等，可采用NLP技术来增加其可用信息，常用的文本分析技术有信号提取、主题建模、聚类、分类等。现在较热的命名实体识别（named-entity recognition , NER）被广泛应用于新的分析应用及生物医学中；多媒体数据分析则是从多媒体数据中提取相关的知识，进而理解数据中包含的语义信息，如音频摘要可提取突出的词句，视频摘要将最重要或最具代表性的序列进行动态或静态的合成。此外还有Web数据、社交网络数据、移动数据等。

总而言之，数据分析是利用分析方法或工具来检验、改变数据，或是通过建立模型的方式，从中获得其意义与价值。现在，很多应用领域都会通过利用相关的数据方法从而获取其预计的结果。即使针对不同的领域，其对数据的需求与数据的相关特性都千差万别，但是，它们仍可能通过一些较为相似的底层技术来支持。

1. 数据解释

在完善的数据分析过程中，数据解释这一步骤极关重要。由于数据量的不断扩大，针对于数据的分析结果往往也会更加复杂多变，以往的数据显示方法已经不能满足当下的需求。因此，为了提高数据的解释、展示能力，现在大多企业都引入了“数据可视化技术”或“人机交互技术”来作为解释大数据最有力的方式。

针对于数据可视化，其核心旨在借助于图形化的方式，更加清楚有力地传送与交流信息。也就是说，数据可视化通过以某种概要形式将有关信息的各种属性及变量提取出来。它是一种不断演变的概念，一般主要是指技术上较为高级的技术方法，而这些技术方法允许利用图形、[图像处理](https://baike.baidu.com/item/%E5%9B%BE%E5%83%8F%E5%A4%84%E7%90%86" \t "_blank)、[计算机视觉](https://baike.baidu.com/item/%E8%AE%A1%E7%AE%97%E6%9C%BA%E8%A7%86%E8%A7%89" \t "_blank)以及用户界面，通过表达、建模以及对立体、表面、属性以及动画的显示，对数据加以可视化解释。

人机交互技术（Human-Computer Interaction Techniques），则是通过计算机的输入、输出设备，以有效的方式来实现人与计算机对话的技术。现在，人机交互技术的领域热点的内在潜力也不断被挖掘。比如智能手机配备的地理空间跟踪技术，可应用于可穿戴式计算机、隐身技术、浸入式游戏等的动作识别技术，还可以应用于[虚拟现实](https://baike.baidu.com/item/%E8%99%9A%E6%8B%9F%E7%8E%B0%E5%AE%9E" \t "_blank)、遥控机器人及[远程医疗](https://baike.baidu.com/item/%E8%BF%9C%E7%A8%8B%E5%8C%BB%E7%96%97" \t "_blank)等的触觉交互技术。

热点技术的应用开发是机遇也是挑战，基于视觉的手势识别率低，实时性差，需要研究各种算法来改善识别的精度和速度，眼睛虹膜、掌纹、笔迹、步态、语音、唇读、人脸、DNA等人类特征的研发应用也正受到关注，多通道的整合也是人机交互的热点，另外，与“无所不在的计算”、“云计算”等相关技术的融合与促进也需要继续探索。

### 1.2.4 大数据时代引起的变革

大数据在现今的各行各业都快速渗透，其影响能力迅速扩大，已经成为当今社会重要的生产因素。目前，各大领域都加快了大数据的应用规模，从而重构未来的核心竞争力。美国麦肯锡咨询机构发布的一份关于大数据的研究报告中指出，大数据已经对金融业、电信行业、政府公共管理、健康医疗等行业产生了重大的影响。2017年4月, 医疗行业发生两件大事, 先是由中国电子信息产业集团有限公司等4家公司发起的中国健康医疗大数据产业发展有限公司成立, 紧接着由神州数码等13家公司发起的中国健康医疗大数据股份有限公司成立, 两大集团都以国有资本为主体, 行业领军企业东软集团、万达信息、易联众、荣科科技成为核心成员单位, 这都表明我国健康大数据行业迎来正规军。[18]许多国家的公共卫生系统现在正在使用先进的医学成像媒体提供电子病历；大数据为流行病学家，医生和卫生政策专家提供了一个很好的机会，可以对数据进行判断，从而更好的对患者进行护理。[19]大数据在生态林业的应用主要从资源监管、生态保护、生态修复、应急服务等方面体现出来，林业生态的建设与信息的相互结合，使其发展的速度十分迅速，同时也为生态林业的发展带来新契机。大数据对教育领域也产生了巨大的影响，其中包括：教育管理思维方式、学习行为、教学评估等。将大数据应用于教育领域之中, 将有利于改革现有的教育理念和教育思维的创新, 促进个性化教育的实现和新的教育方式的重组。有利于学校数据化管理和教育信息化的实现, 促进教育领域的发展。

大数据时代的变革主要在以下三个方面体现：大数据时代的思维变革、大数据时代的商业变革和大数据时代的管理变革。

1. 大数据时代的思维变革

大数据与三个非常重大的思维转变息息相关：第一，如果要对某事物进行数据分析，不再仅仅依靠少量的数据样本，而要分析与其相关的所有数据；第二，人们已从原先追求数据的精准性转变为如今追求数据的复杂多变性。第三，人们更加在意事物彼此之间的相关关系。

在非常多的情况下，使用所有获取的数据变得越来越可能，但因此也会付出一定的代价，也就是说，我们要用较好的措施来处理数据中的错误信息。而大数据时代，要求我们不再过分关注精确性的优劣。我们必须能够接受数据的混乱性及其不确定性，从而开辟新的思维途径。

1. 大数据时代的商业变革

随着大数据时代的到来，大数据已经成为许多公司竞争力的来源，整个行业的结构也因此发生了翻天覆地的变化。掌握着大量数据的大公司通过分析收集到的相关数据，成功实现了商业模式的转型。而小公司可以享受到非固定资产规模所带来的好处。因此，大数据时代 大小公司提供了更加有利的发展，而相反地，针对大多中等规模的公司，如果不能较快适应大数据引发的商业变革，将很可能在行业中遭遇滑铁卢。

1. 大数据时代的管理变革

在大数据时代，我们在生产及信息交流所引起的变革，必须会引起管理变革。因此，建立一个不一样的隐私保护模式，将会更加注重数据使用者对其应用数据的行为约束，并要为其行为承担相应的责任。

大数据时代的到来给商家提供了无限的价值，为我们的生活带来了极大的便利，在给人们思维、商业、管理带来重大变革的同时，也给我们带来了许多不可避免的问题，如社会安全问题、个人隐私问题、国家安全利益等。如何处理与应对大数据时代带来的诸多问题，如何最大程度上保护隐私安全的同时为人们带来利益，成为了当今必然要研究的课题。

## 1.3 大数据时代的安全与隐私保护

大数据隐含着大量的商业价值，目前，各大领域都争先恐后地做着有关大数据的分析与挖掘。企业、运营商，也都在各自所拥有的数据库里发掘信息所潜在价值，从而提高自身的利润或达成其他目的。

然而，科学技术是一把“双刃剑”，人们在享受大数据挖掘为其生活所带来便利的同时，也不可避免地会泄露自身的隐私。大数据所引发的安全问题，与其安全问题所造成的代价同样令人注目。

当前，人们在互联网上的一系列操作都掌握在互联网商家的眼中。例如，购物的习惯、阅读的习惯、搜索的习惯。此外，多数事例表明，即使没有危害的数据，一旦被大量收集，也会暴露用户的个人隐私情况。而事实上，人们所面临的威胁，并非仅仅只是个人隐私泄漏这样简单。大数据，在其存储、处理、传输进程中，也会面临安全风险。

如国内2017年的京东员工泄露隐私事件，50亿条公民的信息被走漏，以韩某为首的犯罪团伙涉嫌侵略社交、游戏、视频直播、医疗等各类公司的服务器，其中犯罪嫌疑人郑某在加入京东前曾在国内多家知名互联网公司工作，并长期与盗卖个人信息的犯罪团体合作。2018年4月，饿了么、百度外卖、美团等外卖平台的个人信息发生泄露。国外，据悉在特朗普当选总统的数据分析公司Cambridge Analytica被发现在未经许可的情况下，获取了约五千万条Facebook用户的数据，并对这些用户精准投放政治广告。2017年5月12日，比特币勒索病毒（WannaCry）在全球范围内爆发，此次勒索事件与以往相比最大的区别在于勒索病毒结合了蠕虫的方式进行传播。同年6月27日，新一轮的勒索病毒Petya袭击欧洲多个国家，该病毒对电脑的硬盘MFT进行加密，使操作系统无法进入，为客户的上网行为带来极大的安全风险。2016年1月，美国最大的有线电视公司时代华纳公开表示，公司旗下近32万用户的邮件和密码信息已被网络黑客窃取。据了解，这些邮件和密码信息很有可能是通过网络钓鱼的方式获得，同时也可能是保存了时代华纳用户数据的第三方合作商信息泄露所致。2016年9月，雅虎公司宣布至少有5亿账户信息被黑客盗取，被盗的内容包括了用户姓名、电邮地址、电话号码、生日、密码等，这也是史上最大单一网站信息遭窃的记录。2018年1月，WhatsApp上匿名卖家提供的一项出售登录凭证的服务可以通过输入任何一个Aadhaar号码(一个12位的唯一标识符，每个印度公民会使用到它)检索印度唯一身份识别管理局(UIDAI)存储的关于被查询公民的诸多类型的信息。这些数据包括姓名、住址、照片、电话号码和电子邮箱地址。据了解，这起数据泄露事件已经损害了在印度注册的11亿公民的个人信息。

在此背景下，大数据安全与隐私保护十分必要，只有加强大数据安全技术的研究，才能确保大数据安全，保护用户的隐私，保证人们在使用互联网过程中数据信息传输的安全性和稳定性，避免给用户带来利益损失，从而活跃市场，推动社会的稳定发展。

本小节将从当今大数据面临的安全、隐私安全及相应的保护技术方面进行阐述。

### 1.3.1 大数据安全

大数据安全（Big Data Security）是一个综合、交叉的学科领域，其内容含盖案例体系结构、安全协议、密码理论、信息分析、案例监控及应急处理等。它是一种主动的包含措施，包括数据本身的安全和数据防护的安全。针对于数据本身，现主要采用现代密码算法的方式来对数据进行主动保护，如数据保密、数据完整性、双向强身份认证等；而针对于数据防护的安全，现主要是采用现代信息的存储手段来对数据进行主动防护，如通过磁盘阵列、数据备份等手段来确保数据的安全。

因此，从宏观上来讲，大数据的安全，应从两个层面出发：第一，要保障大数据的安全，即保障大数据计算过程、数据形态及其应用价值的处理技术；第二，大数据要用于安全，即利用大数据技术从而提升信息系统案例效能和能力。具体参照以下三个方面：

1. 加强大数据安全技术的研究与应用

大数据安全问题的产生与大数据安全技术之间有着必然的关联性，由于大数据安全技术的不合理、不先进，就容易造成大数据安全问题发生，难以保护用户的隐私。对于大数据而言，大数据安全保护技术是大数据安全与隐私保护的直接载体，能够确保数据信息在数据库领域范围的得到有效的处理。为了确保大数据安全，就应当加强大数据安全技术的研究和应用，以先进的大数据安全保护技术为依托来为大数据信息的存储、运输、处理提供安全保护。如身份认证技术，在大数据环境下，通过身份认证技术，用户在使用大数据的时候都需要通过身份认证来获得数据信息的使用权，在身份认证技术的保护下，可以实现最大化的保护用户隐私的目的，避免给用户带来经济损。

1. 加强社交网络中数据信息的监督

社交网络作为人们进行信息交流和沟通的纽带，在大数据时代里，越来越多的人活跃在社会媒体上，而作为社会的一部分，都会涉及到个人信息的部分泄露，对用户的人身安全及财产安全造成威胁。为了避免安全问题的发生，防止用户隐私的泄露，加强社会网络中数据信息的监管十分必要。对匿名的社交媒体信息，要利用信息技术对其进行社会网络匿名保护，确保个人信息安全，避免用户信息泄露而带来巨大的利益损失。同时，在社会网络信息传播过程中，要加强信息的全面监管，保护社会网络用户在交流过程中传输的信息的安全性，避免被他人恶意利用，保护用户的人身安全及财产安全。

1. 做好大数据安全的宣传与隐私保护的宣传工作

人们的安全意识的高低是引起大数据安全问题的一个重要因素。随着互联网的普及，人们利用互联网来进行各种互动，而在以利益为核心价值观的世界例，用户容易受到利益的趋势，而许多不法份子正是利用了用户的这种心理，在互联网页面上参插一些能够吸引用户的小广告，而这些小广告大多待木马病毒，一旦用户点开，就会受到病毒入侵，从而威胁到用户系统安全。为了确保大数据安全，保护用户的隐私，就必须加大大数据安全的宣传，将一些常见的大数据安全风险向广大群众普及，提高他们的认识，同时向用户普及一些大数据安全技术，让用户掌握一些基本的隐私保护技术，从而更好地满足用户的需要，保证用户安全的进行大数据信息传输、处理、存储，避免安全风险的发生。

### 1.3.2 隐私风险及相应保护技术

大数据时代的来临为我们带来了巨大的财富。然而在大数据的搜集、存储、使用等过程中，同时也会伴随着众多安全风险，随着科技的不断发展，新型的隐私问题也在不断地被创造。如何在不泄露用户隐私的前提下，提高大数据的利用率，挖掘大数据的价值，是目前大数据研究领域的关键问题。具体而言，实施大数据环境下的隐私保护，需要在大数据产生的整个生命周期中考虑两个方面：如何从大数据中分析挖掘出更多的价值；如何保证在大数据的分析使用过程中，用户的隐私不被泄露。

下面简要介绍当下一些较为人们熟知的隐私风险及相应的保护技术。

1. 基于位置服务隐私风险

随着无线通信和移动定位技术，如GPS，WIFI等不断发展与完善，基于位置的应用也越来越受到人们的欢迎，用户为了获得自身位置周围的某项特定服务，会向服务商提供自己的位置信息，服务提供商根据用户的精准定位从而提供优质的服务，如美团外卖会推荐给用户周围的饮食，百度地图用于进行导航等。用户在这种服务上获得极大便利的同时，也会暴露自己的个人信息，如日常的移动信息、地址、喜好等。尽管在享受服务的过程中，用户为了保护自身隐私会用一些假名来掩盖自己的真实身份，但是已有的研究表明这样的简单匿名法并不能完全保障用户个人的隐私信息。

针对怎样提供优质的基于位置服务的同时保护好用户的位置信息，研究人员提出了大量的研究算法，如匿名（Anonymity）技术、干扰（Perturbation）技术等。这些技术被称为“语法隐私”保护技术（Syntactic privacy），通过直接改变数据集中的某些属性的值的方式，使得攻击者无法将数据集中的某个记录与其他多个数据记录进行区分，从而保护用户的个人隐私。如k-anonymization技术、l-diversity、t-closeness方法等。

1. 数据发布带来的风险

当企业、政府等机构或是个人收集到其他人的数据后，他们可以对这些数据进行分析，从而提升自身的业务能力或是服务质量。但与此同时，他们也非常有可能会将这些数据共享或发布，从而获取更为精确的、更具价值的信息。如医院每天都会记录大量病患的个人信息，而各大医院为了加强医疗管理与服务，很可能会将数据进行共享，与此同时，其他没有直接获取病人信息的服务部门需要大量的信息来进行相应的医学研究。在此前提下，企业或个人就必须共享、开发这些记录。然而在发布这些记录的同时，又必须时刻保护病人的隐私安全，防止病人隐私的泄露，否则一旦发生隐私信息泄露问题，这将会给企业、分共部门带来经济及名誉上的双重损失。

一般而言，语法隐私类的保护技术较多应用于数据发布技术当中。除上文提到的一些匿名方法外，也会采用抑制、泛化等操作。

1. 数据交互应用带来的风险

在数据交互应用中，数据收集者为了保护数据生成者的隐私，会在真实的查询结果的基础上加入一定量的噪音，然后再将加了噪音的数据反馈给数据用户。不同于非交互式所带来的风险，数据交互过程中，用户始终无法直接接触到原始数据集，数据攻击者通过查询问题，然后得到数据收集者反馈的结果来进行攻击。假设将某人准确的身高作为高度敏感的信息，现在某个数据收集者利用收集的数据查得了人们平均身高，并通过查询得到信息：姓名为Jerry的女生经人们的平均身高低3厘米，从而就可以确定Jerry的真实身高数据，这样，Jerry的隐私信息就被泄露了。因此，攻击者并不关心被攻击的对象是否一定在数据中，而他只要通过查询结果和辅助信息就可以完成攻击。

数据交互技术多采用“语义隐私”保护技术（Semantic privacy），在这种语义隐私保护技术的分类下，数据保护技术的主要目标是满足这样的一种特定要求，即数据集中的单一记录的增加或是删除对于攻击者来说都是不敏感的、无法察觉的。其典型技术为差分隐私技术（Differential privacy），该技术在统计数据库领域已有了广泛的使用，除此之外，也被广泛运用于数据挖掘、机器学习、社交网络、安全通信及经济学等。

## 1.4 小结

随着云计算、云服务、物联网等的迅猛发展，目前，数据正以指数型迅速增长。人们生活的方方面面都离不开大数据，大数据的时代已经来临。合理利用大数据，会为人们的生活带来诸多便捷，但同时也带来了巨大的挑战。

本章节针对国内外大数据的相关研究进行了回顾和总结，围绕大数据的基本概念及特征，简要说明近几年来国内外有关大数据的政策及应用；围绕大数据，提出云计算、云服务，并针对其区别与联系进行了探究；通过对典型大数据的分析流程，分别介绍了数据采集、数据的处理与集成、数据分析和数据解释这四大流程，围绕大数据时代引发的变革提出了隐私安全等问题，简单介绍了当前针对大数据的隐私风险及相应的保护技术。

## 1.5 参考文献

1. 余斌,平新乔,张占斌,白永秀,李建伟,陈彦斌,王宏淼,米本家,刘刚,王佳宁,罗重谱.“十三五”期间的中国经济增长 改革传媒发行人、编辑总监王佳宁对话九位经济学者[J].改革,2015(01):5-40.
2. 国务院印发《促进大数据发展行动纲要》[N]. 人民日报,2015-09-06(001)
3. 工业和信息化部印发《大数据产业发展规划(2016-2020年)》[J].电子政务,2017(02):49.
4. 工业和信息化部印发《云计算发展三年行动计划(2017-2019年)》[J].印刷杂志,2017(05):74.
5. 国家自然科学基金委员会印发《大数据驱动的管理与决策研空重大研究计划2017年度项目指南》
6. 中商产品研究院印发《2018-2023年中国大数据行业发展前景及投资机会研究报告》
7. 田倩飞.美国发布联邦大数据研发战略计划[J].科研信息化技术与应用,2016,7(04):95-96.
8. 中华人民共和国科学技术部印发《瑞士国家重点科研计划大数据专项启动》
9. Barwick H. The “four Vs” of Big Data. Implementing Information Infrastructure Symposium [EB/OL]. [2012-10-02]. http://www.computerworld.com. au/ article/ 396198/ iiiis\_four\_vs\_big\_data/
10. Manyika J, Chui M, Brown B, et al. Big data: the next frontier for innovation, competition, and productivity. McKinsey Global Institute, 2011
11. Cooper M, Mell P. Tackling Big Data. NIST, 2012
12. Naimi A I, Westreich D J. Big Data: A Revolution That Will Transform How We Live, Work, and Think[J]. Mathematics & Computer Education, 2013, 47(17):181-183.
13. Mell P, Grance T. The NIST definition of cloud computing[J]. Communications of the Acm, 2011, 53(6):50-50.
14. Castillo C. Effective web crawling. In: Proceedings of ACM SIGIR Forum, New York, 2005. 39: 55–56
15. Nguyen Q V, Qian Y, Huang M L, et al. TabuVis: a tool for visual analytics ultidimensional datasets. Sci China Inf Sci, 2013, 56: 052105
16. Baah G K, Gray A, Harrold M J. On-line anomaly detection of deployed software: a statistical machine learning approach. In: Proceedings of the 3rd International Workshop on Software Quality Assurance, Portland, 2006. 70–77
17. Moeng M, Melhem R. Applying statistical machine learning to multicore voltage & frequency scaling. In: Proceedings of the 7th ACM International Conference on Computing Frontiers, Bertinoro, 2010. 277–286
18. 刘孝男,付嵘,李连磊.大数据时代,医疗行业信息安全面临的机遇与挑战[J].中国信息安全,2018(07):100-102.
19. Senthilkumar SA, Bharatendara K Rai, Amruta A Meshram, Angappa Gunasekaran, Chandrakumarmangalam S, Big Data in Healthcare Management: A Review of Literature, American Journal of Theoretical and Applied Business. Vol. 4, No. 2, 2018, pp. 57-69. doi: 10.11648/j.ajtab.20180402.14
20. 刘古月,陈剑波.浅析大数据在生态林业上的运用[J/OL].中国战略新兴产业[2018-08-19].https://doi.org/10.19474/j.cnki.10-1156/f.005890.
21. 周爽.基于大数据的教育领域应用研究[J].电脑迷,2018(08):232.

# 第二章 传统的隐私保护方法（陈竑毓，王寒）

大数据环境下信息高度共享，隐私数据的安全问题越来越成为大数据可持续发展的重要需求之一，在数据使用时保护个人隐私信息不被泄露成为亟待解决的问题。隐私保护方法，区别于访问控制和加密技术，最显著的特点就是数据公开发布并且具有可用性。本章针对大数据用户隐私保护问题，介绍了两种传统的保护方法-匿名化技术和安全多方计算。其中匿名化技术主要用于保护数据发布场景下隐私泄露问题，安全多方计算基于不可信“第三方”的假设，在保证输入独立、计算正确的同时，同时不泄露各输入值给参与计算的其他成员。

## 2.1 数据匿名化方法（陈竑毓）

数据发布作为一种有效的信息共享手段，在辅助支持知识决策、科学研究的同时，也导致了个人隐私信息的泄露。传统的数据匿名化技术应用于关系型数据集的发布场景，要求数据发布者不能直接发布收集得到的原始数据集，必须先对其进行匿名化操作，遵循隐私保护匿名化原则。

在本节中将重点阐述以K-匿名模型为基础的数据匿名发布技术，首先简要介绍隐私保护匿名化方法的相关定义；其次，详细介绍与数据匿名化相关的模型与技术，重点阐述基本的K-匿名模型，并总结概括实现匿名化的相关技术，在此基础之上分析常见的针对匿名模型的攻击方式以及应对的匿名模型；最后简要介绍匿名化模型的几个主要的性能指标和度量标准。

### 2.1.1 相关定义

1. 属性分类

数据匿名化方法主要用于研究保护关系型数据集的数据安全，避免隐私泄露。关系型数据集是由代表记录的行和代表属性的列组成的，每条记录包含了多个描述个体相关信息的属性值，将等待发布的数据集中的属性分为以下四类：（1）显式标识符（Explicit Identifier, EID）：能够唯一确定记录的属性，例如身份证号、姓名、手机号等；（2）准标识符（Quasi-Identifier, QID）：能够在攻击者背景知识，即其他外部信息等辅助下识别对应记录的属性的集合，例如，由（邮编，性别，生日）构成的属性集合能以较高概率识别个体。根据攻击者背景知识的不同，准标识符可以相应变化；（3）敏感属性（Sensitive Attributes, SA）：个体需要保护的隐私信息，例如收入、医疗状况等；（4）非敏感属性（Non-Sensitive Attributes, NSA）：不属于上述三类的属性。

1. 链接攻击

链接攻击是间接造成隐私泄露的重要方法。针对多个去除显式标识符处理的发布数据集，看似每个数据集都存在其自身的安全性，即孤立情况下不会泄露任何隐私，但对数据集之间的重叠属性做链接查询就可以重新识别少量甚至唯一的用户个体并获取其隐私信息，很大程度上这些重叠属性组成或包含了准标识符属性。链接查询导致的攻击称为链接攻击，将识别唯一个体，从而导致严重的隐私泄露。文献[1]指出，通过对选民数据集和GIC医疗匿名数据集进行链接攻击，会重新识别匿名数据集中的用户，导致个人隐私的泄露。图1展示了两个数据集发布的内容，其中选民数据集包括Name、Address、Date registered、Party affiliation、Date last voted、ZIP、Birth date、Sex等属性内容，GIC医疗匿名数据集包括Ethnicity、Visit date、Diagnosis、Medication、ZIP、Birth date、Sex等属性内容，其中{ZIP、Birth date、Sex}作为两个数据集的交集，可以看作是个人的QID。因选民数据集中存在Name、Address等个人标识属性，通过两个数据集的链接攻击，则将拥有相同QID的记录映射到同一个体，最终不但识别了个体，并泄露了医疗隐私信息。

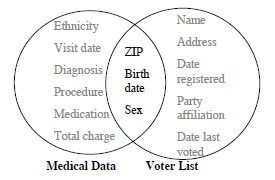


图2-1 链接攻击实例

1. 隐私泄露类型

在数据发布场景下，存在两种形式的隐私信息，一是表中记录对应个体而构成的隐私信息，例如医院发布心脏病科室患者就诊记录表，每条记录对应的用户就是隐私信息，另外一种就是发布表中包含敏感属性列，直接构成隐私信息。攻击者对多个包含用户个人信息的数据集进行链接攻击就可能造成上述两种隐私信息的泄露，将因链接攻击造成的隐私泄露分为身份泄露和属性泄露两种类型。

1. 身份泄露

假设攻击者提前获知攻击目标的准标识符属性同时确定攻击目标存在于数据表中，并对多个发布数据集做了链接查询，进一步查询发现准标识符属性对应唯一或少量记录。此时，攻击者可确定发布数据集中个体的身份信息，从而造成个体身份信息的泄露，称之为身份泄露。身份泄露最为严重，因为一旦发生身份泄露，同时伴随着属性泄露。

1. 属性泄露

发布的数据集中包含敏感属性列，并且攻击者通过链接方式查询数据集，发现对应于某一准标识符属性的多条记录的敏感信息完全相同或很大程度上类似。此时，攻击者就可以以很高的概率推断个体的敏感信息，如果发现攻击目标的准标识符属性与上述一致，就会造成个体敏感信息的泄露，称之为属性泄露。

### 2.1.2 K-匿名模型

K-匿名[1]最早由Sweeney提出，是通过泛化、抑制等匿名化技术对原始数据集中的属性和记录进行处理以满足K-匿名约束，并最终发布K-匿名数据集的方法。简单来说，泛化就是对属性进行抽象提取，例如将连续值19012泛化成[19000-19020]区间或者是1901\*，将分类值“狗”泛化到更抽象的“动物”，抑制就是选择不发布某些属性或记录的方法。

对于任何一条记录，K-匿名约束要求匿名表中至少有K-1条记录在QID上与其保持一致。K-匿名模型为防止攻击者进行身份识别而造成的隐私信息泄露提供了很好的解决方法，其直接应对的是由链接攻击造成的重新识别。

定义2.1 K-匿名：设匿名数据表为，QID是其对应的准标识符，数据表T满足K-匿名约束当且仅当QID在匿名数据表中至少出现K次，即至少有K条记录在准标识符上保持一致。

假设表T(m)在属性集QID上投影得到元组多重集，表T(m)满足K-匿名约束当且仅当对于每个QID得到的，其组内的元组数至少为K（）。

表2-1是一张原始医疗数据表，表2-2是经过匿名处理的满足2-匿名约束的发布表。

表 2-1 原始医疗数据表

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Name | Race | Birth | | Sex | Zip | Disease |
| Alice | Blank | | 1965-3-18 | M | 02141 | Flu |
| Bob | Blank | | 1965-5-1 | M | 02142 | Cancer |
| David | Blank | | 1966-6-10 | M | 02135 | Obesity |
| Helen | Blank | | 1966-7-15 | M | 02137 | Gastritis |
| Jane | White | | 1968-3-20 | F | 02139 | HIV |
| Paul | White | | 1968-4-1 | F | 02138 | Cancer |

表 2-2 医疗数据2-匿名表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Race | Birth | Sex | Zip | Disease |
| Blank | 1965 | M | 0214\* | Flu |
| Blank | 1965 | M | 0214\* | Cancer |
| Blank | 1966 | M | 0213\* | Obesity |
| Blank | 1966 | M | 0213\* | Gastritis |
| White | 1968 | F | 0213\* | HIV |
| White | 1968 | F | 0213\* | Cancer |

数据表K-匿名化就是对QID值进行泛化使原始数据表T(m)生成K-匿名约束表T’(m)的过程，T’(m)称为表T(m)的K-匿名表。

定义2.2 等价类：也称K-匿名组。具有相同QID值的所有元组(行)组成一个匿名组，称为一个等价类。即对于每个QID投影得到的，称之为一个等价类。

例如，表2-2中记录1和2、记录3和4、记录5和6分别构成了一个等价类。

K-匿名模型使得至少K条记录(行)在QID上保持一致，从而保证个体被识别的概率至多为1/K，因此通常用于防止身份泄露。同时，匿名参数K作为K-匿名模型的隐私度量，指定了 可承受的最大信息泄露风险，其值越大，代表隐私保护程度越高，反之亦然。K-匿名模型是通过降低数据可用性来提高隐私保护程度的，在实际应用中需要平衡两者来确定最优K值，文献[2]针对K值的选取进行了讨论。

### 2.1.3 数据匿名化技术

K-匿名模型的提出为企业、研究机构等第三方组织安全发布数据集提供了解决思路，发布数据集满足K-匿名约束，就说这个数据集提供K-匿名保护。问题是如何操作才能让数据集满足K-匿名约束？下面介绍几种常见的实现数据匿名化的技术，其中K-匿名常用的是泛化和抑制。

去标识： 去除记录的显示标识符，但是依据准标识符的再识别攻击也可以唯一确定个体，所以去标识的方法一般用于数据发布的预处理环节；

泛化：对数据进行抽象、概括的方法，例如Age属性的值为20，可以泛化为[20-25]，泛化通过降低数据精度的手段实现匿名，但是不影响数据的统计特性；

抑制：不发布某些属性或者记录，属性列不具备统计特性，可抑制其发布，另外针对数据集中某些记录的属性存在极值的情况，将所有记录泛化到同一层会导致数据过度泛化而失去效用，采用抑制的方法将极值记录删除不发布；

交换：在数据集中多条记录的对应属性值之间进行随机置换的方法，发布经过置换后的数据集，使得攻击者无法将数据记录与个体一一对应；

分割：分割技术不改变QID和敏感属性值，而是通过降低两者之间的关联从而实现隐私保护。将等待发布的数据集分割成包含QID的表QIDT和包含敏感属性的表SAT，两张表通过共同属性GroupID关联。分割技术不改变原始数据，查询精度高，但是在分割表的过程中会产生多余记录；

微聚合：将原始数据集在属性上取值近似的多条记录聚合在一起形成等价组，一般取属性的平均值作为等价组的聚合值，发布的时候用聚合值代替原始值，从而实现了隐私保护。微聚合技术适用于处理数值类型的数据，微聚合和泛化技术都是先形成等价组，后用隐私保护数据代替原始值的思路，两者的区别在于一个是泛化思想，一个是聚合思想。

匿名数据发布主要是为了达到避免身份识别或避免敏感属性泄露或避免高概率推断个体存在于表中的目的，上述提到的匿名化技术分别对应于一种或多种隐私保护目的。在实际应用中，往往将多种匿名化技术进行结合使用，例如基于泛化和抑制的匿名化技术，基于泛化和分割的ANGEL[3]技术等等。

### 2.1.4 针对匿名模型的攻击

匿名模型是指具有特定匿名约束，并应用相应匿名化技术的模型方法，K-匿名模型作为常见的匿名模型，个体被识别概率至多为1/K，因此通常可以防止因身份识别而造成的敏感属性值的泄露。上文提到的链接攻击作为一种针对简单匿名模型的攻击方式，可以有效被K-匿名模型抵抗。然而，K-匿名模型也存在其固有缺陷，未对敏感属性做任何约束，这可能带来隐私泄露。下表列出了几种常见的K-匿名模型无法抵抗的攻击方式。

表 2-3 几种常见的针对K-匿名模型的攻击方式

|  |  |
| --- | --- |
| 几种常见的针对K-匿名模型的攻击方式 | |
| 同质攻击 | 攻击者通过链接攻击无法确定唯一个体，但是等价类组内记录拥有同样的敏感属性，导致敏感属性的泄露。此时攻击者获取敏感属性的攻击方式称为同质攻击。 |
| 相似性攻击 | 攻击者无法确定等价类中的唯一个体，但因组内记录敏感属性值具有极高的相似性，从而攻击者可以以高概率推断出敏感属性。 |
| 背景知识攻击 | 背景知识攻击是攻击者根据所掌握的相关知识来推理隐私的过程。将攻击者提前获得的关于发布数据的信息称为背景知识，例如攻击者可以根据外部信息获得目标对象的多个属性集合，从中攻击者可以获知目标对象的准标识符，从而达到识别个体或推断目标对象的敏感属性的攻击的目的。 |

### 2.1.5 增强K-匿名模型

1. -多样性模型：

针对K-匿名数据集的同质攻击和背景知识攻击，Machanavajjhala[4]等人在K-匿名模型基础上提出了-多样性模型。

定义2.3 -多样性：对于K-匿名表，称其满足-多样性当且仅当对于敏感属性A，表中每个等价类至少包含个“表现良好”的敏感属性值。其中“表现良好”具有下列形式：

1. 不可区分-多样性

发布数据满足不可区分 -多样性当且仅当每个等价类中至少包含种不同的敏感属性值。例如表2中每个2-匿名组都含有2个不同的敏感属性值，因此也是满足2-多样性的；

1. 熵-多样性

发布数据满足熵-多样性当且仅当每个等价类均满足下列公式：

 ( 2‑1 )

其中S为敏感属性A的值域，为敏感属性值s在等价类中出现的概率。公式表明，等价类中敏感属性值分布越均匀，熵值越大，隐私泄露概率也就越低；

1. 递归-多样性

发布数据满足递归-多样性当且仅当每个等价类都满足，其中m表示等价类中不同敏感属性的个数，表示等价类中第频繁的敏感属性值的个数，c是数据发布方给定的一个常量。递归-多样性保证了每个等价类中频率最高的敏感属性值出现频率不至于太高。

满足-多样性约束的数据集在同质攻击和背景知识攻击下，就算攻击者掌握了攻击目标的准标识符属性，并找到其所在的等价类，但是由于敏感属性之间各不相同或者在分布概率上具有离散的特点，使得攻击者无法推断其敏感属性值，从而保护了属性隐私。

1. -closeness：

虽然-多样性模型在敏感属性值上做了约束，但是其仍不能有效抵抗偏斜性攻击和相似性攻击。针对这个问题，Li等人[5]提出了-closeness模型。该匿名模型要求发布数据除了满足-多样性约束之外，还要求所有敏感属性值在每个等价类内的分布和敏感属性值在匿名表中的总体分布之间的差值不得超过阈值t。论文中同时声明距离度量范式Earth Mover Distance（EMD）可计算两个分布之间的差值，并分别定义了分类数据和连续数据的计算公式。

-closeness模型使得所有等价类中敏感属性值的分布尽可能地接近该属性的全局分布，一定程度上解决了针对敏感属性的偏斜性攻击和相似性攻击，但是匿名化的结果是降低了发布数据的可用性，提高发布数据可用性的唯一方法是增大阈值t。

1. -匿名

基于K-匿名模型，Wong[6]等人提出了-匿名模型，其应对的假设是敏感属性S中部分属性值是正常值，不需要对其进行保护，而需要对敏感值进行保护。例如，疾病这个属性是隐含敏感信息的，但是像普通的感冒、发烧则是很普通的属性值，不需要对其进行保护，而HIV则属于敏感值，需要对其进行保护。-匿名模型通过约束敏感属性来实现隐私保护。

假设K-匿名数据集为D，表中准标识符属性集为Q，敏感属性记为S，且，s是敏感属性S的取值，表示一个等价类，表示等价类中敏感属性S取值为s的记录的集合，是用户指定的阈值，且。

定义2.4 -deassociation：数据集D关于准标识符集Q和敏感属性值s是-deassociation的当且仅当每个等价类中s的出现频率小于等于，即；

如表2所示，其2-匿名数据集是关于属性集{Race, Birth, Sex, Zip}以及属性值Cancer 0.5-deassociation的。

定义2.5 -匿名：数据集D是满足-匿名的当且仅当数据集D关于准标识符属性集合Q和敏感属性值s既满足K-匿名约束，又是-deassociation的。

从上述-匿名的定义不难看出，模型参数与选取的敏感属性值有关。如表4所示的3-匿名数据集中，准标识符属性集为{Job, Birth, Postcode}，敏感属性值为HBV时，该数据集满足-匿名。但是在敏感属性值为HIV时，攻击者推断用户患癌概率为50%。基于此类问题，Wong提出一般-匿名模型。

表 2-4 医疗数据集的2-匿名表

|  |  |  |  |
| --- | --- | --- | --- |
| Job | Birth | Postcode | Illness |
| \* | 1975 | 4350 | HIV |
| \* | 1975 | 4350 | HIV |
| Catl | 1955 | 5432 | HBV |
| Catl | 1955 | 5432 | Cancer |
| Catl | 1955 | 5432 | HIV |
| \* | 1975 | 4350 | HBV |
| \* | 1975 | 4350 | Cancer |

定义2.6 -rare：等价类中，对于任一属性X，x表示在属性X上的取值。若属性X的每个取值在任一等价类中均满足，则称关于属性X满足-rare。

例如表2-4，取属性X为Illness，等价类为{3, 4, 5}对应的记录，由于“HIV”、“HBV”、“Cancer”的概率都小于等于0.4，所以称是关于敏感属性X -rare的。

定义2.7 一般-deassociation：数据集D关于准标识符属性集Q和敏感属性S是一般-deassociation的当且仅当对于D中所有等价类，都是关于S - deassociation的。

根据表2-4可知两个等价类是关于Illness敏感属性0.5-rare的，所以表2-4是一般0.5- deassociation的。根据定义可知，数据集关于敏感属性是一般-deassociation的，则该数据集关于每个敏感属性值都是-deassociation的，因此可获知攻击者推断敏感值的最大概率为。

定义2.8 一般-匿名：数据集D是满足一般-匿名的当且仅当数据集D关于准标识符属性集合Q和敏感属性值s既满足K-匿名约束，又是一般-deassociation的。

-匿名通过约束敏感属性中敏感值的比例，使得攻击者最多以的概率推断攻击目标的敏感属性值，这在很大程度上保护了用户的隐私信息不被泄露。

### 2.1.6 K-匿名模型度量标准

隐私保护数据发布一方面要对等待发布的数据进行隐私保护处理，以实现隐私保护，另一方面则要保证数据的可用性，以便匿名发布后的数据仍具有统计效用。衡量K-匿名数据发布算法的性能主要在于以下几方面：一方面衡量其隐私保护性能，隐私保护性能越好，说明算法的安全性更高，例如K匿名算法中K值越大，泛化程度越大，相对的隐私保护性能也就更好；另一方面衡量其信息损失及数据效用性，信息损失与数据效用是两个对立的概念，信息损失越大，数据效用越小，反之亦然。数据效用反映了经过隐私保护技术，原始数据的特征信息保留程度，数据特征信息保留越多，数据效用越高。

匿名模型中给定的参数K、l、等都是用来衡量隐私性能的指标，以下列举几种常见的信息损失及数据可用性度量指标。常见的信息损失度量标准包括不可辨别度量标准、熵度量标准等等。

1. 不可辨别度量标准

不可辨别度量标准（Discernbility Metric, DM）是由Bayardo和Agrawal[7]提出的衡量K-匿名数据集不可区分性大小的度量标准。DM给每条数据分配一个惩罚值，惩罚值大小为匿名数据集中与之不可区分的记录的数目，即对等价类内的泛化记录来说，每条记录的惩罚值大小为等价类的大小|E|；对原始表中按抑制处理的数据来说，惩罚值为原始数据表的大小|D|，则匿名表的惩罚值大小，也就是获取该匿名数据集的总代价为：

 ( 2‑2 )

根据上述说明，对每条记录来说，惩罚值是与之不可区分的记录的数目，即在数据集中不可辨别性的度量。对整个匿名数据集而言，惩罚值越大，表示记录不可辨别性越高，则数据可用性也就越低，但同时隐私保护程度也会相应越大。

1. 熵度量标准

文献[8]提出了熵度量标准，用于度量匿名表的信息损失程度。

给定数据集，针对数据集中每个属性都定义一个概率分布，其概率分布公式计算公式为如下，其中表示数据集中第j个属性取值为a的记录的数目。

 ( 2‑3 )

基于上述概率分布公式，可计算属性的熵为：

 ( 2‑4 )

设为的子集，则定义限制属性范围的条件熵计算公式如下：

 ( 2‑5 )

其中，

 ( 2‑6 )

表示数据集中第j个属性取值为b的记录的数目，表示数据集中第j个属性取值范围在中的记录的数目。

给定数据集，为其匿名数据表，为匿名数据表的准标识符属性，则匿名数据表的熵为：

 ( 2‑7 )

由上述可知，熵反映了数据集内属性值的概率分布，即集中程度。当属性分布越均匀，则熵值越大，表示数据混乱程度越大，可用性越低，但是相对应地攻击者的不确定也就越大，数据的隐私保护程度也就越高。

除了上述提到的两种度量方法，K-匿名模型还有精度度量标准[7]、基于距离的度量[9][10]等等。在选择不同的算法实现K-匿名模型时，可以根据实际需求采用合适的度量标准来衡量算法优劣。

### 2.1.7 匿名化技术存在的不足

大数据环境下，数据发布与数据共享变得日益重要，隐私保护技术越来越受到相关研究人员乃至整个社会的关注。K-匿名及其增强模型，作为传统的隐私保护技术之一，展现了它的价值与锋芒。对于链接攻击、背景知识攻击、同质攻击等安全问题，以K-匿名为基础的匿名化技术无疑做出了相当的贡献。但是匿名化技术在提供隐私保护的同时，也存在其不足的地方。

1. 攻击者背景知识假设

在讨论K-匿名模型及其应对的攻击时，都需要提前假设攻击者拥有的背景知识，再根据假设提出相应的隐私保护模型。因此，K-匿名的不通用性限制了其的研究与发展。

1. 隐私保护程度

虽然K-匿名模型中K值及其他参数可间接用作隐私保护程度的度量标准，但其非直接的表示方式不容易让人有直观的认识，并且K值范围粒度太大，不易于精确表示隐私保护程度。

## 2.2 安全多方计算（王寒）

### 2.2.1相关定义

安全多方计算（Secure Multi-Party Computation）是密码学中的一个重要子领域，其目标是创建一个方法，以便各方共同计算其输入函数，同时保证这些输入的隐私性。1982年，姚期智正式提出了安全两方计算[11]，并且给出了安全两方计算协议并推广到了安全多方计算。安全两方计算也是安全多方计算的一种特殊情况，其仅包含两个参与方实体，这两个参与方均持有各自的秘密输入，并且他们期望共同计算某个功能函数，最终得到各自的输出结果。

定义2.9（安全两方计算）两个互不信任的参与方分别持有各自的秘密输入和，想要通过执行一个两方计算协议，协同计算目标功能函数，并在计算完成后得到各自相应的函数输出和，在此过程中不获得其它任何额外信息。

1987年，Goldreich等人提出了可以计算任意函数的基于密码学安全模型的安全多方计算协议[12]，从理论上证明了可以使用估值电路来实现所有安全多方计算协议。1998年，Goldreich对安全多方计算做了比较完整的总结，并且提出了安全多方计算的安全性定义[13].一直以来，提出了越来越高效的安全多方计算协议，安全多方计算现在已经被认为是解决各种现实问题的实际解决方案，尤其是只需要线性共享秘密的问题，主要是在股份之间进行的没有多少交互的本地操作，如分布式投票，私人竞拍，共享签名以及隐私信息检索等。

简单来说，安全多方计算是指在分布式网络中，多个互不信任的参与方持有各自的秘密输入，通过执行一个多方参与的计算协议，协同计算某个预先设定的功能函数，并在计算完成后得到各自相应的函数输出，在此过程中不获得其它任何额外信息。

安全多方计算问题可以抽象概括成如下数学模型：假设有个参与者，每一个参与者都有一个保密输入。这个参与者共同执行一个协议来计算函数。计算结束后，每个参与者仅能够得到自己的输出，除此之外，不知道其他的输入、输出信息。

例如，假设有三个人Alice，Bob和Tom，他们各自的输入x，y，z分别表示他们的工资。他们想找出三个人中工资最高的人，并且相互不透漏每个人的工资。在数学上表示为：

 ( 2‑8 )

在传统的情况下，可以将自己的输入信息交给安全第三方，由安全第三方完成计算函数，然后将结果发送给各个参与者，比如他们三个人拥有一个共同的可信赖的朋友Charlie，他们可以分别向Charlie报告工资，由Charlie计算出最大工资数，然后把这个数字告诉他们。但是，在现实情况中，很难找到一个可完全信任的第三方。因此，这种方案不能满足实际的安全性需求。所以，安全多方计算的研究主要是针对在无可信第三方的情况下设计一个安全的多方计算协议，让各个参与者进行安全的合作计算，比如他们三个人按照协议要求，在不透漏个人做法，不依靠Charlie的情况下，通过彼此交换信息，仍然可以得到工资最大者。

安全多方计算要求各个参与者在计算中除了得知自己的输入和输出以及由这两者能够推导得出的消息外不能得到更多其他任何有价值的额外消息。一个安全多方计算协议通常具有以下性质[14]：

1. 保密性：所有参与方都不能得到除自己的输入输出外更多的信息，保证了执行完协议后各自的秘密不会被泄露。
2. 正确性：安全多方协议必须按照约定的功能函数计算出正确的结果。
3. 输入独立性：各个参与者的输入都是独立的。
4. 输出传递性：不诚实参与方不能阻止诚实的参与方获得输出。
5. 公平性：所有参与者都应该获得各自的输出。

安全多方计算的参与方或者称为参与者，是指参与协议的各方。在实际情况中，由于某些利益驱使，参与方有可能不按照协议的规则来执行协议，比如输入虚假信息，然后根据得到的输出结果进而猜测其他参与方的输入信息，或者拒绝参与协议。这类不遵守协议规定的行为称为腐败行为。根绝参与方在协议执行过程中的行为是否规范可以将参与方分为三种：诚实参与方、半诚实参与方和恶意参与方。

1. 诚实参与方：是指参与方在协议的执行过程中没有发生腐败行为，完全按照协议的规则和步骤来完成协议，对自己的所有输入输出信息进行保密。
2. 半诚实参与方：半诚实参与方虽然完全按照协议的规则来执行协议，但是可能将自己的输入信息或者输出结果泄露给攻击者。
3. 恶意参与方：恶意参与方按照攻击者的角度来执行协议，可能会泄露自己所有的输入输出，改变输入，篡改中间信息，甚至终止协议。

安全多方计算攻击方或者称为攻击者，是指企图破坏协议正确性或者安全性的人。根据攻击方的计算能力差异可以将其分为拥有无限计算能力攻击者和拥有多项式计算能力的攻击者。一个安全多方计算协议如果对于拥有无限计算能力的攻击者而言是安全的，则称此协议是信息论安全的或者无条件安全的；如果仅对于拥有多项式计算能力的攻击者而言是安全的，则称此安全多方计算协议是密码学安全的或者条件安全的。根据攻击方控制参与方的方式可以将攻击者分为只可以收集参与方信息的窃听攻击者和可以随意更改参与方行为的Byzantine攻击者。根据网络同步状态分类，攻击者可以分为同步网络下的攻击者和异步网络下的攻击者。根据攻击方选择腐败对象的自适应性，攻击者可以分为非适应性或者静态的攻击者和适应的攻击者，前者在协议执行之前就选定了腐败的参与方，而后者可以在协议执行的任何时候腐败任何可以腐败的参与者，也可以选择释放已经腐败了的某些攻击者，以获取更大的攻击能力，有时把具有此类能力的适应攻击者称为移动的攻击者[15,16]。

如果安全多方计算的参与方都是半诚实的或诚实的，那么就称这个模型为半诚实模型。如果参与方中有恶意参与者，称为恶意模型。在恶意模型中，参与者可以不按照协议的步骤进行，如果想要得到正确的计算结果，就需要使用更多的密码协议或者技术。而在半诚实模型下研究安全多方计算要相对容易很多。

安全多方计算的通信模型分为同步和异步模型两种。

1. 同步通信模型：在参与方的环境中存在一个时钟服务器，来统一协议环境中的吋钟。在同一个时钟周期开始吋，所有参与方同时接收发给消息，然后执行那个其他操作，然后发出自己的消息。
2. 异步通信模型：在参与方的环境中不存在时钟服务器，消息的发送直到接受要经过不同的时钟周期。接受到的消息可能存在延迟和乱序的情况。

安全多方计算信道模式分为点到点信道和广播信道。

1. 点到点信道：每对参与奔之问有单独的信道传输。如果攻击者不能窃听两个诚实参与方通信，称为安全信道；如果攻击者能窃听但是不能改信息，称为不安全信道；若攻击者不仅可以窃听而且还能够随意修改信息，称为未经认证信道。
2. 广播信道：每个参与方发送的消息，其他参与方都能收到，同一时刻只能有一个参与方发送消息。广播信道保证了诚实参与方在同一轮中接收到相同的消息，无论这个消息是诚实参与方发送的或齐是被腐败的参与方替换成其他消息。如果任意的两参与方之间存在安全的点到点信道，那么这个广播信道是安全的，称为安全的广播信道。

下面分别介绍安全多方计算的攻击者模型包括半诚实攻击者模型、恶意攻击者模型和隐蔽攻击者模型,其中前两者只介绍两方的情形，多方情形可自然推广或，后者介绍一般的多方情形，两方的形式可直接由此描述出来。

1. 半诚实（或被动）攻击者模型

定义2.10（半诚实模型下的理想模型）在理想模型下，存在诚实的可信第三方。协议执行过程描述如下：

输入：第一方拥有,第二方拥有。

参与方向发送输入：参与者总是将自己的输入发送给。

回应参与方：收到输入对后,计算。发送给第一方;发送给第二方。

输出：诚实的参与方总是输出他从那里得到的信息,半诚实的参与方可能会输出任意一个关于他的初始输入以及他从那里得到的信息的多项式函数。

定义2.11（半诚实模型下的安全计算）记是一个概率多项式时间函数，其中。是计算函数的一个双方协议。

1) 用多项式规模的电路对来表示理想模式下的攻击者。称在现实半诚实模型下是容许的,如果至少存在一个,使（意味着是诚实的）。在理想模式中,当输入对是时,将模式下的联合计算记为。

2) 用多项式规模的电路对来表示现实模式下的攻击者。称在现实半诚实模型下是容许的,如果至少存在一个,使（这里，是能够由推导出来的信息，意味着是诚实的）。在现实模式中,当输入对是时，将模式下协议的联合计算记为。

称协议在半诚实攻击者模型下安全计算,如果存在一个多项式时间可计算的变换,能将现实模式下容许的多项式规模电路对转换为理想模式下容许的多项式规模电路对，使得如下关系式成立，这里。

1. 恶意（或主动）攻击者模型

此模型认为，被攻击者腐败的参与方可以任意的偏离协议,它可以拒绝参与协议、更换自己的本地输入、甚至可以在中途中止协议。

定义2.12（恶意模型下的理想模型）在理想模型下，存在诚实的可信第三方。协议执行过程描述如下：

输入：每方拥有一个输入，用记为。

参与方向发送输入：诚实的参与者总是将输入z发送给。恶意方根据自己的输入，要么拒绝参加协议（什么也不发送给）,要么发送某个给（用长度相同的伪输入替换）。

回应第一个参与方：若收到一个输入对，则他计算并将发送给第一个参与方；否则（即只收到一个输入）就向两个参与方都发送一个终止符（协议终止）。

回应第二个参与方：如果第一个参与方是恶意的,他有可能根据自己的输入以及的回应，决定是否让终止执行。若第一个要求终止执行,则发送终止符土给第二个参与方（协议终止）。否则将发送给第二个参与方。

输出：诚实的参与方总是输出他从那里得到的信息。恶意的参与方可能会输出任意一个关于他的初始输入以及他从那里得到的信息的多项式函数。

定义2.13（恶意模型下的安全计算）设是一个概率多项式时间函数，其中。是计算函数的一个双方协议。

1) 用多项式规模的电路来表示理想模式下的攻击者。称在理想恶意模型下是容许的,如果至少存在一个，使得且（意味着第方是诚实的）。在理想模式中,当输入对是时,将模式下的联合计算记为并定义为：

1. 当且时，即第二方是诚实的，

 (2‑9)

1. 当且时（即第一方是诚实的），

 ( 2‑10)

2) 用多项式规模的电路对来表示现实模式下的攻击者。称在现实恶意模型下是容许的,如果至少存在一个遵守协议的策略。在现实模式中,当输入对时,将模式下协议的联合执行记为并定义为用和作为输入实际执行协议之后所得到的输出对（受到恶意腐败方的攻击策略的影响）。

称协议在恶意攻击者模型下安全计算,如果存在一个多项式时间可计算的变换,能将现实恶意模型下容许的多项式规模电路对转换为理想恶意模型下容许的多项式规模电路对，使得下关系式成立：

 ( 2‑11)

这里。

1. 隐蔽攻击者模型

隐蔽攻击者模型居于半诚实攻击者模型和恶意攻击者模型之间[17]。在隐蔽攻击者模型下，被腐败的参与者可以进行主动的腐败行为从而不是半诚实的攻击者，但它仅在能以一定的概率保证不被抓到的情况下才实行腐败行为从而不是恶意的攻击者。如果它在实际执行中得到了在理想计算情形下所不能拥有的信息，则称某个攻击者在协议的实际执行中成功欺骗。

取常数，称为遏制因子，它表示实际攻击者的欺骗行为将被诚实的参与方以至少的概率发现。越大，攻击者的欺骗行为被抓住的概率越大，对欺骗行为的威慑作用也就越大。特别的，当时,任何欺骗行为都将被抓住。记协议的所有参与方为。表示被攻击者控制的被腐败方的指标集。

定义2.14（隐蔽模型下的现实模型）记是一个方函数性，这里。的一次实际执行中诚实方和攻击者的输出向量记为，其中是输入向量，是的辅助输入，是被腐败方构成的集合的指标集，是安全参数。

定义2.15（隐蔽模型下的理想模型）在理想模型下，存在诚实的可信第三方。协议执行过程描述如下：

1) 输入：每方拥有一个输入（第i方的输入记为），并假设所有输入的长度都为，即。攻击者拥有辅助输入。

2) 参与者向发送输入：任意诚实方发送其输入给。被攻击者控制的腐败方要么发送其输入给，要么发送一个相同长度的伪输入给。这由攻击者根据所以腐败方的输入和辅助输入来决定。将参与者发送给的所有输入记为。

3) 终止策略：如果某腐败方发送给的输入是，则发送给所有诚实的参与方，协议终止；如果某腐败方发送给的输入是则发送给所有诚实的参与方，协议终止。

4) 企图欺骗策略：如果某腐败方发送给的输入是则

1. 以的概率发送给攻击者和所有的诚实方；
2. 以的概率发送和所有诚实方的输入给攻击者,并向攻击者索要所有诚实方的输出并将发送给对应的诚实方。

至此，协议终止。如果没有一个是、、和，则协议继续下列步骤：

5) 回应攻击者：计算，并对所有的，发送给攻击者给攻击者。

6) 回应诚实方：接收到可信方发来的输出后，攻击者发送（对某个）或者continue给。如果收到continue，则发送给对应的诚实方；如果收到（对某个）则它发送给所有的诚实方。

7) 输出：每个诚实方总输出从那里发来的输出,而腐败方没有输出，攻击者输出有关被腐败方的输入和从可信第三方那里接收到的所有信息的任意（概率多项式时间可计算的）函数的函数值。

在如上描述的理想协议的一次执行中,将所有诚实方和攻击者的输出向量记为，其中是输入向量, 是的辅助输入, 是被腐败方构成的集合的指标集, 是安全参数。

定义2.16（隐蔽模式下的安全计算）设，和如上描述。称协议在-遏制的隐蔽攻击者模型下安全计算，如果对实际模式下的任意非一致概率多项式时间攻击者,都存在一个理想模式下的非一致概率多项式时间攻击者,使得对任意、任意均衡的向量和任意辅助输入，下式都成立：

 ( 2‑12)

这里，称向量是均衡向量如果的各分量的长度都相同,即且满足。

### 2.2.2多方安全求和协议

安全总和计算是安全和模型中的一个典型例子，其中，全部参与方都期望得到各部分数据输入的总和，以此来使输入的安全性得到保证。安全求和针对各行各业也有广阔的应用场景，例如，现在有某省的几家通信运营商，如果想得到该省的移动用户总数，但是各家运营商的数据库都不愿意被透露。再比如某人都想得知他所在家族的家产，也会存在上述问题。而安全求和针对上述问题可以得到较好的解决。

安全和协议的描述如下：假设个参与方分别持有隐私数据，。参与方想通过协作计算，在不暴露自己隐私数据的情况下，求得所有参与方的数据和。

Clifton等人提出的安全和协议[19]采用随机化方法计算总和。各参与方被放在一个单向的环内。选择一方作为协议发起方，通过选择一个随机数开始计算，并添加自己的数据。然后总和被传送给下一方。下一方只是将数据添加到收到的金额中，然后将此新金额发送给下一方。重复该过程直到协议发起者接收到所有数据和随机数的总和。由于随机数只对协议发起方是已知的，因此它从总和中减去随机数并将结果通知所有各方。例如，，，，四个参与方执行安全和计算，分别携带数据10，8，7，15。协议的发起方选择一个随机数Random=5并且将这个随机数和它所携带的数据之和发送给。这个综合加上的数据8后，得到一个新的总和发送给，如此反复执行此过程直到接收到45。由于只有知道随机数的值，所以其使用总和减去随机数得到最后结果并发送给各个参与方。该协议正常工作的前提是所有参与方都诚实地遵守协议中规定的每一步，并且除计算结果之外不会尝试了解其他信息。方案的优点在于求和方法简单，通信复杂度与计算复杂度较低，但是不能抵抗相邻两方共谋，如果两个相邻的参与方通过恶意合作来窃取中间方的数据，那么将有可能泄露隐私。



图2-2 随机化方法计算总和的多方安全协议实例

Boneh等人[20]令每个参与者将各自输入的数据拆分为份，再分别发送给其他参与者，针对发送给参与者的份数进行计算其总和并广播，之后每个参与者需再一次地分别计算自身从其他参与者处得到的数据总和。这个方法为共谋问题有效地提供了一个解决方案，很明显地达到了抵抗小于方共谋的策略，然而效率却下降了。

Bhaduri等人[21]在前者基础上，也针对共谋问题的安全求和协议进行了有效地改善，它的核心思想依然是采用[17]中将各参与方依次排成环，再求和的方式。文章还做出假设，在共谋者的求和协议开始前，会通过采用随机的方式向其它参与者发出共谋邀请。其中的诚实参与方会依照自身所得到的共谋邀请来预算出一个正整数k，进而将自身得到的数据分割成k份；与此相反，恶意参与方会接受其邀请，从而加入共谋组。该协议需要不断迭代k次，这样求和协议方能结束。Rashid等人[22]进一步探索安全求和，其最核心的方法也仍然是环形和，在求和的过程中使用不同类型的技术手段将共谋的成功率显著下降。

### 2.2.3多方安全求积协议

#### 2.2.3.1 安全两方求积协议

问题描述：Alice有一个秘密实数，Bob有另一个秘密实数。Alice需要计算获得值，此使得，这里是由Bob秘密随机生成的，且满足以下安全条件：

1. Alice不能从输出结果中推导出的值，也不能从中间计算结果得到任何有关的秘密信息；
2. Bob不能得到的值，即不能单独得到的值，也不能得到任何有关的秘密信息。

下面给出具体的安全两方求积协议[23]，其核心思想是两方均产生个随机数，调用次茫然传输协议，因为茫然传输协议的计算复杂度是线性的，所以该协议的计算复杂度为。假定Alice有一个私密实数，Bob有另一个私密实数。

输入：Alice隐私输入，Bob隐私输入。

输出：Alice得到，Bob得到，且满足。

步骤1：Alice和Bob事先约定两个参数和，使得足够大以至步加法是计算上不可行的；

步骤2：Alice产生个随机数，使得；

步骤3：Bob产生个随机数，使得；

步骤4：对于每一个，Alice和Bob执行下面的子过程；

1. Alice产生私密随机数，；
2. Alice随机生成个实数并将其发送给Bob，且满足；
3. Bob计算；
4. 使用茫然传输协议，Alice取得，而Bob不知道是哪一个；

步骤5：Alice计算最终输出。

安全两方求积协议可推广到多方，假定有个参与方：，每一方都有一个秘密实数；执行协议后，每个参与方均得到一个，满足，且满足以下安全条件：

1. 不能从中单独获得累乘积的值，也不能推导出其他参与方的任何私密信息；
2. 不能从中推导出其他参与方的输出。

文献[24]给出基于同态加密技术的安全多方乘积协议，在通信代价和执行效率上具有明显的优势。

#### 2.2.3.2多方安全点积协议

安全点积协议是解决安全多方计算问题的关键技术之一，它是在不泄露双方向量的情况下，秘密地计算两个向量的点积（内积）。安全点积协议广泛应用于保护隐私的各类协作计算。例如，由于许多数据挖掘算法在很大程度上基于向量和矩阵运算，且大多数数据挖掘（如关联规则挖掘）问题在一定程度上可以被约化为向量点积的计算，因此保护隐私的数据挖掘严重依赖于安全点积协议。

参与方Alice和Bob和分别持有维向量，他们执行协议计算（两向量的点积），协议结束后，Bob获得结果，其中仅由Alice知道。Bob由无法推断出和任意，同样，Alice也无法由推断和的值。要求双方的私有输入对对方保密，并且没有一方能独享计算结果。

文献[25]首次提出点积协议，并分别利用茫然传输和置换协议设计了协议1与协议3这两个解决方案。它们几乎没有信息泄漏，安全性较高，但是协议复杂度比较高，协议不高效，因此适用于安全性要求比较高的场合。文中协议1的通讯复杂度是（是向量维数），计算消耗需要次茫然传输；协议3的通讯复杂度是，计算代价需要次加密，次解密。当很大时，协议3比协议1高效得多。[26]使用了一个半诚实第三方商业服务器，它给参与方分发随机向量和随机数，使得参与方在交互计算过程中隐藏了自己的私有向量，从而实现安全点积。由于第三方的使用，该方案的效率很高，通讯复杂度只有，计算复杂度是。文中进一步指出，为了降低通讯复杂度，商业服务器只需将随机数种子分发给各参与方，由它们自行根据种子来产生随机向量。这样，通讯复杂度就优化至，非常逼近两方点积的最优通讯消耗（点积协议的最好通信代价是指在不需要保护隐私下进行计算所需的通讯成本）。但是，[27]指出这个方案是不安全的。[26]还提出了一个使用可逆矩阵实现安全点积的方案，这个方案引用得比较广泛。

文献[28]使用线性代数技巧提出了一个十分有效与足够安全的点积协议，并用实验结果进行相关验证。该方案优点是若进行多次点积运算，且若计算中某个参与方的私有向量保持不变，由于交互计算过程中有些中间值可反复利用，因而总计算量有所降低，但是[29]认为这个方案不安全。[30]主要技术手段是使用随机数的线性组合来伪装向量元素，尔后通过一些计算来消除随机数对结果的影响。文中指出，这个方法适用于大规模数据场合。对于一些特殊的输入，例如当隐私向量的维数比较小，或者数据的取值范围比较小的时候，协议就不安全。这个方案同时也是某些实用点积协议的代表，它们都在不同程度上泄露了部分输入信息（如向量各分量的总和），以牺牲安全性为代价，从而获得比较好的通信复杂度和时间复杂度。[31]的点积协议基于同态加密并利用，通过密文运算完成明文乘法运算，它分析了两种点积协议方案，说明它们是不安全的，继而提出了一个基于同态加密的可证明的安全点积协议方案，提高了方案的效率，使其可用在大规模数据集上。[32]研究的是布尔点积协议（即向量分量的取值非0即1），在信息论模型下提出了一个点积协议，但是协议只是假设存在半诚实的攻击行为，并且假设如果存在半诚实的第三方，那么提出的协议可以抵抗任何数目的被腐败的参与者。[33]利用系数矩阵转置提出了点积协议，并将其用于隐私保护的数据挖掘中，使用它计算关联集等等。[29]使用置换协议作为子协议，提出了一个新的安全点积协议，同时该文还分析了几种常用的其它安全点积协议。[31]提出了一个基于一个可信的初始化服务器及互不信任的两方之间，模整数的内积协议，并证明该协议在UC下是安全的。

### 2.2.4隐私保护的集合运算协议

在安全多方协议中，隐私保护的集合运算协议是其中较热的研究方向之一，其定义为：个参与方隐私的提供各自的输入，经过相应的集合运算，运算的结果不会泄漏出各自参与方的相关输入。

集合问题的研究是安全多方计算研究的一个重要方面，其中集合相交问题已经在多个方面有了重要的应用，例如，保密的数据挖掘[34]、数据外包[35]、医疗敏感数据分析[36]、个人财产数据及其他隐私数据的安全共享[37]等，例如：

1) 目前有甲、乙两超市，它们想通过合作的方式来挖掘出消费者的消费走势。即他们想得知顾客在甲超市购买产品的同时，在乙超市购买的可能性。然而两个超市为了保障自身的利益与其消费者的隐私权益，都不愿将自身数据库里的详细信息透露给双方。针对这样的场景，就会涉及隐私保护的集合交集问题。

2) 几个城市想监管本省的交通安全，如果将这几个城市一起，通过交通监管可以判断某较有嫌疑的车辆曾经出现在哪几个城市，进而进行推断。针对这类情况，就会涉及到隐私保护的集合并集问题。

设有个参与者分别拥有秘密集合，，现在他们需要计算这些集合的交集。保密的集合相交需要各个参与者知道他们集合的交集，但对其他参与者的集合没有任何信息。最简单的保密求集合相交问题是社会主义百万富翁问题，即有两个元素参与者，他们各自只有一个元素，然后比较这两个元素是否相等。

现有的集合相交问题的解决方案多是基于一些密码学算法，如Kissner[38]提出的方案基于Paillier[39]同态加密算法和多项式求值，Soled[30]提出的方案基于ElGamal[41]同态加密算法和多项式求值，文献[42]给出的方案基于Naor-Reingold的类随机函数，文献[43]给出的方案基于可交换的密码算法，文献[44]给出的方案基于秘密共享[45]等。Freedman[46]利用多项式不经意求值并借助于Paillier同态加密算法和平衡哈希函数提出的一种高效求集合交集的解决方案，其安全性是语义安全的。该方案主要适用于两个参与者的情况，当扩展到多个参与者的情况时其计算复杂性和通信复杂性都非常高；而且该方案只适用于参与者的集合是一个无限大集合子集的情况(其论文中明确指出参与者集合取自于指数大域)，但实际应用中这种情况并不多。

密码学算法与协议的安全性可以分为以下两类：计算性安全和信息论安全(也可以称之为无条件安全)。

针对于前者，无论攻击者是否拥有多项式时间计算本领，该算法或协议总是安全的。语义安全是计算性安全的代表之一，它的安全性基于一些计算困难的环境而作出假设。如较为熟悉的大整数因式分解假设、离散对数假设等。而以上这些困难性假设仅仅针对攻击者的计算能力是有限的前提下才可成立，如果攻击者的计算能力是无限的，那这些假设就不再成立。

针对于后者，即无条件安全算法或协议的安全性，是基于信息论安全的基础上的。这类算法或协议使敌手获取不到攻击途径来得到足够信息，因此可以大大保障菜安全。且无论敌手的计算能力是有否无限，该算法或协议总是安全的。因此，在量子计算的环境下，信息论安全的算法与协议相对更为可靠。

下面将介绍不同模型下的隐私保护的集合运算协议的理论研究，其中将分别考虑攻击者为半诚实模型,攻击者为恶意模型和攻击者为隐蔽模型。这些隐私保护的集合运算协议包括隐私保护的集合交集协议、隐私保护的集合并集协议和隐私保护判定集合是否相交协议。

#### 2.2.4.1 密码学通信模型下的隐私保护集合运算协议

在密码学的通信模型下,隐私保护的集合运算协议大都利用了同态加密体制或者同态门限解密的公钥加密体制。

首先介绍半诚实攻击者模型下的隐私保护的集合运算协议，现已有的大多有关于半诚实攻击者模型下都是利用将根转化为多项式的技术,然后将多项式的系数进行加密后发送,通过选取具有同态性质的加密体制,将加密后的数据进行相关的运算。

文献[46]首次提出了隐私保护的集合交集协议,在[47]中利用了双线性映射提出了一种的隐私保护的集合交集协议,双线性映射是映射，（其中和拥有相同的阶）满足如下性质：1），都有。2）如果是的生成元,那么也是的生成元。3）对于,都有高效的算法计算。

利用集合元素转化为多项式的根以及同态加密技术,每个参与方将接受到的加密的数利用随机数混淆,然后每个参与方进行合作得到一个混淆的多项式,然后,将集合的元素一一带入。

1. **协议1 半诚实攻击者模型下隐私保护的集合交集协议**

输入：有个参与方,其中个恶意参与方可以进行合谋,每个参与方都拥有一个私有的含有个元素的集合,记为,每个参与方都拥有公钥和各自相应的私钥的分享,利用-门限加密体制。

输出：每个参与方都知道，其中采用的加密体制为ElGamal的加密体制。

1) 每个参与方月首先将自己集合中的元素转化为多项式的根,计算，然后将广播给其它的参与方。

对于

[1]随机选择，并且计算。

[2]将所有的相加得到，并将广播出去。

2) 然后每个参与方将所有的相加得到

 (2‑13)

3) 每个参与方将带入到中，然后通过与其他参与方的合作进行解密,如果解密的结果为1,那么很显然就属于交集中的元素。

文献[48]利用可交换的公钥加密体制给出了攻击者为半诚实下的隐私保护的集合并集协议。[49]利用了同态门限加密体制，利用了根转化为多项式的技术，提出了半诚实模型下的隐私保护的集合并集协议。

1. **协议2 半诚实攻击者模型下的隐私保护的集合并集协议**

1）参与方共同协商一个门限同态加密算法,记这个公钥加密算法为。

2）每个参与方利用将集合元素转化为多项式的技术，把粘贴在布告栏上。

3）参与方将粘贴在布告栏上。

4）参与方对于每个粘贴到布告栏。

5) 参与方对于每个，粘贴到布告栏。

6）每个参与方选择一个非零的随机数。

7）每个参与方将粘贴到布告栏，这里参与方是等参与方粘贴后在进行粘贴。

8）参与方将粘贴的元祖进行混淆，然后共同协作进行解密，如果解密出来的值为0，那么参与方进行下一个元组的解密,否则记，，则将放入到并集中。

文献[50]利用同态加密体制，给出了攻击者为半诚实模型的集合是否相交的协议,然后在此基础上利用了零知识证明和UC承诺协议给出了攻击者为恶意模型的集合是否相交的协议。

判定多方集合是否相交的问题的定义：每个参与方都拥有1个全集为。的子集,他们希望能够判断,并且不用泄漏自己的信息给对方。

1. **协议3 半诚实攻击者模型下的集合是否相交的协议**

每个参与方都拥有集合，其中和。每个数据库是以长度为的比特串形式表示，表示为，如果，那么，其中。

1）参与方利用标准的ElGamal同态加密体制执行秘钥产生算法获得。然后发送给公钥和加密的，如下所示：，其中。

2）从那接收到和后，计算，并且将其发送给下一个参与方。

3）参与方从那接收到 ，随机选择一个不为零的随机数，计算，符号代表同态乘运算。然后将数值发送给参与方。

4）参与方利用私钥对接收到的数值进行解密,如果解密出来的值为0，那么可以判定集合是不相交的,否则可以判定集合相交。

显然：，如果对于某个，全部的，那么，否则结果为0，所以当所有乘法的和为0，那么可以判定。

接下来介绍在恶意攻击者模型下的隐私保护的集合运算协议,对于隐私保护的集合运算协议的研究大都是在半诚实攻击者模型的基础上进行一些可验证方案,比如乘法的正确性验证、对于明文的确定性的验证和多项式计算的验证等等。

1. **协议4 恶意攻击者模型下隐私保护的集合交集的协议**

输入：有个参与方，其中个恶意参与方可以进行合谋，每个参与方都拥有一个私有的含有个元素的集合，记为，每个参与方都拥有公钥和各自相应的私钥的分享，利用-门限加密体制，

输出：每个参与方都知道

1）每个参与方首先将自己集合中的元素转化为多项式的根，计算：

，然后将广播给其它的参与方，并且需要证明知道的明文的验证。

2）对于，

[1]随机选择，并且计算，然后将进行广播，并且证明乘法的正确性的验证。

[2]将所有的相加得到，并将广播出去。

3）每个参与方将所有的相加得到：



4）每个参与方将带入到中，并且证明多项式计算的正确性验证，然后对于，通过与其他参与方的合作进行解密，如果解密的结果为1，那么就属于交集中的元素。

1. **协议5 恶意攻击者模型下的隐私保护的集合并集协议**

1）参与方共同协商一个门限同态加密算法，记这个公钥加密算法为。

2）每个参与方将集合中的数值进行零知识证明，并且将粘贴到布告栏。

3）每个参与方利用将集合元素转化为多项式的技术，把粘贴在布告栏上。对于这个多项式的构造需要进行零知识的证明。

4） 参与方将粘贴在布告栏上，对于多项式的乘法需要进行零知识证明。

5）参与方对于每个，粘贴到布告栏，对于粘贴的数值进行零知识证明。

6） 参与方对于每一个，粘贴到布告栏，对于多项式的计算以及乘法进行零知识证明。

7) 每个参与方选择一个非零的随机数，并且对于这个进行零知识证明。

8）每个参与方将粘贴到布告栏，对于粘贴在布告栏上面的值进行零知识证明。这里参与方是等参与方粘贴后再进行粘贴。

9）参与方将粘贴的元组进行混淆，然后共同协作进行解密,如果解密出来的值为0，那么参与方进行下一个元组的解密，否则记，则将放入到并集中。

#### 2.2.4.2 信息论通信模型下的隐私保护的集合运算协议

文献[51]首次提出了信息论模型下的隐私保护集合交集协议：有个参与方相互之间都有安全信道连接,每个参与方都是有无限的计算能力,并且一个半诚实的攻击者能够攻击小于一半数目的参与方。为了考虑问题方便,设每个参与方都有一个含有个元素的集合，表示集合中的第个元素，第个参与方想计算出他们之间的交集。此外,不能泄露出任何的信息给其他的参与方。

1. **协议6 半诚实攻击者模型下的隐私保护的集合交集运算协议**

1） 对于，每个参与方利用集合转化为多项式技术得到一个度为的多项式。。然后在个参与方之间分享多项式系数。

2）对于，每个参与方随机选择个次数为1的多项式，，然后在个参与方间分享多项式系数，从而每个参与方通过以下计算得道新多项式的分享：

 (2‑14)

3）对于，每个参与方通过计算，

得到系数的分享，然后利用秘密重分享技术，将分享转化为分享。

4）对于，每个参与方得到的分享。得到：

 (2‑15)

设，然后通过以下计算得到系数的分享：

 (2‑16)

最后，每个参与方将自己系数的分享发送给其他参与方。

5）每个参与方用恢复出的个系数，然后将自己集合中的元素带入到中，如果有个元素使得，那么。

在恶意攻击者模型下,恶意攻击者能够故意发送错误的秘密分享给别的参与方,所以当发送错误的秘密分享给别的参与方的时候,假设被腐败的参与方人数的时候,可以通过可验证方案防止接收到错误的秘密分享。这种可验证方案可以通过零知识证明或者比特承诺方案。

## 2.3 小结

本章立足传统隐私保护方法，介绍了以K-匿名为代表的匿名化技术和安全多方计算技术，讨论了相关技术的定义、模型以及其提供的隐私保护性能。学界在对这两种技术表示肯定的同时，也发现了不少其存在的问题。

例如以K-匿名为基础的匿名模型应对的假设需要确定攻击者背景知识，同一数据集针对不同的攻击者背景假设，需要不同的模型来应对，除此之外，匿名模型的隐私保护程度并不明显，相关参数并不能表达隐私泄露的概率；现有的安全多方计算问题的解决方案大多是基于某种密码学原语，其计算复杂性和通信复杂性都比较高，并且均有各自的适用范围，没有设计出普遍适用的协议，因此，一种高效的普遍适用的方案有待进一步深入研究。

## 2.4 参考文献

1. LATANYA SWEENEY. k-ANONYMITY: A MODEL FOR PROTECTING PRIVACY[J]. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002, 10(05):557-570.
2. 吴英杰, 唐庆明, 倪巍伟,等. 基于取整划分函数的k匿名算法[J]. 软件学报, 2012, 23(8):2138-2148.
3. Tao Y, Chen H, Xiao X, et al. ANGEL: Enhancing the Utility of Generalization for Privacy Preserving Publication[J]. IEEE Transactions on Knowledge & Data Engineering, 2009, 21(7):1073-1087.
4. Machanavajjhala A, Kifer D, Gehrke J. L -diversity: Privacy beyond k -anonymity[C]// International Conference on Data Engineering. IEEE, 2006:24-24.
5. Ninghui Li, Tiancheng Li, Suresh Venkatasubramanian. t-Closeness: Privacy Beyond k-Anonymity and l-Diversity[C]// IEEE, International Conference on Data Engineering. IEEE, 2007:106-115.
6. Raymond Chi-Wing Wong, Jiuyong Li, Ada Wai-Chee Fu,等. (α, k)-anonymity: an enhanced k-anonymity model for privacy preserving data publishing[C]// ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2006:754-759.
7. R.J. Bayardo, Rakesh Agrawal. Data privacy through optimal k-anonymization[C]// International Conference on Data Engineering, 2005. ICDE 2005. Proceedings. IEEE, 2005:217-228.
8. Iyengar V S. Transforming Data to Satisfy Privacy Constraints[J]. Kdd, 2002:279-288.
9. Lin J. Divergence measures based on the Shannon entropy[J]. Information Theory IEEE Transactions on, 1991, 37(1):145-151.
10. Kullback S, Leibler R A. On Information and Sufficiency[J]. Annals of Mathematical Statistics, 1951, 22(1):79-86.
11. Yao A C. Protocols for secure computation[J]. Foundations of Computer Science Annual Symposium on, 1982:160-164.
12. Goldreich O, Micali S, Wigderson A. How to play ANY mental game[C]// Nineteenth ACM Symposium on Theory of Computing. ACM, 1987:218-229.
13. Goldreich O, Srcure Multi-party Computatiom (manuscript version 1.3)
14. 张华. 安全多方计算及其应用研究[D]. 西安电子科技大学, 2005.
15. Ostrovsky R, Yung M. How to Withstand MobileVtrns Attacks[J]. Proceedings of the Tenth Annual Acm Symposium on Principles of Distributed Computing, 1991:51-59.
16. Franklin M, Yung M. Communication complexity of secure computation (extended abstract)[C]// ACM Symposium on Theory of Computing, May 4-6, 1992, Victoria, British Columbia, Canada. DBLP, 1992:699-710.
17. Aumann Y, Lindell Y. Security against covert adversaries: E cient protocols for realistic adversaries[J]. Journal of Cryptology, 2010, 23(2):281-343.
18. [Secure Multi-party Differential Privacy](http://papers.nips.cc/paper/by-source-2015-1212" \t "_blank)
19. Clifton C, Kantarcioglu M, Vaidya J, et al. Tools for privacy preserving distributed data mining[J]. Acm Sigkdd Explorations Newsletter, 2002, 4(2):28-34.
20. Dan B, Franklin M. Efficient generation of shared RSA keys[J]. Journal of the Acm, 1997, 48(4):702-722.
21. Bhaduri K, Das K, Kargupta H. Peer-to-Peer Data Mining, Privacy Issues, and Games[C]// International Conference on Autonomous Intelligent Systems: Agents and Data Mining. Springer-Verlag, 2007:1-10.
22. Sheikh R, Kumar B, Mishra D K. Changing Neighbors k Secure Sum Protocol for Secure Multi Party Computation[J]. 2010, 7(1).
23. 石润华, 仲红. 一种新型匿名门限秘密共享方案[J]. 山东大学学报(理学版), 2012, 47(11):31-39.
24. 夏超, 仲红, 石润华. 基于同态加密技术的安全多方乘积协议[J]. 计算机工程与应用, 2015, 51(1):76-80.
25. Atallah M J, Du W. Secure Multi-party Computational Geometry[C]// The Workshop on Algorithms and Data Structures. Springer, Berlin, Heidelberg, 2001:165-179.
26. Du W, Zhan Z. A practical approach to solve Secure Multi-party Computation problems[C]// The Workshop on New Security Paradigms. ACM, 2002:127-135.
27. Du W, Han Y S, Chen S. Privacy-Preserving Multivariate Statistical Analysis: Linear Regression And Classification[C]// Siam International Conference on Data Mining. 2004:222--233.
28. Ioannidis I, Grama A, Atallah M. A Secure Protocol for Computing Dot-Products in Clustered and Distributed Environments[C]// International Conference on Parallel Processing. IEEE Computer Society, 2002:379.
29. Du W, Zhan Z. Building decision tree classifier on private data[C]// IEEE International Conference on Privacy, Security and Data Mining. Australian Computer Society, Inc. 2002.
30. Vaidya, J., Clifton, C. Zhu, M. Privacy preserving Data Mining, Vol.19 of Advances in Information Security, Springer, New York, NY,USA. 2006.
31. Goethals B, Laur S, Lipmaa H, et al. On Private Scalar Product Computation For Privacy-Preserving Data Mining[C]// International Conference on Information Security and Cryptology. Springer-Verlag, 2004:104-120.
32. Rajasekaran S. Fast cryptographic multi-party protocols for computing boolean scalar products with applications to privacy-preserving association rule mining in vertically partitioned data[C]// International Conference on Data Warehousing and Knowledge Discovery. Springer-Verlag, 2007:418-427.
33. Vaidya J. Privacy preserving association rule mining in vertically partitioned data[J]. Journal of Computer Applications, 2006, 26(1):639-644.
34. Liang M. Data Mining: Concepts, Models, Methods, and Algorithms[J]. Iie Transactions, 2003, 36(5):495-496.
35. Liu F, Ng W K, Zhang W, et al. Encrypted Set Intersection Protocol for Outsourced Datasets[C]// IEEE International Conference on Cloud Engineering. IEEE Computer Society, 2014:135-140.
36. Drosatos G, Efraimidis P S. Privacy-preserving statistical analysis on ubiquitous health data[C]// International Conference on Trust, Privacy and Security in Digital Business. Springer-Verlag, 2011:24-36.
37. Cristofaro E D, Tsudik G. Practical Private Set Intersection Protocols with Linear Complexity[C]// Financial Cryptography and Data Security, International Conference, FC 2010, Tenerife, Canary Islands, January 25-28, 2010, Revised Selected Papers. DBLP, 2010:143-159.
38. Kissner L, Song D. Privacy-Preserving Set Operations[C]// International Cryptology Conference. Springer, Berlin, Heidelberg, 2005:241-257.
39. Paillier P. Public-Key Cryptosystems Based on Composite Degree Residuosity Classes[C]// International Conference on Theory and Application of Cryptographic Techniques. Springer-Verlag, 1999:223-238.
40. Dachmansoled D, Malkin T, Raykova M, et al. Efficient Robust Private Set Intersection[J]. Lecture Notes in Computer Science, 2009, 2(4):289-303.
41. Elgamal T. A public key cryptosystem and a signature scheme based on discrete logarithms[J]. IEEE Transactions on Information Theory, 1985, 31(4):469-472.
42. Hazay C, Lindell Y. Efficient Protocols for Set Intersection and Pattern Matching with Security Against Malicious and Covert Adversaries[J]. Journal of Cryptology, 2010, 23(3):422-456.
43. Li S, Dou J, Jia X. Secure Two-Party Computation for Set Intersection Problem[J]. Journal of Xian Jiaotong University, 2006, 40(10):1091-1093.
44. Li R, Wu C. An Unconditionally Secure Protocol for Multi-Party Set Intersection[C]// International Conference on Applied Cryptography and Network Security. Springer-Verlag, 2007:226-236.
45. Shamir A. How to share a secret. Communications of the ACM, 1979,22(11): 612-613.
46. Freedman M J, Nissim K, Pinkas B. Efficient Private Matching and Set Intersection[J]. 2004, 3027:1--19.
47. Sang Y, Shen H. Privacy preserving set intersection based on bilinear groups[C]// Australasian Conference on Computer Science. Australian Computer Society, Inc. 2008:47-54.
48. Segre A M, Wildenberg A, Vieland V, et al. Privacy-Preserving data set union[C]// Cenex-Sdc Project International Conference on Privacy in Statistical Databases. Springer-Verlag, 2006:266-276.
49. Frikken K. Privacy-Preserving Set Union[M]// Applied Cryptography and Network Security. Springer Berlin Heidelberg, 2007:237-252.
50. [40] Kiayias A, Mitrofanova A. Testing Disjointness of Private Datasets[J]. Lecture Notes in Computer Science, 2005, 3570:109-124.
51. Li R, Wu C. An Unconditionally Secure Protocol for Multi-Party Set Intersection[C]// International Conference on Applied Cryptography and Network Security. Springer-Verlag, 2007:226-236.

# 第三章 -差分隐私（章国政，边锦）

差分隐私（Differential Privacy，DP）[1]保护模型的思想源自于一个很朴素的观察：当数据集中包含个体Alice时，设对数据集进行任意查询操作（例如计数、求和、平均值、中位数或其它范围查询等）所得到的结果为，如果将Alice的信息从中删除后进行查询得到的结果仍然为，则可以认为，Alice的信息并没有因为被包含在数据集中而产生额外的风险。差分隐私保护就是要保证任一个体在数据集中或者不在数据集中时，对最终发布的查询结果几乎没有影响。

## 3.1 -差分隐私基本知识（章国政）

### 3.1.1 -差分隐私定义

对于一个有限域，为中的元素，从中抽样所得的集合组成的数据集，其样本量为，属性的个数为维度。

对数据集的各种映射函数被定义为查询(Query)，用 来表示一组查询，算法对查询 的结果进行处理，使之满足隐私保护算法的条件，此过程称为隐私保护机制。

设数据集和，具有相同的属性结构，两者的对称差记作，表示 中记录的数量．若，则称 和为邻近数据集(Adjacent Dataset)。

定义3.1 (-差分隐私) 又称精确差分隐私[2]，设有随机算法，为所有可能的输出构成的集合。对于任意两个邻近数据集和以及的任何子集，若算法满足：

 ( 3‑1 )

称算法提供－差分隐私保护，其中参数称为隐私保护预算[3]。

### 3.1.2 隐私预算

隐私保护预算的取值是用来控制算法在两个邻近数据集上输出的概率的比值，它一般体现了随机算法所能提供的隐私保护程度。在实际应用中，通常取很小的值，例如0.01，0.1，或者ln2，ln3等，取值越小，隐私保护程度就越高。当等于0时，保护水平达到最高，此时对于任意邻近数据集，算法都将输出两个概率分布完全相同的结果，这些结果也不能反映任何关于数据集的有用的信息。

在设计算法的时候要合理分配隐私预算，常用的分配策略包括：线性分配、指数分配、自实用性分配、均匀分配等。但常用的分配策略中，存在参数上界选取问题，鉴于此，何贤芒等人[4]针对查询函数下的攻击模型，设计了选取参数的上界公式。计算查询函数的敏感度，并根据设定的查询的容错区间长度及攻击者成功概率，计算出参数选取上界如下所示：

 ( 3‑2 )

在实际应用场景下，隐私预算作为差分隐私保护技术中的主要参数，根据隐私保护要求，如何结合具体需求来达到输出结果的安全性与可用性的平衡，恰当的选取参数*ε*，是使用者不可避免且非常重要的问题。

### 3.1.3 有界/无界-差分隐私

对于任何邻近数据集和，通过删除或替换数据集中任何单个记录的数据， (Differential Privacy)确保对于任何输出都可以以相似的概率得到相同的输出结果。使用的一个重要前提条件是数据和必须是邻近数据集，由前文我们知道邻近数据集是两个数据集之间只有一个元素不同。在这里可以把这两个邻近数据集之间一个元素不同理解为只有单个记录(或元组)不同，但这里仍然有两种选择，这就引出有界和无界的概念[5]。

定义（无界差分隐私[6]） 给定一对数据集和，其中数据集能通过从数据集中添加或移除其中一个元素而得到，对于任意集合，如果下面等式成立：



则随机算法满足无界差分隐私。

定义（有界差分隐私[7]） 给定一对数据集和，其中数据集能通过改变数据集中某一个元素而得到，对于任意集合，如果下面等式成立：



则随机算法满足有界差分隐私。

在有界中，数据集和是邻近数据集的前提是数据集能通过从数据集中用一个新的元素来替换其中一个元素而得到。在无界中，数据集和是邻近数据集的前提是数据集能通过从数据集中添加或移除其中一个元素而得到。当使用有界时，元素数量不同的两个数据集被认为不是邻近数据集。即有界的所有邻近数据集和的大小固定且相等，而无界情况下的邻近数据集大小并不一定相等。因此，对于任意邻近数据集和的大小固定且相等的情况，在有界中，任意的满足，而在无界中，任意的不满足。

通常情况下，选择使用有界或选择无界都是可以的。 但有界有时候也有一定的局限性。更具体地说，有界不是在并行组合下构成的（而无界是的）。在证明算法满足时，经常使用并行组合属性。任何满足无界的算法同时满足有界。因为用一个新的元素来替数据集中原有的元素可以通过先移除一个元素，然后在添加一个元素来得到。

## 3.2 敏感度（章国政）

### 3.2.1 全局敏感度

定义3.2 (全局敏感度[1])设有函数，输入为一数据集，输出为一维实数向量。对于任意的邻近数据集和，

 ( 3‑3 )

称为函数的全局敏感度(Global Sensitivity)。其中，是和之间的1-阶范数距离。

差分隐私机制通过添加噪声来扰乱输出结果从而达到隐私保护的目的，加入噪声过多会影响结果的可用性，过少则无法提供足够的安全保障，敏感度是决定加入噪声量大小的关键参数。

函数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．一些函数具有较小的全局敏感度（例如计数函数，其全局敏感度为1），因此只需加入少量噪声即可掩盖因一个记录被删除对查询结果所产生的影响，实现差分隐私保护．但对于某些函数而言，例如平均值、中位数等函数，则往往具有较大的全局敏感度。

### 3.2.2 局部敏感度

当全局敏感度较大时，必须在函数输出中添加足够大的噪声才能保证隐私安全，导致数据可用性较差．针对这个问题，Nissim等人定义了局部敏感度。

定义3.3 (局部敏感度[8]) 设有函数，输入为数据集，输出为一维实数向量。对于给定数据集和它的任意邻近数据集，则

 ( 3‑4 )

称为函数在上的局部敏感度(Local Sensitivity)。

局部敏感度由函数及给定数据集中的具体数据共同决定。由于利用了数据集的数据分布特征，局部敏感度通常要比全局敏感度小得多。以求中位数函数为例，其局部敏感度为。 局部敏感度与全局敏感度之间的关系可以表示为： 。

### 3.2.3 平滑敏感度（4.1.2 平滑敏感度）

由于局部敏感度在一定程度上体现了数据集的数据分布特征，如果直接应用局部敏感度来计算噪声量则会泄露数据集中的敏感信息。因此，局部敏感度的平滑上界(Smooth Upper Bound)被用来与局部敏感度一起确定噪声量的大小。

定义3.4 (平滑上界[8]) 给定数据集及其任意邻近数据集，函数的局部敏感度为。对于，若函数满足且，则称 为函数的局部敏感度的－平滑上界。

所有满足这一定义的函数都可被定义为平滑上界，将局部敏感度代入此函数中则可得到平滑敏感度(Smooth Sensitivity)，进而用于计算噪声大小。

定义3.5 (平滑敏感度[8]) 给定数据集及，函数

 ( 3‑5 )

称为函数的－平滑敏感度，其中。

### 3.2.4 其他敏感度

设计数查询函数：， 定义敏感度为其中一个重要参数，决定了一些查询结果的准确率。下面给出敏感度的定义：

定义3.6 (敏感度[9]) 设函数，则函数的敏感度为

 ( 3‑6 )

函数中的敏感度可以用来衡量，在最坏的情况下，单个个体的数据对函数的影响程度。

高斯噪音代替拉普拉斯噪音，一阶距离笵数求解的敏感度不再适用，因此，定义敏感度。

定义3.7 (敏感度[9])设函数,则函数的敏感度为

 ( 3‑7 )

## 3.3 噪声机制（章国政）

差分隐私的核心在于其随机算法的设计，在实践中为了使一个算法满足差分隐私保护的要求，对不同的问题有不同的实现方法，这些实现方法称为“机制”。拉普拉斯机制[10](Laplace Mechanism)与指数机制[11](Exponential Mechanism)是两种最基础的差分隐私保护实现机制。其中，机制适用于对数值型结果的保护，指数机制则适用于非数值型结果。其他的还有矩阵机制、k-norm机制、递归机制[13]、函数机制[14]等。

### 3.3.1 拉普拉斯机制

机制通过向确切的查询结果中加入服从分布的随机噪声来实现－差分隐私保护。记位置参数0，尺度参数为b的分布为Lap(b)，其概率密度函数为：

 ( 3-8 )

定义3.5 (机制) 给定数据集D，设有函数，其敏感度为，那么随机算法提供-差分隐私保护，其中为随机噪声，服从尺度参数的分布。

证：若要证明加入噪音后的随机算法满足-差分隐私，只需证明算法满足( 3-9 )式即可。

 ( 3-9 )

设数据集和上函数表示为：

 ( 3-10 )

其敏感度为：

 ( 3-11 )

假设全部为0，那么：

 ( 3-12 )

记输出向量，则有：

 （3-13 )

现在只需要论证：，就有：

 ( 3-14 )

由绝对值不等式：，知，则：

 ( 3-15 )

于是有：

 ( 3-16 )

因此满足差分隐私得证。

### 3.3.2 指数机制

由于拉普拉斯机制只能处理数值型的数据，有其自身的限制，对于非数值型的数据 (如分类操作中的选择问题等) 拉普拉斯机制无法实现差分隐私。为了实现对非数值型数据进行差分隐私处理，Mcsherry等[11]人提出了指数机制。

设查询函数的输出域为，域中的每个值为一实体对象。在指数机制的情况下，将函数 称为输出值为的可用性函数(实用性函数或打分函数)，用来评估输出值的优劣程度。

定义3.6 (指数机制) 设随机算法**的输入为数据集，输出为一实体对象， 为可用性函数，为函数的敏感度。若算法以正比于的概率从中选择并输出，则算法**提供-差分隐私。

证：设邻近数据集和，输出，且。

 ( 3-17 )

指数机制以最大的实用性输出时，实用性函数的敏感度为

 ( 3-18 )

则指数机制以正比于在意一个，其损失参数近似为：

 ( 3-19 )

则(3-17)式变换为：

 ( 3-20 )

由：

 ( 3-21 )

得：

 ( 3-22 )

由和的对称性可知：

 ( 3-23 )

则(3-17)式最终为：

 ( 3-24 )

因此可证指数机制满足差分隐私。

### 3.3.3 矩阵机制

矩阵机制（Matrix Mechanism）是一种在线性计数查询模型上进行差分隐私数据统计发布的方法, 并且具有成熟的理论框架。它通过将查询集转换成负载矩阵，然后寻找最优策略矩阵 实现差分隐私下线性查询的优化。其中查询集是一组线性查询的集合，满足。每条线性查询表示如下：

 ( 3-25 )

其中为数据向量， 为该查询与分量的权重。负载矩阵由每组线性查询的权重组成，并满足：

 ( 3-26 )

矩阵机制是通过寻找策略矩阵的方式求解问题，对于随机算法满足以下公式：

 ( 3-27 )

则称算法满足-差分隐私保护。

其中为阶满秩策略矩阵，为的广义逆矩阵，为阶负载矩阵，为服从拉普拉斯分布长度为p的列向量，查询矩阵敏感度，为隐私预算。

证明：公式( 3-27 )可以表示为：

 ( 3-28 )

其中，由Laplace机制可以证明其满足-差分隐私，因此算法满足-差分隐私。

矩阵机制的均方误差可以表示为：

 ( 3-29 )

矩阵机制是通过直接寻找策略矩阵的方式求解问题，该做法的效率和优化效果都不够理想。低秩矩阵机制[12]（Low-Rank Mechanism）是采用解负载矩阵的斱法来寻找优化策略。该机制将负载矩阵分解成两个矩阵、，即，其中为低秩矩阵下的最优策略矩阵。通过矩阵将原始数据进行及变换，并在变换的结果上添加独立噪声，最后再通过还原矩阵转换正最终的查询结果。对于随机算法满足以下公式：

 ( 3-30 )

其中敏感度与策略矩阵的列笵数相同，即。低秩矩阵的均方误差由下式求得：

 ( 3-31 )

### 3.3.4 K-norm噪声机制

机制适用于线性查询。每个查询计算中元素的一个线性组合，假设把线性查询表达成一个行向量，，并把个查询作为行组成一个矩阵，查询序列的正确结果就是。定义表示维空间中的范数单位球体：，那么的查询核可以表示为，查询核捕捉了查询序列的所有关系。

定义3-（机制）定义，机制输出为，其中加性噪音具有以下概率密度函数：

 ( 3-32 )

其中表示以为单位球体的笵数，也就是说：

 ( 3-33 )

机制可以通过以下两个步骤完成：

1. 从分布采样获得正实数

2. 从中均勾采样获得，并输出。

如果且是一个随机的矩阵，那么机制的误差以极高的概率不超过其它任何差分隐私机制的常数倍。

### 3.3.5 递归机制

在实际应用中，存在许多复杂查询，例如SQL查询结果上的聚合以及子图计数，查询结果却很可能有着过大甚至无限的全局敏感度。Laplace机制虽然较为简单却被广泛应用，但却要求查询结果对个体参与者的全局敏感度必须是有限而且较小的，并不能够处理如此复杂的查询。

递归机制[13]（Recursive Mechanism）是一个基于经验敏感度的查询机制，它可以处理跟个体参与者之间有着复杂关系的查询，并且允许查询有无限大的全局敏感度，并能用于处理任意的子图计数查询并实现节点差分隐私。递归机制基于两条特殊的序列：递归序列和界限序列。下面是对两条序列满足的条件做简单介绍。

定义3-（递归序列）序列作为在上的函数，被称为递归序列当且仅当满足一下条件：

1.对所有的都有

2.递归单调性：对所有相邻的和对任意的都有

 ( 3-34 )

定义（界限序列）对于递归序列和实数，序列作为在 上的函数，被称为的界限序列当且仅当满足以下条件：

1.是递归序列

2.对所有的，所有的和，都有：

 ( 3-35 )

如果，简单的称是的界限序列。

递归机制满足差分隐私，其框架由三个步骤构成：

1. 对单调查询，构造一条递归序列和的一条界限序列，使得对所有的都有都有。

2. 基于，找到一个量，使得近似于或的经验敏感度，并且具有较低的全局敏感度，然后对加入乘性噪音，获得，且满足差分隐私。

3. 基于，找到一个量，使得近似于正确结果，并且具有全局敏感度，然后对加入Laplace噪音，获得，且满足差分隐私。

其中对与敏感数据库和参数，，可以计算求得：

， ( 3-35 )

### 3.3.6 函数机制

函数机制[14]（Function Mechanism，FM）应用于回归模型中。FM是Laplace机制的扩展，它不直接将噪声加入回归结果中，而是通过向目标函数的多项式参数中添加噪音得到扰动目标函数，从而达到隐私保护的目的。

回归分析是对数据集上的目标函数进行最优化处理，最终发布参数模型。揭示了函数和数据集之间的信息，直接发布会造成隐私的泄露，因此，结合差分隐私中的拉普拉斯噪音，提出了函数机制。假设为一个维的权重向量，定义为所有乘积的集合，数据集上的目标函数可以用多项式形式表示为：

 ( 3-36 )

其中，表示变量的系数，。表示维度为时所有乘积的集合，其计算公式如下：

 ( 3-37 )

函数机制添加噪音所需的敏感度，取决于目标函数本身的敏感性，对于任意邻近数据集和，的敏感性可以表示为：

 ( 3-38 )

利用的敏感性及隐私预算，为多项式的每个系数添加Laplace噪音，从而得到扰动后的目标函数及。FM算法流程如下：

|  |
| --- |
| 算法：函数机制（Functional Mechanism）  输入：数据集、目标函数、隐私预算 |
| 1：设置 |
| 2：for |
| 3：      while |
| 4： |
| 5： end while |
| 6： end for |
| 7：让 |
| 8：计算 |
| 9:   输出 |

## 3.4 组合定理（章国政）

一个复杂的隐私保护问题通常需要多次应用差分隐私保护算法才能得以解决。在这种情况下，为了保证整个过程的隐私保护水平控制在给定的预算之内，需要合理地将全部预算分配到整个算法的各个步骤中。这时可以利用隐私保护算法的两个组合性质：顺序组合性、并行组合性。

### 3.4.1 顺序组合定理

性质1 (顺序组合性[15]) 设有算法，其隐私保护预算分别为，那么对于同一数据集，由这些算法构成的组合算法提供差分隐私保护。

证：设算法满足-差分隐私，由定义得：，

 ( 3-39 )

设算法表示为所有算法组合而成的大算法，该算法的输出记为。由于任意两个之间的随机过程相互独立，因而下式成立：

 ( 3-40 )

则算法满足差分隐私保护。

### 3.4.2 并行组合定理

性质2 (并行组合性[15]) 设有算法，其隐私保护预算分别为，那么对于不相交的数据集，由这些算法构成的组合算法提供差分隐私保护。

证：设算法满足-差分隐私，由定义得： ，

 ( 3-41 )

同样设算法表示算法组合而成的算法，该算法的输出记为。由于任意两个之间的随机过程相互独立，因此

 ( 3-42 )

由此可得：

 ( 3-43 )

从而表明并行组合下的差分隐私算法满足-差分隐私。

定义差分隐私保护算法所保护数据集合D中的元素X在集合R上，，

因此有，令为的一种划分，满足 ， 。

又因为 ，得到：

 ( 3-44 )

为元素的定义域，有，，因此下式成立

 ( 3-45 )

即并行组合下的差分隐私算法满足-差分隐私。

## 3.5 数据效用与隐私保护的平衡（边锦）

由于差分隐私通过添加随机噪声的方式来对数据进行隐私保护，所以衡量一个满足差分隐私的算法的效果是一件十分重要的事情。

隐私保护程度表示所设计的方法在满足差分隐私的条件下，确保隐私不被泄露，隐私保护的严密性高。隐私预算代表隐私保护程度，当越小的时候，隐私保护程度越高，在算法中一旦耗尽，那么差分隐私将被破坏，保护算法也失去了自身的意义。

数据可用性表示通过差分隐私保护技术处理过后的数据，数据研究人员在进行数据挖掘时的数据集的效用情况，这是用来衡量发布数据集的质量的一种方式。通常采用评估指标来衡量数据的效用性。其中查询误差衡量方式包括：相对误差、绝对误差、欧式距离以及误差的方差。挖掘算法的评估指标主要有：误差率、准确率、灵敏度、召回率、准确率、对数损失函数、ROC曲线、AUC曲线等。

表3-1 隐私效用权衡方面的工作

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 数据源 | 建模技术 | 隐私内容 | 匿名化技术 | 隐私权衡 |
| 交易 | 表格 | 身份  准标识符  敏感属性 | 泛化  抑制  K-匿名  随机化  差分隐私 | （1）比基于泛化和抑制相比，琐碎的消毒处理提供的相同的效用和更好的准确性。  （2）回报风险权衡，有效边界  （3）率失真理论，效用-隐私权衡区域  （4）在线服务的效用理论分析 |
| 集合 | 表格 | 身份  准标识符  敏感属性 | 同上 | 同上 |
| 社交网络 | 图 | 节点标识  链接强度  链接属性 | K-匿名，泛化，  输出扰动  差分隐私  边权重匿名化 | 最优隐私设置的潜在特质模型（项目反应理论） |
| 位置 | 网格 | 敏感的位置 | 干扰  数据转换  基于PIR的位置隐私 | 隐私保护，实用程序最大化复杂事件处理 |
| 上下文流（移动传感器数据） | 马尔科夫链  隐马尔科夫模型  动态贝叶斯网络 | 敏感的上下文语境 | 概率检查  模拟检查 |  |

## 3.6 计数查询实例（边锦）

计数查询的一般应用形式为“在某数据库中有多少元素是满足属性的？”，之后将一次又一次的返回到这些查询中，有时是以纯粹的形式，有时是以分数的形式（元素在数据库中的比例是多少？），有时与权重相关（线性查询），有时以稍微复杂一点的形式（例如：应用函数，对数据库中的每个元素结果进行求和）。计数是一个极其强大的原始工具，它能够捕获统计查询模型、许多标准数据挖掘任务、基本统计信息中所有可以学习的东西，计数查询的敏感度为1，因为添加或删除一条记录最多改变计数查询结果为1。根据精确差分定理可知，对于计数查询来说实现差分隐私添加的噪声为，也就是说加入的噪声服从，预期的失真、误差与加噪方法、隐私参数有关，与数据库的大小无关。

假设计数查询为一个固定但任意的列表，可以被视为矢量值查询，假设没有与该集合相关的查询信息，那么其矢量查询最坏情况的敏感度为,若要实现差分隐私，对每个查询的正确答案应加入噪声规模为。

目前计数查询在实际生活中医疗方面的应用相对广泛，例如“有多少病人吸烟已经超过15年了？”这就是一个计数查询。请考虑以下均值计数查询案例：“对于一个吸烟超过15年的肺癌患者，平均每天抽多少包烟？”要求属性的平均值满足一定条件的记录。一个有趣的问题是均值是否可以通过向计数添加噪声来计算。 当然，我们实际上并没有输出计数。只输出加噪后的均值。考虑下面的均值加噪算法：

:输入一维数据集，隐私预算,数据范围

 ( 3-46 )

以上算法满足差分隐私吗？对于邻近数据集，，不失一般性，假定比多一个数据值，让，然后，令是服从随机噪声的随机变量，敏感度为，对于任意输出值，我们有： ( 3-47 )

因为是无界的，可以任意大，因此该算法不满足,然而，我们知道真正的均值在 之间，如果我们严格限制输出结果在范围内，我们需确定，这样我们才能讨论算法是否满足精确差分。为此我们提出改进后的均值加噪算法，运用机制对上述均值计数查询案例中的和分别进行加噪处理，对算法加噪结果进行讨论分析。算法如下：

:输入一维数据集，隐私预算,数据范围

 ( 3-48 )

该算法满足差分隐私吗？对于邻近数据集，，不失一般性，假定比多一个数据值，让，然后，令是服从随机噪声的随机变量，且令是服从随机噪声的随机变量，敏感度为，当时，输出结果如算法所示，输出结果。当时，对于任意输出值,且，并设，则对于（3-49）式和（3-50）式，（3-51）式成立。

 ( 3-49 )

 ( 3-50 )

 ( 3-51 )

因为输出值，为数据集比数据集多的一个元素，且，所以，所以可得，最终可得该算法满足差分隐私。

由上述分析可知，机制为差分隐私中常用的加噪机制，但最终结果的输出取决于给函数加噪的方法，以及加噪参数的多少，这些因素都会影响函数的输出结果，以及是否满足精确差分。

表3-2为美国约翰霍普金斯医院胸外科患肺癌患者的资料，根据主治医生对病人的调查了解，抽烟超过15年的肺癌患者每天吸烟情况如下表所示：

表3-2 医疗数据集示例

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 序号 | 姓名 | 性别 | 年龄 | 烟龄 | 吸烟包数 |
| 1 | Tom | 男 | 55 | 30 | 3 |
| 2 | Jack | 男 | 52 | 25 | 2 |
| 3 | Alice | 女 | 48 | 18 | 2 |
| 4 | Edward | 男 | 60 | 35 | 4 |
| 5 | Bella | 女 | 57 | 30 | 5 |
| 6 | Marcus | 男 | 50 | 26 | 3 |
| 7 | Hopper | 男 | 48 | 23 | 2 |
| 8 | Jacob | 男 | 49 | 28 | 3 |
| 9 | Rose | 女 | 45 | 25 | 2 |
| 10 | Girbert | 男 | 42 | 20 | 1 |

根据上述表格所列各位病患的吸烟情况，针对这一科室烟龄较长的患者每天吸烟的包数，计算各位患者在这一科室平均每天的吸烟包数？

该数据集可任意选取邻近数据集均可，只需保证两个邻近数据集相差记录个数为1即可，均可应用于上述两种算法中，=抽烟包数之和，=患者个数，=5,=1,敏感度为，可直接应用机制进行加噪处理，在均值加噪算法中加噪后均值可直接由计算得出，改进后的均值加噪就相对麻烦一些，它是对求和函数、计数个数分别加噪，求比值，计算公式为

 ( 3-52 )

对两种算法运用上述数据集进行均值加噪处理，计算出的结果一定与隐私参数的取值有关，隐私参数的取值不同，得出的加噪均值与原有均值有相对偏差。无论偏差的范围有多大，必须满足加噪后输出的结果在内，这样该算法才一定能满足差分隐私，若输出结果不在内，如上述均值加噪算法所证明的将不满足差分隐私。对算法加噪处理的目的是针对统计输出的随机化方式，使得攻击者无法得到查询结果间的差异，从而能保证数据集中每个个体的安全。

## 3.7 小结

本章通过 -差分隐私，简单介绍了与其相关的基本定义，在此基础之上提出有界/无界的概念和不同类型的敏感度，如全局敏感度、局部敏感度、平滑敏感度等；针对差分隐私要引用的噪声机制，本章简单介绍了拉普拉斯机制、指数机制、矩阵机制、K-norm噪声机制、递归机制和函数机制。并针对差分隐私的基本性质，提出了顺序组合定理和并行组合定理。

此外，立足于差分隐私的特性，本小结也提出了如何衡量数据效用与隐私保护的平衡，并举出了计数查询实例，为下一章节差分隐私的扩展进行了概念支撑。

## 3.8 参考文献

1. Dwork C. Differential privacy[C]// International Colloquium on Automata, Languages, and Programming. Springer, Berlin, Heidelberg, 2006:1-12.
2. Dwork C. A firm foundation for private data analysis[M]. ACM, 2011.
3. Haeberlen A, Pierce B C, Narayan A. Differential privacy under fire[C]// Usenix Conference on Security. USENIX Association, 2011:33-33.
4. 何贤芒,王晓阳,陈华辉，董一鸿. 差分隐私保护参数的选取研究[J]. 通信学报,2015,36(12):124-130.
5. Li N, Min L, Su D, et al. Differential Privacy:From Theory to Practice[J]. Synthesis Lectures on Information Security Privacy & Trust, 2016, 8(4):138.
6. Dwork C. Differential privacy[C]// International Colloquium on Automata, Languages, and Programming. Springer, Berlin, Heidelberg, 2006:1-12.
7. Dwork C, Mcsherry F, Nissim K. Calibrating Noise to Sensitivity in Private Data Analysis[C]// Conference on Theory of Cryptography. Springer-Verlag, 2006:265-284.
8. 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.
9. Dwork C, Roth A. The Algorithmic Foundations of Differential Privacy[M]. Now Publishers Inc. 2014.
10. Dwork C, Mcsherry F, Nissim K. Calibrating noise to sensitivity in private data analysis[C]// Conference on Theory of Cryptography. Springer-Verlag, 2006:265-284.
11. Mcsherry F, Talwar K. Mechanism Design via Differential Privacy[C]// IEEE Symposium on Foundations of Computer Science. IEEE Computer Society, 2007:94-103.
12. Yuan G, Zhang Z, Winslett M, et al. Low-rank mechanism[J]. Proceedings of the Vldb Endowment, 2012, 5(11):1352-1363.
13. Chen S, Zhou S. Recursive mechanism:towards node differential privacy and unrestricted joins[C]// 2013:653-664.
14. Zhang Z, Zhang Z, Yang Y, et al. Functional mechanism: regression analysis under differential privacy[J]. Proceedings of the Vldb Endowment, 2012, 5(11):1364-1375.
15. Mcsherry F D. Privacy integrated queries: an extensible platform for privacy-preserving data analysis.[J]. Communications of the Acm, 2010, 53(9):89-97.

# 第四章 差分隐私扩展（陈竑毓，王寒，章国政）

差分隐私根据不同场景又做出了进一步的扩展，其中比较常见有近似差分隐私，集中差分隐私和本地差分隐私。下面将对这三种进行逐一介绍。

## 4.1 近似差分隐私（陈竑毓）

-差分隐私是一种精确差分隐私，使得两个邻近数据集能够在输出结果上满足-不可区分。但考虑到一方面精确差分隐私过于严格的特性导致产生的噪声过大，容易使数据失真；另外，在特定的混合场景下，不需要对每个属性都进行差分隐私保护。因此Dwork[1]等人又进一步提出-差分隐私，一种近似差分隐私，使得两个邻近数据集在输出结果上满足-近似-不可区分性。

本节首先介绍-差分隐私的定义；其次介绍实现-差分隐私的高斯机制，在这基础之上介绍-差分隐私的近似最优机制；最后简单论述-差分隐私的组合定理。

### 4.1.1 近似差分隐私定义

定义4.1 -差分隐私：设有随机算法，为所有可能的输出结果构成的集合。对于任意两个邻近数据集和，若算法满足

 ( 4-1 )

则称算法提供-差分隐私保护。其中参数称为隐私保护预算，是单个记录影响输出结果的上限，是隐私损失超出限度的概率，即近似差分的解释为，对于所有的邻近数据库和，以至少的概率将隐私损失控制在内。

### 4.1.2 高斯机制

高斯机制是实现-差分隐私的基本机制，同Laplace机制一样，是适用于数值型查询结果的差分隐私保护机制，实现原理是在查询结果上添加满足高斯分布的噪声以达到扰动作用，其定义如下：

定义4.2 给定数据集D上的查询函数，令，为服从独立同分布的高斯随机变量，则在输出结果添加高斯随机变量的算法：提供-差分隐私保护。

对于高斯噪声，其分布如图4-1所示，由图中可知，形状参数越小，图形越瘦高，数据分布越集中，由逆累计分布函数计算得到的输出值也就越小，即参数和参数越大，随机算法中加入的噪声越小。此时与-差分隐私基本要求一致，即隐私预算越小，隐私保护程度越高，数据扰动程度越大。

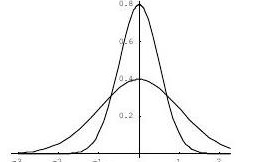


图 4-1 高斯分布概率密度与随机变量的关系

高斯机制与Laplace机制都是适用于数值型查询结果的隐私保护方法，在隐私预算消耗完全之前，添加服从Laplace分布或者高斯分布的不同噪声，就说该查询方法满足对应的差分隐私。表4-1及图4-2简单对高斯机制与Laplace机制相关内容做了对比。

表 4-1 高斯机制与Laplace机制对比

|  |  |  |
| --- | --- | --- |
|  | Laplace机制 | 高斯机制 |
| 表达式 |  |  |
| 敏感度 | L1敏感度 | L2敏感度 |
| 噪声分布 |  |  |
| 联系 | Laplace分布=高斯分布+指数分布 |  |

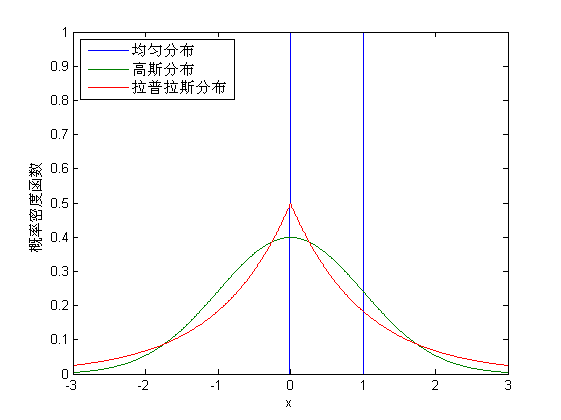


图 4-2 高斯分布与Laplace分布图像对比

前文给出了近似差分隐私和高斯机制的定义，下文将首先证明高斯机制满足-差分隐私的定义，为在算法中应用高斯机制提供理论基础。证明思路分为三部分：第一步先给出满足-差分隐私保护的条件，即证明在何种条件下能以的概率将隐私损失限制在内；第二步需要证明当高斯噪声正态分布的形状参数，即满足什么条件时，可以满足第一步中的条件；最后证明高斯机制能够满足-差分隐私。

证明：

证明过程先证一维查询函数下的高斯机制，后拓展至多维查询函数。

第一步，给出满足-差分隐私保护的条件。

设为查询函数，其中是实数集，此时有查询函数输出结果差值，即若，则。

已知服从高斯分布的连续型随机变量的概率密度为，而高斯噪声服从位置参数为0，形状参数为的高斯分布，则查询函数在邻近数据集和输出相同结果的概率比值为，其中和分别表示需要添加的高斯噪声，表示高斯噪声输出。

此时实际隐私损失计算公式为：

 ( 4-2 )

当实际隐私损失时，求得添加的噪声，可知，当高斯噪声时，隐私损失被限定在内。则根据近似差分定义以及邻近数据库对称原理，得到。进一步化简得到，（1）。

上述证明过程表明要满足差分隐私，则高斯噪声超过的概率要严格小于。

第二步，证明当高斯噪声正态分布的形状参数，即满足什么条件时，可以做到超过的概率要严格小于。

根据标准正态分布的尾概率不等式得到高斯噪声输出不等式（2）。当时，有（3）。

化简（3）式得到：

 ( 4-3 )

将代入以上不等式，可得：

 ( 4-4)

假设，因为和都是随着增加而增加，且前者增加幅度小于后者，这里不求的最小临界值，只求能够满足不等式的值，所以让满足非负，单纯计算。令（随着c的增大而增大），首先求非负情况下的范围。将代入可得：

 ( 4-5)

再计算，代入得：

 ( 4-6)

因为且，则。要让，只需保证。

 ( 4-7)

，则。

上述证明表明当，即时，高斯噪声超过的概率严格小于。

第三步，证明当时，定义满足差分隐私。

令高斯噪声的输出结果为，其中

 ( 4-8)

同理，令查询函数输出结果为，其中

 ( 4-9)

可证得：

 ( 4-10)

证明高斯机制在一维实数查询函数下满足-差分隐私。

证明了高斯机制在特定情况的查询函数之后，证明其在通用查询函数，即多维实数查询下的隐私性，只需要将维的输出结果映射到1维即可。

设****，其中是实数集，此时查询函数输出结果差值为，即若，则。对于固定邻近数据库和，使得，则此时输出相同结果造成的实际隐私损失的计算公式为：

**** ( 4-11)

其中，和分别为需要添加的维高斯噪声，假设。

化简公式，得到：。

建立维空间坐标系，令，其中，。

假设不失一般性情况下，平行于，则得到下列表达式：

 ( 4-12)

 ( 4-13)

因为平行于，所以，所以：



此时，隐私损失表达式近似为，确保严格小于等于，将维结果映射为1维变量，接下去证明过程同一维查询函数。

上述证明过程证明了在数值型输出上添加高斯噪声，即可满足-差分隐私，为在算法上实现近似差分隐私保护提供了理论基础，为大数据背景下的隐私保护提供了另一种解决方案。

### 4.1.3 近似最优机制

上一小节介绍了实现-差分隐私的一个基本机制-高斯机制，本节针对-差分隐私的近似最优机制问题[2]进行展开讨论。

-差分隐私的近似最优机制问题是指在差分隐私保护程度确定的情况下，使得数据效用最大化。这之中需要考虑的内容包括效用最大的定义、查询函数的类型、隐私预算的值，只有综合考虑几部分要素，才能得出对应场景下的近似最优机制。

在下文介绍差分隐私下的近似最优机制时，为方便理解，首先假设查询函数为单值整数查询类型，再扩展到多维实数查询函数类型。介绍顺序为先在单值整数查询函数背景下，对损失最小/效用最大进行形式化定义，之后对单值整数查询函数在特殊场景-差分隐私下的近似最优机制问题进行相关分析，在这之上总结一般参数差分隐私下的最优机制，最后将单值查询函数扩展到多维实数查询函数进行分析。

1. 损失最小/效用最大的形式化定义（单值整数查询前提）

实现差分隐私保护的基本方法是在查询函数的输出结果上添加噪声。令表示查询函数在数据集上的输出，差分隐私机制即噪声添加机制遵循以下格式，其中表示机制中添加到查询输出的噪声：

 ( 4-14)

在单值整数查询函数中，为了让差分隐私机制的输出结果有效，令 。

用表示添加的噪声的概率分布函数，令表示，则对于一个集合，用表示。

根据-差分隐私的定义公式，首先将关于机制输出结果的差分隐私约束转变到对噪声约束的概率分布上：

 ( 4-15)

其中。

同时根据-差分隐私定义可知对于任意集合有，并且，其中表示查询函数的全局敏感度，则结合关于噪声约束的概率分布公式可得：

 ( 4-16)

假设此时噪声的损失函数为。效用最大化的目标是使得损失最小，也就是使得添加噪声的平均期望最小，即：

 ( 4-17)

为方便理解，规定损失函数是对称的，并且当时，，当时，单调递增。

1. (0,δ)差分隐私下的最优机制

首先考虑-差分隐私在时的特殊情况。-差分隐私要求邻近数据集上的查询输出的条件概率分布的总变化率被控制在内。

对任意整数而言，取子集为，并且令，则噪声的概率分布必须满足以下公式：

 ( 4-18)

发现上述表达式存在的特殊场景，以下证明分别针对特殊场景及一般场景两种情况讨论。

1. 当即全局敏感度为1：

化简噪声的概率分布公式，得：。又因为对称损失函数在时单调递增，易知此场景下均匀分布的噪声机制满足最优机制需求。为了方便理解，假设为整数值。

定理4-1：如果，添加噪声的平均期望为：

 ( 4-19)

此时噪声的最优概率分布为：

 (4-20)

上述分析表明在单值整数查询函数、-差分隐私背景下，只要查询函数的全局敏感度，差分隐私隐私保护近似最优机制为均匀噪声机制，且给出其噪声的最优概率分布表达式。

1. 当即全局敏感度大于等于2：

回顾当前待优化问题及其约束条件：

 (4-21)

既然是一个对称函数，不妨多给定几个假设，用来计算的下界。现在假设噪声的概率分布是对称函数，并且，对于任意，都有。此时，可计算噪声的平均期望的下界为：

 (4-22)

定理4-2：如果应用在-差分隐私框架下的查询机制的代价函数满足公式：

 (4-23)

就有。

定理4-2表明在-差分隐私框架下，只要全局敏感度不为1的查询机制的代价函数满足定义公式，就可以得出其损失的下界。

接下来仍考虑均匀分布的噪声机制，即机制的噪声分布函数满足：

 (4-24)

定理4-3：在均匀概率分布噪声机制下，噪声的平均期望具有上界，为：

 (4-25)

定理4-3表明在-差分隐私框架下，全局敏感度不为1的均匀噪声查询机制其噪声平均期望上界为定义表达式所示。

通过比较噪声平均期望的下界与均匀噪声分布机制下噪声平均期望的上界，能够得出一些有趣的结论。

为了使噪声平均期望下界能被使用，现在假设当前的代价函数为，称之为代价函数。经验证，发现当，代价函数满足定理4-2的前提公式，因此在代价函数背景下，可应用定理4-2。代入可得

 (4-26)

此时噪声平均期望上下界差值为，则差值是一个常数，与隐私参数无关。因噪声平均期望上下界处于同一数量级，且差值为固定值，则均匀噪声机制是代价函数背景下的近似最优机制。

总结上面的分析可知，在-差分隐私框架、单值实数查询机制敏感度大于1、且指定代价函数的背景下，近似最优机制为均匀噪声机制。

总结上面的分析步骤，采用相同方法对后续其他代价函数以及其他查询机制和其他差分隐私约束情景进行分析：

第一步，确定差分隐私框架，当前讨论处于-差分隐私框架假设，可不进行特殊假设，后续针对一般参数-差分隐私框架需要对特定情况进行假设才能正确讨论在特定代价函数下的情况；

第二步，确定查询机制，当前讨论处于单值整数查询机制，后续针对多维实数查询，需要进行更多假设才可以得出正确结论；

第三步，确定代价函数，计算其噪声平均期望的表达式。如果满足类似定理2的前提条件，则计算其下界，之后提出某个噪声机制，利用假设计算其上界。

第四步，通过比较噪声平均期望的上下界进行验证。在某种假设下，差值为固定值或者比较值处于同一量级，则可以证明所提的噪声机制为当前所有假设下的最优机制。

同样步骤针对代价函数展开，经过验证可知当，代价函数满足定理4-2的前提公式，因此在代价函数背景下，可应用定理4-2。代入可得

 (4-27)

发现当，有，则证明此时在-差分隐私框架并且、单值实数查询机制敏感度大于1、且指定代价函数的背景下，近似最优机制为均匀噪声机制。

和代价函数相同，代价函数在当，有。此时在相关前提下，近似最优机制仍是均匀噪声机制。

只要代价函数满足以下公式，就有，对后续证明所提的均匀机制是否为近似最优机制具有很大帮助。

1. 扩展理论

和上述分析方法类似，文献[3]给出了-差分隐私单值整数查询函数的最优机制为均匀噪声机制和离散Laplace机制。-差分隐私多维实数查询函数，即直方图查询的最优机制为均匀噪声机制。

### 4.1.4 组合定理

这一节主要介绍-差分隐私的三个组合定理，分别是-差分隐私下的顺序组合定理[4]，后处理查询类型的组合定理[1]，以及高级组合定理[5]。

三类组合定理是针对应用场景更贴近现实情况而提出的。顺序组合定理，每个查询针对特定数据库，且每个查询之间独立，因而简单对每个查询的隐私预算累计求和；后处理类型的适应场景是当前的查询结果会对后续查询机制造成影响，此时组合查询的隐私预算仍为各查询的累计求和，在下文后处理类型的证明中，每个查询隐私预算相同，具有特殊性，但其组合性质不变；顺序组合对隐私预算累计求和容易导致总隐私预算过大，高级组合则希望使用更一般的组合以及希望降低隐私预算，具体是通过增加损失超限概率的方法来降低隐私预算。

1. 顺序组合定理

设算法提供-差分隐私保护，其中。令组合算法为，则组合算法提供-差分隐私保护。

证明：

算法满足-差分隐私保护，

。

将不等式稍作处理，右侧进行交运算，和的交等于交的和，将隐私损失概率和损失超限概率进行独立运算。

 (4-28)

对于，有

(4-29)

由此计算可得：

(4-30)

上述证明表明算法满足-差分隐私保护。

1. 后处理类型组合定理

令提供-差分隐私保护，对于，使得也提供-差分隐私保护。其中，对于所有给定的，则对于所有邻近数据集D和D’以及，有。

证明：

首先证明当k=2时，上述定理成立。

对于任意，定义，可知。

对于所有的，有，其中表示值域的微小增量上的输出结果。

根据定义，。对不等式右侧进行交运算，可得到下列运算内容。

 (4-31)

因为P是连续函数，可得到下列不等式，此处积分公式含义是累加求和：（计算过程涉及P(T2,T1)=P(T2,s1)\*P(T1)、第三步过程δ和T1事件独立）

 (4-32)

同样，可将定理推广到。

1. 高级组合定理

上述组合定理，一方面是个查询机制都针对同一数据库使用，另一方面个查询机制独立不相关，具有特殊性或查询机制具有相关性，但隐私预算提前给定，会造成隐私预算的浪费。高级组合定理是将应用场景一般化，提出了-fold自适应组合。-fold自适应组合首先根据攻击者信息最大化假设，即用户信息分布于多个数据库，且都可以被攻击者访问以组成特定的生成数据库，攻击者根据前一次的查询结果可以自适应地采用最优机制，并且使用最优生成数据库。

利用顺序组合定理可以证明个机制的顺序使用满足-差分隐私。假设现在需要组合机制满足-差分隐私，那么则需要每个查询机制满足-差分隐私，当过大时，会导致每个机制的隐私预算过小，导致添加噪声过大，会造成输出结果严重失真。文献证明：-fold自适应组合下的组合机制将隐私超限概率放宽，就可以将隐私预算下降到，小于。

定义4-4：对于所有的，个-差分隐私机制满足-fold自适应组合下的-差分隐私。此时，。

给出证明之前，先了解几个相关概念。

1. KL散度（相对熵）：

将取自相同域的两个随机变量和的KL散度定义为：

 (4-33)

可知KL散度用于衡量两个分布的平均离散程度，且结果大于等于0，当且仅当两个分布相同时，KL散度等于0，此外，和的分布相似度越高，KL散度越小。

1. 最大散度：

将取自相同域的两个随机变量和的最大散度定义为：

 (4-34)

和的-近似最大散度被定义为：

 (4-35)

可知最大散度用于衡量两个分布的最大离散程度。

1. 最大散度与差分隐私的联系：

机制满足-差分隐私当且仅当对于邻近数据库和，有和成立；

机制满足-差分隐私当且仅当对于邻近数据库和，有和成立。

1. 统计距离

随机变量和的统计距离被定义为：

 (4-36)

最大散度和和统计距离都是衡量两个分布的最大离散程度，但可以理解成一个是相对离散，一个是绝对离散。

1. 引理：

引理4-1：假设随机变量和满足和，则；

引理4-2（Azuma鞅不等式）：令表示实数随机变量，且对于所有的，有。当所有，有成立。则对于每个，可得到：

 (4-37)

引理4-3：

P1：成立当且仅当存在随机变量，有及成立；

P2：和同时成立当且仅当存在随机变量和，有、及成立。

利用上述的定义和引理可以证明-差分隐私的高级组合定理，证明思路先从的高级组合开始，再拓展到高级组合。

当时，个查询机制的高级组合满足-差分隐私，其中。

实验的输出结果格式应该为，其中表示的抛硬币的内容，表示机制的输出结果。令：

 (4-38)

在这里只要证明，就可以得到

 (4-39)

也就证明了。根据差分隐私和最大散度的联系的第二条，证明组合机制满足-差分隐私。

下面证明：。

首先，令表示实验0中输出结果的视图，令表示实验1中输出结果的视图。则对于固定值，有

 (4-40)

对于每一个前缀，此时。分析的期望值和最大可能值，一旦前缀确定，则第个机制，以及其使用的邻近数据库和，以及相应的参数也被确定。因此比较大小，只需要比较和的大小。

因此，。对于-差分隐私而言，值被限定在范围内，则

 (4-41)

此时，根据引理4-1，求均值。

 (4-42)

此时，取代入Azuma鞅不等式，并且根据前面分析内容可得到，代入得

 (4-43)

则证明的高级组合满足-差分隐私。

将高级组合拓展到通用场景，即证明-差分隐私，其中。

若，使得，并且W和是组合机制的版本。对于每一个前缀，有，则以及。

根据上述证明，可得。又因为，根据引理4-3的P1内容，可得，此时证明组合机制满足-差分隐私。

## 4.2 本地差分隐私（王寒，边锦）

在精确差分隐私下，是以最坏情况下的隐私保护作为衡量的依据。即保证在绝对信任的第三方前提下进行隐私保护。具体来说，数据在客户端没有经过任何处理，直接上传至可信第三方，然后在第三方对数据进行差分隐私干扰，实现用户的个人隐私保护。但是在实际应用过程中，所谓的真正可信的第三方数据收集平台是不存在的，而且即使第三方真的可信，在需要数据交互的情况下，对应的用户就可以通过权限查看到别人的隐私数据。而且由于众包服务的兴起，使得数据的复杂度进一步的提升，仅仅依靠可信第三方的精确差分隐私保护手段的不足以保护个人的隐私信息。因此，提出已不可信第三方为前提的本地差分隐私，即假设第三方完全不可信的情况下进行差分隐私。

### 4.2.1 定义与性质

#### 4.2.1.1本地差分隐私

本地差分隐私[6]的保护模型充分考虑了客户端数据中用户隐私的可能性。在用户数据上传至第三方之前，对每一个客户端中的数据进行隐私保护，然后将隐私保护后的数据上传至第三方。这样，第三方得到的数据为处理后的隐私数据，在对数据进行统计分析的同时，更好的保证用户的隐私信息不被泄露。为了保证数据可用性，本地差分隐私需要在海量数据下进行实现，由于随机噪音的不确定性，需要最终通过聚合大量正向和负向的扰动结果来抵消添加在其中的噪声，从而得到有效的统计结果。

定义（本地差分隐私）已知拥有n个客户端，每个客户端分别对应一条记录，且任意的两条记录不相等。设有随机算法，如果在任意两条记录上的取值相似，且满足以下不等式，称随机算法满足本地差分隐私**错误!未找到引用源。**

 (4-44)

而在文献**错误!未找到引用源。**中，孟晓峰等人也给出了关于本地差分隐私的定义：给定*n*个用户，每个用户对应一条记录，给定一个隐私算法*M*及其定义域*Dom(x)*和值域*Ran(M)*，若算法*M*在任意两条记录*t*和*t`*()上得到相同的输出结果()满足下列不等式，则*M*满足本地差分隐私。

 (4-45)

简单来讲，本地化差分隐私技术通过控制任意两条记录的输出结果的相似性，使第三方服务无法通过数据本身来判断用户的敏感信息。在本地差分隐私中，每个用户将各自的数据进行扰动后，再上传至数据收集者处，而任意两个用户之间并不知晓对方的数据记录，即本地差分隐私中并不存在全局敏感性的概念，因此拉普拉斯机制和指数机制并不适用。目前，本地差分隐私主要采用随机应答技术来确保隐私算法满足本地差分隐私。

#### 4.2.1.2 本地化与中心化差分隐私对比

本地化差分隐私保护技术是在中心化差分隐私保护技术的基础上提出的，其继承了中心化差分隐私保护技术上的组合特性，同时又对其进行了扩展，利用随机响应的扰动机制抵抗不可信的第三方数据收集者带来的隐私攻击。下面对两种技术进行对比。

表4-2 中心化差分隐私与本地化差分隐私对比

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 组合特性 | 第三方是否可信 | 噪声机制 | 应用场景 |
| 中心化差分隐私 | 定义在邻近数据集上 | 可信第三方 | 拉普拉斯机制、指数机制 | 第三方数据收集者 |
| 本地化差分隐私 | 定义在其中的两条记录上 | 不可信第三方 | 主要是随机响应技术 | 每个用户个体 |

#### 4.2.1.3 随机应答

在在社会调查中，人们往往不愿意回答一些敏感性问题，例如是否存在逃税、赌博、吸毒行为等等，直接调查此类问题将导致拒绝或错误回答。Warner在1965年首先提出了随机化应答**错误!未找到引用源。**(Randomized Response，RR)方法，能够在收集敏感信息同时有效保护回答者的隐私。其基本特征是被调查者对所调查的问题采取随机回答的方式，避免在没有任何保护的情况下直接回答敏感性问题，从而最大限度地保证被调查者的隐私不会被泄漏，这样既对被调查者的隐私和秘密加以保护，而且又能获得所需要的真实资料。详情见3.3.3

#### 4.2.1.4组合原理

本地差分隐私由传统差分隐私扩展而来，本身也拥有并行与串行两种组合性质，但是与传统差分隐私相比，主要区别在于传统差分隐私的定义范围在两个兄弟数据集上，而本地差分隐私的组合定理的定义范围在两条记录上。下面给出定义

组合定理之并行性质，给定*n*个算法，并且算法都满足差分隐私。那么在数据集中任意相交两条记录下，算法的组合满足差分隐私保护。

组合定理之并行性质，给定*n*个算法，并且算法都满足差分隐私。那么在数据集中任意不相交两条记录下，算法的组合满足差分隐私保护。

### 4.2.2 随机化应答

Warner在1965年首先提出了随机化应答**错误!未找到引用源。**(Randomized Response，RR)方法，目的是为了解决进行社会研究是涉及敏感问题带来的困难，例如，使用传统的直接问答方式“请问你有没有过盗窃？”，大多数的受访者肯定会回答“没有”，或者拒绝回答，因为没有人愿意直接承认自己曾经盗窃，这就使得研究结果的效用大大下降。

RR方法能够在收集敏感信息同时有效保护回答者的隐私。其基本特征是被调查者对所调查的问题采取随机回答的方式，避免在没有任何保护的情况下直接回答敏感性问题，从而最大限度地保证被调查者的隐私不会被泄漏，这样既对被调查者的隐私和秘密加以保护，而且又能获得所需要的真实资料。

首先介绍一下Warner模型：假设具有敏感属性的一类*A*与不具有敏感性特征的一类，我们的目的是估计具有敏感属性的人在总体中所占的比例。在简单随机有放回抽样下从总体中抽取容量为*n*的样本，然后对该样本进行随机化回答调查．所使用的随机化装置描述如下：外形相同的卡片上分别写有“你有敏感属性吗？”与“你没有敏感属性吗？”而后以预定的比例充分混合后放入一盒子中，由被调查者依简单随机抽样的抽签方式随机抽出一张卡片，再根据卡片上的问题和自己的情况进行回答。只回答“是”或“否”，回答完后仍把卡片放回盒子里。

设是具有敏感属性的人所占的比例，p代表写有问题“你有敏感属性吗？”的卡片所占的比例，则

 (4-46)

假设所有抽样者回答结果真实情况下，有

 (4-47)

设调查结果中有个人回答“是”，有个人回答了“否”，则当时，得到的极大似然估计，然后就可以得到校正后的统计值。

这里举出一个具体的例子，假设总体受访者可以分为A和B两类，研究者若要知道A占人口的比例，设计了一个分为A和B的转盘，转盘上有一个指针，受访者在访问人员不知情的情况下转动转盘，如果指针停在A位置，A类受访者便回答“是”，若停在B位置，则B类受访者回答“否”。由于访问员不知道指针停留位置，无论受访者回答“是”或“否”，他都不会知道受访者的真正类别，因此受访者的隐私得到保障。

经随机抽样搜集足够资料后，研究人员将会直到回答“是”的比例，并且，受访者在转盘上转到A的概率和A在转盘上所占的面积比例相等。假设A类受访者在人口所占的比例为，则回答“是”的概率等于A类人把指针转到A的概率加上B类人把指针转到B的概率：

 (4-48)

假设，，代入算式得，故求得A类人在所有受访者中占25%。因为是研究人员自行设定，若设定为0.5，则因分母为0而不能算出的值；若设定为1，则相当于直接发问；若设定在0.5和1之间，则可以计算出。

最终，根据总人数*n*、回答“是”的人数和扰动概率*p* ,即可得到真实有敏感属性人数的计值，为保证其满足本地化差分隐私，根据定义，隐私预算的大小将设定为:

 (4-49)

随机化应答是基于客户端的一种隐私保护方案，它不依赖于受信任的第三方服务器，并将数据控制权交还给客户端。具体保护的方式为：假设给定一个客户端的值x，客户端执行随机算法向非信任服务器报告扰动值y。有n个客户端，每一个客户端都有一些关于敏感属性X的私有值。不受信的服务器需要学习个体私有数据的某些统计属性，但是客户端不愿意透漏个人信息。为了确保隐私，每一个客户端只向服务器发送的扰动后的。服务器收集来自所有个体的扰动信息，然后通过一些重建程序来恢复统计属性。

### 4.2.3 RAPPOR机制

RAPPOR机制用于从用户的客户端软件收集统计数据，并使用随机应答技术提供强大的隐私保护。RAPPOR的目的是在隐私保护的同时允许通过大量客户收集客户端的值和字符串的统计信息，例如其类别，频率，直方图和其他集合统计信息。

RAPPOR（Randomized Aggregatable Privacy-Preserving Ordinal Response）作为单值频数统计的代表性方法，针对的是离散型数据。主要思路通过对变量中的不同取值进行编码-解码等技术的转换，来满足随机应答对二值变量的要求。即通过对0/1串上每一个位进行随机应答，来实现隐私保护。具体步骤如下：

首先，利用Bloom Filter技术对数据集的每个属性的每个字符串取值进行编码。具体实现为利用每个字符串取值生成一个长度为h的向量B，且B的取值为0或1，同时记录字符串与 Bloom 串的映射关系矩阵，然后利用随机应答技术对向量B的每一个位进行扰动，得到永久性随机响应(Permanent Randomized Response, PRR)的结果向量B`。扰动的方式按照以下公式进行，其中表示概率的取值：

 (4-50)

然后，再对向量B`的每一个位进行第二次扰动得到瞬时性随机响应(Instantaneous Randomized Response，IRR)结果向量S，其中,第二次扰动的方式按照以下公式进行，其中与分别表示B`的取值为0和1时S为1的概率：

 (4-51)

在每个用户得到最终的扰动结果S后，会上传至第三方进行统计并进行校正，然后结合通过 Lasso 回归进行矩阵映射**错误!未找到引用源。**，完成每个属性对应的频数统计，具体证明如文献[9]所示。

RAPPOR机制的主要特点体现在如下两个方面：

该机制通过将随机响应应用于具有强-差分隐私保证的Bloom过滤器来收集有关任意字符串集合的统计信息。

RAPPOR机制在客户端本地执行，并且不需要可信的第三方。

整体来看，RAPPOR在Bloom Filter过程中，可能会产生一定的误差，可以通过对相关参数的调整来降低误差的影响；而对于传输方面，每个客户端需要传输长度为*h*的向量，会导致整体的传输代价增加；还有利用随机应答技术干扰时本身的误差，这也是不可避免的。目前，谷歌在Github上已经对RAPPOR代码进行了开源[https://github.com/google/rappor]，并给出了可视化工具方便使用者进行进一步的分析研究。

### 4.2.4 极值机制

极值机制作为某一类机制的统称，可以看做最优解机制的一个集合（the family of optimal mechanism）。对于极值机制在差分隐私下的应用，目前的研究较少，主要分为几何机制和梯度机制两种，由于梯度机制可以看做是一均匀概率分布的几何混合体，所以可以将梯度机制看做一种特殊的几何机制。Ghosh**错误!未找到引用源。**等人利用计数查询，证明了在贝叶斯框架的最小期望成本下，保证差分隐私的最优机制是几何机制（geometric mechanism）；Geng**错误!未找到引用源。错误!未找到引用源。错误!未找到引用源。**等人针对最坏情况下的输入，提出了一系列针对不同的差分隐私模式下的梯度机制（staircase mechanisms），并证明了在任意敏感度下，这些机制中的单调效用函数是最优的。而对于本地差分，Peter等人**错误!未找到引用源。**分别针对高与低两种不同的隐私，分别结合二进制和随机应答，分别实现对应的满足本地差分隐私的梯度机制。

下面对于本地差分下的极值机制进行简单的介绍：

首先，给出了一个最优解的特征方程：假设随机变量，隐私输出变量，当随机变量在任何时候满足分布，则隐私输出的分布则满足，其中，那么就得到最优的特征方程：

 (4-52)

其中，是所有满足-本地差分隐私机制的集合。

接下来给出最优解的表达式，设有隐私机制*Q*，对于任意*y*与兄弟数据集*x，x`*，都存在与的比值，即对于所有， ，都满足：

 (4-53)

简单来说，这是一个复杂的非线性过程：在Q中最大化一个凸函数；且Q的维度可能是无限维，则最优私有化机制可能会产生一个无限大的输出字母集合。

最后给出极值机制的定义：对于任何，任何一对分布和，以及任何一个*f-divergence*模型，都存在一个最优机制，在所有 -本地差分隐私的情况下，使得最优特征方程中的*f-divergence*最大化。例如

 (4-54)

其中，对于所有，都有。

总的来说，对于最优解的机制都可以称为极值机制。举例来说，梯度机制也是极值机制的一种，从图中可以看出，最终得到的对数似然比的绝对值只能取两个极值中最优的一个值（见图4-3）。因此，也可以说梯度机制是满足本地差分隐私的。

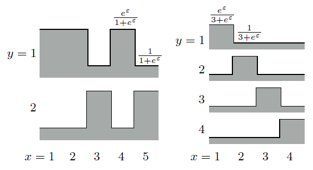


图4-3 梯度机制实例

### 4.2.5 基于信息压缩和扭曲机制

目前，随机化响应是本地化差分隐私的主流扰动机制。除此之外，还有基于信息压缩和扭曲的扰动机制。随机响应技术的扰动机制主要以一定的概率分布衡量输入和输出的关系，而基于信息压缩和扭曲的扰动机制主要从信息损失的角度考虑输入和输出的关系。

Xiong S[22]等人提出了压缩输入域的扰动机制Compression。每个用户对应一个维的元组，其中，，称为输入域。对输入域压缩得到。记中的元组为。压缩过程保证元组与的对应关系满足本地化差分隐私保护，其中压缩率的定义如下：

 (4-55)

对输入域进行压缩会使得与之间产生一定的偏差，该偏差的定义如下：

 (4-56)

为了对与之间的偏差程度进行限制，引入了信息扭曲度，即：

 (4-57)

其中，为偏差数学期望，和表示与的分布。最终把问题转化为了给定扭曲度和压缩率下求解最小的凸优化问题。其中，表示隐私保护程度，表示数据统计结果的可用性，通过压缩率将两者直接联系起来。

### 4.2.6 本地化差分隐私的方法分析

随着社会的文明和进步，人们开始高度重视精神需求，最主要的体现是越来越重视个人隐私权的保护。目前，以第三方为主的隐私泄露问题成为主要的隐私安全问题，搜索引擎搜完，网站马上跳出关键词广告，B站账号口语登陆360快视频等事件，都能看出个人的隐私安全受到严峻的考验。在隐私意识不断增强的背景下，如何保证用户的隐私安全成为了热门的研究方向。

因此，在不可信第三方的前提下，基于传统差分隐私提出了本地化差分隐私。本地化差分隐私不同于以往的差分隐私算法，主要在用户客户端进行隐私保护，然后将隐私保护后的数据上传至第三方，这样大大减小了第三方隐私泄露的可能性，而且在客户端进行隐私保护，也让用户可以自定义自己的隐私需求，使整个隐私保护过程更加直观。目前，本地化差分隐私的研究热点主要应用于数据统计，具体为进行单值频数发布，多值频数发布和均值发布等。

#### 4.2.6.1 本地化差分隐私的单值频数发布

基于本地化差分隐私的单值频数发布是指每个用户只发送一个变量取值的情形，用户将数据发送给数据收集者后，数据收集者根据已有的或统计得到候选值列表，统计其中每一个候选值的频数并进行发布，其主要是通过编码-解码技术以及概率扰动技术发布属性候选值的频数来保护数据隐私，例如RAPPOR**错误!未找到引用源。**，然后针对RAPPOR方法的主要存在两个方面的缺陷：用户和数据收集者之间的传输代价比较高和数据收集者需要预先采集候选字符串列表以进行频数统计，分别提出了S-Hist**错误!未找到引用源。**和O-RARRPR**错误!未找到引用源。**方法。在S-Hist方法中，每个用户对字符串进行编码后，随机选择其中一个比特位，利用RR进行扰动后，将其发送给数据收集者，大大降低了传输代价。而O-RARRPR基于RAPPOR方法的编码和解码方式，针对变量取值未知的情况，对每个字符串首先利用哈希函数进行一次值的映射，后续的扰动步骤直接对哈希值进行处理，而不关注字符串本身。

#### 4.2.6.2 本地化差分隐私的单值频数发布

多值频数发布是指每个用户发送多个变量取值的情形，用户将数据发送给数据收集者后，数据收集者根据已有的或统计得到候选值列表，统计其中每一个候选值的频数并进行发布。其在此基础上进一步利用采样技术和降维技术等提高了数据的可用性，例如RAPPOR-unknown**错误!未找到引用源。**，是RAPPOR方法的一个改进，基于n-gram思想，在客户端对数据数据进行扰动，并从字符串中随机抽取若干长度相同的子串，并将扰动结果与子串相关信息一起上传至服务器。而对于集值数据，提出了LDPMiner**错误!未找到引用源。**方法进行频数发布，LDPMiner可以看做是RAPPOR 和 S-Hist 方法的一个组合方法，由于每个客户端需要发送多个数据项，通信代价较高，因此采用随机采样技术使每个用户只发送其中的一个数据项，其中把采样技术和 RAPPOR 方法的组合叫做 sampling RAPPOR，而把采样技术和 S-Hist方法的组合叫做 sampling SH，然后分别利用 RAPPOR 方法和 S-Hist 方法来进行数据扰动，通过采样技术简化了隐私预算的分配，缩小了需要噪音添加数据项的集合。

#### 4.2.6.3 本地化差分隐私的均值发布

针对本地化差分隐私的均值发布研究工作还较少，其主要思想一般是在无偏估计的前提下对连续值进行离散化，然后通过大量的正负噪音对结果进行抵消。目前提出的算法有MeanEst**错误!未找到引用源。**，Harmony-mean**错误!未找到引用源。**。两者的区别在于MeanEst算法根据变量的个数平均分配，而Harmony-mean方法则通过采样技术，把全部分配给采样变量，在一定程度上降低了计算开销。

## 4.3 集中差分隐私（章国政）

差分隐私是一种数学上严格的隐私定义，适用于大数据集的分析，并有正式的关于隐私预算度量。此外，差分隐私算法将一个参数(通常为，在算法的任何执行中给定的限制允许的隐私预算，并提供具体的隐私与效用之间的平衡)作为输入，能为数据个体提供一定的隐私保护。差分隐私的优势之一是有能力推断出在多重分析上的累计隐私预算，并给出单个分析的值。在基本定律的界限内，可以适当选择的值，同时释放任何给定数量的评估统计量，然而，这个界限并不严密。

因此，为了能够得到更精确的分析结果且减少隐私预算，在传统差分隐私基础上提出了集中差分隐私。与传统差分隐私相比，集中差分隐私拥有更高的数据准确性和更少的噪声，并且在组合机制上基本没有隐私损失。

### 4.3.1 定义与性质

集中差分隐私是对传统差分隐私的一种改进，主要的实现是在传统差分隐私的基础上添加次高斯随机变量和次高斯分布两个概念。因此，在介绍集中差分隐私定义之前，需要先来介绍一下集中差分隐私定义中需要用到的一些基础理论知识。

#### 4.3.1.1 隐私损失随机变量

定义4.1 （隐私损失随机变量-- Privacy Loss Random Variable）假设有随机算法M，在一对邻近数据集中对于每一次的输出结果O，都有隐私损失且其值的大小为：

 (4-58)

在集中差分隐私中，当算法*M*在*x*上运行时，实值随机变量通过从抽样并计算出隐私损失结果。这个随机变量可以取正值或者取负值，对传统差分隐私算法而言，它的大小通常用来界定，对于近似差分隐私算法而言，除了的概率，它的大小也通常用来界定。

#### 4.3.1.2 次高斯相关理论基础

著名数据家高斯在1809年发表了关于正态分布的著作，从而得名了高斯分布。后来，它被公认为正态分布的同义语。我们知道，正态分布在概率论中起着举足轻重的作用，在各种分布中，它居于首要地位，我们实际中遇到的一些变量，他们的分布近似于正态分布。吴尚文等[23]通过从高斯变量定义出发，通过建立一些重要的不等式，研究有界随机变量的性质得出数学期望为零的有界随机变量是次高斯的。下面来介绍一些次高斯随机变量的相关理论基础知识。

定义4.2 （次高斯随机变量）对于一个常量和随机变量，如果等式成立，那么随机变量是次高斯随机变量。

定义4.3（次高斯集中引理）当，且随机变量是次高斯随机变量，则以下等式成立：

， (4-59)

定义4.4（次高斯和）对于，是（联合分布）实值随机变量且，则随机变量是次高斯的。当时，随机变量是次高斯的。

定义4.5（次高斯随机变量方差）任意次高斯随机变量的方差为。

定义4.6（次高斯分歧和不可分辨性）对任意两个随机变量和，如果满足，且(集中)分布被定义为次高斯分布，次高斯参数最多为。则次高斯分歧成立，如果两个次高斯分歧和都成立，则称随机变量对和是次高斯不可分辨。

#### 4.3.1.3 集中差分隐私

根据以上隐私损失随机变量和次高斯的相关理论基础，给出下面集中差分隐私的定义：

定义4.7 （集中差分隐私[24] (Concentrated Differential Privacy,CDP)）

如果随机算法对所有的邻近数据集和都满足，则称算法满足集中差分隐私。也就是说，如果隐私损失随机变量的平均值是，且该隐私随机变量与平均值的差值所产生的（居中的）随机变量是具有标准（尺度参数)的次高斯分布，则算法满足集中差分隐私。即：

 (4-60)

其中，集中差分隐私保证期望隐私损失的值是，且损失超过其平均值的概率的界限是。

### 4.3.2 组合定理

集中差分隐私和传统差分隐私在组合定理上都具有相同的性质，但事实上，组合的集中差分隐私基本上没有隐私损失。在组合过程中，集中差分隐私既包括在同一数据库上重复使用各种集中差分隐私算法，并对其进行模块化构造；也可以在不同的数据库中重复使用（各种）集中差分隐私算法，且这些数据库之间可能存在交集。这两种组合情况下，集中差分隐私算法在相同的“隐私预算”下能够提供更大的效用。

K个集中差分隐私机制(在相同的数据库或不同的数据库上)的组合由一系列随机变量来定义的。随机变量是敌对选择和自适应选择集中差分隐私机制的结果。在序列(现实)中，随机变量是通过运行机制在包含个人信息(例如Bob的数据)的数据库(对手选择的数据库)上采样得到的。在序列(替代现实)中，随机变量是通过运行机制在同一数据库上采样而得到的，但是Bob的个人数据被数据不同个体(Alice，敌对选择的)的数据代替。敌对选择和自适应选择对选择机制和数据集的要求是序列(Bob-真实)和序列(Alice-真实)的选择结果“非常接近”，特别是隐私预算是次高斯的。

更详细的解释，假设有一个游戏，在这个游戏中，一个庄家翻转一个硬币在符号和之间进行选择，并且一个对手自适应的选择一系列相邻数据库和一个满足的机制，并且这个机制将在(如果庄家选择)中运行或在中(如果庄家选择)运行，最后返回数据结果(其中)。对手的选择是完全自适应的，因此庄家不仅可以依赖于任意的额外知识，而且可以依赖于步骤中观察到的情况。对手的目标是最大化隐私预算，而大量的隐私预算与庄家决定选择哪个符号或的能力有关。

定理4.1 集中差分隐私组合定理

在集中差分的组合形式下，对于任意整数，隐私损失预算，尺度参数以及序列，都满足以下公式：

 (4-61)

证：

假设为一系列随机变量且隐私损失随机变量。

实值随机变量通过从抽样并计算输出隐私损失结果。

对手在第步选择的机制和数据集取决于敌手当时的观点。敌手的观点来源于选择随机性和敌手迄今为止观察到的结果。假设和分布表示序列和序列中的随机变量，则对于任意的和随机串有：

 (4-62)

对于每一个前缀，假设，并分解随机变量为。一旦前缀确定，则对手输出的下一对数据库和以及机制就确定了。则对于任意值和来源于的分布，有：

 (4-63)

由机制的CDP属性可知，满足次高斯。

由随机变量的次高斯属性可知，即隐私损失随机变量等于随机变量的和。

由期望的线性得以及由定义4.4可知，随机变量是次高斯的。

### 4.3.3 集中差分与传统差分

Dwork等人提出，任何满足差分隐私的算法也都满足集中差分隐私，并在文献[2]给出了证明，下面将针对以上结论介绍两个相关定理。

定理4.2 如果算法满足那么它的期望隐私预算只是(对于足够小的)，任何算法的隐私损失随机变量都满足参数为的次高斯。

定理4.3 对任意满足-差分隐私的算法，都满足-集中差分隐私。

### 4.3.4 集中差分与近似差分

集中差分隐私和近似差分都是对传统差分隐私的改进，但是相比较而言，集中差分隐私能添加更少的噪声并得到更精确的分析。

集中差分隐私与差分隐私相比，有两个优点：

1) 提高准确性

集中差分隐私是针对大规模计算的情况而提出的。传统上，为确保高概率情况下累计损失较小，允许为每个查询设置的损失非常小(),即使对于任何单个查询本身的隐私预算(例如甚至)可能都不是很重要的。这正是集中差分隐私中精确的灵活性：不关心单个查询的损失，而关心累计损失的概率界限。

2) 组隐私

即使对于一群个人数据不同的一对数据库，组隐私也会造成隐私损失。例如，在健康调查中，被调查人不仅希望自己的健康信息是被保密的，而且希望全面保护自己的家庭信息。任何差分隐私自动确保对任何大小为的组满足差分隐私，并且期望隐私损失增长倍。在确定的条件下，满足所有差分隐私机制并添加高斯噪声的任何集中差分隐私对大小为的组满足集中差分隐私。

当这种情况，集中差分隐私明显弱于近似差分隐私，因为即使期望损失非常小，前者的隐私预算超过ε的概率可以是保持不变的常量，而后者的隐私预算超过ε的概率中非常微小，只有δ。另外，差分隐私的期望隐私预算是没有界限界定大小的。

下面我们针对组隐私在集中差分隐私中的一些基础理论进行介绍。

### 4.3.5 组隐私

在本节中，我们将讨论集中差分隐私中的组隐私、如何限制组的期望隐私预算和如何限制组的次高斯标准。在前面的内容中，我们提到了保证集中差分隐私的任意机制也为组提供了集中差分隐私。这一点将在下文中论述。

定理4.4：组集中差分隐私 假设是满足集中差分隐私的算法，和表示两个有且仅有个不同记录的数据集，如果的大小由一个足够小的常数来界定且，则以下等式成立：

 (4-64)

其中，由于，所以隐私预算随机变量期望值大约为，且次高斯标准约等于。

组集中差分隐私定理的证明Dork等人[2]已经给出，并证明了组隐私中两个重要的性质。下面仅给出限制期望隐私损失的组隐私和限制次高斯标准的组隐私的性质。

#### 4.3.5.1 限制期望隐私损失的组隐私

对于实数范围内的次高斯随机变量分布，和，如果满足和且，则以下等式成立：

 (4-65)

#### 4.3.5.2 限制次高斯标准的组隐私

对于实数范围内的次高斯随机变量分布，和，假设和，如果满足和，则对任意的实数，都满足以下等式：

 (4-66)

其中，从分布到分布(集中)差分隐私随机变量是次高斯的，且次高斯标准以为界限。

## 4.4 小结

本章主要介绍关于差分隐私的扩展模式，主要从近似差分，集中差分和本地差分三个方面进行。首先介绍了近似差分的定义以及相关的实现机制，以及近似差分模型下的组合定理；然后介绍了集中差分的定义以及与差分隐私之间的关系，并提出特有的隐私保护形式组隐私；最后介绍了本地差分隐私的定义与性质，以及相关的实现机制。

## 4.5 参考文献

1. Dwork C, Kenthapadi K, Mcsherry F, et al. Our Data, Ourselves: Privacy Via Distributed Noise Generation[J]. Lecture Notes in Computer Science, 2006, 4004:486-503.
2. Geng Q, Viswanath P. The Optimal Mechanism in $(\epsilon,\delta)$-Differential Privacy[J]. 2013.
3. Geng Q, Viswanath P. Optimal Noise Adding Mechanisms for Approximate Differential Privacy[J]. IEEE Transactions on Information Theory, 2016, 62(2):952-969.
4. Dwork C, Lei J. Differential privacy and robust statistics[C]// ACM Symposium on Theory of Computing. ACM, 2009:371-380.
5. Dwork C, Rothblum G N, Vadhan S. Boosting and Differential Privacy[C]// IEEE, Symposium on Foundations of Computer Science. IEEE Computer Society, 2010:51-60.
6. Duchi J C, Jordan M I, Wainwright M J. Local privacy and statistical minimax rates. Foundations of Computer Science (FOCS), 2013 54th Annual IEEE Symposium on. IEEE, 2013: 429-438. [doi: 10.1109/FOCS.2013.53]
7. 叶青青,孟小峰,朱敏杰,霍峥.本地化差分隐私研究综述[J/OL].软件学报:1-26[2018-03-20].https://doi.org/10.13328/j.cnki.jos.005364.
8. Warner S L. Randomized response: A survey technique for eliminating evasive answer bias. Journal of the American Statistical Association, 1965, 60(309): 63-69.
9. Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological), 1996: 267-288.
10. Ghosh A, Roughgarden T, Sundararajan M. Universally utility-maximizing privacy mechanisms[J]. SIAM Journal on Computing, 2012, 41(6): 1673-1693.
11. Q. Geng and P. Viswanath. The optimal mechanism in differential privacy. arXiv preprint
12. arXiv:1212.1186, 2012.
13. Q. Geng and P. Viswanath. The optimal mechanism in differential privacy: Multidimensional setting.arXiv preprint arXiv:1312.0655, 2013.
14. Q. Geng and P. Viswanath. The optimal mechanism in -differential privacy. arXiv preprint arXiv:1305.1330, 2013.
15. Kairouz P, Oh S, Viswanath P. Extremal mechanisms for local differential privacy[C]//Advances in neural information processing systems. 2014: 2879-2887.
16. Erlingsson Ú, Pihur V, Korolova A. Rappor: Randomized aggregatable privacy-preserving ordinal response. Proceedings of the 2014 ACM SIGSAC conference on computer and communications security. ACM, 2014: 1054-1067. [doi: 10.1145/2660267. 2660348]
17. Bassily R, Smith A. Local, private, efficient protocols for succinct histograms. Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing. ACM, 2015: 127-135. [doi: 10.1145/2746539.2746632]
18. Kairouz P, Bonawitz K, Ramage D. Discrete distribution estimation under local privacy. Proceedings of the 33rd International Conference on Machine Learning, New York, NY, USA. 2016:2436-2444.
19. Fanti G, Pihur V, Erlingsson Ú. Building a RAPPOR with the unknown: Privacy-preserving learning of associations and data dictionaries. Proceedings on Privacy Enhancing Technologies, 2016, 2016(3): 41-61.
20. Qin Z, Yang Y, Yu T, Khalil I, Xiao X, Ren K. Heavy Hitter Estimation over Set-Valued Data with Local Differential Privacy. Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. ACM, 2016: 192-203. [doi: 10.1145/2976749.2978409]
21. Duchi J C, Jordan M I, Wainwright M J. Local privacy, data processing inequalities, and statistical minimax rates. ar Xiv preprint ar Xiv:1302.3203, 2013.
22. Nguyên T T, Xiao X, Yang Y, Hui S C, Shin H, Shin J. Collecting and Analyzing Data from Smart Device Users with Local Differential Privacy. ar Xiv preprint ar Xiv:1606.05053, 2016.
23. Xiong S, Sarwate A D, Mandayam N B. Randomized requantization with local differential privacy[C]// IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2016:2189-2193.
24. 吴尚文. 次高斯变量[D]. 湖北大学, 2004.
25. Dwork C, Rothblum G N. Concentrated Differential Privacy[J]. 2016.

# 第五章 面向数据发布的差分隐私方法（边，章，乐）

## 5.1 隐私数据发布需求（章国政）

信息技术的飞速发展使得各类数据的发布、采集、存储和分析变的方便快捷。例如，医院电子病例记录病人基本信息、疾病信息及药品购买记录；人口普查记录市民的家庭住址以及收入情况；金融业务服务会记录客户私有信息及其交易行为等。而这些数据的收集和发布直接给个人隐私造成威胁。一方面，如果数据拥有者直接发布隐含的敏感信息，而不采用适当数据保护技术，将可能造成个人的隐私泄露。另一方面，对发布后的数据进行分析也给数据的隐私带来了威胁。例如，采用数据挖掘和机器学习技术对医疗病例记录和搜索日志进行挖掘，可以获得病人所患何种疾病以及用户搜索的行为模式等敏感信息。

隐私数据发布研究的问题是如何在满足差分隐私的条件下保证发布数据或查询结果的精确性，研究内容主要集中在发布机制和算法复杂度的调整上，研究方法主要是基于计算理论和学习理论的定量分析。差分隐私保护数据发布根据环境不同可分为两种，即交互式数据发布和非交互式数据发布。

数据发布一直是差分隐私研究的核心内容，其研究进展直接影响到差分隐私在其它相关领域的应用。虽然目前差分隐私在数据发布上获取了很大进展，但仍然有些关键问题需要进一步解决：

1）高敏感度查询问题。差分隐私针对低敏感度查询的效果较好，例如敏感度为1的计算查询，其噪声方法仅为。但实际应用中，会遇到很多高敏感度查询，例如查询最大值，其敏感度可能远远大于1。这时加入的噪声往往会覆盖原有数据，造成数据可用性急剧下降。

2）计算复杂度问题。大部分数据发布机制的计算复杂度都是非高效的，超过的多项式阶，甚至达到指数阶。高复杂度限制了差分隐私在实际应用中的效率，成为目前需要解决的问题之一。

## 5.2 发布模式（章国政）

### 5.2.1 发布策略

满足差分隐私情况下的数据发布旨在向公众公布汇总信息而不披露数据中任何单个人的记录。 这个问题可以表示如下：如果一个数据集拥有者有一个数据集并且要在此数据集中进行一个查询集，他/她必须在数据不泄露个人敏感信息的情况下回答每个查询中的每个查询。在这个发布场景中涉及到两个发布设置：交互式和非交互式：在交互式设置中，单个查询的查询结果是在前一个查询的查询结果被公布之后才公布查询结果。在非交互式设置中，数据集拥有者一次得到所有查询结果，可以在拥有查询结果集知识的情况下提供答案。在非交互式设置中，查询函数之间的相关性引起更高的敏感度，则非交互式比交互式需要添加更多的噪声。

对于大小为的数据集，为了达到一定的准确度，机制最多只能回答查询集中最多个子线性查询。为了简化问题，大多数关于交互式设置的论文都假设这些查询是相互独立的。这些弱点使拉普拉斯机制在需要回答大量查询的场景中不切实际。为了解决机制这种弱点，提出一些新的机制旨在限制添加噪声的情况下发布各种类型的隐私数据。这些新机制设计侧重于查询的数量、输出的准确性和计算效率。其中，查询次数意味着机制可以回答的最大查询次数，表明该机制的能力。 发布输出的准确性通过随机答案和真实答案之间的预期误差来衡量。此外，如果回答查询的运行时间是关于和的多项式，则发布机制被认为是有效的。

### 5.2.2 交互式发布

在交互式环境下，用户向数据管理者提出查询请求，数据管理者根据查询请求对数据集进行操作并将结果进行必要的干扰后反馈给用户，用户不能看到数据集全貌，从而保护数据集中的个人隐私。交互式数据发布主要是针对各种类型（transtions、直方图、数据流、图数据集）的输入数据集进行操作。

交互式数据发布问题可表述为：给定数据集和查询集合，需寻求一种数据发布机制，使其能够在满足差分隐私保护的条件下逐个回答中的查询，直到耗尽全部隐私保护预算。发布机制的性能通常由精确度来衡量。交互式数据发布即是要满足一定精确度的条件下，以给定的隐私保护预算回答尽可能多的查询。

交互式隐私保护数据发布的研究主要集中在发布机制和基于直方图的发布方法上，两者的区别在于，前者直接对数据集进行操作来响应查询，而后者先根据数据集建立直方图分布，然后根据直方图分布来响应查询。

最早用于交互式数据发布的差分隐私保护机制是Dwork等人提出的机制，在该机制下，根据查询函数的敏感度和隐私保护预算产生服从分布的噪声，并添加到每个查询结果中，但缺点是查询的数据有限，与数据集记录数为次线性关系。另外，在干扰针对连续属性的查询结果时会产生较大的噪声。

之后，Roth和Roughgarden[1]提出了中位数机制(Median)。相对于机制，中位数机制能够在相同预算的情况下提供更多数量的查询。在该机制下，查询被分为“难查询”和“易查询”两类，其中，“易查询”的结果可以根据“难查询”的结果来确定。因此“易查询”就无需消耗任何预算。他们的研究证明 ，给定域和个查询，“难查询”的数量级为，其它数量级的查询均为“易查询”。“难查询”的结果通过独立的噪声进行干扰，而“易查询”的结果则用之前查询结果的中位数来确定。中位数机制的缺点则在于其算法的时间复杂度会随着数据集数量的增长呈指数增长，同时，其样本复杂度也是超多项式的。

Hardt等人[2]提出了另一种有效的机制，即PMW（Private Multiplicative Weights）。该机制的理论来源是机器学习中的加权多数算法（Weighted Majority Algorithm），该算法通过投票机制来构建一个复合算法。与之相似的，PMW也采用了一种投票机制来减少隐私预算的消耗，使得该机制能够在给定的隐私预算的情况下回答更多的查询。具体的方式为，PMW把数据集在数据域上的分布视作一个直方图，首先将每个频数设为相同，然后等待查询，通过将每个查询的结果加上噪声之后和上一次查询结果相比较，若差异小于设定的阈值，则发布上一次查询结果的值，此步骤不消耗隐私保护预算。只有当差异大于此阈值时，才会发布新的查询结果，并调整直方图中相应频数的值。由于很多查询不消耗隐私预算，所有此机制能比普通机制回答更多的查询。

Gupta等人[3]提出一种更为通用的迭代数据集生成架构（IDC Framework），并进一步证明之前的Median 和 PMW 机制都是此架构的特例。在该架构下，对于一个查询集合，首先在定义域空间中任意选择一个数据集，作为初始数据集假设，然后用次数据集回答所有的查询，若发现有某个查询结果和真实结果之间的差别大于预定义的阈值时，则根据此查询结果来更新数据集假设，使更新后的数据集能在阈值范围内回答此查询。这个迭代过程循环进行，直到所有的查询结果和真实结果的差异不再大于阈值，由于迭代次数会少于总查询数量，所以耗费的隐私保护预算会比普通机制更少，从而降低噪声总量。

另外还有一些针对特殊查询的发布机制，如布尔连接查询（Boolean Conjunction Query）、半空间范围查询（Halfspace Range Query）等。

交互式发布机制的研究方法是应用计算理论来分析各种发布机制在保证差分隐私的前提下查询结果的精确度，这些研究的结论是差分隐私在数据挖掘、机器学习等领域应用研究的理论基础。

### 5.2.3 非交互式发布

在非交互式环境下，数据管理者针对所有可能的查询，在满足差分隐私的条件下一次性发布所有查询的结果，或者，数据管理者发布一个原始数据集的“净化”版本，这是一个不精确的数据集，用户可对该版本的数据集自行进行所需的查询操作。

非交互式数据发布问题可表述为：给定数据集和查询集合，需寻求一个数据发布机制，使其能够在满足差分隐私保护的条件下一次性回答F中的所有查询。

早期的差分隐私保护研究认为数据发布很难再非交互式环境下实现差分隐私，DInur等人[4]曾提出一个数据集如果精确回答了超过次线性个查询，那么用户就能够以很高的概率还原出原始数据集。因此，如果要在非交互式环境下回答查询，或者发布一个被净化处理的数据集，必须在发布的内容中加入大量噪声，但这会极大地破坏其可用性，所以，早期的研究大多集中在交互式环境下的查询数据发布。但随着应用要求的提高，单纯的交互式环境在查询数量和应用方式上存在许多局限，从而促进了非交互式环境下数据发布的隐私保护研究。

非交互式数据发布的研究主要集中在批查询、列联表发布、基于分组的发布方法以及净化数据集（Sanitized Dataset）发布方法上。

批查询（Batch Query）：数据管理者针对所有可能的查询，一次性对外发布所有查询的结果，这种模式称为批查询。在批查询模式下，由于每个查询之间彼此相关，删除数据集中任一记录有可能会导致多个查询结果发生改变。因此，基于差分隐私保护的批查询函数具有比单一查询高得多的敏感度。目前，批查询主要通过映射和变换查询集合来降低总敏感度，从而降低噪声量。比较有代表性的有Xiao等人[5]提出的小波变换方法（Privelet）、Hay等人[6]提出的层次查询方法等。

列联表发布:列联表是对数据集中记录按照变量进行分类时所列出的频数表，它是非交互式数据发布的一种特殊形式。例如数据集包含个记录，由个布尔变量组成，列联表就是对数据集按照个可能的组合值进行统计计数所形成的表格。事实上，在数据分析研究中发布的内容通常并非列联表本身，而是按照多个变量的组合值进行统计所得的计数，也称为边缘频数（Marginal）。差分隐私保护在列联表发布中的应用主要有两种方法。其一是向列联表的每个单元格中加入噪声，用户可以根据被干扰的频数计算每个边缘频数。这种方法能够维持所有边缘频数之间的一致性，但边缘频数的累计噪声也会较大。另外一种方法是先计算出拟发布的边缘频数，然后对它们加入噪声后在发布。由于每个边缘频数只添加了一次噪声，所以数据可用性更好。但因为每个边缘频数都是独立地添加噪声，因此所发布的各边缘频数之间可能会违背数据上的一致性。

基于分组的数据发布：基于分组的数据发布方法将早期的匿名泛化技术应用到非交互式环境下来实现差分隐私保护。K-anonymity及其衍生模型（如l-diversity模型和t-closeness模型）属于典型的基于分组的隐私保护方法。研究表明，这些模型对攻击者所掌握知识假定过少，并不能提供足够的安全保障。但一些学者也认为差分隐私保护对攻击者所掌握的指数假定过多（假定攻击者掌握了数据集中除攻击目标之外所有其它个体的信息），对安全的要求过于严格。基于分组的数据发布延续了k-anonymity模型的思想，并用差分隐私的要求来控制分组及匿名化处理的整个过程。从已有的研究来看，隐私保护预算在算法中的合理分配与充分利用是基于分组的数据发布需要继续研究的问题。

净化数据集发布：净化数据集是对原始数据集进行隐私保护处理后发布给用户的数据集。直接发布一个满足差分隐私的净化数据集来让用户进行任意的查询，一直被认为是一个十分困难的问题。因为这会引入非常大的噪声，并覆盖整个数据集的原始数据[7]。但随着差分隐私保护理论研究的深入，研究者发现如果将学习理论引入差分隐私保护，可以在一定程度上解决非交互式数据发布中精确度较低的问题。学习理论拓展了非交互式发布的研究领域，证明了在保证精确度的情况下发布针对某类查询的净化数据集是可行的。但研究的难点在于如何降低计算复杂度，如何处理数值型数据以及如何提供更广泛的查询类型等问题上。

### 5.2.4 基于差分隐私保护的数据发布方法分类比较

数据发布一直是差分隐私研究的核心内容，其研究进展直接影响到差分隐私在其它相关领域的应用。本节对基于差分隐私保护的数据发布方法进行分类并分析了各自的特点（如表5-1所示）。可以看出，虽然目前差分隐私在数据发布上获取了很大进展，但仍然高敏感度查询、计算复杂度等问题需要进一步解决。

表5-1基于差分隐私保护的数据发布方法分类比较

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 方式 | 方法 | 方法描述 | 典型机制或方法 | 优点 | 缺点 |
| 交互式 | 交互式查询 | 对原始数据集查询产生结果，加噪后发布 | Laplace,  Exponential,  Median,PMW,  k-norm，IDC | 容易实现，可满足所有的查询类型 | 噪声较大，查询次数有限 |
| 直方图发布 | 由原始数据集产生加噪后的直方图，根据直方图响应查询 | KD-Tree,DPCube,  LP,Boost1,Boost2,NoiseFirst,  StructureFirst,Privelet,GS,EFPA,  P-HPartition,FPA,Privelet,  Constrained Inference | 敏感度小，分析简单，噪声可以控制在较小范围内 | 查询类型受限制，查询次数有限 |
| 非交互式 | 批查询发布 | 数据管理者针对所有可能的查询，一次性对外发布所有查询的结果 | Privelet,  Hierarchical ranges,  Matrix mechanism，  Low-rank mechanism,  Quad-tree,  Adaptive Mechanism | 容易实现，可满足所有查询类型 | 噪声较大，但可以采用不同机制降低到一定程度 |
| 列联表发布 | 对数据集中的记录按k个属性的排列组合产生k维频数表，家噪声发布 | Fourier basis,  Non-uniform strategy,  Correlated Row | 可满足大部分查询类型 | 高纬度列联表噪声大 |
| 分组发布 | 对原始数据集进行泛化处理并发布 | Safe  k-anonymization,  DiffGen,  DT-Diff | 结合泛化和差分隐私方法，容易实现 | 隐私保护预算分配主观性大 |
| 净化数据集发布 | 对原始数据集加入噪声后产生净化数据集并对外发布 | Exponential Searching,  Boosting ,  Recursive,  Threshold learning,  Learning theory approach | 可满足多种查询类型，查询次数可达到n的指数阶 | 时间复杂度高，实现困难，噪声大 |
| 树划分发布 | 基于先转换或者压缩原始数据，再对转换后的数据添加噪声的发布策略，考虑如何设计支持数据划分的索引结果，并依据索引结构发布隐私数据 | Quad-Post,  kd-strandard,  k-noisemean,  Diffpart,Hybrid-Bus,  n-gram,DP-tree | 可满足多种查询类型，查询精度高 | 防止树自身泄露隐私困难，如何分配隐私预算困难 |
| 网格划分发布 | UG[8],AG[8] | 能够比较合理地给出空间数据的格划分粒度，并且能够均衡噪音误差与均匀介绍误差 | 仅适用于二维数据发布 |

### 5.2.4 适应式发布

适应式查询：

我们考虑回答关于受到差分隐私影响的敏感数据集的查询问题。这些查询可以通过离线、在线、自适应等三种方式在较大集合的允许查询中选择查询，离线查询：一次选择查询，并且差分私有机制在同一批次中回答查询。在线查询：一次选中全部查询，但该机制仅以流方式接收查询，并且必须在查看下一个查询之前回答前面的每个查询。适应性查询：一次选择一个查询，差分隐私机制必须在选择下一个查询进行查询之前回答每个查询。特别是，每个查询可能取决于以前查询的答案。在自适应模型中，许多不同的差分隐私机制与在线查询模型一样高效。

自适应查询模型：查询没有预先确定，每个查询依赖于前个查询的查询结果。

差分隐私看起来非常适合自适应查询模型。比如一个简单的差分隐私算法，其采用个个体的数据集并在错误为的情况下回答自适应模型中的规模为的统计查询，只需简单的通过仔细校准的噪声独立扰动每个答案即可。Hardt和Rothblum [9]的私有乘法权重算法可以在很多情况下以强大的准确性保证来回答任意的，自适应选择的统计查询的指数数量，而[10]表明私有乘法权重的准确率几乎是最佳的，即使对于简单的查询系列也是如此。

## 5.3 面向直方图发布的差分隐私（13边锦，24李映乐）

直方图作为近似估计数据发布的主要技术之一，常常用于获取数据的概要信息，实现数据的分析与查询。直方图就是将数据集按照某属性分割成许多不相交的小数据集，采用桶数据形式，来描述数据集的整体轮廓特征。直方图较好的展示了数据集中指定属性值的分布情况，对直接发布的直方图进行线性查询，在获取特定属性相关统计类信息的同时，极有可能泄露个人的敏感信息。

以医疗数据为例对直方图中存在的隐私泄露问题进行描述。医疗机构中依据病人胃癌患病情况统计发布的数据如表（5-2）所示，该数据集发布的胃癌患者年龄分布等宽直方图如图（5-1）所示。

表5-2胃癌患者医疗记录

|  |  |  |  |
| --- | --- | --- | --- |
| 姓名 | 年龄 | | 是否患有胃癌 |
| 张三 | | 42 | 1 |
| 李四 | | 37 | 1 |
| 王五 | | 45 | 0 |
| 赵六 | | 39 | 1 |
| 钱七 | | 48 | 0 |
| …… | | … | …… |

图5-1 胃癌患者年龄分布直方图

攻击者在掌握了病人李四的年龄后，可以依据一定的背景知识，推断出病人李四是否患有胃癌。比如，攻击者知道了李四的年龄位于[35,40]之间，且知道了除李四之外的数据集发布的直方图对应[35,40]桶计数值为1，表示年龄在35-40之间，数据集中有1人患有胃癌，由数据集发布的直方图对应[35,40]桶计数值为2，可以推断出李四患有胃癌。

因此，对数据采用直方图等技术进行统计分析，在直方图发布前，需要对直方图中相应的统计信息进行保护。与传统隐私保护技术相比，差分隐私可以忽略攻击者的背景知识，通过向发布的直方图中添加适量的、额外的噪声进行扰动，从而达到隐私保护的作用。

下面根据原始直方图结构重组方式，来分类阐述直方图的发布方法。

### 5.3.1 层次树变换

按照层次树结构重组直方图的发布方法，是立足于原始直方图的基础之上，在为每个桶计数添加噪音之前，先对直方图的自身结构进行重新组织，然后再对重组之后的结构添加噪音，这种操作不但能够提供精确的长范围计数查询结果，而且也能够减少噪音误差。

该类方法的主要目的是采用树结构对原始直方图进行组织，利用树的层次特征精确地相应较长范围的计数查询，其中代表方法,Hay[11]等人建立了差分隐私区间树，并进行了同方差加噪与一致性调节。

利用树对等宽直方图进行重新组织，其中为该树的扇出，该方法根据树的高度决定重新构造直方图敏感度，其中表示桶的个数，因此，树中每个节点添加拉普拉斯噪声的大小为,例如图5-2中的节点[40,50]的计数为1，噪音量为。为了能够精确地响应范围计数查询，采用了查询语义一致性约束与最小二乘法相结合的后置处理技术，对树中的噪音计数进行约束推理。

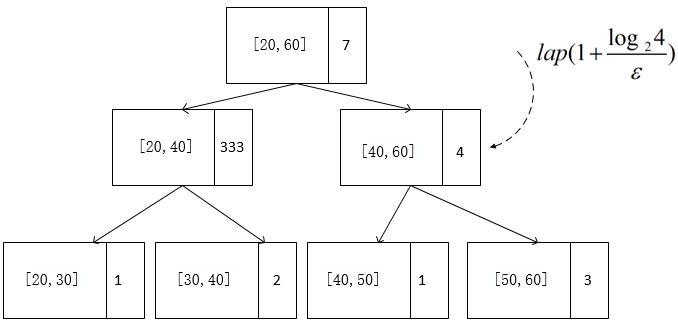


图5-2 变换

的优点在于能够比较精确地响应较长范围的计数查询，而它们的不足在于查询

敏感性比较高，实际的性能比较差．通常直方图实际的桶个数会非常大，层次树的中间结点数目可能达到指数级别，这样直接导致响应长范围计数查询的质量很差．而且方法仅适用于一维直方图的发布，并没有讨论层次树的扇出对最终发布精度的影响。扇出的选择决定树高度以及每层应分得的隐私预算。因此，如何基于层次树结构重组方法，发布多维直方图以及设定合理的扇出都是未来的研究方向。

现有基于区间树结构的差分隐私直方图发布方法大多采用同方差加噪方式， Xu[12]等人提出了StructureFirst，优化直方图划分策略，并在划分区间后的构造区间树中进行了同方差加噪。采用同方差加噪方式的发布方法有效提升了发布数据的可用性和算法效率。而实际上通过研究可以发现，差分隐私直方图发布加噪的另外一种方式，异方差加噪，可进一步提高发布精度。Wahbeh等人[13]提出了通过迭代方式在区间树进行层次间隐私预算分配的方法，有效降低了查询误差，但其在相同层次的节点中仍使用相同的隐私预算，因此仍具有进一步优化的空间。Peng等人[14]提出了DPtree（differential private tree），通过异方差加噪，能够对多维数据进行发布并提高查询精度，但该方法采用了完全K叉树结构，限制了树结构调节的灵活性。

针对以上问题，康健等人[15]提出了一种异方差加噪下面向任意区间树结构的差分隐私直方图发布算法LUE-DPTree。首先根据区间计数查询的分布，计算区间树中节点的覆盖概率，并据此分配隐私预算，实现异方差加噪；接着经分析指出该异方差加噪策略适用于任意区间树结构，且从理论上证明了在任意区间树结构下进行异方差加噪后，仍可在一致性约束下利用最优线性无偏估计进一步降低区间计数查询的误差。

在给定区间树的情况下，我们知道需要通过运用噪声机制使每个节点满足-差分隐私，若对任意节点，均有，则称作同方差加噪方式；若存在节点，使得，则称作异方差加噪方式。

1） 假设对区间树进行区间计数查询时，所有查询区间的出现概率相等，那么可由CNCP算法计算节点覆盖概率。CNCP算法具体步骤：

输入：待计算节点及其父节点。

输出：区间树中所有节点的覆盖概率。

1. 若为根节点：
2. 若为其他节点：
3. 对所有，执行;

2）为最小化区间计数查询误差的期望，区间树中差分隐私预算分配方案需满足：，当然，为节点的节点系数。节点系数CNP的计算步骤详见[15]

2）计算出节点系数,后，可通过分配每个节点的差分隐私预算并进行异方差加噪，步骤如下：

输入：待加噪节点，路径隐私预算和。

输出：区间树所有节点的加噪计数值。

1.  //分配隐私预算
2.  //添加噪声
3. 对所有，执行：；

对于任意树结构差分隐私区间树的构造算法,构建步骤如下详见[15]

5）分配新节点编号和子区间

; //继续构建子树

}

最后将待计算发布计数值的节点作为输入，运用算法计算区间树所有节点的参数。计算出参数后，通过算法计算优化后的最终发布值。具体算法实施即可参考文献[15]，最终在异方差加噪下，面向任意区间树结构进行最优线性无偏估计优化的差分隐私直方图发布算法。

具体实施步骤：

输入：原始直方图，差分隐私参数

输出：异方差加噪，任意树结构的差分隐私区间树

1.  //构建区间树
2.  //节点覆盖概率计算
3.  //节点系数计算
4.  //异方差加噪
5.  //最优线性无偏估计优化
6.  //计算发布值

针对实验过程中运用不同特征的数据集，若要更好地提高数据发布质量，需要建立不同的树结构，包括任意区间树结构。算法仅适用于完全叉树的情况，使得人们无法通过改变树结构来降低查询误差。而算法则可以适用于任意区间树结构，同时，由于任意树结构放宽了对数据集长度的限制，提高了算法的适用范围，使得算法具有更高的实用性。使得人们可以有更大的调整空间，能够寻找更佳的构建方式来进一步提高直方图的发布质量。

### 5.3.2 聚类变换

针对于基于聚类变换的直方图发布的差分隐私方法，DWork等人[16]最早提出了一种通过将拉普拉斯随机噪声添加到原始直方图的每个小区计数来发布差分私有直方图的方法，它被认为是一种基线策略。Xu等人[17]进一步地提出了差分私人直方图的两种与聚类相关的实现算法：NoiseFirst算法和StructureFirst算法。其中，NoiseFirst算法基于基线策略，而StructureFirst算法使用动态规划技术构造最优直方图，以确定待合并的容器的极限。虽然算法利用聚类降低了直方图中噪声的大小。但以上算法仅仅涉及了聚类的有关知识，并没有从实质上针对聚类变换本身提出有关直方图发布的算法应用。

目前，基于聚类变换的直方图发布的差分隐私方法主要以基于分区和基于分层这两种聚类方式为主。

针对基于分区的算法，Tarek Benkhelif [18]等人先发布一个匿名数据表，并以同时满足不同的隐私要求为目标，提出了解决方案结合了联合聚类与合成数据生成以产生匿名化数据。但该算法的复杂度较大，运行速度较慢，且不能针对维数较大的直方图发布。

Xiao Y所提出了DPCube算法[19]，该算法使用两阶段分区的策略方式。算法首先基于域（而不是数据）的基于单元的划分用于生成细粒度的等宽单元直方图（如在基线LPA策略中），其给出了原始数据分布的近似值，然后根据细胞直方图释放合成数据集Dc。其次，在Dc上执行基于kd树的多维分区 获得统一或接近统一的分区。结果分区键用于对原始数据库进行分区并获取每个分区的嘈杂计数，从而得到v-最优直方图。最后，给定用户发出的查询，估计组件使用v-最优直方图或两个直方图来计算查询的答案。具体而言，由于第二多维分区步骤仅访问专用合成数据集Dc*，*因此DPCube保留差分隐私 对隐私保证没有影响。第二步使用基于kd树的空间分割策略，旨在产生接近均匀的分区，以最小化分区内的估计误差。它从覆盖整个空间的根节点开始。在每个步骤中，启发式地选择该维度上的当前分区范围内的拆分维度和拆分值，以将空间拆分为子空间。该算法重复直到满足预定义的要求。在这里，可以使用信息熵和方差等几个度量标准来衡量分区的一致性，并决定是否分割当前分区并选择最佳分割点。

Qing Q[38]研究了在边差分隐私下发布私有图的节点强度直方图的问题。提出了两种基于序列感知和局部密度的聚类方法来聚合直方图。算法首先研究在边缘差分隐私下发布加权图的节点强度直方图的问题。给定图，其目标是在满足边缘差分隐私的同时释放尽可能接近G的真实分布的节点强度直方图。但该目标存在的一个关键挑战是真正的分布情况太稀疏。如果最大节点强度远大于节点数，则较大数量的区间的值将为零。这将意味着当查询间隔较大时，噪声累积将导致查询结果的准确性较低。因此，为了解决这个问题，提出了以下解决方案。首先，作者通过设置边缘权重的上限t来减少直方图的区间数。注意，给定更大的t，可以保留更多的权重信息。同时，较低的t意味着节点强度的分布可以较少稀疏。然后作者生成权重有界图的节点强度直方图。此外，作者私下学习了分区B的分区，并用每个分组的总和替换每个分区的值。最后，为了减少噪声，变换和聚合引起的误差，作者使用指数机制输出节点强度直方图的准确估计。

在聚类变换过程前，序列感知聚类方法已经保留了差异隐私并减少了噪声和近似的误差。 但是由于初始候选者是随机选择的，因此这样的方式这不能保证稳定的结果。基于此，讨论并采用基于观察到群集中心在相邻仓中的波动较小的策略。即如果簇bin具有更多相邻的簇bin，其值与其类似，则它更适合作为簇中心。 作者认为这个簇bin具有更高的局部密度。

通过与其相邻bin 之间的累积平均距离来测量。累积平均距离公式如下：

其中：

该算法在聚类变换方面的步骤如下：

|  |
| --- |
| 基于密度的直方图聚类算法 |
| 输入：原始直方图，桶数k，隐私预算  输出：集群桶B  1对任意的 ，**循环**  2 计算bin 的邻近距离d(i,j)  3给该距离加Laplace噪声处理  4 **循环结束**  5 对任意的 ，**循环**  6 计算局部密度  7**循环结束**  8选择具有最高局部密度的前k个区域作为聚类中心  9将每个hi与最近的聚类中心相关联，并将分成k组  10 **返回B** |

针对基于分层的聚类算法，G Acs[20]等人提出了一种使用可分层次聚类方案来压缩直方图的方法。属于同一聚类的直方图库具有相似的计数，因此可以用它们的平均值来近似。最后，只释放具有较小灵敏度的噪声聚类中心。

文章考虑直方图作为时间序列，并开发一个增强的傅立叶扰动算法（EFPA），给定实数或复数的*n*维向量，傅里叶变换将*x*转换成复数的*n*维向量

，得出的公式:

(5-1)

利用*F*的逆离散傅立叶变换可以无损地恢复*X*：

(5-2)

, (5-3)

因此该算法的核心思想是属于同一个簇的直方图箱具有相似的计数，因此可以用它们的平均值近似，仅释放具有较小灵敏度的噪声聚类中心就足够了。该算法面临的挑战是如何在差分隐私下对直方图区域进行分层划分。但遗憾的是，此算法仅涉及一维和二维直方图发布，并且不适合于多维数据的发布，且算法的时间复杂度较大。

### 5.3.3 傅里叶变换

采用傅里叶变换发布直方图，该类方法是利用基于有损压缩的离散傅里叶转换（Discrete Fourier Transform，DFT）来发布直方图，其代表方法是、、等。

Rastogi等人[21]提出的算法，主要是对个查询均添加噪声没有意义这一点进行改进，通过重构个查询答案将傅里叶系数作为个查询的数目，采用分布式扰动方法，最后发布离散傅里叶查询结果。支持长区间查询，但当时，数据可用性较低。

算法主要以提高长序列的查询答案的准确性为目的，该算法主要基于离散傅里叶变换进行的。

维序列的离散傅里叶变换是线性变换，给出另外一个维序列，带有

列元素定义为，同样可以计算的逆公式为

，而且。

算法主要步骤：

|  |
| --- |
| 输入：序列,参数 |
| 1）：计算  2）：计算  3）：返回 |

将表示为的前个元素，被称为个最低频的傅里叶系数，它们表示中的高级趋势。从中可以获得和的近似值，表示为通过将个零附加到而获得的长度的序列,计算,显然，可能与不同，因为忽略最后的傅立叶系数可能会引入一些误差。我们用表示缩写重建后第个位置的误差值为。

 [22]算法是在基础上改善的算法实现有损压缩，保证了数据发布的准确性。

 [22]是该类方法的典型代表，是基本傅里叶扰动算法，对于给定一个直方图，首先采用傅里叶变换DFT，直方进行转换操作，即，然后，从中选出个系数作为，是为了缩小查询敏感度，避免高频度桶对发布质量的影响，即剔除个其它的系数，再后，对中的个系数添加拉普拉斯噪音，使变成，并在中补充个0，即，最后对执行DFT的逆操作IDFT，即得到待发布的直方图。

FPA方法对直方图进行了相应压缩，生成的误差由噪音误差和重组误差构成，该方法中的选取决定了误差的大小，较大会导致噪音误差增加，而较小的会导致重组误差增加，因此，值的选取至关重要。

EFPA为改进后的傅里叶变换方法，其算法的主要步骤如下：

|  |
| --- |
| 输入：长度为原始直方图,是奇数  输入：隐私预算  输出：长度为的加噪后的直方图 |
| 1:  2:  3：计算，对于所有的，并且  4：选择概率正比于  5：  6：  7: |

EFPA方法的提出[23]，主要是解决FPA方法中的值选取问题，通过调整相应参数来均衡重构误差与噪音误差。EFPA是基于指数机制设计了一种自适应的挑选方法，依据FPA的发布误差设计了一种更有效的打分函数，

 (5-4)

并将带入指数机制中，最终获得的挑选概率。

FPA方法可以支持较长范围计数查询，精度高，扩展性差，EFPA方法支持单位长度范围的查询，查询精度低，且两种方法只适用于一维直方图的转换。

### 5.3.4 小波变换

基于小波变换的直方图发布的差分隐私方法是通过引入小波变换的数据发布技术，该技术不仅保证了差异隐私，也为所有范围计数查询提供准确的查询结果，即对每个属性的谓词是范围的计数查询。特别地，基于小波变换的直方图发布的差分隐私方法，可以保证任何计数查询都都可以用多对数英寸的噪声方差来回答。这显着提高了噪声方差由Dwork等人的方法提供。

在文献[24]中，Xiao等人提出使用小波变换的隐私保护方法回答直方图查询。该算法将Haar小波变换应用于数据集的频率矩阵。哈尔小波本质上构建了数据集的二叉树，其中每个节点（或“系数”）表示其右子树中节点的平均值与左子树中节点的平均值之间的差值。具体而言，给定矩阵*M*的Haar小波系数，*M*中的任何条目都可以容易地重建。设为基系数，为分解树*R*中在第*i*层的祖先，则有

(5-5)

如果是的左子树，则是1；反之，若是的右子数，则为-1。

隐私预算分为不同级别，该方法将噪声添加到与其灵敏度成比例的每个变换系数。然后这些系数用于通过应用逆小波变换来重新生成数据集的匿名版本。但由于大多数小波变换隐私保护技术都是基于启发式概念设计的，因此算法没有可以证明隐私保证。

文献[25]提出了基于epsilon-differential的算法来生成隐私保护聚类的扰动数据。算法使用Haar小波变换（HWT）进行分析。具体而言，对于二维数据集，此方法在应用小波变换时使用标准分解。将该数据集视为频率矩阵，该方法首先对每一行应用Haar小波变换。结果是每行的尾数系数矢量。使用细节系数矩阵作为输入，该方法对列进行变换。然后将噪声添加到每个单元中，与该单元中的系数的灵敏度成比例。该算法使用Haar小波不但提高了扰动数据的维数，且显著提高了聚类算法的效率；而且通过Haar小波噪声获得了更具有合适效用水平和差分隐私保证的扰动数据。

针对于高维数据情况下大数据不能有效地运行，文献[26]提出了一种基于Daubechies-2小波变换（D2WT）并保留差分隐私概念的新的差分小波变换（PPC）算法。算法使用D2WT对数据进行噪声前的原始数据预处理。通过对原始数据应用D2WT，得到的数据不仅与原始数据相比具有较低的维度，而且由于较少的噪声添加，可以提供具有高效率的差分隐私保证。

Lin C等人[39]提出了一种针对身体传感器网络中敏感大数据的差分隐私保护方案。该方案通过构造树结构以减少错误并提供远程查询，并利用Haar小波变换方法用于将直方图转换为完整的二叉树，从而更方便地添加噪声干扰，极大地减少了计算开销。

具体而言，算法首先要对数据集进行处理。一般而言先通过结构变形进行结构变换以降低灵敏度。然后将噪声添加到数据集中以在变形之后获得数据集的释放。因此算法建立了一个直方图用于记录添加噪声后的变形。从而可以减少错误并可以提供远程查询。通过分析和比较，算法采用树形层次结构，并将每组直方图用作树的叶节点，根据从左到右的直方图的顺序对节点id进行编号。然后构建完整的树。由于灵敏度Δf可以过删除或向数据集添加数据而创建的，而该参数这也导致差异隐私保护的关键错误。因此要确保相同的直方图在树中具有相同数量的叶节点。如果使用不同的节点度，它将生成不同高度的树。因此，灵敏度Δf等于树的高度。

然后，作者将说明如何构建树结构以及如何基于直方图设置相应的值。首先，对于完整的二叉树，Haar小波变换用于分组直方图。 Haar小波函数是不连续的，类似于阶梯函数，其定义为：

比例函数定义为：

对于完整的二叉树，范围的左节点被视为x（0,1 / 2），而范围的右节点是x（1 / 2,1）。通过Haar小波变换，则可以将直方图转换为完整的二叉树。

假设现在有n组直方图，通过使用Haar小波变换方法将直方图转换为完整的二叉树并优化结构。并通过添加拉普拉斯运算，完整二叉树的每个节点将生成具有差分隐私组的n行。从n组发布的数据不仅可用，而且还可以阻止攻击者获取所需信息。

因此，对于一组数据集，医学调查中的每个id号对应于心率。首先，将直方图数据优化为若干组，每组与相应的年龄范围匹配。每个心跳数据点对应于适当的范围，频率范围将加1.此外，每个组表示结果直方图中频带的原始数据的频率。

接下来要获得树结构。首先建立Knode结构用于存储原始数据集，其次将组的宽度除以创建节点的函数，以将每组中的原始数据转换为直方图。每组存储频率节点的值是直方图中每个组的原始数据。在树的所有非叶节点的初始化中：创建，，，，其中存储完整二叉树中叶节点的频率的平均值和，和是小波系数。 因此，令为小波系数的，的左子树中的所有叶节点的平均值表示为a，并且的右子树中的所有叶节点的平均值表示为b，则值计算如下：

然后为每个小波系数的二叉树添加噪声。 在这里使用拉普拉斯方程：

进一步地，通过小波添加噪声可以计算为：

其中，，h表示 完整二叉树的高度，i表示位于二叉树的层。

然后执行拉普拉斯变换以分配噪声或每个系数，在接收到完整二叉树的所有小波系数之后，计算每个叶节点中记录的值，则有：

这里的是小波系数的值，并且的值与的左右子树相关联。 从基于上述公式的高度为h的完整二叉树的第一层开始，如果叶节点属于的左子树，则 = 1; 如果它低于右子树，那么= -1; 如果它不是子树，将在没有处理的情况下忽略它。

该算法步骤如下：

|  |
| --- |
| 基于小波变换的直方图聚类算法 |
| 1将原始数据转换为，每个都存储对应的频率  2 存储完整二叉树的所有叶子的平均值  3haar小波方法用于为每个分配值  4基于二叉树结构为非叶节点添加拉普拉斯噪声  5 获取小波系数后，找到每个根叶节点的路径  6 通过诱导每个叶节点的小波系数将每个节点的频率保存到文件中 |

## 5.4 面向划分发布的差分隐私（李映乐）

我们每天都会与位置感知设备进行交互。这些设备包括支持GPS的手机和平板电脑，以及导航系统。每台设备都可以向集中式服务器报告多个位置数据。这种位置信息通常称为地理空间数据，如果进行了适当的处理和分析，可以有很大的好处。但是如果将这些数据直接进行发布共享，则可能会对研究和其他用途产生重大影响。

差分隐私如今已成为隐私保护数据发布的事实标准，因为它能够提供强大的最差情况下的隐私保证。给定一个数据集和数据集中的元组所在的二维域，将每个元组视为二维空间中的一个点。即先数据将域分割成单元，然后以满足差分隐私的方式获得每个单元的噪声计数。差异性的私人概要由这些细胞的边界和它们的嘈杂计数组成。该概要可用于生成合成数据集，也可用于直接回答查询。

一般来说，在回答查询时，在这种差别私人提要方法中有两个错误来源。第一个来源是为了满足不同隐私而添加的噪声。该噪声具有预定义的方差并且独立于数据集，但取决于使用多少个单元来回答查询。 第二个来源是数据集本身的性质。假设数据点是均匀分布的，当发出只与某个单元部分相交的查询时，那么必须估计相交单元中有多少个数据点。 这个误差的大小取决于数据集中点和分区的分布。

目前面向划分发布的差分隐私主要分为基于树结构与基于网格结构的划分。下面将分别从这两个方面进行叙述。

### 5.4.1 基于树结构的划分

针对于基于树结构的划分分布下的差分隐私，直接添加二维和空间数据集的大多数方法都使用递归分区，这些方法执行数据域的递归二进制分区。如肖等人[27]提出了面向划分发布的适应标准空间索引方法，即KD树中的树结点按照某个维度递归分割。为了降低误差，Xiao等人[19]使用启发式方法来选择分割点，使得两个子区域尽可能接近均匀。此外，该算法主要分为两大步骤，首先基于数据域的统一分区生成合成数据，然后生成基于合成数据的kd树分区。

Cormode等人[28]提出了几种基于KD树的替代方案。他们没有使用均匀性启发式，而是沿着分区维度的中间值分割节点。其中，树的高度是预先确定的，隐私预算随着分布发层次的不同而不同。且一部分隐私预算用于选择中位数，另一部分则用于获取噪声计数。同时，该论文也提出了基于四叉树的方法，叶节点通过每个范围的中点，水平和垂直地成四个相等的区域。该论文中所提出的第三种方法结合了上述两种方法，他们称之为“KD-混合”。此方法对前几个分区级别使用四叉树，且将其他一些优化方法，如Hay等人 [29]提出的约束推理和隐私预算的优化分配应用在当前的算法中。他们的实验表明，第三种算法优于基于KD树的方法。

但以上算法只针对于一维或二维数据，Qardaji等人[30]针对多维数据集，提出了一般递归分区框架。在每个级别的递归中，沿着维度执行分区，从而导致数据点的最平衡分区。该方法所采用的平衡分区具有产生相似尺寸区域的效果。但当该算法应用于二维数据集时，上述方法与构建基于噪声中值的KD树非常相似。

上面的递归分区方法基本上在数据点的表示上构建了一个层次结构。而大量文献中已经提出了几种方法来改进对这种层次结构的计数查询。Hay[3]等人提出了层次方法的约束推理的概念，以提高范围查询的准确性。这项工作主要是在一维数据集的背景下开发的。使用这种方法，可以将所有查询的时间间隔转换为二进制树，其中单位长度间隔是树叶。计数查询然后在树中的所有节点上发布。约束推理利用了一致性要求，即父母的数量应该等于所有儿童数量的总和以提高准确性。

此外，Wu等人[31]采用区域均匀性判断，从下到上对四叉树分解的结果进行调整和组合，算法使用了一致性约束来对后处理节点进行处理，使得范围查询的准确性在一定程度上得到了改善。但是噪声的大小仍然受树的深度控制。

针对现有分层方法对深度的依赖，Zhang等人[32]提出了PrivTree方法，通过引入偏差变量来分割子域，避免了预定义的分割深度的限制。的定义如下：

(5-6)

其中，表示节点的精确点计数，用来决定节点是否应该分裂的阈值，表示私有树的衰减因子，表示从节点*V*到根的距离。之后，算法将Laplace噪声引入偏差变量进行保护，即：，若大于阈值，则树继续分裂。因此算法不受树的高度的限制，拆分节点*v*的决定只取决于。

算法的流程如下所示：

表5-3 PrivTree算法

|  |
| --- |
| PrivTree (*D,*) |
| 1用根节点初始化四叉树  2设置*dom()=*，标记为未访问的节点  3**若**存在未访问的节点 *v* **循环**  4 标记 *v* 已访问  5 引入衰减因子，计算偏差变量  6 如果过小，则做出调整：  7 计算添加噪音后的偏差变量  8 **如果**  9 分裂*v*，并将其子节点添加到  10 标记 *v* 为未访问的子节点  11去除所有点计数，返回四叉树 |

由此可知，该算法脱离了四叉树受树高*h*限制的这一弊端,不再担心递归问题，且依赖于启发式选择适当的*h*值，可以产生更加细粒度的分解，显著提高数据的适应性。

在PrivTree算法基础上，Li等人进一步地提出了基于四叉树的差分隐私混合分解算法。算法在分解的第一层，采用自适应的密度网格进行聚类，在第二层及之后的层级中采用基于改进的差分隐私四叉树分解算法，通过引入偏差向量的方式消除了四叉树对遍历深度的依赖，针对四叉树的每一层采用几何方式分配不同的隐私预算，并提出了四叉树分解的停止条件。算法的程序流程图如下图所示：

具体而言，为了充分利用位置大数据的时间和空间的冗余特性，首先提出了一种基于差分隐私的层次分解算法。算法通过结合发布数据的实际分布从水平和垂直两个方向来确定密集区域。其中网格由水平和垂直坐标的密集和稀疏间隔的交集构成。通过比较所形成的网格与给定的密度阈值，进而对网格单元进一步分类。

网格数目的定义如下：

其中， 表示位置数据在水平方向上的所有位置信息， 表示在垂直方向上的所有位置信息。 表示集体内的元素数量， 是在水平方向上所划分的区间长度，是垂直方向上划分的区间长度，、 是权重系数。 和 分别代表水平方向与垂直方向的网格数目。

根据水平与垂直方向的密度阈值，进而求得区间相交所形成的网格密度，将其与给定的密度阈值进行判断。若，则将其视为稀疏区域；反之，若，则将其视为密集区域。

密度阈值 定义如下：

图5-1 算法程序流程图

其中， 代表网格个数。

一般来说，四叉树的预算分配采用统一分配方式，该方式在四叉树中沿着根到叶路径，为每个节点分配相同的隐私预算，然而，随着四叉树不断地向下延伸，数据的敏感程度也随之不断增强，如果为每一层都分配相同的隐私预算，则对于敏感数据的保护强度较弱，同时也会带来了较大的噪声干扰，影响算法的查询精度。此外，针对于稀疏，甚至空白区域，总体的隐私预算将会有一部分被舍弃。而为每一层分配相同的隐私预算，无疑增加了多余的计算量。因此本算法采用几何隐私预算分配策略，证明几何预算分配方法的查询精度优于同一分区级别下统一分配的查询精度。该分配方式通过误差分析推导而出，四叉树分割后的每层得到的隐私预算为：

其中， 表示四叉树的层数。

接下来采用PrivTree算法，考虑到四叉树中每一层中注入的噪声量必须与最大递归深度成正比，因此，输入参数 必须在算法能够确定正确的噪声量的前提下被固定。然而， 的选择是一个严重的问题。如果选择过小，则算法的精确度不高，而过大则会引入太多的无关点，为算法带来大量冗余。且不能直接调整的大小，否则其本身的信息会被泄露。因此，算法通过引入偏差向量，完全消除了算法对的依赖，不再担心递归深度，显著提高了算法的效率。

偏差计数定义如下：

其中，表示当前节点所在网格内所有节点的数目，表示当前节点所在层数，参数为衰减因子，是一个较小的正数，其定义如下：

表示该四叉树的扇出，即当前节点下最大的子节点个数，与递归深度无关。为拉普拉斯因子。

接着判断偏差计数与 之间的大小，将两者中较大的值重新赋值给偏差计数，也就是说：

之后生成噪声计数，判断加噪后的偏差计数是否大于给定阈值，若大于，则对该节点进一步划分。否则，该节点作为四叉树的叶子节点。

本算法通过引入偏差计数，使得拆分节点的条件完全取决于偏差计数本身，而与树本身的高度无关。算法的最终效率将不再受输入参数 限制。既避免了输入参数的隐私泄露风险，也进一步提高了算法的查询精度。

因此，算法针对密集区域，算法首先创建四叉树的根节点，令，并将标记为未被访问。之后进行多次迭代，每次迭代中判断是否还有没被访问的节点*v*，若存在这样的节点，将其标记为已访问的同时，计算得到偏差计数。对偏差计数进行大小调整，并加入拉普拉斯噪声，其中的值可由四叉树的扇出求得。得到加噪后的偏差计数后，再进一步判断当前节点*v*是否满足拆分条件，即判断是否成立。若满足该条件，则对当前节点*v*进一步拆分，通过生成*v*的子节点并将它们作为未被访问的节点插入到四叉树中；反之，若，则*v*作为四叉树的叶节点。直到四叉树中所有节点全被访问到，算法结束。其具体步骤如下：

|  |
| --- |
| 算法2 |
| 输入：数据集，隐私预算，网格个数，预算分配因子  输出：查询结果*Q*  步骤：  1. 计算密度阈值  2. 判断当前区域内数目与密度阈值的大小。若小于阈值，则视为稀疏区域，直接加入参数为的拉普拉斯噪声；若大于阈值，则进入第三步  3. 创建四叉树的根节点，并将其标记为未被访问  4. 如果存在没有被访问的节点，则开始循环：  5. 将该节点标记为已访问，得到当前遍历次数*i*并计算得到偏差计数  6. 对偏差计数进行大小调整，并加入拉普拉斯噪声  7. 得到加噪后的偏差计数  8. 判断当前节点*v*是否满足拆分条件。若是，对*v*进一步拆分，作为未被访问的节  点插入到四叉树中，否则作为四叉树的叶子节点  9. 循环结束 |

该算法虽然较好地满足了位置大数据查询的要求。然而算法不能较好地应用于实时动态变化的位置流数据。数据相关分解方法主要基于空间数据点的实际分布。

Inanet al [33]根据Kd-树结构提出了一种新的分割几何空间的方法。该算法减少了统一的假设误差。且由于每层的分割线的确定取决于数据的实际分布，而这些分割线需要一些隐私预算来对中间值进行保护。算法首先使用依赖于数据的Kd-树方法来划分原始数据，然后使用数据无关的四叉树方法再次分割上述结果。实验结果表明，综合Kd混合算法在响应距离查询中可以得到更准确的结果。

### 5.4.2 基于网格结构的划分

以往的简单的均匀网格划分方式，是在数据域上应用等宽的网格，然后在网格单元上发出独立的精确查询，但是这样的方法往往会由于数据集本身的特性而存在大量缺陷。

基于此，研究者们进一步提出一种新颖的自适应网格方法。这种方法的动机是需要在密集区域上有更精细的粒度分区，同时还需要对稀疏区域进行粗分区。自适应网格在数据集上覆盖粗粒度网格，然后根据其噪声计数进一步划分每个单元。然后使用这两个分区级别来回答数据集上的查询。

目前最为常用的方法为均匀网格（UG）方法[34]，这种方法将数据域划分为大小相等的个网格单元，然后获得每个单元的噪声计数。但是UG算法的主要缺点是它地对数据集中的所有区域的处理方式都相同，也就是说，密集区和稀疏区都以完全相同的方式进分分区。而在实际情况下，当一个区域密集时，希望使用更细粒度的分割，同样，当一个区域稀少时，希望在那里使用更粗糙的网格。

基于这一观察，作者提出了一种自适应网格（AG）方法[34]。AG方法首先在数据域上放置一个粗糙的网格，创建个第一级单元，然后使用隐私预算为每个单元添加一个计数查询，其中，，决定如何分割两个不同级别单元的隐私预算。借鉴Hay M [35]中提取的方法，AG算法产生第二级单元，产生网格。对于每个单元，设作为小区的噪声计数，AG使用基于自适应地选择的小区尺寸来划分小区，进一步创建叶单元。

具体而讲，假设是第一级单元的噪声计数，令为在第二级单元进一步划分而成的噪声计数。则可以利用的和及的加权平均值得到一个更为精确的计数，从而使得的噪声误差的标准偏差达到最小。其计算公式如下：

(5-7)

然后，通过在所有节点之间均匀分配差异，将该值传播到叶节点。

(5-8)

UG和AG方法在概念上简单，且易实现，同时两个算法都适用于非常大的数据集。针对于这两种基于差分隐私的网格划分算法，Qardaji et al [1]在此基础上提出了一种数据无关的分解方法，首先使用UG将二维空间分散到大小相等的网格单元中。算法针对每个单元的统计数据，将的隐私预算与拉普拉斯噪声相结合的方式，对数据进行差分隐私保护。通过这种自适应分区策略，AG方法避免了第二层上的过度分区和欠缺分区等问题。但是，该算法尚有不足，如针对密集区和稀疏区，算法在第一层中的划分方式完全一致等。

此外，Darakhshan等[36]设计了蜂窝网络移动数据的DP-Where方法，采用统一的网格结构，添加受控噪声的方式修改WHERE模型来实现差异隐私。总体而言，与数据无关的分解方法通常无法平衡噪声误差和均匀假设误差。在数百万数据的实际位置信息集上，无法获得理想的分割结果，范围查询的准确性有限。

Huang等人[37]提出的Kd-PPDP方法[37]利用二次和误差测量当前网格的均匀性，并在AG方法的两阶段划分之后合并相邻网格以减少噪声组合。该算法基本思想是: 针对二维空间数据，先进行粗粒度划分，进行噪声扰乱，根据扰乱后粗 粒度划分结果利用 KD-树思想启发式地对原始数据 进行重新划分，添加(0*<<*1)噪声，得到第一层划分;再对第一层的每个划分区域自 适应地进行的细粒度划分，而后同第一层 划分方法一样，添加噪声，得到第二层划分;最后对两层结果进行一致性调节。利用 kd-树重新划分，实质是对相邻近似区域进行合并的过程，牺牲少量均匀假设误差来减少噪声的叠加。该方法的比较和合并过程需要遍历所有网格的划分边界，但该算法的计算复杂度较高。

关于如何选择网格的关键决策，等人提出了一种用于发布分类数据的新的差异隐私PrivPfC[40]方法，从而解决了为分类分析构建差异私有概要的问题。PrivPfC算法的基本思想是在一个步骤中私下选择一个网格，该网格将数据域划分为多个单元格。 通过使用具有新颖质量函数的指数机制来完成该选择，该函数通过直方图分类器最大化正确分类记录的预期数量。 PrivPfC与以往的方法的不同之处在于它在一个步骤中选择高质量的网格，而以前的方法使用迭代过程，因此遭受两个弱点。首先，迭代过程必须在所有迭代中划分隐私预算，导致在每次迭代中做出的选择具有显着的噪声。其次，迭代过程是一个贪婪的过程，即使不考虑噪声也往往导致次优的全局选择。且指数机制需要生成包含高质量网格的候选网格池，而不会使候选池太大，这会影响运行时间和准确性。此外，人们需要一种能够有效识别高质量网格并同时具有低灵敏度的质量功能。

具体而言，给定具有k个不同类标签的数据集D，L = {1,2，...，k}，网格g和用于将拉普拉斯噪声添加到计数的参数，网格质量通过直方图分类器的预期正确分类记录的数量来测量，其定义如下：

(5-9)

这里的，是具有类标签i的单元c中的数据点的数量，并且是类i单元c中的主要类的概率，即具有最高噪声计数。

因此，该算法提出两种基于网格结构划分的面向划分发布的差分隐私思想。其一是在给出允许算法考虑多少网格的上限时通过候选网格进行枚举。其二是新的质量功能，可以选择高质量的“网格”。该质量函数考虑了注入噪声对分类精度的影响，适应隐私参数并且具有低灵敏度。

## 5.5 小结

本章在上一章提出差分隐私相关知识的基础上，首先介绍了数据发布的需求，在此需求的基础之上提出了数据发布的三种模式：交互式发布、交互式发布和适应式发布。然后介绍了层次树变换、聚类变换、傅里叶变换以及小波变换等四种面向直方图发布的差分隐私方法，最后分析了基于树结构划分和网络结构划分的两种面向划分发布的差分隐私方法。

## 5.6 参考文献

1. Roth A, Roughgarden T. Interactive privacy via the median mechanism[C]//Proceedings of the forty-second ACM symposium on Theory of computing. ACM, 2010: 765-774.
2. Hardt M, Rothblum G N. A multiplicative weights mechanism for privacy-preserving data analysis[C]//Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on. IEEE, 2010: 61-70.
3. Gupta A R A. Ullman J. Iterative construction sand private data release[J]. Proceedings of the 9th Theory of Cryptography Conference (TCC). Taormina, Italy, 2012, 339: 356.
4. Dinur I, Nissim K. Revealing information while preserving privacy[C]//Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems. ACM, 2003: 202-210.
5. Xiao X, Wang G, Gehrke J. Differential privacy via wavelet transforms[J]. IEEE Transactions on Knowledge and Data Engineering, 2011, 23(8): 1200-1214.
6. 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.
7. Dwork C, Naor M, Reingold O, et al. On the complexity of differentially private data release: efficient algorithms and hardness results[C]//Proceedings of the forty-first annual ACM symposium on Theory of computing. ACM, 2009: 381-390.
8. Li N, Yang W, Qardaji W. Differentially private grids for geospatial data[C]// IEEE International Conference on Data Engineering. IEEE Computer Society, 2013:757-768.
9. Hardt M, Rothblum G N. A Multiplicative Weights Mechanism for Privacy-Preserving Data Analysis[C]// IEEE, Symposium on Foundations of Computer Science. IEEE Computer Society, 2010:61-70.
10. Bun M, Ullman J, Vadhan S. Fingerprinting codes and the price of approximate differential privacy[C]// ACM Symposium on Theory of Computing. ACM, 2014:1-10.
11. 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.
12. Xu J, Zhang Z, Xiao X, et al. Differentially private histogram publication[J]. The VLDB Journal—The International Journal on Very Large Data Bases, 2013, 22(6): 797-822.
13. Qardaji W, Yang W, Li N. Understanding hierarchical methods for differentially private histograms[J]. Proceedings of the VLDB Endowment, 2013, 6(14): 1954-1965.
14. Peng S, Yang Y, Zhang Z, et al. DP-tree: indexing multi-dimensional data under differential privacy[C]//Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data. ACM, 2012: 864-864.
15. 康健,吴英杰,黄泗勇,陈鸿,孙岚.异方差加噪下的差分隐私直方图发布算法[J].计算机科学与探索,2016,10(06):786-798.
16. Dwork C, Mcsherry F, Nissim K. Calibrating Noise to Sensitivity in Private Data Analysis[M]// Theory of Cryptography. Springer Berlin Heidelberg, 2006:637-648.
17. Xu, J., Zhang, Z., Xiao, X., Yang, Y., Yu, G., Winslett, M.: Differentially private histogram publication. VLDB J. 22(6), 797–822 (2013)
18. Tarek Benkhelif, Françoise Fessant, Fabrice Clérot, et al. Co-clustering for Differentially Private Synthetic Data Generation[J]. 2017:36-47.
19. Xiao Y, Xiong L, Fan L, et al. DPCube: Differentially Private Histogram Release through Multidimensional Partitioning[J]. Transactions on Data Privacy, 2012, 7(3):195-222.
20. Acs G, Castelluccia C, Chen R. Differentially Private Histogram Publishing through Lossy Compression[C]// IEEE, International Conference on Data Mining. IEEE Computer Society, 2012:1-10.
21. Rastogi V, Nath S. Differentially private aggregation of distributed time-series with transformation and encryption[C]// ACM SIGMOD International Conference on Management of Data. ACM, 2010:735-746.
22. Acs G, Castelluccia C, Chen R. Differentially private histogram publishing through lossy compression[C]//2012 IEEE 12th International Conference on Data Mining. IEEE, 2012: 1-10.
23. Rastogi V, Nath S. Differentially private aggregation of distributed time-series with transformation and encryption[C]//Proceedings of the 2010 ACM SIGMOD International Conference on Management of data. ACM, 2010: 735-746.
24. X. Xiao, G. Wang, and J. Gehrke, “Differential privacy via wavelet transforms,” IEEE Transactions on Knowledge and Data Engineering, vol. 23, pp. 1200–1214, 2011.
25. Dishabi M R E, Azgomi M A. Differential privacy preserving clustering based on Haar wavelet Transform[J]. Intelligent Data Analysis, 2014, 18(4):583–608.
26. Dishabi M R E, Azgomi M A. Differential privacy preserving clustering using Daubechies-2 wavelet transform[J]. International Journal of Wavelets Multiresolution & Information Processing, 2015, 13(04):150609203158003.
27. Xiao Y, Xiong L, Yuan C. Differentially Private Data Release through Multidimensional Partitioning[C]// VLDB Conference on Secure Data Management. Springer-Verlag, 2010:150-168.
28. G. Cormode, M. Procopiuc, E. Shen, D. Srivastava, and T. Yu, “Differential private spatial decompositions,” in ICDE, 2012.
29. M. Hay, V. Rastogi, G. Miklau, and D. Suciu, “Boosting the accuracy of differentially private histograms through consistency,” Proc. VLDB Endow., vol. 3, pp. 1021–1032, September 2010.
30. W. Qardaji and N. Li, “Recursive partitioning and summarization: A general framework for privacy preserving data publishing,” in Proceedings of the 7th ACM Symposium on Information, Computer, and Communications Security. ACM, 2012.
31. Wu, J., Lu, Q., Cai, J., et al.: Differential privacy two-dimensional data partitioning publication algorithm based on quad-tree. J. Huazhong Univ. Sci. Technol. (Nat. Sci. Edn.) 44(3), 99–104 (2016)
32. Zhang, J., Xiao, X., Xie, X.: PrivTree: a differentially private algorithm for hierarchical decompositions. In: Proceedings of the 36th ACM International Conference on Management of Data, pp. 155–170. San Francisco, CA, USA (2016)
33. Cormode, G., Procopiuc, C., Srivastava, D., et al.: Differentially private spatial decompositions. In: Proceedings of the 28th International Conference on Data Engineering (ICDE), pp. 20–31. IEEE, Washington (2012)
34. Qardaji W, Yang W, Li N. Differentially private grids for geospatial data[J]. 2012:757-768.
35. 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.
36. Mir, D.J., Isaacman, S., Caceres, R., et al.: DP-where: differentially private modeling of human mobility. In: Proceedings of the 2013 IEEE International Conference on Big Data, pp. 580–588 (2013)
37. Huang, S., Chen, T., Lu, Q., et al.: Differential privacy two-dimensional dataset partitioning publication algorithm based on kd-tree. J. Shandong Univ. (Eng. Sci.) 45(1), 24–29 (2015)
38. Qian Q, Li Z, Zhao P, et al. Publishing Graph Node Strength Histogram with Edge Differential Privacy[C]//International Conference on Database Systems for Advanced Applications. Springer, Cham, 2018: 75-91.
39. 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.
40. Su D, Cao J, Li N, et al. PrivPfC: differentially private data publication for classification[J]. The VLDB Journal—The International Journal on Very Large Data Bases, 2018, 27(2): 201-223.

# 第六章 面向数据挖掘和机器学习的差分隐私方法

## 6.1扰动方式（陈竑毓）

数据挖掘/机器学习算法在对数据进行处理的过程中，存在隐私泄露的可能，极大地威胁到用户的隐私安全。为了保护用户隐私以及敏感信息，考虑在数据挖掘过程或机器学习算法中添加扰动内容，得到一个差分隐私学习算法并最终实现隐私保护的目的。

扰动方式主要包括输入扰动、输出扰动、目标扰动以及样本和聚合框架，这些方法分别是在挖掘或学习的不同阶段添加随机性，但其共有目的是保护数据隐私安全，防止泄露用户隐私。本节将简单介绍这些扰动方式。

### 6.1.1 输入输出扰动

输入输出扰动是实现机器学习差分隐私性质的基本方法，其通过对算法的输入数据集添加扰动，或者为算法的输出参数添加扰动的方式实现差分隐私。

通过对输入数据进行扰动，可以改变用户的原始数据，从而导致攻击者窃取的数据是扰动后的而非真实数据，基于这些数据攻击者无法分析出用户的敏感信息，也就此达到隐私保护的目的。

对于输出扰动，机器学习算法首先在原始数据集上进行训练以得到模型最优解，后在输出的模型参数上添加Laplace噪声，或者在分类模型的推断结果上实现指数机制来实现对原始数据的隐私保护。

对于机器学习算法下的输出扰动，文献[1-4]使用Laplace机制实现了对输出结果的扰动，文献[5,6]使用指数机制实现了对推断结果的扰动。而文献[7-10]等则是基于迭代或者多步骤的学习算法，用Laplace机制和指数机制实现了对输出参数的隐私保护。

### 6.1.2 目标扰动

目标扰动，简单来说就是针对于某一个目标进行扰动。当攻击者对某一算法模型进行模拟训练后，就可能通过自己模拟的算法框架推算出用户本身的敏感信息。而在目标扰动中，可以针对算法模型这一目标进行扰动，也就是说，通过对算法模型中某些参数进行干扰，进而对算法模型的特征空间进行隐私保护，使攻击者无法获取算法模型中参数的正确取值，这样自然也就无法模拟出正确的算法框架，最终保证对整个模型框架进行隐私保护。文献[11,12,13,14]进行了目标扰动的示例。

### 6.1.3 样本与聚合框架

样本与聚合框架首先将数据集划分成许多个子集，接着在各个子集上进行学习得到输出，最后将各个子集上得到的输出进行聚合。其中对聚合步骤添加噪声以实现对数据集的隐私保护，框架的适用前提是查询函数能够用数据集中的小样本进行推断。文献[15,16,17]就是基于样本与聚合框架展开研究的。

机器学习算法在绝大部分输入输出扰动和目标扰动方式都需要确定样本空间的大小及样本范围，因为上述两者会对敏感度造成影响。而样本和聚合框架则没有这个限制，但是基于样本和聚合机制的差分隐私机制一般采用差分隐私。在实际应用中，如果样本空间没有限制并且希望样本能够满足差分隐私，可以对样本简单进行裁剪预处理。

## 6.2 面向无监督学习的差分隐私（李映乐，边锦）

在现实生活中常常会遇到缺乏足够先验知识等问题，因此难以采用人工标注类别，或者是在实际生活中进行人工类别标注的成本太高。因此很自然地希望计算机能完成这些工作，或至少提供一些帮助。根据类别未知(没有被标记)的训练样本解决模式识别中的各种问题，称之为无监督学习。

无监督学习在训练过程中样本没有标签，在现实世界中，大部分样本是不带标签的，所以无监督学习要比监督学习应用更广泛。其中，深度学习是由多层神经网络组成的，需要一层一层地抽取主要特征，忽略次要细节，因此，深度学习中采用的无监督学习方法需要满足三个条件：

1）可以从多维空间中抽取主要特征映射至低维空间；

2）具有递归性；

3）算法不能太过复杂，否则深层架构的计算量太大。

而针对于无监督学习，也会存在隐私安全的问题，这就需要我们在算法的基础上进行隐私保护。下面我们以无监督学习中的代表：聚类和和频繁模式挖掘为例，进行面向无监督学习的差分隐私介绍。

### 6.2.1 聚类

聚类(Clustering)分析是数据挖掘中面向无监督学习的常用手段，能够从大量的数据中抽取出其分布模式，根据数据自身特点划分为不同的聚簇，在同一聚簇的数据之前相似度越高越好，而不同聚簇内的数据间差异则是越大越好。

针对基于差分隐私的聚类算法，国内外已取得大量成就。目前基于差分隐私的聚类算法主要分为基于分区、层次、密度、网格、模型这五种类型，此外，还有结合两种聚类算法的混合算法等。下面以六个小节的形式分别论述。。

#### 6.2.1.1 基于差分隐私的分区聚类算法

针对基于差分隐私的分区聚类算法，主要以k-means算法为主，k-means算法是基于分区的聚类算法中最具代表性，也是聚类算法中最为基础的算法。Dwork[18]最早将k-means与差分隐私结合起来(DP k-means算法)，算法的主要步骤如下：

|  |
| --- |
| 基于差分隐私的k-means算法 |
| 1输入*d*维空间的*n*个点  2随机选取*k*个点，返回d维空间内*k*个添加了噪声的新点作为初始中心点  3将每个样本点：  4 划分到最近的中心点，将样本集合{}划分成共*k*个集合  5 对于，计算集合内点的和，和数目  6 对和与数目分别添加噪声得到，更新为集合的新中心点  7迭代完所有的样本点，直到点到集合的划分不再弯化或迭代次数达到上限 |

然而受限于当时的技术，该方法针对所有的数据点都加入了噪声进行保护，虽然提高了数据信息的安全性，但聚类数和初始中心点的选择仍是盲目的、随机的，因此算法的实用性较差，且聚类结果的可用性对噪声不具有鲁棒性。

随着差分隐私的应用领域越来越广泛，大量关于k-means的改进算法也相继被提出。 Li Y[19]等人提出了IDP k-means算法，通过将拉普拉斯噪声添加到每个子集的总和与数目，从而对数据隐私进行保护。算法虽然比DP k-means具有较高的聚类可用性，但该算法在初始中心的选择上仍是随机的，因此聚类精度仍旧不高。Acs G[20]等人提出了一种基于k-means的差分隐私直方图方法，大大提高了范围查询的准确性。然而，它们的方法不能应用于异常值的处理。Yu[21]等人提出了一种孤立点消除差分隐私k-means算法，算法根据数据点的分布密度选择初始中心点，并将拉普拉斯噪声添加到用于隐私保护的原始数据。算法证明了满足差分隐私定理，且实验结果表明该算法有效地保持了数据的隐私性，提高了聚类结果的准确性。

以上算法只是对k-means算法直接进行改进，也有不少采用与其他算法结合的方式来更好地确定中心值与聚类数的选择。如Shang T[22]等人将一种优化的树冠算法与DP k-means算法相结合，并将其应用到Hadoop平台上。算法利用MapReduce得到的中心点的个数和集合来实现DP k-means算法。因此，改进的树冠算法可以优化Hadoop平台上的中心和簇数的选择。

但如以上所述，大多数的方法都是交互式的方法，只是开发定制的差分隐私算法将其运用于各种数据挖掘任务。而非交互式方法的目的是开发差分私有算法，可以输出输入数据集的概要，然后可以用于支持各种数据挖掘任务。Su D[23]等人基于这个目的，提出了非交互式下的基于差分隐私的k-means算法(EUGKM算法)，算法结合扩展的统一网格，证明针对于高维数据具有良好的聚类效果。

除了k-means算法外，目前大量学者也将差分隐私与其它分区算法，如k-medoids算法等结合起来，对数据进行更有效地安全保护。马银方[24]等人提出基于差分隐私的KDCK-medoids动态聚类算法，算法针对于动态数据，在差分隐私保护技术的基础上将KD-树优化选取出的k个聚类中心和增量数据相结合建立新的KD-树，然后采用近邻搜索策略将增量数据分配到与其相应的聚类簇中，从而完成最终的动态聚类。实验证明不论针对维数较小的动态数据集还是多维动态数据集，该算法都能在实现隐私保护的同时高效地处理增量数据的动态聚类问题。

#### 6.2.1.2 基于差分隐私的层次聚类算法

虽然k-means算法是种较好的聚类方法，但该算法始终存在k的选择及初始聚类中心点等问题。为了避免这些弊端，层次聚类算法被提出。层次聚类，顾名思义，就是一层一层地进行聚类。一般而言，层次聚类主要分为两种，分裂法和凝聚法。

而针对于层次聚类算法，最具有代表性的是BIRCH算法，Yao Zhang[25]等人也第一次把差分隐私与BIRCH算法相结合，提出了Diff-BIRCH算法，该算法依据差分隐私模型，并借鉴了概率分配的思想，并通过采用异方差的加噪方式对数据进行保护。算法进一步提出改进后的BIRCH算法，并同时采用线性回归及迭代等运算方式，解决了DP-BIRCH算法中存在的不一致的约术性问题。

算法的主要流程如下：

|  |
| --- |
| ASDP-HPA算法 |
| 输入：流入滑动窗口的数据集，历史滑动窗口生成的直方图，历史隐私保护分配预算，隐私保护预算，参数，敏感度，直方图桶个数  输出：ASDP-HPA在滑动窗口下发布的直方图 |
| Begin  1： //对数据集进行预处理  2： //生成滑动窗口待发布的直方图  3： //ARIMA模型预测滑动窗口待发布的直方图 //  4：；//初始化ASDP-HPA算法在滑动窗口下生成的直方图  5： //计算隐私预算，根据数据集可调整参数  6： //计算抽样概率  7:  //计算抽样个数  8： //在0-之间随机生成个不相同的整数  8： for  do //遍历  9: if  then //若为抽样点  10:  //若被抽中，直接发布加噪值  11: else if  then //若为非抽样点且满足条件  12:  13: else //若为非抽样点且不满足条件  14:  15: end  16: end  17: end  18：Return  //输出直方图  End |

#### 6.2.1.3 基于差分隐私的密度聚类算法

基于分区和基于层次的聚类算法是把距离作为两个样本或者两个簇之间相似度的评价指标，因此这两种聚类算法指在发现球状簇，对于任意形状的聚类簇表现则相对较差。而基于密度的聚类算法可以发现任意形状的簇，且在假设聚类结构能够通过样本分布的紧密程度来确定，针对于噪声点也可以进行有效处理。常见的基于密度的聚类算法有DBSCAN、OPTICS算法等。

Wei-Min[26]等人将基于密度的聚类算法与差分隐私相结合，提出基于差分隐私保护的DP-DBSCAN聚类算法，将足够高密度的区域划分为簇，并在具有噪声的空间数据库中发现任意形状的簇。其中，Eps表示对象半径；MinPts表示给定对象半径范围内的样本点数。算法的主要思想如下：

|  |
| --- |
| 基于差分隐私的DP-DBSCAN算法 |
| 1假设和是*n*维空间数据集*D*中直接密度可达的两个点  2针对每个点*x*和*y*  3 计算*x*和*y*之间的点密度：*dis=*  4 分别在各维中添加随机噪声，使得*dis`=*，其中，  5 重复上述过程直到所有点都已包含在任何一个簇中或被标记为噪声 |

该算法虽然能够有效地保护个人隐私，且适用于不同规模及不同维度的数据集。但是由于DBSCAN算法本身存在对参数输入敏感的问题，因此会影响聚类的效果及噪声添加，进而影响数据的实用性。基于此Wang H[27]等人提出DP-OPTICS算法，该算法通过对稀疏型数据集进行压缩处理，确定隐私预算的上界，有效平衡了敏感信息的隐私信和数据可用性之间的联系，有效地解决了DBSCAN对参数输入敏感的问题。ALC Mendonca[11]等人提出了DiPCoDing算法，该算法用来计算记录之间的相关性在统计数据库中的方式。算法结合了基于差分隐私的DBSCAN聚类和高斯混合模式，旨在分组相似的记录，以减少差分隐私的敏感度，从而减少添加到查询答案的噪声量，以提高数据的实用性。

#### 6.2.1.4 基于差分隐私的网格聚类算法

基于密度的聚类算法虽然能发现任意形状的簇，但这类算法一般时间复杂度都比较高。而基于网格的聚类算法可以有效地减少算法的计算复杂度，但不可避免地，这一类算法同样对密度的参数选择较为敏感。基于网格的方法，一般是把对象空间量化为有限个数目单元，从而形成一个网格结构，所有的聚类操作都将在这个网格结构上面进行。基于网格的方式最大的优点是处理速度快，且处理速度与数据对象的大小无关的，只与量化空间中每一维的单元数目有关。常见的算法有STING、CLIQUE、WAVE-CLUSTER算法等。

L Chen[29]等人提出了四个基于差分隐私的小波聚类算法，这些算法采用一种通用的技术，使得在不同的小波变换，如哈希变换、双正交变换下，实现小波域上的差分隐私。文章首先提出了一种基线算法(Baseline)，算法通过在数据发布上实现差分隐私实现波形差异隐私，但是在大多数情况下，它不能产生精确的波形聚类结果。然而当该方法用于生成合成数据集时，噪声负计数被重置为零计数，导致数据分布从根本上彻底改变，并且进一步导致差分隐私波形结果的严重偏差。文章接着提出私人量化算法(Private Quantization)，即在关键步骤上执行差分隐私，但算法在量化步骤中要选择适当的网格大小，网格大小会影响波形聚类结果的精确度，进而影响差分隐私波形结果。因此文章相继提出两种改进算法。改进之一的PrivTHR算法提出了分步加噪方式，修正了因为小波变换而改变的值，提高了算法的效用性。改进之二的PrivTHREM算法则采用指数机制来从排序列表中导出代替中心点的替代技术，也可以很好地提高算法的聚类效果。

基于文献[29]进一步提出改进，在量化阶段提出压缩稀疏行的存储方式，有效节省了空间开销；在簇识别阶段用广度优先搜索算法代替以往的连通方法，提高了聚类的速度。改进的算法流程如下所示：

|  |
| --- |
| 算法 *-* |
| 输入：数据集，压缩行数组，数组大小，网格数目，密度阈值，小波变换序列*W*，隐私预算，预算分配因子  输出：一组差分隐私下的簇  步骤：  1. -  2. //获得引入Laplace噪声的计数矩阵  3. //构建压缩行存储空间，创建只储存非空单元的查找表  4. //小波变换得到  5. //得到只保留正值的序列  6. //统计零值数据并加噪  7. //丢弃以后的值，以减少加噪后的重要网格个数的失真  8. //根据密度阈值得到参数  9. //获得序列中的前个值作为输出  10. //广度优先算法得到重要网格  11. //将重要网格构成簇输出  12.  13. //保证算法整体满足差分隐私  14.  15.  16.  17. //得到中的零值个数  18. //零值个数的噪声估计  19.  20. |

该算法在原算法的基础上，时间复杂度与空间复杂度均降为，且比原算法更高效，且有效地解决了内存空间占用率较高，聚类精度差等问题。

#### 6.2.1.5 基于差分隐私的模型聚类算法

基于模型的聚类算法主要是指基于概率模型的方法和基于神经网络模型的方法，尤其以基于概率模型的方法居多。这种方法的优点是不再针对于类而精准地进行划分，而是以概率的形式表现出来，即每一类的特征用参数来进行表达；但因此也会导致必然的不足，如针对分布数量较多但数据量很少的时候，此类算法的执行效率不高，其中，基于模型的聚类算法中最具典型就是高斯混合模型（GMM，GaussianMixtureModels）。

目前，基于差分隐私的模型聚类算法并没有单一地去研究，基于模型的聚类方式通常与其他形式结合起来。如ALC Mendonca[28]等人提出了DiPCoDing算法，就是结合了基于差分隐私的DBSCAN聚类和高斯混合模式，显著减少添加到查询答案的噪声量，提高数据的实用性。Wu Y[30]等人提出了DPGMM算法，即一个在差分隐私下使用高斯混合模型的参数密度估计算法。该算法在每次迭代的M步得到估计的参数之后，添加两个额外的步骤，以实现差分隐私和精确计算。

#### 6.2.1.6 基于差分隐私的混合聚类算法

随着数据挖掘任务的数据集规模化、数据类型复杂化、数据特征高维化等情况的出现，算法需要各方面性能更优的挖掘算法来支持，进而出现了很多改进的混合聚类算法。通过组合不同类型的聚类算法，可以实施分阶段的聚类，或者是在组合中实施对算法的执行约束条件、迭代控制机制等操作，以实现各种复杂数据挖掘任务的聚类要求。

针对于差分隐私的混合聚类算法，Chen T[31]等人提出了一种结合匿名化理论的聚类差分隐私保护方法。该算法首先利用DSPCAN密度聚类算法对原始数据集进行初始化，实现第一次聚类，然后结合微聚合技术来实现数据集的二次聚类。最后在聚类数据记录中加入拉普拉斯噪声，对不满足差分隐私模型的数据进行校正，保证了数据的有效性。Gao Z Q[32]等人提出了一种新颖的、有效的混合k-均值聚类，在SARK中保护差分隐私，即差分隐私混合k-均值（DPHKMS）。在大数据计算平台Apache Spice的框架下，我们将粒子群优化算法和布谷鸟搜索相结合，更好地实现了聚类中心选择。此外，在Laplace机制下实现了DPHKMS，并在理论上证明了在确定的隐私预算分配下满足差分隐私。

### 6.2.2 频繁模式挖掘

频繁模式主要包括频繁项集、频繁子序列或子结构，而频繁项集是频繁模式中最主要的研究内容。是生成关联规则的重要一步，也是其他频繁模式的依据。

频繁项集挖掘最初应用于事务数据库中发现关联规则，并没有考虑数据记录内项之间的关系，是最简单的频繁模式挖掘类型。[33]方法是差分隐私下方法的典型代表。该方法借鉴截断频率思想提出了2种满足差分隐私的频繁项集挖掘策略。挖掘策略如表1所示：

表6-1 TF方法的2种挖掘策略比较

|  |  |  |
| --- | --- | --- |
| 步骤 | PT-Exp策略 | PT-Lap策略 |
| 1 | 以概率从C中无放回执行k次选择操作得到top-k频繁项集 | 为c中所有项集添加大小为拉普拉斯噪声后重新获取top-k频繁项集 |
| 2 | 对k个项集的真实计数分别添加大小为拉普拉斯噪声 | 对k个项集的真实计数分别添加大小为拉普拉斯噪声 |

在 Dwork[34]、Mc Sherry[35]等人提出了差分隐私，拉普拉斯机制以及指数机制等概念之后，回顾历年来差分隐私下频繁项集挖掘方法对比分析详见表6-2：

6-2表 差分隐私下频繁项集挖掘方法对比分析

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 方法名称 | 主要技术 | 主要优点 | 主要缺点 | 分配 | 算法性能 |
| TF | 截断频率 | 实现简单，精度高 | 处理较大k值或l值时性能与效率比较差，未考虑记录本身长度带来的影响 | 指数机制、拉普拉斯机制，均分为两份，平均分配 |  |
| Privbasis | 基映射 | 性能和效率优于TF | 难以兼顾隐私保护与模式可用性的不足，未考虑记录本身长度带来的影响 | 拉普拉斯机制，平均分配 |  |
| DP-topkP | 一致性约束 | 一致性约束输出结果，精度较高 | 处理较大k值或l值时性能与效率比较差，未考虑本身长度带来的影响 | 指数机制、拉普拉斯机制，  均分为两份，平均分配 |  |
| SmartTrunc | 事务截断 | 规约数据集，精度高 | 扩展性比较差 | 拉普拉斯机制，平均分配 |  |
| DiffPart | 分类树划分 | 发布精度高，扩展性高 | 仅支持计数查询，没有考虑过不同项之间的语义关联 | 拉普拉斯机制，平均分配，自适应分配 |  |
| DiffMR | 一致性约束 | 支持MaprReduce分布式查询，实现简单 | 只支持计数查询，处理较大K值或l值时性能与效率较差 | 指数机制、拉普拉斯机制，  线性分配 |  |
| DAT | 事务截断策略 | 发布精度高，扩展性高 | 仅支持计数查询，没有考虑过不同项之间的语义关联 | 拉普拉斯机制，平均分配 |  |

频繁项集挖掘是数据挖掘技术的一个重要的分支，并且频繁项集挖掘技术已经被应用到了多个领域，诸如推荐系统，生物信息分析等。下图是频繁项集挖掘所面临的一个重要的挑战就是数据的不确定性。导致数据不确定性的原因很多，例如实验误差，噪声和数据的不完整性。传统的基于确定数据的频繁项集挖掘算法很难直接被应用于挖掘不确定数据的频繁项集。因此，基于不确定数据的频繁项集挖掘算法已经越来越受关注。针对记录用户隐私信息的不确定数据不被攻击者所窃取，为了确保从确定数据集中挖掘长度为  的前个最频繁项集满足差分隐私，在2015 年，Maruseac [36]等人在差分隐私理论和抽样理论的基础上提出了挖掘高置信度关联规则算法。在2016 年，Xu等人[37]提出了基于差分隐私的频繁项集挖掘算法。

为了从不确定的数据集中挖掘出基于期望支持度的前个最频繁的频繁项集，并且保证挖掘结果满足差分隐私，Ding等人[38]提出 FIMUDDP 算法（frequent itemsets mining for uncertain data based on differential privacy）。FIMUDDP 利用差分隐私的指数机制和拉普拉斯机制确保从不确定数据中挖掘出的基于期望支持度的前 K 个最频繁的频繁项集和这些频繁项集的期望支持度满足差分隐私。该算法一共分为两个阶段：第一个阶段中，FIMUDDP 会得到不确定数据集的基于期望支持度的前K个最频繁的频繁项集。在这一步，差分隐私的指数机制被应用去确保输出结果满足差分隐私。在第二阶段，利用差分隐私的拉普拉斯机制得到不确定数据的前K个频繁项集的噪声期望支持度，从而确保基于不确定数据的前K个最频繁的频繁项集的期望支持度满足差分隐私。令FIMUDDP 的总隐私保护预算为。和是分配给 FIMUDDP第一步和第二步的隐私保护预算。设=0.5。最终利用6 个现实世界的数据集和2 个动态生成数据集，验证了FIMUDDP 的有效性。

频繁序列挖掘其目的是找出相对时间或其他顺序频繁出现在数据集中的模式，是关联规则、相关性分析、分类、聚类等多种数据挖掘任务的基础。随着大量序列数据的收集和存储，频繁序列模式可以为诸如市场分析、需求预测等数据分析任务提供大量有价值的信息。然而，频繁序列模式本身和其相应支持度计数都有可能泄露隐私信息。

在2011年，Chen等人借鉴DiffPart[39]方法的树划分思想，第一次提出了结合自顶向下前缀树(prefix tree)发布轨迹数据集的方法STM-Full，支持频繁序列挖掘。为了弥补STM-Full 方法的不足，文献[40]提出了一种基于变长n-gram模型发布序列数据集的方法N-gram。为了有效提高序列可用性，文献[41]提出了一种两阶段方法PT-Sample，该方法以字符串序列数据集为例，采用字符串指纹匹配方法对原始数据集执行分箱离散化，极大地降低候选频繁序列集合的规模，有效提高了top- *k* 频繁序列的准确性。差分隐私下频繁序列挖掘方法对比分析详见表6-3:

表6-3 差分隐私下频繁序列挖掘方法对比

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 方法名称 | 主要技术 | 主要优点 | 主要缺点 | 分配 | 算法性能 |
| STM-Full | 前缀树划分 | 支持多维数据依赖计数查询，精度高 | 忽视轨迹时间戳信息，实用性差，前缀树冗余 | 拉普拉斯机制，线性分配 |  |
| Prefix-Hybrid | 混合粒度前缀树划分 | 支持多维数据依赖计数查询 | 忽视轨迹时间戳信息，实用性差，前缀树冗余 | 拉普拉斯机制，划分为两份，混合分配 |  |
| N-gram | 变长N-gram模型 | 支持多维数据依赖计数查询，精度高 | 容易受到序列维度影响，有一定信息丢失 | 拉普拉斯机制，自适应分配 |  |
| PT-Sample | 数据变换和推理求精 | 频繁前缀树和序列模式挖掘，精度高 | 数据变换过程中存在信息丢失 | 拉普拉斯机制，划分为两份，混合分配 |  |
| PFS | 数据集采样策略，事务长度限制策略，阈值下降策略 | 支持多维数据依赖计数查询，精度高 | 序列树木的增加导致噪声规模的增大 | 拉普拉斯机制，划分为五份，混合分配 |  |

针对之前算法在重构序列数据库时未考虑不同候选序列重要程度的差异性，导致算法准确度低，挖掘结果效用性差。李艳辉等人[42]提出一种满足差分隐私(DP)的频繁序列模式挖掘(DP-FSM)算法。该算法主要从两个方面提高挖掘结果的准确度。首先，基于候选集中不同候选序列重要程度的差异性，提出了一种启发式的方法设计打分函数来区分不同候选序列的重要程度，从而能在序列截断时使具有高优先级的候选序列优先保留; 其次，针对挖掘频繁序列模式的误差问题，提出一种阈值修正策略，通过判断某序列是否频繁和判断某序列是否用于产生候选序列而设置不同的阈值以减小误差。

频繁子图挖掘在许多实际应用中具有巨大的潜在价值，如语义网络、行为建模、生物网络分析、化学化合物分类和大分子分析等。给定图,,当且仅当且,称图为图的子图。如果两个图和具有相同的拓扑结构，则称这2个图是同构的，由于子图同构问题是NPC 问题，使得子图计数操作非常困难，进而导致在频繁子图挖掘过程中输出域无法直接得到。因此，如何快速有效地构建输出域是差分隐私下的频繁子图挖掘亟需要解决的问题。

肖恪等人[43]提出了满足差分隐私的频繁子图挖掘算法Naïve, Naïve在FSG算法的基础上引入拉普拉斯噪音对子图的支持度进行扰动，得到噪音支持度，通过噪音支持度与阈值比较得到频繁子图。Naïve算法在预处理阶段使用二分估值法得到最大子图，之后算法扫描数据库，得到候选１子图集合。之后在候选１子图的支持度上添加符合分布的噪音得到噪音支持度，通过对比噪音支持度和阈值，得到频繁１子图集合。之后挖掘频繁子图时，首先根据频繁子图集合得到候选子图集合。对中的每个子图的真实支持度上添加的拉普拉斯噪音，选出噪音支持度等于或大于阈值的子图作为频繁子图，之后重复这个过程，直到得到所有频繁子图。

## 6.3 面向监督/半监督学习的差分隐私（章国政）

### 6.3.1 分类

分类的目的是使用分类函数或者分类模型将数据库的数据项映射到不同的类别中去。分类计数可以用来提取重要的数据模型并预测数据趋势。待测数据集通过分类模型计算，得到一个结果。然后通过这个结果判定这条待测记录属于给定类别之中的某一个。分类在预测金融、市场、健康状况等未来趋势时有非常重要的作用。分类的第一步是输入训练集。训练集由样本组成，每一条记录可以表示为一个特征向量，特征向量的每一个维度的值可以代表一个属性，剩下的最后一个维度代表类别标签。分类时，输入进知识库的是已知类别的样本集。样本向量形式化定义为：。其中表示属性，y表示类别。分类的第二步是分析输入数据，找到样本的特性，形成分类模型。分类模型一般用函数的形式来表示。然后将待测目标代入函数当中进行计算，得到计算结果，然后通过对计算结果的分析来判断待测目标应该属于哪个类别。这就是分类预测作用的体现，可以判断未知类别的待测目标。尽管预测的正确率不能达到百分百，但是对大多数待测目标能够保证有一个相对合理的估计。

分类方法主要有五种，基于决策树的分类算法、基于SVM的分类算法、基于K-近邻的分类方法、基于贝叶斯的分类方法、基于关联规则的分类方法。

**基于决策树的分类算法：**决策树（Decision Tree）是常用的分类模型。ID3是经典的学习算法，C4.5是对ID3算法的改进，它通过计算信息增益（Information Gain）寻找具有最大信息量的属性，构造决策树。SuLQ-based ID3[44]实现了差分隐私保护。它主要思想是在每次计算属性的信息增益时，加入Laplace机制的噪声计算值，生成决策树，但加入噪声会导致预测准确率的下降。文献[45]对SuLQ-based ID3进行了改进。该算法使用Partition算子将查询数据集分割成不相交的子集，利用其计算时并行组合性的特点，提高隐私保护预算的利用率，该算法直接利用噪声计数值评估信息增益标准，在使用ID3算法生成决策树。由于计算信息增益的计数值需要对每个属性进行计算，所以需要将整个隐私预算分配到每次查询中，导致每次查询的隐私预算较小，当数据集较大时将引入了大量噪声。为了解决以上问题，Friedman和Schuster基于指数机制提出了Diff-ID3[46]，指数机制在一次查询中同时评估所有属性，减少了噪声和隐私预算的浪费。但该算法是用来处理离散数据的。为了处理连续型数据，Friedman和Schuster进一步提出了DiffP-C4.5[46],指数机制可以挑选属性分割点。该算法在处理大量查询时分类精度有所降低。DiffGen[47-48]算法通过信息增益或者子类中最高频类对候选集打分，利用指数机制选取候选集。然后自顶向下把数据记录划分到子类中去，最后对每个子类添加Laplace噪声。DiffGen算法在属性类型较少时，能够获得满意的保护效果。但当属性类型较多时，分类树过大，造成算法效率下降。DT-Diff[49]类似Diff-Gen算法，也使用了泛化技术和指数机制对属性进行分割，但提出了特性模型方选择策略，通过构造属性模型，对记录进行分组，然后添加噪声发布结果。DT-Diff能够充分利用隐私预算，提高分类准确率。但隐私预算分配的主观性比较大。文献[50]将RDT（Random Decision Tree）用于分布式数据挖掘，提出了水平和垂直类型数据的并行处理算法。文献[51]通过实验证明打分函数，例如信息增益、最大算法和基尼指数只能提供相同的精度。还通过实验证明了使用差分隐私保护的随机森林可以获得与经典的随机森林同样的精度。文献[52]为RDT中的差分隐私提供了理论证明和一个通用的RDT差分隐私算法。近几年的学术研究表明基于差分隐私的RDT相比于传统的差分隐私决策树，精度和效率都得到提升。文献[53]和文献[54]都提出了一种面向随机森林的差分隐私保护算法，前者通过随机森林算法在建立决策树的过程中，先用指数机制对属性选择和属性划分过程进行扰动，然后向叶子节点添加Laplace噪音以满足差分隐私的需求。后者在每一棵树决策树的构建过程中采用指数机制选择分裂点和分裂属性，并根据Laplace添加噪声。这两种方法都在满足差分隐私的基础上便捷地实现了分类并保持了较高的分类准确度。

**基于SVM的分类算法：**SVM（Support Vector Machine）是一种常用的分类技术，基于统计学习理论，主要解决小样本、非线性及高维模式识别。核函数是SVM的关键，低维空间向量难于划分，而通常的解决方式是将其映射到高维空间，而映射过程本身会增加计算复杂度，引入核函数能够解决这个问题。SVM的输入训练集和目标函数最优解都有可能泄露敏感信息。文献[55]提出了PRIVATE SVM算法，首先使用SVM算法、核函数和损失函数计算空间向量，其中是2-纬特征映射，。最终得到响应向量。为噪声参数，其中，正则参数为近似参数，n是数据集大小。加入Laplace后的噪音能够满足差分隐私保护要求，并且计算复杂度在最坏时为，当数据集较大时，算法时间复杂度较高，噪音大。文献[56]针对机器学习中的隐私泄露问题，提出了Objective SVM算法。虽然损失函数是连续但不可微分的，可以采用双重可微函数（凸函数）或者Huber Loss 解决，对目标函数加入Laplace噪声，得到扰动目标函数。该算法只能用于特定的目标函数，不具有通用性。

**基于K-近邻的分类方法：**K近邻方法基于类比学习。这种分类方法不包含参数。它的有效性基于统计，主要是模式识别方向。这种方法可以从理论上精确证明出它的准确率服从正态分布。这种方法的思想是：首先定义相似度，最简单的方法就是把相似度定义成为欧氏距离。然后算待测与样本的相似度。把相似度进行降序排列，选择前K 个。然后分析这K 个样本属于哪一类最多，然后就把待测目标归为那一类。但该方法存在一定局限性。尤其当样本数量比较大而且维度比较高的时候。第一，系统中存储着太多样本，很多样本甚至在计算之后被证明对分类不构成任何影响。在面对待测目标的时候，如果把相似度定义为欧氏距离，那么随着维度的增加，相似度的值也会越来越大，给系统运行造成不便。第二，如果把相似度定义成为欧氏距离，也就是默认了各个维度对分类的影响度相同。现实生活中，可能会有极少的情况符合这种假设，在更多的情况下，这种假设是不成立的。因此该方法准确率有待观察。研究人员进行了探索，终于提出了几个改进方法。这些改进方法的中心思想无一不是缩小样本集。缩小样本集的方式主要就是对样本集进行预处理，当然，这种方法只需要用在当样本集比较大的情况下。通过对样本集的预处理可以让系统不再存储那么多不必要的样本，提高分类的效率。目前最主要的方法有：产生原型法、修改原型法、编辑处理法、多重分类结合法 。

**基于贝叶斯的分类方法：**贝叶斯分类是统计分类方法，这种方法的核心是利用概率统计对待测目标进行分类。这种方法巧妙地运用了贝叶斯定理来对待测目标的属性进行预测。在分类时，利用贝叶斯公式分别计算待测目标属于样本类别的概率，然后分析到最大的概率并将待测目标划入那相应的类别当中。但贝叶斯方法也存在缺点，因为这种方法假设属性与属性之间相互独立。现实生活中，可能会有极少的情况能够符合这种条件。但是在更多的情况下，这种假设是不成立的，分类准确率因此受到影响。研究人员对朴素贝叶斯分类方法进行了改进，提出了 TAN 方法。TAN 方法的第一步就是探索属性之间的相关性。从而能够减少假设的计算。这种方法的思想其实并不复杂。就是在朴素贝叶斯的基础上增加了对属性相关性的考虑。这种方法的具体实现过程就是：第一步，构造一个图。第二步，将相互关联的属性节点用有向边连接。第三步，子节点为属性，根节点为类别建立属性与类别的相关关系。举例来说，假如属性存在一条指向属性的边，这就表明能够改变对类别的影响。TAN方法也存在局限性。这种方法单纯地把两个属性间的相关性放入模型进行计算。但多个属性之间的相互影响却没有被考虑进去。贝叶斯网络方法和 BAN 方法很好地解决了这个问题。

**基于关联规则的分类方法：**挖掘关联规则就是发现大量数据中项集之间有趣的关联或相关联的过程。关联规则挖掘用于分类问题取得了很好的效果。要找到关联规则所需频繁项集有很多种方法。应用最广泛也是最著名的方法那就是Apriori 算法，该方法是单维单层布尔关联规则。这种方法的核心是迭代时逐层搜索方法被其利用。用 k-项集搜索(k+1)-项集。第一步，制定1L，在其中装满 1-项集合。第二步，制定 2L，2L中装满被1L找到的2-项集合。第三步，制定 3L，以此类推。直到不能找到频繁 k-项集。扫描数据库一次就能找到一个 k-项集。只需要两步就能从k-项集获得(k+1)-项集：链接和剪枝。最后一步是分析库中每个项的支持程度。这一步可以通过分析数据库完成。在这一步当中，支持度较低的项已经被自动剔除。所以，可以这样说：通过上面几步，能够找到所有的频繁项集。目前来讲，Apriori 算法适用范围非常广泛。它几乎能够涵盖所有高效的关联规则方法。Apriori 算法在数据交换时有一个验证加入的环节。这个环节存在剪枝过程当中。这个环节是算法性能的缺陷。因为这个环节的存在，可能会要求多次扫描大数据库。很多候选集也因此而产生。再有就是有重复分析数据库的隐患。这两点是 Apriori 算法的两缺点。因此，经过改进后的 LIG方法被研究人员所提出。这种方法计算 1-项集不单单只是计算 1-项集。对应项的相关区间也被给出。这样在第二步计算 2-项集时的范围就会被缩小。而在此基础上提出的 FP 算法干脆直接利用频率增长。分析频繁项集的方法被放弃。这种方法的第一步是分析频繁项并且对其排序。然后构造 FP 树进行规则挖掘。这种方法的优点是：第一，在挖掘过程中不会遗漏关键信息也不会掺入冗余信息。第二，第二算法效率比 Apriori 提升将近一倍。但是这种方法也有缺点。那就是因为 FP 树结构受到制约，所以不能处理大规模数据集。

### 6.3.2 线性回归

线性回归是用于建模因变量与若干独立变量之间的线性关系的基本统计方法。 它已被广泛用于包括拟合预测模型和分析因变量与一个或多个自变量之间的关系等实际应用中。

线性回归模型的输入是一个数据集，其中是一个d维实向量，是实数标量值。 输出结果是一个d维向量。最优化目标函数是

 (6-1)

换句话说，线性回归表示d-维值的线性回归函数的结果值y，使得预测的y值的平方误差的总和最小化。

线性回归的全局敏感度是无界的。例如，给定一个数据集，其中x是一维数据，包含N-1个（0,0）点以及一个（1/N,0）点。最优线性函数为。向数据集中添加一个额外的点(1,N)，结果最优线性函数变为。因此，根据全局敏感度给线性函数的参数添加噪声会完全丢失数据所有可用性。

为了解决现有的回归分析解决方案要么限于非标准类型的回归要么无法产生准确的回归结果的问题，zhang等人[57]提出了一种函数机制，其主要思想是通过扰动优化问题的目标函数而不是输出结果来保证敏感信息的泄露问题，并通过函数机制向线性回归模型中的模型参数直接添加扰动，使线性回归过程满足精确差分隐私。Kasiviswanathan等人[58]在流数据方面考虑了私有增量回归，将持续发布与ERM（Empirical Risk Minimization）技术相结合，分析了几种算法的风险界限，以线性回归模型为例，他们不断更新梯度函数的噪声，以最小化损失函数。

### 6.3.3 逻辑回归

逻辑回归也是学习线性系数向量的方法;通过使用逻辑函数根据这些系数的内积和一个数据点的独立变量来估计因变量的概率。

输入数据集为，其中是一个布尔域为的d维实向量。输出y=1且概率为的预测函数。模型参数是通过使最优化目标函数值最小得到的。

最优化目标函数： （6-2）

其中，损失函数为，是正则化参数。

文献Chaudhuri and Moonteleoni[59]表明，逻辑回归上的输出摄动方法的敏感度大小为，其中是正则化参数，N是数据集大小。这意味着当不使用正则化参数时，此敏感度边界将变为。

Chaudhuri[60]等人提出了一种满足差分隐私保护的logistic回归算法。首先，他们证明了如果直接利用Laplace机制，在输出的回归模型上加入噪声，其分类准确度会随着正则化参数的减小而降低。所以他们提出了一种新的算法，先将噪声加在目标函数的参数中，然后通过标准的logistic回归算法进行参数估计来获取模型。此模型相对于直接利用Laplace机制，具有更高的分类准确率。Chaudhuri等人[61]将文献[60]的思想拓展到正则ERM中损失函数和正则因子满足特定的凸函数特性和可微特性的隐私保护问题上，提出了一种为正则ERM机器学习算法设计的隐私保护的新方法—目标函数加扰。并提出了在通用的机器学习算法中调节参数以保护隐私的方法，从而提供了衔接各个训练过程的隐私保障。应用这些成果模拟生成隐私保护的正则化逻辑回归和支持向量机，在权衡了隐私保护和学习效果方面的综合表现都优于现存的输出加扰方法。

### 6.3.4 支持向量机

支持向量机作为数据挖掘的代表性算法，它的基本模型是定义在特征空间上的间隔最大的线性分类器，利用核函数将输入从输入空间映射到特征空间，并在特征空间里建立有一个最大间隔超平面进行分类操作。

近年来，关于支持向量机的差分隐私技术有了很大的发展，从[62]中的经验风险最小化的隐私保护到[63]中对于支持向量机的算法的本身的输出函数进行了隐私保护，都是最基本的隐私保护算法。但是对于多分类算法来讲，很少有差分隐私框架的支持。Lin等人[64]提出的PPSVC算法，主要利用后处理的方法破坏支持向量的属性值，然后通过泰勒公式转换并输出一个近似于原始支持向量机分类器的属性充当原始值; Li等人[65]提出了关于RFB核的差分隐私支持向量机算法，主要改动思想是在添加拉普拉斯噪音之前，先对决策函数先进行转换;针对分布式系统下的众包数据分类应用，Sun[66]等人提出P3SVM算法，主要基于全局随机简化核，通过高斯扰动计算全局随机矩阵，对垂直数据进行分类，该算法主要用于生成仅需要单个线性方程组解的线性或非线性分类器；Yu[67,68]等人也提出了PPSVM算法，从分布在多方水平分区数据和垂直分区数据中构建全局支持向量机隐私保护模型，该模型的优点在于不会将每方的数据公开给他人。

**支持向量机的隐私保护**

对于支持向量机来说，本质是基于经验风险最小化来实现的凸函数问题。目前，支持向量机大体上分为线性与非线性两种，对于线性支持向量机，它的计算部分类似于一个单层的神经网络，本质上可以看作一个矩阵乘积。而对于非线性支持向量机，它主要采用了核机制，并结合拉格朗日乘子法与KKT条件，使数据样本能够在再生核希尔伯特空间可分。

该章节主要针对非线性的支持向量机情况下的分类问题，目前对于支持向量机的隐私保护方式思路主要有以下三种：

第一种思路主要是给输出直接添加干扰。作为最原始的隐私保护方式，通常直接给予法向量w直接进行噪音的添加，即 。这种方法虽然简单方便，但是可能在某种情况下，因为法向量的敏感度过高可能会导致相当大的噪音量，对于分类精度来讲会造成相当大的影响，因此，这种方式不作考虑。

第二种思路主要给目标函数添加干扰。通过对目标函数内部的某一部分进行噪音的添加，来完成隐私保护工作。即，其中w`为添加噪音后的结果。这种方式从整体效率上来看优于前一种方式。尤其在多分类情况下的数据集，能够取得较好的分类效果。

第三种思路主要给输入输出进行干扰。这种方式主要针对的是平移不变核的核函数，目的是对输入数据进行一定的转换，来达到隐私保护的目的。比如通过对高斯核函数进行傅里叶转换后得到 ，然后在进行相应的运算。

由于第一种思路的分类精度较低，无法满足实验数据的可用性问题，因此将分别利用第二种和第三种思路提出对应的隐私保护算法。算法的主要思想是通过对法向量*w*的扰动，来达到隐私保护的目的，而法向量的生成，则需要依靠相应的核函数进行计算。因此，首先要确定分类算法所需要的核函数。

目前最常用的可能会用到三种核函数。第一种核函数为线性核，具体表现形式为。对于类别多的大数据问题，线性分类器的分类速度将会有很大的优势，而且也具有很高的分类准确率。第二种核函数为多项式核，具体表现形式为，在模型训练规律波动较大的情况下，可以得到较好的分了效果。第三种核函数为高斯核，具体表现形式为。作为平移不变核的一种，具有将样本从低维度映射到高维空间的特性，可以从理论上将样本数据分类成为无限维。

接下来以高斯核为例，简单讲解几种算法的实现。

从第二种思路角度来看，主要通过对核函数进行拉普拉斯噪音的添加来实现，即，通过干扰核函数的方式间接的干扰法向量*w*的取值。

从第三种思路的角度来看，通过利用泰勒公式****对高斯核函数进行部分替换，以求取得核函数最终的最相似解，具体核函数分解式如下：

 （6-3）

而最终得到的法向量公式如下：

 （6-4）

这种方法通过对法向量*w*计算公式中的核函数进行泰勒公式的转换，来得到法向量*w*的最相似解。

为了进一步增强隐私保护效果，我将上述两种思路进行结合，改进了支持向量机隐私保护算法，即对泰勒公式转换后的核函数添加了Laplace噪音，希望能够得到数据更高的隐私保护。算法的表现形式为：

 （6-5）

从目前的实验效果来看， 虽然进一步增强了隐私保护，但是对于整体的分类精度并没有造成太大的干扰，符合实验的预期。

## 6.4 面向深度学习的差分隐私（陈竑毓）

近年来，深度学习成为机器学习中最成功的方法之一，深度学习的概念是在2006年由Hinton等人提出，其基本思想是应用一个多层架构从高维数据中提取特征，并以此来解释数据的隐含规律。含多隐层的多层感知器是一个基本的深度学习模型，可很好地解决非线性问题。在此基础之上，针对不同问题，出现了例如卷积神经网络、循环神经网络、递归神经网络等多种网络模型。当前，深度学习已经被广泛应用于图像分类，语音识别，自然语言理解，社交网络分析，生物信息学、语言翻译和搜索引擎等多个领域。

深度学习在多个领域都取得了非常惊人的成果，并且通过移动终端和互联网开始全面影响人类的工作和生活，但是其带来便利的同时，也存在着隐私安全问题。例如，文献[69]指出企业收集的用户图片、语音和视频数据在用于共享目的时会导致隐私泄露，基于一种恶意攻击手段，利用模型的输入与输出可以造成隐私泄露；文献[70]提出一种模型反演攻击，通过面部识别系统可以恢复图片；文献[71]假设攻击者不但可以使用深度学习模型，而且具有模型参数、模型结构等模型全知识。针对上述隐私问题，可以采用差分隐私解决方案，采用深度学习加差分隐私的结合技术，一方面可以将模型应用于具体的分类、聚类、预测等实际任务，另一方面能够保证数据隐私不被泄露，做到数据效用与隐私保护的平衡。但是将扰动直接作用于模型参数会导致模型性能变差，甚至不可用，因此采用扰动训练过程的方式来实现差分隐私是一种较好的方法。

本节先介绍基本的深度学习模型及其优化方法，之后通过扰动训练过程的方式提出两种隐私保护解决方案：基本的拉普拉斯方法及其改进方法。

### 6.4.1 深度学习模型

深度学习本质上是构建含有多个隐藏层的神经网络架构模型，通过大规模数据进行训练，得到大量更具代表性的特征信息，从而对样本进行分类和预测，提高分类和预测的精度。深度网络结构包含多层，其中每一层由不同数量的神经元组成。每层中的神经元接收来自前一层的有限数量的神经元输出及其相关权重。训练过程的目的是调整这些神经元之间的权重以适应训练样本，实际训练过程中常采用梯度下降（Gradient Descent，GD）方法及其变体来实现这一目标，这些优化方法将在下一小节进行介绍。

假设是一个训练集，其中代表每个样本的标签。深度学习模型采用多层网络流行架构，包括一个输入层，一个输出层，多个隐藏层，图6-1是一个典型的含3个隐藏层的多层结构。其中底层是模型输入层，用于处理样本输入，中间是三个隐藏层，用于提取输入的抽象特征，最上层是网络的输出层，用于输出样本结果，例如分类标签。

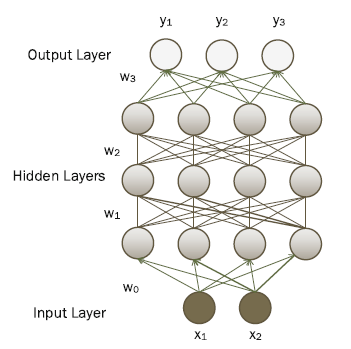
图中每一层的一个圆代表一个神经元，每个神经元接收来自前一层有限数量的神经元的输出及连接权重。每个神经元通过对输入值应用非线性激活函数来计算输出，假设当前神经元接收来自上一层四个神经元的输出结果作为输入，表示为，与四个神经元的连接权重用表示，则当前神经元的输出表达式为，其中表示激活函数。图6-2展示了一个神经元的输入与输出过程。

图 6-1 深度模型结构

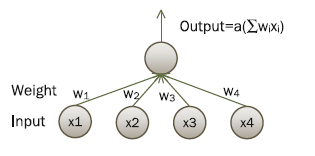


图 6-2 神经元的输入与输出

激活函数通常是一个用于转换输入的加权和的非线性函数，常用的有Sigmoid激活函数、tanh激活函数和ReLU激活函数。具体地，

Sigmoid激活函数： （6-6）

tanh激活函数： （6-7）

ReLU激活函数：

 （6-8）

Sigmoid函数能把输入的连续实值压缩到0到1之间，而tanh激活函数是Sigmoid函数的变形，其输出范围是[-1, 1]。ReLU激活函数则是一个分段的线性函数，当输入小于0时，输出一直为0。

上文所述是一个典型的全连接神经网络，其特点是当前神经元与上一层所有神经元相连。全连接神经网络针对具体的应用场景会出现其固有缺陷，因此针对不同的需求，出现了不少其他的深度学习结构，例如针对图像处理问题，全连接神经网络会导致参数数量过多，因此提出了卷积神经网络；在序列预测及文本翻译问题上，全连接神经网络不能处理前后输入之间的关联，因此提出了循环神经网络；其他还有长短时记忆模型、GRU、递归神经网络等结构模型。

在深度学习模型设计中，连接权重是由目标函数来确定的。深度学习的训练过程其实是一个优化连接权重的过程，通过不断缩小目标函数的损失来实现对权重的优化更新。对于目标函数有几种不同的定义，最普遍的是在训练样本集上的平均损失，表示为。正因为目标函数的非凸特性，常采用梯度下降方法来计算连接权重。

### 6.4.2 梯度下降及其变体

在深度学习模型中，需要调整的参数即连接权重可能有成千上万个，因此权重的学习过程可以被认为是一个非线性的优化问题，其中目标函数用于衡量问题的误差程度。当下最流行的调整权重的算法之一是梯度下降算法，其基本思想是对于一个给定的权重，如果在其上增加一个微小增量，则可以使用一个梯度向量来表示误差会增加或减少的量，然后在反方向调整该权重。具体地，用公式（6-9）解释上述过程。

 （6-9）

其中，表示学习率，表示训练集大小，表示由训练集样本计算得到的梯度值。

对于公式中的取值，可将其分为批量梯度下降（Batch Gradient Descent, BGD）、随机梯度下降（Stochastic Gradient Descent, SGD）和小批量梯度下降（Mini-batch Gradient Descent, MBGD）三种方法。

当取值为训练集大小，即传统的批量梯度方法，使用这种方法存在以下问题：进行参数更新时将计算整个数据集梯度，并且一次迭代训练过程将只进行一次更新，因此在处理大型数据集时速度很慢并且难以控制；因为每次训练过程都采用所有数据集，所以训练过程存在冗余的权重更新。上述缺点导致批量梯度下降方法在训练大数据集时更新效率低下，甚至导致内存溢出，因此出现了SGD方法。

当取值为1时，即是随机梯度下降方法。随机梯度下降对每个训练样本计算梯度并进行参数更新，使得执行速度更快。SGD方法这种利用一个样本近似替代所有样本用于调整权重的方法，因计算得到的并不是准确的梯度，因而在更新权重时会存在一定问题。但是对于最优化问题，凸问题，虽然不是每次迭代得到的梯度都向着全局最优方向，但是整体上方向仍朝着全局最优解，最终结果往往是在全局最优解附近。相比于批量梯度下降方法，随机梯度下降方法更快收敛，虽然不是全局最优，但结果却在接受范围之内。公式（6-10）给出其更新公式：

 （6-10）

表示每次迭代随机选中样本计算得到的梯度值。针对SGD方法无法收敛到全局最优解的问题，提出了另一种称为小批量梯度下降的变体。

当取值介于1到训练集大小之间的某个合适值上，即为小批量梯度下降方法。MBGD方法每次读取小批量个样本，在其上计算梯度并进行参数更新。使用小批量梯度方法具有以下优点：可以减少参数更新的波动，最终得到效果更好和更稳定的收敛；大数据集场景下，可以提高运行效率。因此，在实际的应用中，常采用小批量梯度下降方法来训练神经网络。下文中也将采用MBGD方法作为深度学习模型的优化方法。

上述两小节介绍了深度学习模型的基本概念，以及其训练过程使用的优化方法。深度学习因其解决复杂任务的能力，被广泛应用于多个领域，但像其他机器学习模型一样，深度学习模型容易受到多种类型的攻击。例如，企业收集的用户图片、语音和视频数据在用于共享目的时会导致隐私泄露；利用模型反演攻击可以通过脸部识别系统反向恢复图片；攻击者不但具有深度学习模型的使用接口，还具有模型参数、模型结构等背景知识；攻击者还可以利用植入的数据进行恶意攻击，从而造成隐私泄露。

差分隐私可以解决上述深度学习场景下的隐私安全问题。深度学习隐私保护的目标包括保护训练数据集以及训练参数两部分，根据不同阶段的扰动方法，可以将扰动过程设置在数据集输入阶段、模型参数输出阶段、以及模型训练阶段。例如，文献[72希望通过扰动图像数据的输入特征来生成对抗性样本以实现抵抗攻击者的反演攻击，文献[73,74]都是以通过扰动输入数据生成对抗性样本并实现保护数据隐私的目的。文献[75]提到的FedAvg算法是对模型输出进行扰动并最终进行平均计算的方法。

但是对于输入扰动，一方面难以分析其敏感度，另一方面对于整型输入，添加整数噪声会导致扰动太大，添加小数噪声则破坏原有结构；对于输出扰动，直接将噪声添加到模型参数上，会影响学习算法的性能，甚至导致学习算法不可用；因此，针对深度学习的差分隐私，现有的主要方法集中在模型训练阶段，包括：（1）将噪声添加到现有优化算法的执行过程中；（2）扰动给定优化算法的目标函数。可以理解成方法（1）在计算得到的梯度值上进行扰动，而方法（2）则是先将扰动添加到目标函数上后计算梯度。利用这两种方法可以避免粗粒度的扰动方式，将隐私增强的方法应用在内部的训练过程，而不是应用到模型的输入输出上。但是，在训练过程中进行扰动仍存在以下问题：

**高敏感度：**在深度学习模型训练过程中，扰动操作中目标函数的敏感度很高；

**有限的隐私预算：**迭代的训练过程将隐私预算分配到不同部分，但是过小的隐私预算会产生较大噪声，对最终结果产生较大影响。

因此，具体的扰动方式要根据模型结构以及数据集特点具体调整。现有文献提出了几种常规的方法，旨在保证训练过程中为训练数据提供差分隐私保护，同时能够很好地保证深度学习模型的性能。文章接下来将重点介绍基本的拉普拉斯方法以及其改进的隐私梯度下降方法。当然，在这之外，还有包括分布式隐私梯度下降算法等，具体内容可参考文献[69]。

### 6.4.3 基本的拉普拉斯方法

深度学习差分隐私中基本的拉普拉斯方法是将噪声添加到梯度下降过程中，即将拉普拉斯噪声添加到每个迭代步骤的梯度计算中，如公式（6-11）所示：

 （6-11）

计算出加噪的梯度，后用更新公式更新权重，具体公式如（6-12）：

 （6-12）

本文中深度学习的优化方法均采用小批量梯度下降方法，其具体步骤用以下伪代码给出：

|  |
| --- |
| 算法：基本的拉普拉斯方法 |
| 输入：训练数据集，学习率，噪声规模，批量大小，隐私预算 |
| 输出：模型参数 |
| * 1. 随机初始化模型参数   2. 分配隐私预算   3. 迭代训练，   4. 从训练集中随机抽取批量样本   5. 计算梯度：对每个样本，   6. 添加噪声：   7. 梯度下降：   8. 结束迭代   9. 输出： |

上述算法给出了基本拉普拉斯方法的步骤，通过最小化目标函数训练模型以得到模型参数。

第1步：给模型的连接向量，即权重集合随机赋予初值；

第2步：假设训练过程共需迭代次，则将隐私预算平均分配到个迭代步骤中，每一次分配到的隐私预算为；

第3步：开始迭代，对于第次迭代，从训练集中随机抽取批量数据；

第4步：计算中每个样本的梯度；

第5步：利用步骤4计算得到的每个样本的梯度，计算小批量平均梯度并添加噪声以得到噪声梯度，具体计算步骤是先求和，再加噪，最后计算平均值，相对于先求均值，再加噪的方式来说，噪声粒度更合适；

第6步：利用上述噪声梯度更新权重，更新方向为梯度的负方向，更新步长与学习率及计算得到的梯度相关；

第7步：结束迭代过程，输出最大迭代次数下的连接权重作为模型参数。

显而易见，基本的拉普拉斯方法在步骤5中进行加噪，让模型参数满足-差分隐私，从而实现对深度学习模型及其训练数据集的差分隐私保护，但是这个基本方法存在其固有缺陷。一方面，当迭代次数设置过大，会导致每次迭代过程分配到的隐私预算都非常小，从而会导致非常大的噪声，例如，设置最大迭代次数为1000，训练集上设置的总隐私预算为1，则每一次迭代过程分得的隐私预算为0.001，而这会产生较大噪声，实际上，最大迭代次数为1000在实际训练中很常见；另一方面，因为不对梯度做约束，所以实际计算中，敏感度的值可能很大。在计算噪声量的过程中，敏感度值较大，隐私预算值较小，则整体产生的噪声量将较大，将对原有数据集产生较大偏移，从而导致模型性能变差，严重者将导致模型无法收敛。基于上述缺点，将基本的拉普拉斯方法直接应用于深度学习模型上是不现实的，因此，提出基本方法的改进方法，能在一定程度上解决上述困难。

### 6.4.4 改进的隐私梯度下降方法

文献[71]基于基本的拉普拉斯方法，提出了改进方法。同基本方法类似，噪声被添加到目标函数上，但是针对三个细节点进行了改进以提升模型性能，包括：

**范式裁剪：**通过对目标函数进行范式裁剪，来限制函数敏感度；

**组批量：**将几个批量组合在一起形成组，再将噪声添加到组；

**隐私组合：**使用moment accountant作为目标函数的最优隐私组合。

以下给出改进隐私梯度下降方法的伪代码：

|  |
| --- |
| 算法：改进的隐私梯度下降方法 |
| 输入：训练数据集，学习率，噪声规模，组大小，梯度阈值 |
| 输出：模型参数 |
| * 1. 随机初始化模型参数   2. 迭代训练，   3. 以概率从训练集中随机抽取组样本   4. 计算梯度：对每个样本，   5. 梯度裁剪：   6. 添加噪声：   7. 梯度下降：   8. 结束迭代   9. 输出： |

从上述伪代码可以看出，同基本拉普拉斯方法相比，改进方法新加了几个参数，包括噪声规模，组大小，以及梯度裁剪阈值，这些参数用在加噪、组合批量数据、范式裁剪等阶段。

第1步：给模型的连接向量，即权重集合随机赋予初值；

第2步：假设训练过程共需迭代次，对于每次迭代，以的概率随机抽取个样本作为组数据，可以将组数据看做为几个批量的合并，步骤2主要在于减少噪声的影响；

第3步：对于组数据中每个样本，计算其梯度值；

第4步：利用梯度裁剪阈值对梯度进行裁剪，这一步骤的目的是为了防止梯度爆炸，同时减小梯度的敏感度，采用的方式是当梯度超过阈值时，将梯度更新为阈值；

第5步：利用步骤4计算得到的每个样本的梯度，计算组数据的平均梯度并添加噪声以得到噪声梯度，具体计算步骤是先求和，再添加高斯噪声，最后计算平均值，同样地，相对于先求均值，再加噪的方式来说，噪声粒度更合适；

第6步：利用上述噪声梯度更新权重，更新方向为梯度的负方向，更新步长与学习率及计算得到的梯度相关；

第7步：结束迭代过程，输出最大迭代次数下的连接权重作为模型参数；

第8步：基于隐私会计方法计算整个过程的隐私损失。

对比基本拉普拉斯方法，改进的方法将噪声规模控制在一定范围，不会导致学习算法性能的明显下降，接下来将对其改进的细节点进行详细介绍。

**范式裁剪：**影响差分隐私深度学习算法性能的一个较为关键的因素是目标函数的高敏感度，为了限制这种影响，改进的隐私梯度下降方法通过预定义梯度裁剪阈值，使用范数对梯度进行裁剪，裁剪规则是对于超过阈值的梯度，将其更新为阈值，用公式解释如下：

 （6-13）

因此在具体的隐私裁剪步骤中，梯度向量的计算公式为，同时梯度的敏感度被限定在。对于多层神经网络，每层结构可以设置不同的裁剪阈值。

**组批量：**基本的拉普拉斯方法通过损失函数计算批量样本的梯度并加噪求均值，因此，批量数目越小，同等规模的噪声对数据的扰动越大。为了减小噪声扰动的幅度，改进方法将几个批量样本组合在一起形成组，然后将噪声添加到组中。实际使用中，每个组都以的概率独立选择每个样本进行创建。

**隐私组合：**改进的隐私梯度下降方法使用敏感度为的高斯机制，根据高斯机制的定义，当时，每个组满足-差分隐私。训练集存在个组，则模型最终输出满足-差分隐私，这是一个相对松散的上界。

改进的隐私梯度下降方法使用moments accountant方法，在给定合适的噪声规模以及裁剪阈值下，能将隐私上界提高到。其存在以下定理：

**定理：**存在常量和，给定样本概率为，迭代次数设置为，对于任意的，，改进的隐私梯度下降方法满足-差分隐私仅当

 （6-14）

## 6.5 小结

本章主要介绍了面向数据挖掘和机器学习的差分隐私方法。首先介绍了数据挖掘和机器学习过程中噪声扰动方式，然后详细介绍了聚类、频繁模式挖掘等无监督学习的差分隐私方法以及分类、线性回归、逻辑回归、支持向量机等监督/半监督学习的差分隐私方法，最后从深度学习模型、梯度下降及其变体、改进的梯度下降方法等详细介绍了面向深度学习的差分隐私方法。

## 6.6 参考文献

1. Vaidya J, Shafiq B, Basu A, et al. Differentially private naive bayes classification[C]//Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2013 IEEE/WIC/ACM International Joint Conferences on. IEEE, 2013, 1: 571-576.
2. Mir D J, Wright R N. A differentially private graph estimator[C]//Data Mining Workshops, 2009. ICDMW'09. IEEE International Conference on. IEEE, 2009: 122-129.
3. Sala A, Zhao X, Wilson C, et al. Sharing graphs using differentially private graph models[C]//Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference. ACM, 2011: 81-98.
4. Ji Z, Elkan C. Differential privacy based on importance weighting[J]. Machine learning, 2013, 93(1): 163-183.
5. Chaudhuri K, Sarwate A, Sinha K. Near-optimal differentially private principal components[C]//Advances in Neural Information Processing Systems. 2012: 989-997.
6. Vinterbo S A. Differentially private projected histograms: Construction and use for prediction[C]//Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, Berlin, Heidelberg, 2012: 19-34.
7. Hardt M, Roth A. Beyond worst-case analysis in private singular vector computation[C]//Proceedings of the forty-fifth annual ACM symposium on Theory of computing. ACM, 2013: 331-340.
8. Kapralov M, Talwar K. On differentially private low rank approximation[C]//Proceedings of the twenty-fourth annual ACM-SIAM symposium on Discrete algorithms. Society for Industrial and Applied Mathematics, 2013: 1395-1414.
9. Jagannathan G, Pillaipakkamnatt K, Wright R N. A practical differentially private random decision tree classifier[C]//Data Mining Workshops, 2009. ICDMW'09. IEEE International Conference on. IEEE, 2009: 114-121.
10. Friedman A, Schuster A. Data mining with differential privacy[C]//Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2010: 493-502.
11. Zhang J, Zhang Z, Xiao X, et al. Functional mechanism: regression analysis under differential privacy[J]. Proceedings of the VLDB Endowment, 2012, 5(11): 1364-1375.
12. Chaudhuri K, Monteleoni C. Privacy-preserving logistic regression[C]//Advances in Neural Information Processing Systems. 2009: 289-296.
13. Chaudhuri K, Monteleoni C, Sarwate A D. Differentially private empirical risk minimization[J]. Journal of Machine Learning Research, 2011, 12(Mar): 1069-1109.
14. Rubinstein B I P, Bartlett P L, Huang L, et al. Learning in a large function space: Privacy-preserving mechanisms for SVM learning[J]. arXiv preprint arXiv:0911.5708, 2009.
15. Nissim K, Raskhodnikova S, Smith A. Smooth sensitivity and sampling in private data analysis[C]//Proceedings of the thirty-ninth annual ACM symposium on Theory of computing. ACM, 2007: 75-84.
16. Czerniak J, Zarzycki H. Application of rough sets in the presumptive diagnosis of urinary system diseases[M]//Artificial intelligence and security in computing systems. Springer, Boston, MA, 2003: 41-51.
17. Dwork C, Lei J. Differential privacy and robust statistics[C]//Proceedings of the forty-first annual ACM symposium on Theory of computing. ACM, 2009: 371-380.
18. Blum A, Dwork C, Mcsherry F, et al. Practical privacy:the SuLQ framework[C]// Twenty-Fourth ACM Sigmod-Sigact-Sigart Symposium on Principles of Database Systems. ACM, 2005:128-138.
19. Yang L I, Guangzhou. Research on Differential Privacy Preserving k-means Clustering[J]. Computer Science, 2013, 59(1):1–34.
20. Acs G, Castelluccia C, Chen R. Differentially private histogram publishing through lossy compression[C]//Data Mining (ICDM), 2012 IEEE 12th International Conference on. IEEE, 2012: 1-10.
21. Yu Q, Luo Y, Chen C, et al. Outlier-eliminated k-means clustering algorithm based on differential privacy preservation[J]. Applied Intelligence, 2016, 45(4): 1179-1191.
22. Shang T, Zhao Z, Guan Z, et al. A DP Canopy K-Means Algorithm for Privacy Preservation of Hadoop Platform[J]. 2017.
23. Su D, Cao J, Li N, et al. Differentially private k-means clustering[C]//Proceedings of the Sixth ACM Conference on Data and Application Security and Privacy. ACM, 2016: 26-37.
24. 马银方, 张琳. 基于差分隐私保护的KDCK-medoids动态聚类算法[J]. 计算机科学, 2016, 43(s2):368-372.
25. Zhang Y, Li S. Privacy Preserving BIRCH Algorithm under Differential Privacy[C]//Intelligent Computation Technology and Automation (ICICTA), 2017 10th International Conference on. IEEE, 2017: 48-53.
26. Wei-Min W U, Huang H K, Computer S O. A DP-DBScan clustering algorithm based on differential privacy preserving[J]. Computer Engineering & Science, 2015.
27. Wang H, Lina G E, Wang S, et al. Improvement of differential privacy protection algorithm based on OPTICS clustering[J]. Journal of Computer Applications, 2018.
28. Mendonça A L C, Brito F T, Linhares L S, et al. DiPCoDing: A Differentially Private Approach for Correlated Data with Clustering[C]// International Database Engineering & Applications Symposium. ACM, 2017:291-297.
29. Chen L, Yu T, Chirkova R. WaveCluster with Differential Privacy[J]. Computer Science, 2015, 11(2):191–198.
30. Wu Y, Wu Y, Peng H, et al. Differentially private density estimation via Gaussian mixtures model[C]// Ieee/acm, International Symposium on Quality of Service. IEEE, 2016:1-6.
31. Chen T, Kang H. Research on Clustering-Differential Privacy for Express Data Release[C]// International Conference on Information and Communications Security. Springer, Cham, 2017:427-437.
32. Gao Z Q, Zhang L J. DPHKMS: An Efficient Hybrid Clustering Preserving Differential Privacy in Spark[C]//International Conference on Emerging Internetworking, Data & Web Technologies. Springer, Cham, 2017: 367-377
33. Bhaskar R, Laxman S, Smith A, et al. Discovering frequent patterns in sensitive data[C]//Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2010: 503-512.
34. Dwork C, McSherry F, Nissim K, et al. Calibrating noise to sensitivity in private data analysis[C]//Theory of Cryptography Conference. Springer, Berlin, Heidelberg, 2006: 265-284.
35. McSherry F, Talwar K. Mechanism design via differential privacy[C]//Foundations of Computer Science, 2007. FOCS'07. 48th Annual IEEE Symposium on. IEEE, 2007: 94-103.
36. Maruseac M, Ghinita G. Differentially-private mining of moderately-frequent high-confidence association rules[C]//Proceedings of the 5th ACM Conference on Data and Application Security and Privacy. ACM, 2015: 13-24.
37. Xu S, Cheng X, Su S, et al. Differentially private frequent sequence mining[J]. IEEE Transactions on Knowledge and Data Engineering, 2016, 28(11): 2910-2926.
38. 丁哲,秦臻,秦志光.基于差分隐私的不确定数据频繁项集挖掘算法[J/OL].计算应用研究,2018(07):1-7[2018-06-21]. [http://kns.cnki.net/kcms/detail/51.1196.TP.2 017](http://kns.cnki.net/kcms/detail/51.1196.TP.2%20017) 0 7 2 7.21.18.

074.html.

1. Chen R, Mohammed N, Fung B C M, et al. Publishing set-valued data via differential privacy[J]. Proceedings of the VLDB Endowment, 2011, 4(11): 1087-1098.
2. Chen R, Acs G, Castelluccia C. Differentially private sequential data publication via variable-length n-grams[C]//Proceedings of the 2012 ACM conference on Computer and communications security. ACM, 2012: 638-649.
3. Bonomi L, Xiong L. A two-phase algorithm for mining sequential patterns with differential privacy[C]//Proceedings of the 22nd ACM international conference on Information & Knowledge Management. ACM, 2013: 269-278.
4. 李艳辉,刘浩,袁野,王国仁.基于差分隐私的频繁序列模式挖掘算法[J].计算机应用,2017,37(02):316-321+340.
5. 肖恪. 满足差分隐私的频繁子图挖掘算法研究[D].北京邮电大学,2017.
6. Blum A, Dwork C, McSherry F, et al. Practical privacy: the SuLQ framework[C]//Proceedings of the twenty-fourth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems. ACM, 2005: 128-138.
7. McSherry F D. Privacy integrated queries: an extensible platform for privacy-preserving data analysis[C]//Proceedings of the 2009 ACM SIGMOD International Conference on Management of data. ACM, 2009: 19-30.
8. Wang Q, Jiang Y, Lin D. Linear complexity of Ding-Helleseth sequences of order 2 over GF (l)[J]. Cryptography and Communications, 2016, 8(1): 33-49.
9. 赵格, 康海燕. 基于差分隐私的电子商务隐私保护数据发布[J]. 物流工程与管理, 2014 (11): 119-122.
10. Mohammed N, Chen R, Fung B, et al. Differentially private data release for data mining[C]//Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2011: 493-501.
11. Zhu T, Xiong P, Xiang Y, et al. An effective deferentially private data releasing algorithm for decision tree[C]//Trust, Security and Privacy in Computing and Communications (TrustCom), 2013 12th IEEE International Conference on. IEEE, 2013: 388-395.
12. Vaidya J, Shafiq B, Fan W, et al. A random decision tree framework for privacy-preserving data mining[J]. IEEE transactions on dependable and secure computing, 2014, 11(5): 399-411.
13. Patil A, Singh S. Differential private random forest[C]//Advances in Computing, Communications and Informatics (ICACCI, 2014 International Conference on. IEEE, 2014: 2623-2630.
14. Bojarski M, Choromanska A, Choromanski K, et al. Differentially-and non-differentially-private random decision trees[J]. arXiv preprint arXiv:1410.6973, 2014.
15. Li Z, Li S. Random forest algorithm under differential privacy[C]//Communication Technology (ICCT), 2017 IEEE 17th International Conference on. IEEE, 2017: 1901-1905.
16. 穆海蓉, 丁丽萍, 宋宇宁,等. DiffPRFs:一种面向随机森林的差分隐私保护算法[J]. 通信学报, 2016, 37(9):175-182.
17. Edemskiy V. About computation of the linear complexity of generalized cyclotomic sequences with period p n+ 1[J]. Designs, Codes and Cryptography, 2011, 61(3): 251-260.
18. Cai H, Zhou Z, Yang Y, et al. A new construction of frequency-hopping sequences with optimal partial Hamming correlation[J]. IEEE Transactions on information theory, 2014, 60(9): 5782-5790.
19. Zhang Z, Zhang Z, Yang Y, et al. Functional mechanism: regression analysis under differential privacy[J]. Proceedings of the Vldb Endowment, 2012, 5(11):1364-1375.
20. Kasiviswanathan S P, Nissim K, Jin H. Private Incremental Regression[J]. 2017:167-182.
21. Chaudhuri K, Monteleoni C, Sarwate A D. Differentially private empirical risk minimization[J]. Journal of Machine Learning Research, 2011, 12(Mar): 1069-1109.
22. Chaudhuri K, Monteleoni C. Privacy-preserving logistic regression[C]// International Conference on Neural Information Processing Systems. Curran Associates Inc. 2008:289-296.
23. Chaudhuri K, Monteleoni C, Sarwate A D. Differentially Private Empirical Risk Minimization[J]. Journal of Machine Learning Research Jmlr, 2011, 12(2):1069.
24. Lindell, P.: Privacy preserving data mining. J. Cryptol. 15(3), 177–206 (2002)
25. Saranya, K., Premalatha, K., Rajasekar, S.S.: A survey on privacy preserving data mining.In: International Conference on Electronics and Communication Systems, pp. 1740–1744.IEEE (2015)
26. Lin K P, Chen M S. On the design and analysis of the privacy-preserving SVM classifier[J]. IEEE Transactions on Knowledge and Data Engineering, 2011, 23(11): 1704-1717
27. Li H, Xiong L, Ohno-Machado L, et al. Privacy preserving RBF kernel support vector machine[J]. BioMed research international, 2014, 2014.
28. Sun L, Mu W S, Qi B, et al. A new privacy-preserving proximal support vector machine for classification of vertically partitioned data[J]. International Journal of Machine Learning and Cybernetics, 2015, 6(1): 109-118.
29. Yu H, Jiang X, Vaidya J. Privacy-preserving SVM using nonlinear kernels on horizontally partitioned data[C]//Proceedings of the 2006 ACM symposium on Applied computing. ACM, 2006: 603-610.
30. Yu H, Vaidya J, Jiang X. Privacy-preserving svm classification on vertically partitioned data[C]//Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer Berlin Heidelberg, 2006: 647-656.
31. Shokri R, Shmatikov V. Privacy-preserving deep learning[C]//Proceedings of the 22nd ACM SIGSAC conference on computer and communications security. ACM, 2015: 1310-1321.
32. Fredrikson M, Jha S, Ristenpart T. Model inversion attacks that exploit confidence information and basic countermeasures[C]//Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security. ACM, 2015: 1322-1333.
33. Abadi M, Chu A, Goodfellow I, et al. Deep learning with differential privacy[C]//Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. ACM, 2016: 308-318.
34. Papernot N, McDaniel P, Jha S, et al. The limitations of deep learning in adversarial settings[C]//Security and Privacy (EuroS&P), 2016 IEEE European Symposium on. IEEE, 2016: 372-387.
35. Goodfellow I J, Shlens J, Szegedy C. Explaining and Harnessing Adversarial Examples[J]. Computer Science, 2014.
36. Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks[J]. arXiv preprint arXiv:1312.6199, 2013.
37. McMahan H B, Ramage D, Talwar K, et al. Learning Differentially Private Recurrent Language Models[J]. 2018.

# 第七章 面向复杂数据类型的差分隐私方法

## 7.1 图数据的差分隐私（陈竑毓）

### 7.1.1 隐私保护技术在图数据中的应用

#### 7.1.1.1 图数据隐私泄露风险

社交网络是图数据的典型代表，在下文中，我们将以社交网络为例来描述图数据。图7-1展示了原始社交网络图（图1a）使用“天真”匿名机制转换成匿名社交网络图（图1b）。

 (a)原始社交网络图 (b)匿名社交网络图

图7-1 “天真”匿名机制

“天真”匿名机制简单地用随机伪身份（即数字）替换与每个顶点相关的身份信息（即姓名）。其优势在于它可以对发布的社交网络数据进行高效用性的分析。而且，对于无背景知识的攻击者来说，“天真”匿名可以保护个人的隐私。然而，攻击者可以利用各种关于图的结构信息来重新识别匿名图中的个体[1]，[2]。例如，攻击者可以通过识别代表目标个体的目标顶点或从发布的数据中推断出关于个体的一些有用信息来破坏个体的隐私。这就导致了社交网络数据发布面临三大重要的隐私风险[3]：

* 身份披露风险：如果可以从发布的匿名图中识别出特定人员，就会发生身份泄露。例如，一个已婚和成功的政治家如果在犯罪圈中被识别出来，对于这个人和他的政治生涯是毁灭性的。
* 边披露风险：边披露揭示了社交网络用户之间的关系。例如，医患关系是必须得到保护的私密信息。这是因为如果知道某个医生是艾滋病专家，那么关系披露可能会导致患者被确认为感染艾滋病毒。
* 内容泄露风险：当发布的社交网络数据中与每个顶点或边相关的敏感属性（如年龄，性别，性取向）受到损害并且受损属性可直接归因于匿名图中的特定人员时，会发生内容泄露。

一般来说，这些风险可以分为主动，半静态和被动[1]。在主动攻击中，在发布之前将新的顶点和边添加到社交网络中。在半被动攻击中，在发布之前只有边被添加到社交网络图中。在发布之前不对被动攻击做出社交网络图的修改，并且攻击利用辅助数据源[4]。

现今已经提出了各种防止内容泄露攻击的方法[5]，[6]和链接披露攻击[7]，[8]，[9]。社交网络链接匿名化已经在[10]，[7]，[11]，[8]，[12]中讨论过。加权社交网络图的匿名化在[13]，[14]中讨论，其中边权重可以用来表示某些信息，如友谊程度。最近出现了基于用户的个人隐私请求提供隐私保护服务的方法[15]，[16]。例如，Yuan等人[16]提出了一个框架来实现个人隐私的要求，考虑了三种不同的条件：节点上的标签，节点的度和边上的标签。

#### 7.1.1.2 传统的图数据隐私保护技术

本节针对不同背景知识可能导致的隐私泄露，提出了相应的图数据隐私保护方法。主要分为匿名化、数据扰动和推演控制。

* 匿名化方法，其主要思想是：攻击者基于目标背景知识在匿名化社会网络数据中进行匹配识别时，至少有个候选符合，即目标的隐私泄露概率小于，匿名化方法具体分为：随机图编辑、概率图编辑、-匿名化和聚类泛化；
* 数据扰动[17,18,19]的主要思想是：对社会网络进行随机化修改，使得攻击者不能准确地推测出原始真实数据，数据扰动方法具体分为数值扰动[17,19]和图结构扰动[18]；
* 推演控制[20]的主要思想是：对于不同隐私预测模型，通过对图数据进行针对性地修改，使得攻击者采用预测模型不能推演出隐私信息，从而达到保护社会网络隐私的目的。

1. 匿名化方法
2. 随机图编辑模型：该模型通过随机添加，删除或转换边来编辑图形[21]，[22]，[23]，[7]，[24]。通常，这是通过以下方式实现的：

* 随机修改：随机选择两个未连接的顶点并在它们之间添加一条边，随后在两个连接的顶点之间随机选择一条边并将其删除。
* 随机转换：随机选择图中的一对边，，并将他们转换成，。

1. -匿名化方法：-匿名化模型通过确定性地编辑图的顶点和边（添加/删除）来提供匿名性。-匿名的主要思想是：给定一个原始图和一个整数，生成原始图的-匿名图，使得具有背景知识的攻击者将不能重新识别具有大于概率的目标个体。
2. 基于聚类的方法：聚类策略[21]，[25]将图抽象为超级顶点和超级边，以减轻顶点重新识别攻击，而不是像在-匿名和随机化方法中那样修改图结构。基于聚类的匿名化问题可以非正式地表述如下：给定一个原始图，将个顶点和个边划分成最小数量的超级集群，同时保持每对节点的结构信息彼此不可区分。大致上可以分为两步：第一步，对顶点进行划分，使得超级集群包含沿其边的至少个顶点的子集，并具有相似的特征，如度或同构邻域。第二步，基于它们的关系执行聚类和边泛化使顶点不可区分。集群被推广到超级集群。超级群集可以通过超级边连接。边可以被分类为内部边和外部边。
3. 概率隐私保护模型：概率隐私保护技术[26]，[27]，[24]描述了图匿名化方法。这些方法通过利用边概率的语义将确定性图转换为不确定图。概率隐私保护问题可以非正式表述如下：给定一个原始的确定图，产生一个非确定图，使得具有背景知识的攻击者将不能重新识别目标顶点或推断来自的任何敏感信息。
4. 数据扰动方法

图数据扰动隐私保护方法的基本思想是：通过对图数据进行随机化修改，使得攻击者不能准确推测出原始真实数据，从而起到保护社会网络数据隐私的作用[28,17,18,19]。本节从数值扰动[17,19]和图结构扰动[18,28]等方面介绍基于数据扰动的社交网络隐私保护技术。

1. 数值扰动

社交网络中记录着大量的数值信息，对这些数值信息进行随机的扰动和修改可以使攻击者无法原始的真实数据。数值扰动的方法常用于权重社交网络中的边权重隐私保护。

文献[29]提出了一种网络权重隐私保护算法。该算法利用无向有权图表示社交网络，把边权重序列作为一个无归属直方图处理，将包含敏感信息的权重加入拉普拉斯噪音使得满足差分隐私保护要求。为了减少噪音量，对直方图中具有相同计数的桶合并成组，根据组间-不可区分性保证了差分隐私保护要求，通过对原始的权重序列进行一致性推理保持了网络最短路径不变。算法的加噪部分如下：

将分成两个部分，对加噪，对于每一个权重，如果，则，否则。其中，为隐私预算，存储权重集合中每个元素出现的次数：，存储集合中每个元素出现的次数：，为不可区分性参数。

1. 图结构扰动

通过随机进行图数据扰动和修改，可以阻止攻击者获知原始图结构，从而保护社交网络数据隐私[18]。图扰动的主要方法是随机添加、删除边和交换边端点等。

1. 推演控制

推演控制是指对于不同隐私预测和推演模型，针对性地修改社交网络，使得攻击者采用预测模型不能推演出隐私信息，达到保护社交网络隐私的目的。文献[20]给出相应的推演控制技术来防止隐私泄露。

在文献[20]中，首先提出了基于共同邻居数目的敏感关系预测模型，并定义了两种链接推演攻击：单步链接推演攻击和级联链接推演攻击。单步链接推演是指对于图上的所有无边连接的结点对执行链接推演操作；级联链接推演是指在图上执行多次单步链接推演操作。为了阻止链接推演攻击，提出了一种基于链接世系溯源的防推演机制来切断敏感链接的推演路径，在保护社交网络中敏感关系的同时，保持了图数据可用性。推演控制技术能够有效地防止特定预测模型导致的隐私泄露，由于其针对性地修改社交网络，可以保持图数据的高可用性。但是推演控制技术的隐私保护能力有限，对于图数据隐私保护不具有通用性。

### 7.1.2 图数据差分隐私保护技术

#### 7.1.2.1 节点隐私

节点差分隐私目的是通过保护节点和与它相连接的边来限制推理网络中节点的存在性。在节点差分隐私中，邻近图的定义是：对于两个图和，如果通过添加或者删除一个节点和与之相邻的边能够得到，那么称这两个图为邻近图。图7-2展示了原始社交网络图中添加一个新的节点9（2b），原始社交网络中删除节点7（2c）。

 (a)原始网络图 (b)添加新节点9 (c)删除节点7

图7-2 节点差分隐私的邻近图

定义 （-节点差分隐私）：设有随机算法，为所有可能的输出构成的集合。对于任意两个邻近图和以及的任何子集，若算法满足：



则称算法提供-节点差分隐私保护，其中参数是一个较小的正值，用来权衡隐私和精确度。

从直观上看，攻击者区分的输出和的输出的能力是非常有限的。因此，差分隐私模型表明，任何单个节点的添加或删除都不会严重影响数据分析的输出分布[66]。这意味着，图和将在统计数据上产生类似的分布。需要注意的是，图中的每个节点所包含的信息都对应于一个特定的节点和与该节点相邻的所有边。因此，节点及其邻边都应保持隐藏。

现今已经提出了各种方法来实现节点差分隐私。Hay等人[30]介绍了节点差分隐私的概念，并提出实现节点差分隐私的一些困难。Raskhodnikova和Smith [31]提出了一种节点差分隐私算法，用于释放图的度分布的近似值。该算法基于Lipschitz扩展和广义指数机制。同样，Borgs et al [32]采用与指数机制相结合的Lipschitz扩展方法来引入节点差分隐私算法。

Kasiviswanathan等[66]分析了几种节点差分隐私算法后，提出了分析实际网络中算法的准确性的方法。同时还讨论了使用线性规划进行Lipschitz扩展的功能，例如子图计数。他们定义了一个非严格映射，为此需要解决线性程序以获得输出。为了解决由于Lipschitz扩展引起的缺陷，Kasiviswanathan等人[66]提出了投影方法，其背后的核心思想是将输入图投影到最大度数低于某个阈值的图集上。输入图中度数高于阈值的任何节点都将被丢弃。第二种方法引入了一种称为“天真”截断的投影函数，它简单地丢弃图中的度数较高的节点。投影的局部敏感度根据图中具有预定义度的顶点的数量来限定。算法步骤如下：

输入：，，，度分布查询

输出：

1. 随机截断参数：；
2. 用计算和平滑上界；
3. 输出。

算法中有三个主要步骤。第一步是确定截断参数。由于我们可能不知道图中的最大度数或者度数非常大，因此，使用来近似。第二步通过丢弃度数大于的节点来创建截断图。最后，将满足柯西分布的噪声添加到度分布中。

Blocki等人[21]提出了用于分析稀疏图的节点差分隐私算法。他们将低敏感度投影用于给定最大度数的低度图集，同时对局部敏感度提供平滑上界。就投影图的结构而言，投影的局部敏感度根据图的与S的距离来约束。引入投影函数以增强Lipschitz扩展的弱点。 但是，需要线性规划来计算输入图的投影。

#### 7.1.2.2 边隐私

边隐私是一种比节点隐私更弱的概念[66]它已经比节点隐私保护得到了更广泛的研究。直观上讲，边隐私要求是否需要隐藏某一条边，但整体关系模式可以公开。在边差分隐私中，邻近图的定义是：对于两个图和，如果通过添加或者删除两个顶点之间的边之后能够得到，那么称这两个图为邻近图。图7-3展示了原始社交网络图中添加一条边（4，8）（3b）,原始社交网络中删除一条边（3，4）（3c）。

 (a)原始网络图 (b)添加新边 (c)删除一条边

图7-3 边差分隐私的邻近图

对于图和，其中，，，分别是图和的顶点集合和边集合，如果：



图的查询函数满足差分隐私保护则称满足-边差分隐私保护。

因此，边差分隐私确保算法的输出不会揭示图中特定的边。Hay等人[30]提出了一种用于发布社交网络图度分布的边差分隐私算法。他们的算法是对[33]中提出的全局敏感度方法的扩展，使用一种后处理的方法在差分隐私的输出结果中去除一些噪声。通过对真实和合成网络的经验分析，Hay等人[30]表明度分布估计显示出低偏差和方差，并推荐用于分析普通网络分布，例如幂律分布。然而，Hay等人提出的方法[30]不发布图。在某些情况下，网络数据的结构特性非常令人感兴趣，发布匿名图成为一种更好的选择。此外，该方法只能提供度分布的近似估计。

Nissim等人[34]在估计最小生成树的成本和图中三角形数量的情况下考虑边差分隐私。他们将查询在特定数据集上的局部敏感度定义为查询答案在图中添加或删除边时可以更改的数量。该机制添加的随机噪声量被校准为敏感度的局部变体，称为平滑敏感度。Nissim等人[34]也讨论了计算各种领域统计数据平滑敏感度的算法。

Block等人[35]提出了一种用边差分隐私分析图数据的算法。在边邻接中，假定相邻接图只在一个边或仅一个顶点的属性中不同，作为全局敏感度[33]和平滑敏感度[34]的替代方案，他们引入了限制敏感度的概念，以减少添加到输出的噪声量，从而提高差分隐私数据分析的准确性。对于边差分隐私，Blocki等人[35]提出了一个非常简单的投影算子，它具有不变的边敏感度。另外，Blocki等[36]将Johnson-Linden Strauss变换应用于近似图中的切割查询中。他们表明Johnson-Linden Strauss变换允许发布边差分隐私图。

Rastogi等人[37]针对一般子图计数发布问题，结合边差分隐私来抵抗贝叶斯攻击者。首先需要计算局部敏感度的高概率上限，然后将噪声添加到与计算的上限局部敏感度成比例的数据中。所提出的方法主要基于以下假设：边的存在不一定更有可能使其他边存在。该假设与社交网络中边之间的正相关性相矛盾（例如，拥有共同朋友的两个人更可能是朋友）。此外，关于攻击者先验知识的假设限制了隐私定义的适用性。

Qin等人[67]针对现有的局部差分隐私和图生成技术不能够很好的保留重要的图结构问题，提出了一种本地化的边差分隐私算法LDPGen，这是一种新颖的多阶段技术，它基于用户与整个人群的不同分区的连接逐渐将用户聚类。每次用户报告信息时，LDPGen都会仔细注入噪声以确保本地差分隐私。我们在此过程中导出最优参数，以将结构相似的用户聚集在一起。一旦获得了良好的用户聚类，LDPGen就会调整现有的社交图生成模型来构建合成的社交图。

LDPGen设计由三个阶段组成：初始分组，分组细化和图形生成。前两个阶段涉及用户的数据收集。接下来介绍该算法的三个阶段：

阶段一：初始分组。将用户随机分为k0个相同大小的组（k0是系统参数），并将此分组方案*ξ0*传达给所有用户。之后，每个用户根据*ξ0*形成它的度向量，根据为阶段一分配的隐私预算添加噪声，并将噪声度向量发送给数据管理员。当数据管理者从所有用户那里收到噪声度矢量，它就使用（i）从收集的度数矢量中获得的所有用户的估计度分布，和（ii）隐私预算，计算具有k1个分区的新分组方案*ξ1* 分配到阶段二。

阶段二：分组细化。用户现在再次报告基于*ξ1*的噪声度向量以及为阶段二分配的隐私预算。 数据管理员进行另一轮用户分组，并根据用户的第二轮噪声程度向量将用户划分为k1个簇。阶段二可以重复多次，每次都会根据前几轮收集的程度向量逐步细化分组。

阶段三：图生成。数据管理者根据BTER模型生成合成社交图。数据管理员从阶段二中获得的用户分区开始，并使用收集的度向量生成边。然后，数据管理者从用户收集度矢量，并根据细化用户分组。之后在第二阶段，用户分组再次进行另一轮数据收集。最后，在阶段三中，基于用户分组和度向量生成合成图。

### 7.1.3 社交网络差分隐私保护技术

#### 7.1.3.1 节点隐私

节点差分隐私目的是通过保护节点和与它相连接的边来限制推理网络中节点的存在性。在节点差分隐私中，邻近图的定义是：对于两个图和，如果通过添加或者删除一个节点和与之相邻的边能够得到，那么称这两个图为邻近图。图7-4展示了原始社交网络图中添加一个新的节点9（4b），原始社交网络中删除节点7（4c）。

 (a)原始网络图 (b)添加新节点9 (c)删除节点7

图7-4 节点差分隐私的邻近图

定义 （-节点差分隐私）：设有随机算法，为所有可能的输出构成的集合。对于任意两个邻近图和以及的任何子集，若算法满足：



则称算法提供-节点差分隐私保护，其中参数是一个较小的正值，用来权衡隐私和精确度。

从直观上看，攻击者区分的输出和的输出的能力是非常有限的。因此，差分隐私模型表明，任何单个节点的添加或删除都不会严重影响数据分析的输出分布[66]。这意味着，图和将在统计数据上产生类似的分布。需要注意的是，图中的每个节点所包含的信息都对应于一个特定的节点和与该节点相邻的所有边。因此，节点及其邻边都应保持隐藏。

现今已经提出了各种方法来实现节点差分隐私。Hay等人[30]介绍了节点差分隐私的概念，并提出实现节点差分隐私的一些困难。Raskhodnikova和Smith [31]提出了一种节点差分隐私算法，用于释放图的度分布的近似值。该算法基于Lipschitz扩展和广义指数机制。同样，Borgs et al [32]采用与指数机制相结合的Lipschitz扩展方法来引入节点差分隐私算法。

Kasiviswanathan等[66]分析了几种节点差分隐私算法后，提出了分析实际网络中算法的准确性的方法。同时还讨论了使用线性规划进行Lipschitz扩展的功能，例如子图计数。他们定义了一个非严格映射，为此需要解决线性程序以获得输出。为了解决由于Lipschitz扩展引起的缺陷，Kasiviswanathan等人[66]提出了投影方法，其背后的核心思想是将输入图投影到最大度数低于某个阈值的图集上。输入图中度数高于阈值的任何节点都将被丢弃。第二种方法引入了一种称为“天真”截断的投影函数，它简单地丢弃图中的度数较高的节点。投影的局部敏感度根据图中具有预定义度的顶点的数量来限定。算法步骤如下：

输入：，，，度分布查询

输出：

1. 随机截断参数：；
2. 用计算和平滑上界；
3. 输出。

算法中有三个主要步骤。第一步是确定截断参数。由于我们可能不知道图中的最大度数或者度数非常大，因此，使用来近似。第二步通过丢弃度数大于的节点来创建截断图。最后，将满足柯西分布的噪声添加到度分布中。

Blocki等人[21]提出了用于分析稀疏图的节点差分隐私算法。他们将低敏感度投影用于给定最大度数的低度图集，同时对局部敏感度提供平滑上界。就投影图的结构而言，投影的局部敏感度根据图的与S的距离来约束。引入投影函数以增强Lipschitz扩展的弱点。 但是，需要线性规划来计算输入图的投影。

#### 7.1.3.2 边隐私

边隐私是一种比节点隐私更弱的概念[66]它已经比节点隐私保护得到了更广泛的研究。直观上讲，边隐私要求是否需要隐藏某一条边，但整体关系模式可以公开。在边差分隐私中，邻近图的定义是：对于两个图和，如果通过添加或者删除两个顶点之间的边之后能够得到，那么称这两个图为邻近图。图7-5展示了原始社交网络图中添加一条边（4，8）（5b）,原始社交网络中删除一条边（3，4）（5c）。

 (a)原始网络图 (b)添加新边 (c)删除一条边

图7-5 边差分隐私的邻近图

对于图和，其中，，，分别是图和的顶点集合和边集合，如果：



图的查询函数满足差分隐私保护则称满足-边差分隐私保护。

因此，边差分隐私确保算法的输出不会揭示图中特定的边。Hay等人[30]提出了一种用于发布社交网络图度分布的边差分隐私算法。他们的算法是对[33]中提出的全局敏感度方法的扩展，使用一种后处理的方法在差分隐私的输出结果中去除一些噪声。通过对真实和合成网络的经验分析，Hay等人[30]表明度分布估计显示出低偏差和方差，并推荐用于分析普通网络分布，例如幂律分布。然而，Hay等人提出的方法[30]不发布图。在某些情况下，网络数据的结构特性非常令人感兴趣，发布匿名图成为一种更好的选择。此外，该方法只能提供度分布的近似估计。

Nissim等人[34]在估计最小生成树的成本和图中三角形数量的情况下考虑边差分隐私。他们将查询在特定数据集上的局部敏感度定义为查询答案在图中添加或删除边时可以更改的数量。该机制添加的随机噪声量被校准为敏感度的局部变体，称为平滑敏感度。Nissim等人[34]也讨论了计算各种领域统计数据平滑敏感度的算法。

Block等人[35]提出了一种用边差分隐私分析图数据的算法。在边邻接中，假定相邻接图只在一个边或仅一个顶点的属性中不同，作为全局敏感度[33]和平滑敏感度[34]的替代方案，他们引入了限制敏感度的概念，以减少添加到输出的噪声量，从而提高差分隐私数据分析的准确性。对于边差分隐私，Blocki等人[35]提出了一个非常简单的投影算子，它具有不变的边敏感度。另外，Blocki等[36]将Johnson-Linden Strauss变换应用于近似图中的切割查询中。他们表明Johnson-Linden Strauss变换允许发布边差分隐私图。

Rastogi等人[37]针对一般子图计数发布问题，结合边差分隐私来抵抗贝叶斯攻击者。首先需要计算局部敏感度的高概率上限，然后将噪声添加到与计算的上限局部敏感度成比例的数据中。所提出的方法主要基于以下假设：边的存在不一定更有可能使其他边存在。该假设与社交网络中边之间的正相关性相矛盾（例如，拥有共同朋友的两个人更可能是朋友）。此外，关于攻击者先验知识的假设限制了隐私定义的适用性。

## 7.2 连续数据的差分隐私（边锦，章国政）

数据流作为一种实时动态数据类型，越来越多的被用作研究分析，快速而又准确的对数据流上聚集的信息进行发布与挖掘，可以获取有价值的信息，并利用这些信息来帮助人们更好的生活，但若不采取适当的保护措施，直接发布或者挖掘包含个人隐私的信息，将可能造成敏感数据的泄漏，从而给数据所有者带来危害。

差分隐私模型与传统的隐私保护模型相比，可以忽略攻击者的背景知识，使攻击者在已知除目标数据以外的其他所有数据的情况下，仍然不能确定目标数据是否真实，在极大程度上保护了用户的隐私。因此，随着大数据时代的到来，在数据流的发布和挖掘等领域，如何结合差分隐私技术，实现隐私保护与算法效用性之间的平衡，具有重要的研究意义。

### 7.2.1 时间序列及动态窗口

随着大数据时代的到来，云计算和信息共享技术的飞速发展，众多领域每天都要接收、处理、发布数以万计甚至无限的数据，区别于传统静态的数据，以流的形式出现的数据统称为数据流。数据流可以采用无限集合的形式来描述：，其中表示时刻到达的数据。数据流作为一种新的特殊数据形态，以不同速率持续地流入和流出处理模型，具有海量、连续、时序、潜在无限、快速持续到达等特点。

作为一种实时动态数据类型，数据不断更新且高速到达，其数据量远大于可用内存，因此对其处理及应用并非是针对整个数据流，而是引入数据流处理模型，将待处理的数据局限于某个窗口模型中。目前，在数据流研究领域，根据数据处理的时间范围，可分为三种数据流处理模型：快照模型，界标模型，滑动窗口模型。

快照模型（Snapshot Model）：预先设定窗口的起始时刻，终止时刻，根据设定好的时间范围，处理该时间段内对应的数据项，即处理时间到之间的数据流。快照模型的缺点是不能对即将到达的数据进行及时处理，只适用于实时性要求较低的场合，现各领域下生成的数据流大多具有较强的实时性，因此该模型不能满足大多数的应用。

界标模型（Landmark Model）：该窗口待处理的数据为从预先设定的起始时刻到当前时刻之间的数据项。随着时间的推移，窗口内的数据量也会不断增加，界标窗口模型未能及时对过期数据进行处理，因此采用随机抽样、直方图、小波分析以及哈希技术等数据分析方法，近似模拟能反映整体数据特征的概要数据，实现数据的发布和挖掘。

滑动窗口模型（Sliding Window Model）：该窗口的起始时刻和终止时刻会随着时间的变化，不断动态更新，新的数据进入窗口内，旧的数据不断被删除，保证了窗口中要保存和处理的数据都是最新到达的一些数据。任意相邻两个滑动窗口采样时，存在重叠数据，较好的实现了对历史及近期数据的处理。但由于数据流的实时变化及无上界增加，需要考虑内存空间的利用及磁盘的存取，这使得滑动窗口大小的设定成为一个难题。

针对数据流具有的数据流量大、连续、难以存储和恢复等特性，结合滑动窗口机制，并依据数据流在不同时间段内流速的变化，采用时间衰减模型，定义滑动窗口的调整阈值及调整大小，实现滑动窗口内，数据所占内存空间利用率的有效提高。

为了更好的实现滑动窗口的动态分配，首先定义基本窗口、滑动窗口及相关变量。

定义3-1（基本窗口）给定一时间，时间跨度，在时间段内到达的数据流数据组成一个时间相关基本窗口。

定义3-2 （滑动窗口）一个连续的时间基本窗口序列构成一个时间滑动窗口，每个滑动窗口对应的窗口单元大小为，记为个连续基本窗口构成时间滑动窗口，其时间跨度为：。

设滑动窗口的窗口单元大小为，假设理想的情况下，数据流的传输速度是均匀的，滑动窗口内的平均流速(Average Velocity, AV)，为最优数据流速，即内存利用率最高。而实际情况下，滑动窗口内的流速(Real Velocity, RV)，随着数据流的实时变化而随之改变。

基于数据流实际流速与最优流速比值，引入衰减因子，定义窗口调整值的大小，如公式所示：

 （7-1）

为了避免滑动窗口的频繁变化，基于窗口单元大小及最优流速，定义界定阈值，如公式所示：

 （7-2）

滑动窗口的窗口单元，根据调整后为，对应公式如所示。

 （7-3）

由于数据流如流水一般源源不断的产生于网络中，并且随着时间推移不断地出现新的数据，需要利用梗概数据结构存放部分数据，采用折中的方法权衡挖掘结果和实时响应，普遍只能挖掘出近似结果。针对海量性、有序性、动态变化性、实时性、高维性的动态数据流，白川平[38]提出了在传统的滑动窗口基础上提出了基于动态权值滑动窗口的频繁项集挖掘算法FIMDWSW（Frequent Itemsets Mining in Dynamic Weighted Sliding Window）和 FIMDWSW-Imp（Frequent Itemsets Mining in Dynamic Weighted Sliding Window - Improvment），在提出的算法模型中，用户给定挖掘窗口的数量、大小，根据窗口大小和挖掘时间动态调整不同挖掘窗口的权值。并通过实验验证FIMDWSW-Imp算法能更快地判定候选项集是否频繁项集，且在不同数量的事务和不同数量的滑动窗口下，FIMDWSW-Imp算法时间复杂度都低于FIMDWSW算法。

### 7.2.2 基于事件的差分隐私

随着互联网技术的迅速发展，引发了以隐私保护的方式发布实时统计数据的研究热潮。大量的应用程序受益于统计数据的持续监测。例如，交通服务实时发布每个区域的汽车数量，以便实现最佳路线计算。此外，医院会定期发布患有某些疾病的患者数量，这可能有助于及时发现疾病暴发。此外，社交网络不断报告当前正在讨论某个主题的用户数量，从而可以列出针对性主题的热门话题。然而，无意披露此类统计数据可能会泄露个人的隐私，例如上班族访问的地点，患者所患的疾病类型以及用户所传达的政治观点。差分隐私是一种流行的提供具有强大理论保障的数据统计发布保护隐私的范例。该框架需要在数据发布之前扰乱数据，以隐藏参与统计分析的个人的敏感信息。

最近，差分隐私框架已被应用于流式场景，其中统计信息在某些时间戳不断发布。这些统计数据是根据用户事件计算的，即用户在特定时间戳处采取的“行动”，这些行为有助于公布的数据。例如，假设交通服务定期发布每个位置的乘客数量。然后，在特定位置的通勤者的存在是在某个时间戳发生的事件。在这些设置中存在差分隐私的两种定义;事件级别和用户级别[39]。前者保护任何单个事件，而后者隐藏整个流中的任何用户的所有事件。例如，事件级隐私保护单个位置访问，而用户级隐私保护所有通勤者的所有位置访问。隐私级别会影响所使用的摄动量，这与敏感信息对统计数据的贡献成正比。有关流上差分隐私的先前工作主要集中在无限流上的事件级隐私[40-43]，以及有限流上的用户级隐私[44]。第一类在用户引起连续时间戳事件的实际场景中无用，该时间戳集体披露了敏感信息。例如，虽然事件级隐私保护任何单个位置访问，但它不保护连续时间戳中遍历的用户路径。第二类在大多数实际环境中的适用性有限，数据必须无限期地发布。例如，假设流量报告服务在先验已知时间间隔后关闭是有限制的。并且为无限流提供用户级隐私需要无限的扰动，从长远来看破坏了数据的实用性。

定义 (w-事件隐私) 设A是一个以任意大小的流前缀作为输入的机制。 令S为A的所有可能输出的集合。如果对于所有集合S⊆S \*、所有w-邻近流前缀、所有t，满足下式，我们就说A满足w事件差分隐私，

 (7-4)

当前最先进的数据流隐私保护隐私保护模式是w事件隐私保护。但是它忽略了如果数据集的一部分元素随时间变化而另一些基本上稳定下来，那么在指定的时间戳中处理所有用户数据将带来额外的噪声并降低数据的效用的情况。Mian Cheng等人[45]提出了一种新的隐私保护方法，称为G事件，它遵循常规使用的w事件差分隐私。根据差异计算将每个时间戳的统计结果分组，然后将高差异组比类似组更频繁地发布并保证所有更大变化的结果都会通过增加噪声来发布，变化较小的结果将与相应的最新发布的统计数据近似。最后通过实验表明，此方法提高了数据的实用性。

有些文献将差分隐私框架应用于无限的“事件”流（即由用户生成的数据项）计算统计数据并定期发布的设置，旨在隐藏整个流中的单个事件。在大多数实际情况下，敏感信息是从连续时间点发生的多个事件中揭示出来的。Georgios Kellaris等人[46]提出了w-事件隐私超越无限流的新概念，它保护了连续时刻内发生的任何事件序列。首先解释w事件隐私并在此基础上提出新的w-事件隐私机制。然后提出了一种滑动窗口方法，捕获各种各样的w事件私有机制。遵循提出的方法机制构建了每个时间戳的单独子机制，每个机制花费控制摄动的某个隐私预算（预算越高，摄动越低）。然后，当w时间戳的任何窗口内的机制的预算的总和最多是总隐私预算ε时，获得w事件隐私。并提出了新的、灵活的、动态的预算分配新方案 ——预算分配（BD,Budget Distribution）和预算吸收（BA, Budget Absorption）。当统计数据在连续时间戳中可能变化不明显的情况下，这两种机制有效地分配预算。

### 7.2.3 基于滤波原理的时间序列隐私保护

时间序列是某种统计指标按时间顺序排列所形成的序列，作为数据存储和发布的重要形式，时间序列普遍存在于多个领域。对时间序列进行数据挖掘可以带来很多便利，但由于时间序列包含个人敏感信息，数据挖掘的结果可能会泄露个人隐私，因此，如何发布个人数据且保证敏感信息不被泄露成为研究者关注的焦点。

时间序列作为一种典型的相关性数据，目前对其已有一些差分隐私保护方法，主要分为2类：1) 基于建模的方法；2) 基于变换的方法。基于建模的方法通过建立相关性模型重构敏感度函数，相关性模型主要有马尔可夫、贝叶斯等概率模型和系数矩阵模型。在基于建模的方法中，比较典型的是建立相关性序列发布概率模型。Cao等人[47]针对相关性序列中的异常数据会导致全局敏感度增加，提出了一种马尔可夫相关性模型，通过该模型中的一步转移概率检测并去除异常数据，从而减小加入到原始序列中的噪声水平，但由于检测结果不够精准。Yang等[48]提出的贝叶斯模型假定当前数据发布的概率与之前所有数据发布的概率有关，以此来提高检测结果的精确度。除了建立概率模型的方法外，Zhu等[49]利用相关性矩阵表示时间序列的相关性，并将相关性矩阵作为权重重新计算敏感度函数来代替全局敏感度，从而降低全局敏感度带来的额外噪声。

基于变换的方法主要有2种，一种将相关性的时间序列变换为另一个域的独立序列进行处理，代表性算法有离散傅里叶变换和离散小波变换；另一种提取时间序列的相关性特性并用一组独立特性表征，代表性算法有主成分分析等数据特征提取方法。在基于变换的方法中，主要通过将相关性的时间序列变换为另一个域的独立序列进行处理。Rastogi等人[50]采用DFT把时间序列变换到独立域，在变换系数中加入噪声之后进行离散傅里叶逆变换（IDFT)，从而实现对相关性时间序列的加扰，然而DFT是一种全局变换，对原始相关性时间序列局部特征的描述不够精确。作为一种改进方法，Xiao等[51,52]给出了DWT来拓展DFT方法的应用范围,相比于DFT，DWT可以保留序列更多的属性。此外，在处理高维相关性数据时，Jiang 等[53]利用主成分分析等高维数据降维方法提取相关性时间序列的主要特征，并利用正交分解等方式将相关性特征分解为几组不同的独立特征，然后在独立特征中加入噪声。相比于基于建模的方法，基于变换的方法在噪声水平上表现较好，可以保证较高的数据可用性，但这些方法损失了较多的相关性特征，对相关性时间序列的隐私保护强度不足。

现有的研究在相关性时间序列的差分隐私保护方面仍存在以下问题：1) 基于相关性建模和数据变换的方法假定时间序列的相关性符合一定的模型和规律，但隐私保护安全性的研究重点是“攻击—防御”体系的建立，目前的方法在攻击端没有针对具体的攻击模型展开研究；2) 由于不同的方法基于不同的模型和规律，且没有统一的攻击模型，它们的隐私保护强度无法进行横向比较和度量。

针对以上2个问题，熊等人[54]针对相关性时间序列差分隐私的攻击模型。根据信号处理中滤波的原理，设计一个线性滤波器以滤除相关性时间序列差分隐私保护机制加入的噪声，并计算现有方法在滤波前后隐私保护强度的变化，从而为它们的隐私保护强度提供统一的度量。攻击原理变化如图7-6所示：

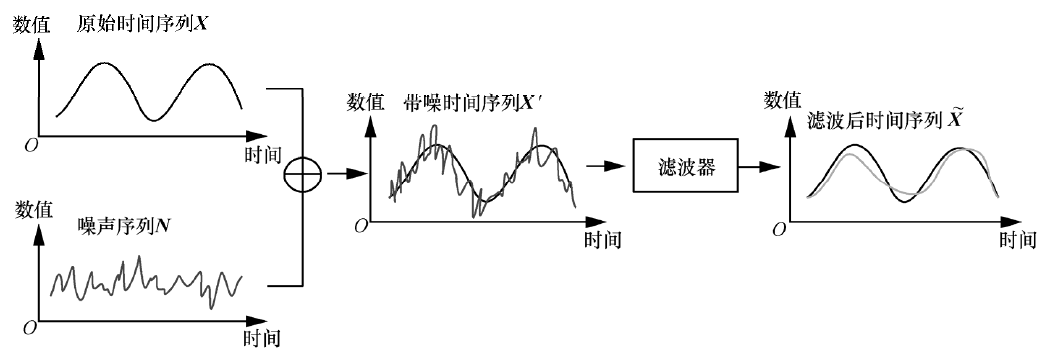


图7-6 攻击模型原理

由图1可知，带噪序列是通过机制在原始时间序列中加入噪声序列得来，即

 (7-5)

其中，。当通过冲激响应为的线性滤波后，由信号与系统中滤波的相关知识可知滤波后的序列为

 (7-6)

因此，滤除的噪声系列为

 (7-7)

根据维纳-霍夫方程，维纳滤波器冲激响应矢量的求解式为

 (7-8)

其中，是的自相关函数，是与的互相关函数，

因此，维纳滤波器的冲激响应矢量为

 (7-9)

由于噪声序列是白噪声序列，其自相关函数为

 (7-10)

因此得到带噪系列的自相关函数，以及原始时间序列与带噪序列的互相关函数，求解式为

 (7-11)

 (7-12)

其中，是一个列向量，是方阵。代入式（5）即可得到维纳滤波器的冲激响应矢量。

该攻击模型的工作过程为了将带噪时间序列中噪声进行滤除，因此将原始时间序列***X***设为输入，将带噪时间序列***X***′设为输出；并计算带噪时间序列***X***′的自相关函数***R***，以及原始时间序列***X*** 和***X***′的互相关函数***P***；然后根据***R***和***P***设计一个最佳滤波器冲激响应矢量***H***，以尽量多地滤除***X***′中的噪声；使用该最佳滤波器滤除***X***′中独立同分布的噪声，得到滤波后的时间序列***X***并返回。整个工作过程中中，最重要的部分是滤波器冲激响应矢量***H***的求解。

对4个原始时间序列数据集X分别采用WT、DFT、CIM、贝叶斯和马尔可夫5种相关性时间序列差分隐私保护方法，设置隐私预算,并计算4个时间序列数据集和其邻近数据集查询结果的概率密度函数，查询结果的精确度可用均方误差（MSE）衡量。

 (7-13)

*MSE* 越低，数据可用性越高。

### 7.2.4 基于差分隐私的流式直方图发布

目前,静态环境中已经存在许多满足差分隐私的直方图发布方法，但现有的基于分组的静态发布方法无法适应于数据流环境，其主要原因包括 3 点：（1）数据流的动态性要求所发布的直方图要连续更新，否则无法概要全部统计信息；（2）数据流的实时性要求直方图应及时发布，一旦新的计数值到达,需要立即对其保护处理,然后发布；（3）最后一个主要原因是，现有的静态方法无法一次性把所有的数据载入内存进行处理.此外,这些方法也无法较好地均衡数据流中直方图的近似误差与拉普拉斯误差。

针对上面的问题，张啸剑等人[55]提出了一种有效的数据流直方图发布方法SHP (streaming histogram publication)。通过连续分割每个滑动窗中的桶计数，使其构成不同的分组，根据不同的范围计数查询敏感性，采用拉普拉斯噪音添加机制以实现差分隐私保护，并基于当前的滑动窗，依赖于一种自适应抽样方法对下一时刻的计数进行预测，对比预测误差值与给定的阈值大小，重新估算待发布的直方图。

SHP算法利用大小固定的滑动窗模型处理数据流。一个长度为T的数据流，由定义在T个时间戳上的数据点构成，其中每个数据点包含第和第个时间戳之间的动态信息。采用滑动窗口机制对数据流进行建模，每个滑动窗口由3个参数定义：开始时间、当前时间和窗口大小(所包含数据点的个数)。给定滑动窗，设，其中包含w个数据点，为了能够比较精确地响应中的范围计数查询，H通常被分割成大小不同的分组。通常，所有滑动窗口均使用固定窗口大小。滑动窗每次向前滑动一个时间戳时，SHP均会发布一个满足差分隐私的直方图。

SHP算法流程如下所示：

|  |
| --- |
| **输入:长度为的数据流，滑动窗口，隐私参数，中分组个数，采样频率.** |
| **输出:隐私处理后的直方图** |
| 初始化第1个直方图  **1.**  窗口滑动阶段  **2.**  **3. for** **do**  **4.**  **5. end for**  **6. for each** new arriving count **do**  **7. if**  is sampled **then**  **8.**  **9. else**  **10.**  **11. end if**  **12. Insert**  **or** into the current window ;  **13.**  **14.**  **15. end for**  **16. Compute** for every group ；  **17. Return** |

SHP算法根据给定一个数据流D和相应的参数，首先利用Partition方法对第1个滑动窗中的数据流实现初始化，并发布该窗口中的直方图，如算法流程第1行所示。

然后，滑动窗基于时间戳向前滑动，SHP算法分两种情况处理当前窗口中的数据流：一是先利用Adaptive-Sampling方法，以si为抽样频率在当前窗口中进行随机抽样，如算法流程第2行~第5所示；二是对新来的计数值进行处理，如算法流程第6行~第15所示。如果新来的数据值恰好被抽样，则直接对该值添加拉普拉斯噪音，算法流程第7行~第8所示；否则，基于已经抽样的数据值，利用Estimation方法对新值进行估计，如算法流程第10行；在当前窗口中，结合抽样值与估计值再次利用Partition方法对当前的数据进行分割，如算法流程第12行~第14所示。

最后，发布当前窗口中的直方图，如算法流程第16行~第17所示。

之后针对差分隐私下基于分组的直方图发布问题，张啸剑等人[56]提出了一种满足差分隐私的精确直方图发布方法DiffHR,通过分析直方图桶计数序列的排序有助于提升发布精度，利用Markvon链蒙特卡洛方法中的Metropolis-Hastings技术与指数机制，提出了一种有效排序方法，通过不断置换2个随机选取的桶以逐渐逼近正确排序，基于抽样排序后的直方图，提出了一种基于懒散分组下届的自适应贪心聚类算法，并且可有效均衡近似误差与拉普拉斯误差。

Li等人[57]针对流式直方图发布的隐私保护问题，提出自适应抽样的差分隐私流式直方图发布算法(Adaptive Sampling based Differential Private Histogram Publication Algorithm for Data Stream, ASDP-HPA)，该算法基于动态调整的滑动窗口，采用差分隐私技术下的拉普拉斯机制进行加噪，并为了解决发布的流式直方图可用性低、发布误差大等问题，依赖抽样方法，并结合自回归积分滑动平均模型，重新估算待发布的数值。

ASDP-HPA算法利用ARIMA模型，预测下一个滑动窗口待发布的直方图，并对预测值进行检验。通常采用和方差（Sum Squared Error，SSE）来度量产生的误差。算法采用自适应抽样对待发布的直方图进行加噪，其中抽样频率的定义公式为：

 （7-14）

若到达的不是一个抽样点，则需要对ARIMA预测生成的误差值与加入Laplace噪音生成的误差值进行比较，即当预测值满足以下条件：

 （7-15）

则发布预测值，否则发布加入Laplace噪音的估计值。若到达的是抽样点，则直接发布加入Laplace噪音的估计值。

ASDP-HPA算法在滑动窗口下进行直方图发布的主要流程如下：

|  |
| --- |
| ASDP-HPA算法 |
| 输入：流入滑动窗口的数据集，历史滑动窗口生成的直方图，历史隐私保护分配预算，隐私保护预算，参数，敏感度，直方图桶个数 |
| 输出：ASDP-HPA在滑动窗口下发布的直方图 |
| Begin  1： //对数据集进行预处理  2： //生成滑动窗口待发布的直方图  3： //ARIMA模型预测滑动窗口待发布的直方图 //  4： //初始化ASDP-HPA算法在滑动窗口下生成的直方图  //  5： //计算隐私预算，根据数据集可调整参数  6： //计算抽样概率  7:  //计算抽样个数  8： //在0-之间随机生成个不相同的整数  8： for  do //遍历  9: if  then //若为抽样点  10:  //若被抽中，直接发布加躁值  11: else if  then //若为非抽样点且满足条件  12:  13: else //若为非抽样点且不满足条件  14:  15: end  16: end  17: end  18：Return  //输出直方图  End |

## 7.3 高维数据的差分隐私（王寒，陈竑毓）

真实世界中的很多数据都是高维的，例如航天遥感数据、生物数据、网络数据以及金融市场交易数据，这些数据都包含很多属性或特征。尽管高维数据比低维数据拥有更多的信息量，但在实际应用中对高维数据进行直接操作会有一系列困难，其处理面临两个问题：一是维数灾难问题，维数膨胀给高维数据中模式识别和规则发现带来极大挑战；二是维数增长带来的“维数福音”问题，高维数据中蕴藏的丰富信息中可产生解决问题的新的可能性。为了解决高维数据所面临的问题，一种有效的方法是对其进行降维。笼统地说，降维是指为高维数据获取一个能忠实反映原始数据固有特性的低维表示。降维可分为特征选择和特征提取两大阵营，前者是从原有属性或特征集中一句某种准则选择一组特征子集，后者是将原空间变换到一个维数更低的空间并将数据映射到该空间。两者对比，特征选择仍然属于原有特征集，而特征提取会产生不属于原有特征集中的新特征，比如，是原有特征的线性组合。

目前降维已经成为机器学习、模式识别、数据挖掘、信息检索等多个研究领域共同关注的核心课题，并已有大量的有关降维的方法。本文基于特征提取阵营的主成分分析法（Principal Component Analysis, PCA）以及其他降维方法，介绍其对高维数据进行降维的原理，并且针对其存在的隐私泄露问题，提出使用差分隐私解决方法应对隐私泄露问题。

### 7.3.1 高维数据及降维

高维数据拥有成百上千甚至更多个特征，其用维度来表示特征的个数。这种类型的数据比较难处理，一方面随着维数的增加，要使高维数据具有价值，则需要更多的样本数据；另一方面，高维数据往往是比较稀疏的。

由于上述问题的存在，使得直接处理高维数据会出现不少问题，例如维数灾难问题，指的是随着维数的增长，高维网络数目呈现指数增长，使得不可能在这样一个离散的多维网格中用穷举搜索去优化一个具有很多变量的函数。很多在低维数据集具有优良性能的聚类、分类算法，在高维数据中性能将大大降低甚至完全失效。再比如空空间问题，由于样本集太小，使得高维空间中样本点及其稀疏，进而导致整个高维空间很空，其他例如不适应、算法失效等问题都发生在高维数据场景下。

为解决上述存在的问题，对高维数据进行降维是一种有效的解决思路。高维数据可以实现降维的原因在于数据的原始表示常常包含大量冗余并且有些特征之间的关联程度很高，因此可以通过去除噪声、冗余特征，来获取数据的本质特征，即实现高维数据的降维处理。

一般来说，高维数据降维可以分为特征选择和特征提取。

特征选择是从原始数据特征中选择一个满足条件的最优特征子集来建立模型的过程，又称选维。即设有数据，包含个特征，，特征选择就是要选出个最具代表性的特征构成新的特征矢量，其中。特征选择的特点是不会丢失单个重要特征的信息，但是如果子集个数较小，原始特征又很不相同的情况下，特征选择会导致信息丢失。特征选择又分为过滤器模型，包裹模型和混合模型三类，SFG、WSFG、BBHFS是其中常见的三种模型。

特征提取是通过映射函数将特征从原始数据空间映射到新的特征子空间的过程。对于特征矢量，变换函数，经变换降维以后得到维新的特征矢量，其中。特征提取可以在不丢失原始特征空间很多信息的情况下，缩小原始特征空间的规模，降低特征维度。特征提取又可分成线性变换和非线性变换两种，其中主成分分析（Principal Component Analysis，PCA）、线性判别分析（Linear Discriminant Analysis，LDA）和局部线性嵌入（Locally Linear Embedding，LLE）是其中几个常见算法。

本节基于高维数据的降维处理背景，引出特征提取方法解决高维数据降维问题，并在3、4小节中介绍几类满足差分隐私的降维解决方案。

### 7.3.2 特征提取

高维数据降维是要将高维数据转换为地位数据而不失其重要信息。特征选择只选取一部分特征用于模型构建，而丢弃其余的特征不予考虑。而特征提取是通过一个变换过程将数据从数据高维特征空间映射到新的低维子空间，从而实现降维，一定程度上而言，特征提取保留原始数据的信息更多。

给定，特征提取是要寻找这样一个映射函数：，其中是原始样本点的表达，是样本点映射后的低维向量表达，，使得变换以后的特征向量可以保持的大部分信息，或者保持的流行结构。总的来说，特征提取是一种新特征的构建过程，这些新特征满足：1）是原始特征的线性组合；2）彼此正交；3）具有数据中的最大方差和。

特征提取算法依据映射函数的类别分为线性和非线性两类，其中线性特征提取假设高维数据存在于低维的线性子空间，使用矩阵分解将原始数据投影到子空间。给定数据集，存在一个投影矩阵和一个映射矩阵，其中，利用特征向量的正交属性，可以得到。其中线性特征提取典型算法有PCA、LDA、LPP等。非线性特征提取有多种形式，可以考虑将低维数据映射到高维空间以寻找特征之间的非线性关系，也可以使用流形。对于寻找特征之间的非线性关系，使用映射函数将特征映射到高维空间发现特征之间的线性关系，再映射回低维空间，特征关系又可以被看做是非线性的。对于这种方法，常见的算法有核主成分分析Kernel PCA。对于流形，假设数据点存在于比原始数据更低维的嵌入的非线性流形上，按图论的观点，构造样本点的近邻图及权值矩阵，最小化流形重构误差函数，计算流形的低维嵌入，常见的算法有LLE、LE等。

3、4小节介绍几个常见的特征提取算法，并介绍其差分隐私解决方案。具体地，第3小节详细介绍主成分分析算法及其差分隐私解决方案，第4小节简要提及几个算法及其差分隐私解决方案。

### 7.3.3 主成分分析

主成分分析是一种使用广泛的线性降维方法，其主要思想是使用尽可能少的特征去表示原始数据，形象表述，是通过线性变换将原始数据从原来的坐标系转换到新的坐标系，转换坐标系时，以投影后的方差大小作为新坐标轴的选取标准，因为数据的最大方差会给出数据最重要的信息。第一个新坐标轴，即第一主成分选择的是原始数据中方差最大的方向，第二个新坐标轴选择的是与第一个选择的坐标轴正交的候选集合中方差最大的方向。以此类推，重复选择新坐标轴，最终大部分方差都包含在最前面的几个新坐标轴中，忽略余下的坐标轴，以此达到降维的目的。过程可以简述为，PCA通过找到个正交方向使得高维数据在这个方向的投影具有最大方差，利用这个方向的投影来表示原始数据，通常比原始数据的维数小得多。

对于上述降维的解决思路，可以用特征值与特征向量解释。例如，对对称矩阵而言，可以求得正交的特征向量，即对所代表的空间进行正交分解，使得中的向量可以由它在各个特征向量上的投影长度来表示，这个投影长度即为对应的特征值。据此，求出投影长度前大的分量，丢弃余下分量，可以尽最大可能保存矩阵包含的信息，同时能够大大降低矩阵维度，也就实现了降维操作。

设原始数据集，低维空间维数为，现给出PCA算法的具体步骤如下：

1. 对所有样本进行中心化：；
2. 计算样本的协方差矩阵；
3. 对协方差矩阵做特征值分解；
4. 取最大的个特征值所对应的特征向量分别作为列向量组成变换基；
5. 计算降维结果：。

以上过程将原维向量降维到维向量。即PCA方法通过计算数据矩阵的协方差矩阵，然后得到协方差矩阵的特征值及特征向量，选择特征值最大的个特征所对应的特征向量组成的矩阵，就可以将数据矩阵转换到新的空间中，实现数据的降维。

上述步骤能够有效将高维数据进行降维，从而方便算法及模型进行处理，但是在这过程之中，其存在隐私泄露过程，差分隐私机制可以很好地解决这个问题。

下文将介绍两种不同的差分隐私机制来保证PCA算法实现过程原始数据的隐私安全，差分隐私PCA机制通过迭代选择特征向量并进行扰动来保证隐私安全。迭代方法基于频谱分解，即确保如果从协方差矩阵中除去前i-1个特征向量对应的分量，则第i个最大特征向量成为剩余的特征向量中最大的特征向量。因此，选择最大的K个特征向量的过程可以通过重复找到最大的特征向量并去除与所选择的特征向量相关的分量。下文提到的两种方法都使用这种思想，但是区别在于如何选择第一个特征向量。

文献[58]提出了一种-差分隐私机制，这个机制使用幂方法，规定如果v与第一个特征向量不正交，则认为可以收敛于协方差矩阵A的第一个特征向量。机制随机选择一个单元长度的向量开始，并不断迭代，每次迭代向量的更新过程为，其中表示第次迭代过程添加的高斯噪声。因为选择的随机向量与第一个选择的特征向量正交的可能性极小，根据上述规定，所以经过固定次数的迭代之后，向量会收敛于第一个特征向量，但考虑到添加的噪声，向量不会完全收敛，因此可以看作非常接近于第一个特征向量。因此，这个机制输出经过若干次迭代的向量作为当前的最大特征向量，然后依照同样的方法寻找下一个最大的特征向量。幂方法对于输出第一个特征向量具有效用保证，但其对于总的K个特征向量不能完整其效用，对于每个在协方差矩阵上使用幂方法输出的特征向量，其与第一个特征向量间的距离（，其中表示第一个特征向量）是数量级的。

文献[59]提出了一种PCA的-差分隐私机制。对于任意单位长度向量，协方差矩阵的第一个特征向量满足最大的性质，因此文献[2]提出的机制使用作为指数机制中的打分函数来从集合中选择第一个特征向量，并让其满足差分隐私，并且这个选择算法规定需要在可接受的时间完成计算。除此之外，文中还给出了机制的两个效用性证明，对于任意的以及隐私预算，如果协方差矩阵的第一个特征向量满足，则第一个特征向量具有以下性质：；若对于任意的以及隐私预算，如果协方差矩阵的第一个特征向量满足，用表示第个特征向量，用表示阶的矩阵输出，则上的最大特征向量在很大概率上小于的值。

不是所有的差分隐私PCA方法都必须使用迭代方法，文献[60]提出了一种能够同时计算个最大特征向量的-差分隐私机制PPCA。该机制利用协方差矩阵的前个特征向量是的矩阵其中的列向量这个性质，使用作为指数机制的打分函数，并设置输出对象满足，即输出值只能从满足的矩阵集合中选择，这里选择Gibbs采样器方法来选择矩阵。文献[60]对第一个特征向量提供了保证，对于任意，如果样本大小，则以的概率输出第一个特征向量和真实向量的内积大于。

### 7.3.4 降维聚类

数据降维是把数据从高维空间通过线性或非线性映射投影到一个低维空间，在降维过程中最大限度地保持原始数据的本质特征，实现高维数据的低维表示。本节介绍2个降维算法和差分隐私结合的隐私保护技术，包括随机投影和线性判别分析。

#### 7.3.4.1 随机投影

随机投影是一种简单且计算有效的降维方法，通过交易受控量的误差来减少数据的维数，从而缩短处理时间并缩小模型尺寸。控制随机投影矩阵的尺寸和分布，以便近似维持欧氏空间内成对映射点的距离，使变换降维后的数据具有原始数据的重要特性信息。利用随机投影实现低失真嵌入的依据是Johnson-Lindenstrauss引理[61,62]。

在随机投影中，使用随机将原始*d*维数据投影到*k*维()子空间{\ displaystyle k \ times d}维矩阵*R*，*R*的行具有单位长度。使用矩阵表示法：如果{\ displaystyle X\_ {d \次N}}是*N*维的原始集合，则是将数据投影到较低的*k*维子空间。随机投影在计算上很简单，形成随机矩阵“R”并投影{\ displaystyle d \ times N}数据矩阵*X*到*K*维{\ displaystyle O（dkN）}。如果数据矩阵*X*是稀疏的，每列大约有c个非零项，则该操作的复杂度是{\ displaystyle O（ckN）}。

Faraz Ahmed[63]等人提出了一种发布网络数据的隐私保护机制。该方法首先使用随机投影将邻接矩阵的每一行投影到低维空间，然后使用随机噪声对投影矩阵进行随机干扰，最终发布扰动的矩阵。随机投影的方法降低了发布矩阵的维数，有效的避免了发布一个大型密集矩阵，并且随机投影的方式降低了数据维数，本身就具有隐私保护的效果，这增强了整个算法的隐私保护效果。该算法的具体步骤算法1所示。

输入：邻接矩阵，随机投影数，随机噪声变量

输出：扰动的随机矩阵

1. 计算一个随机投影矩阵*P*，
2. 计算一个随机扰动矩阵*Q*，
3. 计算投影矩阵
4. 计算随机扰动矩阵

#### 7.3.4.2 线性判别分析（LDA）

线性判别分析技术主要用于对数据集的分类中，他将带标记的数据按照类间离散度最大，类内离散度最小的方式进行映射，得到投影矩阵。基本思想是将高维的模式样本投影到最佳鉴别矢量空间，以达到抽取分类信息和压缩特征空间维数的效果，投影后保证模式样本在新的子空间有最大的类间距离和最小的类内距离，即模式在该空间中有最佳的可分离性。

把每一条原始数据的记录看作符合高斯分布的独立同分布的随机变量，那么类的样本均值为：



总体样本均值为：



类间离散度矩阵：



类内离散度矩阵：



目标函数：

越大越好

得到：



就是将特征分解后的特征向量，这样构成的投影是最佳投影。

戚名钰[64]等人提出了一种基于LDA的差分隐私数据发布算法，具体流程如下：

输入：原始数据集，差分隐私保护预算，分类数目

输出：发布数据集

1. 将原始数据集归一化，使对任意，满足
2. for i=1 to p

计算，得到向量

end

1. 计算类间离散矩阵



1. 计算类内离散矩阵



1. 计算映射矩阵



1. 计算，其中，，选取中前个特征向量，组成矩阵（行列）
2. 计算，得到投影矩阵
3. 对投影矩阵中的元素

for i=1 to c

for j=1 to n

计算，构成矩阵

end

end

1. for i=1 to p

计算，得到向量

end

1. 输出矩阵

Xiaoqian Jiang[65]等人提出了一种差分隐私LDA算法。由于LDA只使用不同类别数据的均值和方差-协方差矩阵，因此必须尽可能多地保留它们的信息。其基本思想是：首先得到这些嘈杂的统计数据，然后从中生成数据。假设所有p变量都是[0，1]中的数值变量，该机制详述如下：

1. 对于分类给定一个数据集或者矩阵，假设数据集中的每一条数据都被限制为[0,1]，得到两个分类和；
2. 在噪声总和上增加噪声，然后将它们除以两个分类的大小，整个数据集得到和；
3. 在其统计数据和噪音统计数据之间进行近似最小化L1距离抽样。

## 7.4 地理位置数据的差分隐私（李映乐，章国政）

无处不在的移动定位设备的部署和高速无线网络的广泛使用，极大地收集了个人在道路网络上的大规模轨迹数据。而这些轨迹数据通常由一系列三重位置信息序列（纬度、经度、时间戳）所组成，且该序列可以捕获用户的各种旅行信息，包括用户的运动模式、旅行路径和旅行目的地等。

虽然通过收集这些地理位置数据可以帮助分析个人移动用户的轨迹，以全面了解用户的个人旅行建议，但对不同移动用户扫描的历史轨迹数据的汇总分析可以提供更为普遍的旅行建议，比如某一城市的十大必去地点，或是哪一个购物中心在这个地区最受欢迎等等。

地理位置轨迹数据提供了准确的、有用的兴趣点推荐信息，但敏感轨迹信息的暴露同时也会给用户造成隐私安全风险，特别是定位信息，它一般被表示为二维地理区域，且这些数据通常与语义相关，如大学、购物中心和医院。移动用户与该位置之间的关联的披露可以揭示关于用户的健康状况、生活方式和社会及政治信仰等私人信息。例如，如果对手推断出用户和治疗中心之间的关联，则可以显示用户的健康状态。

因此，针对地理位置数据，我们要研究的课题，是如何平衡个人隐私，同时使个人的位置数据可运用于挖掘和分析。在发布基于个人信息的信息时，我们需要处理两个关键的挑战。首先需要限制这种固有的标识位置相关信息，但不会对数据的潜在效用产生不利影响；其次，我们需要确保个人隐私可扩展到大数据。

目前，地理位置数据又可以进一步细分成历史轨迹数据和实时位置数据。下面，我们将从这两个方面分别叙述。

### 7.4.1 历史轨迹数据的隐私保护

定位技术与普适计算的蓬勃发展催生了历史轨迹大数据，轨迹大数据表现为定位设备所产生的大规模高速数据流。及时、有效地对以数据流形式出现的轨迹大数据进行分析处理，可以发现隐含在轨迹数据中的异常现象，从而服务于城市规划、交通管理、安全管控等应用。然而，受限于轨迹大数据固有的不确定性、无限性、时变进化性、稀疏性和偏态分布性等特征，传统的隐私保护技术不能直接应用于轨迹大数据的安全保护。且静态轨迹数据集的隐私保护方法通常假定数据分布先验已知，忽视了轨迹数据的时间特征，也不能评测轨迹大数据中动态演化的异常行为。

针对于上述问题，Li C[68]等人提出了一种针对用户的兴趣点推荐的差分隐私轨迹分析算法，其目的在于最大化地推荐结果的准确性，同时保护暴露的轨迹的隐私。该算法首先将原始轨迹数据集转换成具有代表用户和兴趣点的二分图，然后提取表示二部图的关联矩阵。并且在针对产生推荐点数的有序列表的转接数据进行超链接诱导主题搜索（HITS）之前，执行扰动关联矩阵的后处理。实验证明，在满足所需的差分隐私保证的情况下，ITIM是高效的、可扩展的。Wang S[69]等人则提出了一种新的基于差分隐私的分级消减方法（DPHLS），为了保护轨迹中的各个位置点，并最小化隐私所需的噪声幅度，该方法引入了*"*数据集；且算法通过一种个性化的分层机制来有效降低灵敏度，提出利用历史位置历史以灵活和可控的方式混淆用户位置。算法可以成功地扩展到大数据场景，同时利用数据集（轨迹）和保护个人用户隐私。

除此之外，针对传统的*k-*匿名轨迹数据发布技术无法有效的保护用户隐私，基于差分隐私的（*k-*）匿名轨迹数据发布隐私保护算法也被提出。该算法基于（）匿名，考虑轨迹上各时刻不确定阈值在不同路网区域上路段的变化，构造具有自适应的不确定阈值集合，进而提高*k*条轨迹之间的不可分辨性；并且采用添加拉普拉斯噪声的方法对*k*条轨迹进行差分隐私扰动，从而可以抵御具有强大背景知识攻击者的攻击。

轨迹差分扰动是以位置差分扰动为基础，对轨迹集合中的条轨迹在时刻个位置组成的位置数据集合进行扰动，其中，，位置数据集合表示为，。现对中的每个位置坐标添加拉普拉斯噪声，从而确保这个位置点中任意两个位置点产生相同扰动的概率总是被限定在常数因子内，从而确保具有相近位置知识的攻击者很难推理出真实的位置点。本文对位置数据集中的个位置都进行添加拉普拉斯噪声扰动，下面为添加拉普拉斯噪声的轨迹差分扰动算法，其中拉普拉斯噪声为。

|  |
| --- |
| 算法2 ：添加拉普拉斯噪声的轨迹差分扰动算法 |
| 函数1：*D GetSourceData*(*Trails,T*)  输入：轨迹集合*Trails*的数据*Trails=*{}，时间集*T*={}  输出：需要添加噪声的原始数据集*D*={}  步骤：  1. 对每一个  2. 对每一个*Trails*  3. //表示第j条轨迹在时刻的位置坐标  4. 结束遍历  5.  6. 结束遍历  7. 返回  函数2： *LapDifferential*(*D,*)  输入：原始数据集*D*={}，常数因子  输出：添加Laplace噪声处理后的数据集  步骤：  1. 对每一个  2. 对每一个  3. +[-bsign()ln(1-2||)]  4 +[-bsign()ln(1-2||)]  5  6. 结束遍历  7.  8. 结束遍历  9. 返回 |

### 7.4.2 实时位置数据的隐私保护

随着信息技术的发展，基于位置的服务（Location-based services）由于其具有使城市变得更加智能的潜力而受到人们的普遍关注。位置数据主要来源于车联网、移动社交网络、新微博等网络平台，具有种类多、数量大、复杂而又稀疏（低密度性）的大数据特点。一方面，位置数据技术为本文提供了便捷的位置服务，带来生活上的便利，如车载导航、智能定位服务等；另一方面，位置服务业会泄露个人隐私，如攻击者根据某人的轨迹，推测出某人出现在某地的频率。位置数据隐含了移动对象、空间坐标、当前时间和某些独有特点等，离散而又价值高，匿名或者基于匿名的大部分隐私保护方法主要是对非位置数据的保护，而不适合位置数据的隐私保护，其主要原因有二点：一是位置数据的复杂性是数据难于表示，二是对背景知识的敏感，随着攻击者对某人有关位置数据认知的增加，可以分析出其隐私。所以很多隐私保护方法要么是不能保护数据，如匿名方法；要么就是数据的效用性不足，如加密方法。

隐私保护的关键是保护数据的同时又具有较高的数据效用性，如何兼顾数据的保护和数据的效用性，是研究的难点。Samarati等人[70,71]首次提出了k-匿名方法以及大量基于k-匿名的方法,在很多保护系数不高的领域起着很大的作用。文献[72-74]研究表明，匿名方法并不能很好的保护那些来源广泛、数量大的数据；文献[75]提出了加密的私密性保护方法，能够完全保护数据，但数据的可用性不足。传统的位置数据隐私保护技术发展经历了三个阶段，由文件[76]提出的“知情与同意”方法发展到针对单个位置的匿名化处理，再到对轨迹数据的匿名化处理。由于企业因无法承担预先请求用户对某位置数据创新性用途后通知用户同意的损失，所以“知情与同意”方法的效用性不足。

启发式隐私度量方法[77]、基于概率推测[77]和基于隐私信息检索方法[77]是保护位置数据的常用方法，启发式隐私度量方法主要是对于一些隐私保护不严格的用户进行保护，如k匿名[78]、t紧密型[79]、m不变性[80]、1多样性[81]等方法，基于信息检索的隐私保护方法能对用户实现完美差分隐私保护，但由于数据发布就会存在或多或少的隐私信息，所以这方法将导致没有数据可以发布，而基于概率推测的方法在一定条件下能保护数据和实现数据较好的效用性。但这三种方法都是适用文献[77]提出的统一的攻击模型，是在一定背景知识的前提下对位置数据进行保护，但随着攻击者对背景知识的增加，这些方法就不能有效地保护数据了。

文献[82]应用压缩感知理论提出了一种感知机制，是一种能解决效用性不足的方法，但它破坏了数据之间的联系，而文献[83]提出了一种基于压缩感知机制的一种事务数据发布策略（TDPS）算法，它通过建立查询树把事务中的项与项联系在一起来提高效用性，但是它缺乏效用性上的严谨定义，文献[84]提出了DP-topkP方法，在效用上有较为严谨定义，但忽略了事务数据间的联系和单独处理数据的低效处理方式，算法可用性不高。

杨理皓等[85]人提出的 LQ-Trie-DPK 算法为一种较为严谨的差分隐私保护方法，既能保证数据的高可用性，还能增强算法的可用性，满足位置数据保护的需求。首先是建立位置搜索树（Location query tree LQ-Trie)，其次遍历该树，选取位置访问频率大于 min\_Count（指定的最小访问值）的 N 个模式记录，再根据指数机制在 N 个模式记录中提取其中的 k 个频繁模式记录，最后基于差分隐私保护的拉普拉斯机制对位置访问频率加入噪声，最后返回加噪的位置搜索树 LQ-Trie。

## 7.5社区发现差分隐私保护（王寒）

社交网络通常公开社区结构，其中节点组与同一组中的其他节点共享许多连接，而与其他节点共享的连接相对较少。该特性捕获关于组织甚至网络演化的有价值的信息。在过去的十年中，针对日益复杂的网络，提出了大量的社区检测算法。然而，针对该过程中的隐私保护问题却日益突出。因此，本节将针对若干社交网络社区发现算法的差分隐私保护进行阐述。

### 社区发现算法概述

（1）Louvain社区发现算法

Louvain算法是一种基于模块度优化的社区发现算法，其目标是最大化整个社区网络的模块度。它能够在很短的时间内发现大型复杂网络中的社区结构，且不需要人为设定划分社区的数目，同时，与其他的社区划分算法相比，它能够检测到不同分辨率的社区，从而不会遗漏一些小社区结构，因此，Louvain社区划分算法很适合本文的大型复杂网络的情况。Louvain算法的步骤如下：

1. 将网络中的每一个节点都视为一个社区；
2. 对于每一个节点，将节点从当前社区尝试加入到其邻居节点所在的社区，并计算加入前后的模块度增量，其计算公式为：

 (7-16)

其中，表示起点和终点均在社区内的边权重之和，表示指向社区内的边权重之和，表示节点的权重度数和，表示节点与社区内节点之间的边权重之和。找出最大的节点的社区编号，如果，则把节点加入到当前社区，否则节点仍在原社区不变；

1. 重复上述过程，直到所有节点不再发生移动，此时社区划分已经达到最优，对网络进行聚合，生成一个新的网络；

（2）Kernighan-Lin（KL）算法

KL算法是一种基于图分割和贪婪优化思想的试探性二分算法。KL算法在划分的过程中需要引入一个增益函数，表示两个社区内部的所有边数和社区之间连接边数的差。KL算法在不同结构的网络中都比较适用，精度较高，但是需要预先知晓网络最终划分的社区数目以及每个社区所包含的节点个数，算法复杂度较高，不太适用于大规模的网络。

KL算法的步骤主要是：

1. 在提前预知每个社区包含的节点个数的情况下，将网络中节点随机划分为两个社区，并且计算当前网络的增益函数的值；
2. 交换网络中属于不同社区的所有节点对，计算增益函数的变化量；
3. 选取具有最大值的节点对进行交换，若增益函数变化最大量，算法结束，否则交换相应节点，重复第二步过程，即计算采用交换节点对的方法来获得最大值。

（3）Girvan-Newman（GN）算法

GN(Girvan-Newman)算法是一种经典的基于边介数的分裂层次聚类算法。GN算法引入了一个边介数参数，网络中一条边的边介数是指经由该边的最短路径的数目。通过不断计算并删除边介数最大的边，逐步实现对网络的社区发现，最终得到合理的社区结构。由于算法需要不断重复计算边介数，因此，GN算法的时间复杂度较高。

GN算法划分步骤如下：

1. 分别计算网络中各个连边的边介数，获取全部连边的边介数；
2. 比较网络中存在连边的边介数，从中找到边介数最大的边并删除；
3. 若网络中所有连边全部被去除，则划分结束，否则重新计算网络剩余边的边介数，并返回第二步循环执行。

（4）K-means算法

基于划分的K-means算法是通过不断更新每个社区的聚类中心，迭代计算目标函数值，直到满足收敛条件。K-means算法思想简单、应用广泛且收敛速度快，需要预先设定目标社区数量K值，并且依赖于初始社区中心的选取。K-means算法的基本思想包括：

1. 从网络中随机选取K个数据对象作为初始的社区聚类中心；
2. 分别计算网络中剩余的数据对象到每个社区中心的距离，把数据对象分配到与之距离最近的聚类中心所代表的社区中；
3. 重新计算出K个新的社区中心；
4. 与上一次得到的K个社区中心进行比较，若社区中心不再发生变化或达到预设的阈值，则算法结束，否则转到步骤2重新划分社区结构。

（5）LPA算法

LPA算法基于标签传播思想提出的。该算法最初为每个节点分配唯一的标签，接着选择邻居节点中标签出现频率最高的标签来更新当前节点的标签，利用标签更新策略重复迭代更新节点标签，直至网络中各节点的标签不再发生变化或者达到最大的迭代次数，从而具有相同标签的节点则被规划到同一社区中。LPA算法的步骤如下：

1. 为网络中的每个节点分配唯一的标签，此标签表示节点所属的社区标识；
2. 随机排列网络中全部节点，从而生成节点更新序列表；
3. 按照生成的序列表依次更新节点标签。在节点标签过程中，选择邻居节点中出现次数最多的标签来更新当前节点的标签。若邻居节点中标签出现频率最高的标签不止一个时，则随机选择一个标签进行更新；
4. 重复步骤2和步骤3，直至网络中每个节点的标签不再发生变化或者达到最大的迭代次数；
5. 将具有相同标签的节点划分到同一个社区中，输出社区结构发现结果。

### 差分隐私社区发现算法

针对社区发现过程中存在的隐私保护问题，Nguyen等人[86]提出了一种基于Louvain社区发现算法的差分隐私保护算法。该算法首先给定组大小, LouvainDP通过随机通过节点集并将每个连续的个节点分组为一个超节点，从具有节点的超图开始。这种排列防止了中节点排序可能存在的偏差。超边的集合可以从中计算出来。范围大小为，非0超边的噪声数是。然后通过one-side过滤估计阈值和0通过的数量。对于每个非零超级边，如果噪声值不小于，则添加一个几何噪音并且将超级边添加到。对于的0超级边，从分布中得到一个积分权值，如果，则将权值为的加上。该算法的伪代码如XX所示。

Mülle等人[87]提出了一种基于社区发现的隐私保护方法，该方法保证了社区发现结果满足1-边差分隐私。PIG通过对依赖于差分隐私参数的图的邻接矩阵进行扰动来扰动输入图。PIG由两个步骤组成：输入图的扰动和应用于受扰动图的图聚类算法。PIG的思想是：如果相邻图的扰动版本相同，则使用图聚类方法计算相同的聚类结果。

该算法是边采样和边翻转相结合的方法。例如边随机化。它作用于输入图的邻接矩阵A。aij指的是第i行和第j列中A的条目，因此如果节点i和节点j之间存在一条边，则包含该信息。边的存在用值1表示，而缺失用0表示。由于输入图是无向的，邻接矩阵是对称的。由于self循环的定义在无向图的情况下没有意义，因此在受扰动的条目中，相应的条目是0，并且仍然是0。

扰动包括以下机制：对于邻接矩阵中的每个条目，首先确定是否应该执行保留或随机化。扰动方法是在网络图转换成的邻接矩阵上进行边的随机化，表示中的第行和第列的值，其中，表示边存在，则表示边缺失。对于邻接矩阵中的每个值，通过引入隐私参数来进一步确定是否应该执行保留或者随机化，并以的概率选择保留，同时以的概率选择随机化。因此，越大，则表明矩阵中的随机项就越多。在保留的情况下，未扰动图的邻接矩阵中的值保留在扰动网络图中。随机化的情况下，均以0.5的概率来使得扰动图中的值为0或者1。

Dong等人[88]基于层次随机图（HRG），借助社区发现将差分隐私应用于图的边到不同程度。基于分而治之的原理，算法为每个社区构建一个HRG，并将子级HRG合并为一个完整的HRG，消除处理大规模图形时算法的低效率问题。

在发布的社交网络中，用户经常关注某些特定领域。对于经常查询的部分数据，需要加强隐私保护，而对于用户不感兴趣的部分，则重点关注数据的可用性。

算法将社区检测与HRG模型相结合。作者认为，在发布的社交网络中应该更多地关注这些特定社区而不是其他社区。此外，根据最优HRG的属性，这些社区集中在由其他内部节点连接的一些子树中。不难发现同一社区的节点通常集中在同一个子树中。因此，这得到了MCMC采样的收敛保证和HRG的特性的支持。由于HRG的内部节点表示图中的边，所以社区子树内的节点存储较大的概率并且可以携带较大的噪声。相反，这些子树之间的节点存储较小的概率并且可以仅携带较小的噪声，使得HRG可以更准确地恢复原始图。

对于大型图，构建完整的HRG非常耗时，即使使用MCMC采样，也需要每次计算HRG的成本。算法利用分而治之的思想来解决这个问题。算法首先将原始社交网络图分成若干社区，针对每个社区构建HRG。然后扰乱每个社区的HRG并将它们汇总成一个完整的HRG。最后，将完整的HRG恢复为复杂的图形并发布数据。

## 7.6 动态社交网络的隐私保护（王寒）

社交网络通常情况下能够用一个简单的无向图结构来表示，其中表示网络图的个节点集合，表示个节点之间的边的集合。

在实际中，社交网络通常随着时间变化，新的个体会不断地加入到网络中，也可能从网络中消失，个体间的关系也会发生相应的变化，这种不断发生演化的社交网络称为动态社交网络。表示在时刻的动态社交网络，其中，表示时刻的节点集合，表示时刻的边的集合。

### 7.6.1 传统的动态社交网络保护方法

为了保护动态社交网络的发布隐私，Jyothi[89]等人提出了一种k隐私保护模型，其中表示隐私保护级别，表示以的时间间隔来监视攻击者所花费的时间。算法对网络数据的每次发布都进行匿名化处理，使得攻击者无法通过对多次搜集发布的信息进行分析来重新识别受害者。

该算法将动态问题作为边权匿名化问题。边的权值表示两个端点的共同朋友数。为了保留网络的特性，算法保留原始图的顶点集，并且只允许添加新的顶点。为了在动态友谊网络攻击的顺序发布中保留大部分特征，我们将网络表示为时间戳图，它可以保留社交网络的大部分功能。为了获得更好的性能，算法收集了每个以前发行版本的共同好友序列的边，并将其汇总到一个名为GS-Table的表中。在每次匿名化之前，我们增量地更新GS-Table，并根据GS-Table直接生成当前网络的适当版本。这样避免搜索所有可能的隐私违反通过发布和生成动态的基础上的以前的版本将是有效的。

算法包括两个部分，GS-Table的构造和更新以及匿名化方法。首先构造了GS-Table，然后根据互好友序列的前缀对边进行排序。在GS-Table中，根据他们的共同的朋友来进行边排序。因此，相邻的序列组将共享一个共有的朋友序列个数前缀，且边属于最接近一致的组。当图信息更新时，算法利用一些小的边集，有效地增量地维护GS-Table。在每次实例更新当前网络的GS-Table后，根据其在GS-Table中的共同好友数对边缘进行匿名化。在匿名化过程中，GS-Table可以快速找到具有相似数量的互好友序列的边。假设一条边必须匿名，首先检查是否让属于一个现有的k-匿名序列一致组。如果是，则将与最近的k-anonymous一致的GS-Table组合并。否则创建一个新的具有的k-匿名序列一致组和具有相同一致组的其他k-1边。

针对动态社交网络中的社交关系隐私，Liu等人[90]提出了一种LinMirage动态社交网络隐私保护算法。首先对社交图进行聚类，然后分别匿名化群内链接和群集间链接。以保留社交图的关键结构特性的方式混淆链接。该算法不仅可以适用于静态的社交网络，而且通过对时间实例持续聚合社交图，也使得该方法适用于社交图的演化动态，包括节点或者连接的添加和删除。

该算法大体上可以分为两个部分：动态聚类和选择性扰动。算法首先将和划分成子图，将每个图分为不同的社区。为了避免聚类过程中的随机性（保证一致性）并减少计算复杂度，将两个图形动态聚类在一起，而不是单独聚类。在动态聚类时，选择某个节点的m跳范围内的所有节点，将这些节点进行聚类。利用动态聚类方法对基于的进行聚类后，则基于和扰动的扰动。首先比较在和中检测到的社区，并将它们分类为已更改或未更改。未改变的分类不要求社区完全相同，但顶点/连接之间的重叠超过阈值。选择性扰动的主要思想是：保持不变社区中的连接的扰动过程与之前快照中的扰动相同。通过这种方式可以最大限度地保护这些不变连接的隐私。对于改变的社区，则采用独立扰动的方式扰乱它们的连接，与前一时间戳中的扰动无关。对于独立扰动，算法采用Mittal等人的静态扰动方法。静态扰动删除了原始图中的所有边，并且用从开始的跳随机游走中选择的假边替换每个边。更大的扰动参数对应于更好的隐私并导致更差的效用。最后，算法需要将每一个子图进行连接以发布完整的社交网络。在连接之前首先寻找需要连接的两个子图的边缘节点，然后通过一定的概率来随机连接这些边缘节点。

### 7.6.2 不确定图的隐私保护

最近的方法不是通过一般化或添加/删除边缘来匿名化社交图来满足给定的隐私参数，而是利用不确定图的语义来实现隐私保护。对于一个简单图，将表示为中所有对无序顶点的集合，即。不确定图是一对，其中是一个函数，该函数将已经存在的概率分配给所有可能的边。这些技术通过将确定性图形转换为不确定形式来对其进行匿名化。

最初的不确定图隐私保护方法由Boldi等人提出。该方法基于在社交图中注入不确定性并发布由此产生的不确定图。从概率的角度来看，添加不存在的边对应于将其概率从0更改为1，而删除现有边对应于更改它的概率从1到0。它们不是仅考虑二进制边缘概率，而是允许概率取范围[0,1]中的任何值。因此，每个边缘与不确定图中的特定概率相关联。然而，他们提出仅将不确定性注入到顶点对的小候选子集中并假设其他顶点对不存在，即。如果分布除以的至少个顶点的熵大于或者等于，即。

Nguyen等人提出了基于不确定邻接矩阵的广义混淆模型，其保持预期节点度等于原始图中的那些，以及匿名化方法的隐私和效用量化的通用框架。同时Nguyen等人还提出了基于最大方差的第二种方法，以实现隐私和数据效用之间更好的权衡。其还通过评估统一空间中典型方案的隐私和效用分数来描述图表匿名化的量化框架。

重要的是要强调必须定义（或重新定义）统计和度量以应用于此类图形，因为几乎所有这些都设计为使用二进制边缘图并且不能直接应用于不确定图。因此，Boldi等人提出了基于度的统计计算，例如边数，平均度，最大度和度方差，同时还提出通过对特定不确定图引起的可能edge-binary图空间中的一些图进行采样来计算基于最短路径距离和聚类系数的统计量。

### 7.6.3 动态社交网络差分隐私保护

针对动态社交网络的差分隐私保护技术，Yan等人[91]提出了基于分形的差分隐私聚类数据流直方图发布方法DP-FC，算法利用为具有多个属性的数据流计算的分形维数来聚类具有类似时间的数据流，之后通过添加拉普拉斯噪声发布聚类结果。DP-FC平台的结构由两个主要部分组成：数据分析模块和差分隐私模块。数据分析模块用于分析数据。数据分析模块的第一个组件是滑动窗口部分，它接收数据流并在滑动窗口中绑定它们的数据。令为一组数据流源。作者主要关注不断输入的数据流，并称这些记录为个体，用于缓冲分形维数计算数据的滑动窗口。该窗口被划分为计数周期，每个周期都有许多事件，这些事件也表示其移动步骤。以作为数据流的形式。一旦有足够的数据实现窗口（W）。现在可以处理这些数据了。然后，将之前打开的信息发送给负责执行分形维数分析的模块。分析模块应用SID-Meter模块对进行增量计算。分形维数分析部分中最重要的部分是基于分形的聚类算法，它可以对包含多个属性的数据进行迭代。属性计数在此期间完成。在DP-protection阶段，输入集群结果并输出可发布的数据。

由于将静态社交网络数据隐私发布算法直接应用于动态社交网络会导致图形结构的执行时间过长，信息丢失率大，因此，Liu等人[92]对基于MCL的差分隐私网络数据发布进行了改进，设计了一种满足ε-差异隐私的动态社交网络数据发布算法DDPA。DDPA算法是静态社交网络数据发布中隐私保护算法的改进。MDPA算法为整个网络图增加了噪声，但DDPA算法为改变的网络边缘权重增加了噪声。算法的基本思想是识别随着迭代次数增加而改变的边缘权重信息，并添加满足ε的隐私保护预算。因此，DDPA算法大大降低了算法的执行成本，降低了权重信息的丢失率。

Zhu等人[93]考虑到隐私的分布式在线优化问题，其中一组代理旨在最小化本地凸成本函数的总和，同时每个代理都希望单个代理的本地成本函数保持差分隐私。为了解决这个问题，作者提出了在时变有向网络上的差分隐私分布式随机次梯度在线优化算法。为了克服定向网络拓扑引起的不对称性而引入了平衡权重的定义。同时提出了基于权重平衡的差分私有分布式随机子梯度在线算法来解决隐私分布式在线优化问题。对于差分隐私分布式在线算法，将每个代理的最佳点的估计添加到每轮中的随机噪声向量，作者假设随机噪声向量是独立于拉普拉斯分布绘制的。然后，每个代理计算其邻近代理的噪声估计和该轮中其自己的估计的加权平均值。最后，每个代理沿着本地成本函数的次梯度的负方向更新其估计。

Song等人[94]通过了解敏感社交网络随时间变化的动态性，考虑在网络上持续发布统计数据的同时保护其参与者的隐私的问题。他们表明，如果整个图形序列中任何节点的最大度数存在公认的上界，则差分序列具有较低的灵敏度，差分序列即后续图形中计算的统计量的差值序列。作者指出整个差分序列的灵敏度仅取决于公认的上限，而不取决于序列的长度。这意味着可以以相对较高的准确度发布差分序列的隐私版本，这可以用于以高隐私准确性权衡来持续释放目标统计。同时，作者推导出差分序列对许多常见图形统计的敏感性，例如度分布，高度节点的数量，以及固定子图的计数等。

Wang等人[95]研究了隐私保护社交网络中实时时空数据发布的问题，设计了一个在无限流上的在线聚合监控框架并且具有w-event隐私保证的隐私保护实时时空数据发布算法RescueDP。RescueDP包括自适应采样，自适应预算分配，动态分组，扰动和过滤四个部分，同时将数据动态考虑在内，并应对数据稀疏性带来的挑战。

自适应采样机制是根据数据动态调整采样率，在所选采样时间标记处扰乱统计数据，并在最后采样时间标记处用扰动统计量近似非采样统计量。然后，作者提出了一种自适应预算分配机制，该机制在任何连续的w时间戳上的采样点处动态分配整个隐私预算e。抽样机制和自适应预算分配的主要思想是：可以在没有任何预算分配的情况下近似非抽样统计数据。因此，给定一个固定的ε，可以将更多预算分配给任何连续的采样点w个时间戳，可以减少拉普拉斯噪声引入的扰动误差，并增加释放统计数据的效用。作者还提出了一种动态分组机制，该机制可以减少引入具有较小统计值的区域的扰动误差。动态分组机制不是根据空间相关性对区域进行分组，而是根据其变化趋势的相似性将具有较小统计值的区域组合在一起。该机制的优点在于：随着统计变化，组随时间动态变化，并且远离空间域的区域也可以组合在一起以抵抗噪声。利用分组策略和自适应预算分配，我们应用拉普拉斯机制来扰动统计的真实值，最后采用过滤机制来提高发布数据的准确性。

文献最后进一步提出了一种增强的RescueDP方案，利用神经网络准确预测每个区域的统计数据，并提出一种新的采样和动态编程的动态分组机制，在提高发布数据的准确性的同时保持 w-event隐私。

Wang等人针对动态社交网络的隐私保护问题，提出了一种基于差分隐私的动态社交网络发布算法。该算法结合动态社区发现、差分隐私和社区连接有效解决了动态社交网络中的隐私保护问题。算法首先对社交网络进行动态社区发现，然后，通过Jaccard系数相似度判断该时刻的社区相对于上一时刻是否发生变化，对于演化的社区将其转换成邻接矩阵的形式并进行选择性差分隐私处理，对于未发生演化的社区则保留上一时刻的已存在的扰动方式。算法最后采用动态社交网络连接算法将社区进行连接生成完整的社交网络。

动态社区发现算法是根据Louvain社区发现算法的一种改进，算法的基本思想是使用最后一次划分来初始化算法，并通过已经发生改变的节点和边的信息来使得迭代的节点数量最小化。首先，采用与传统Louvain算法相同的初始化方式；然后，对于接下来的每一步都使用最后一个划分的修改版本，即删除网络图中接近变化的节点或者将节点移动到各自的社区，其中，“接近”是指与边直接相邻或具有到相邻节点的路径，该路径小于或等于给定阈值。最后，将修改后的划分和所有受影响的节点列表传递给算法，在算法的每一次迭代中，仅仅处理上述提及的受影响的节点。

通过Jaccard系数相似度的值来比较和中检测到的社区，并将它们分成两类，即发生演化的社区和未发生演化的社区。针对未发生演化的一类社区，算法的主要思想是保持未发生演化的社区中连接的扰动过程和前一时刻的社区的扰动过程相同。针对发生演化的这类社区，算法主要通过将社区转换成邻接矩阵并进行扰动，该扰动与上一时刻网络图的扰动无关。选择性扰动算法的伪代码如算法XXX所示。

|  |
| --- |
| 算法：邻接矩阵扰动算法 |
| 输入：社区，隐私参数 |
| 输出：扰动后的社区 |
| 1 将社区转换成邻接矩阵；  2 对于所有的且，循环执行：  3 如果选择的概率是，则：  4 令中的为；//保留  5 否则：  6 如果选择0的概率是，则：  7 令中的为；//随机化  8 否则：  9 令中的为；  10 判断结束；  11 判断结束；  12 返回；  13循环结束；  14返回扰动后的社区； |

最后，算法需要将划分的社区进行社区连接，以发布完整的社交网络图。该阶段的基本思路是选择所有社区中的其中两个社区进行递归连接，直到得到一个完整的单个网络图。本章借鉴了最小合并算法，该算法通过优先连接机制将两个网络图连接起来，很好的保留了参与合并的两个网络图的原始结构。优先连接机制最早由Barabási等人提出，该机制表明在一个网络中，如果某个节点的连接越多，那么这个节点接收新连接的可能性就越大。作者证明了其算法满足差分隐私保护，并且分析了算法对网络结构的影响和算法的复杂度。用实验表明算法具有一定的效用性。

## 7.7 本地化差分隐私在社交网络中的应用（王寒）

传统的集中式差分隐私需要将数据上传到第三方数据收集者，然后进行差分隐私处理。这是基于一个假设，即存在一个可信的第三方数据收集者，它可以确保用户的敏感信息不被泄漏。但是在实际中，第三方数据采集者也有可能泄露用户的敏感信息，使用户的隐私仍然得不到保障。为了解决这个问题，Kasiviswanathan S P等人提出了本地化差分隐私技术。本地化差分隐私首先在用户端对数据进行隐私扰动处理，然后将处理后的数据发送给数据收集者进行统计分析。这样使得个人的敏感信息对数据收集者也是保密的，因此也不再需要可信第三方数据收集者的假设，使得敏感信息更加安全。本节将阐述本地化差分隐私在社交网络中的应用。

Qin Z等人[67]针对现有的合成图生成技术噪声注入过多且无法保留重要的图结构和图属性问题，提出了一种基于本地化差分隐私的网络图生成模型LDPGen。该模型迭代地将节点划分为组，并向组中的度向量添加噪声来满足LDP，最后，应用图生成模型生成一个具有代表性的综合网络图。该模型能够更加全面地评估生成的图的属性，满足复杂的数据挖掘任务。Zhang Y等人[96]针对RR算法应用于合成图中出现的图的密度过大问题，对RR算法进行了优化，提出了LDPGM-ORR算法。该算法通过利用图的噪声度进行聚类分组，并结合安全多方计算将扰动分组发送到第三方，第三方针对每组的邻居列表进行随机响应来满足LDP，最终通过ORR算法来获得扰动网络图。Ren X等人[97]针对高维众包数据在实现LDP时存在的效率低下的问题，提出了一种本地化差分隐私高维数据发布算法LoPub。该算法首先结合RAPPOR算法[98]对元组进行扰动并将其发送给数据收集者，然后基于EM和Lasso回归从分布式用户收集和构建高维数据，并通过识别相关属性将属性分割成几个紧凑的低维属性簇来实现降维。最后，通过联合概率分布对每个低维数据集进行采样，从而合成一个近似数据集进行发布。Wang Z等人[99]研究了在第三方服务器不可信的情况下，具有强隐私保护的实时众包数据发布问题，提出了一种基于分布式代理的隐私保护框架DADP。DADP在用户端和第三方服务器端引入了多个代理，用户端将采集到的数据通过匿名技术随机上传到某个代理，代理对收到的数据进行聚合并对统计数据进行LDP扰动，最终通过第三方服务器完成所有代理的扰动信息聚合以供发布。Qin Z等人[100]针对Heavy hitter查询，提出了基于LDP的频数发布方法LDPMiner。该方法首先将随机采样的数据项通过H-Hist方法进行扰动后发送给数据收集者，收集者将频数最高的个项的集合返回给用户端，用户端对返回的数据利用RAPPOR算法处理后再次发送给数据收集者，最终得到个数据项的频数。

接下来将对LoPub算法进行详细阐述。LoPub算法主要由四种机制组成：本地化数据保护，多维分布估计，降维和数据合成。本地化数据保护也称为本地随机化器，它在编码数据记录上采用RR，确保每个用户输出满足LDP。特别地，我们将每个属性值本地化变换为随机位串。然后，将清理后的数据发送到中央服务器并在中央服务器上进行聚合。在多维分布估计中，作者提出了三种多维联合分布估计方案以推导出联合和边际概率分布。其首先扩展了基于EM的分布估计方法。然而，这种直接的扩展的方式不考虑高维数据中的稀疏性，这将导致分布估计的高计算复杂度。为了加速估算，提出了一种基于Lasso的方法，该方法成本略有降低。最后，提出了一种在准确性和效率之间取得平衡的混合方法。维度降低是通过识别相关属性并将属性分割成若干紧凑的低维属性集群来开发用于降维的技术。考虑到异构属性，采用互信息来测量相关性，形成无向依赖图。然后根据从依赖图构建的联结树来分割属性。同时作者还提出了一种启发式修剪方案，以加快相关识别过程。最后，根据属性聚类的连通性和每个属性聚类上的估计联合（或条件）分布对每个低维数据集进行采样，从而合成近似数据集。

本地随机化器设计包括三个步骤：（1）假设第个用户有一条记录。第个用户通过散列函数的集合将每个编码为Bloom过滤器，特别是对于；第个用户使用的散列函数将映射到长度的位串；表示位串的第位。（2）每一位根据RR规则翻转为0或者1：



其中，是量化LDP的随机化水平的一个参数，用于控制隐私水平。（3）第个用户连接来获得一个位向量，并且将其发送给服务器。

EM-Based分布估计主要包括以下五个步骤：（1）设置均匀分布作为初始先验概率。（2）在本地化转换中，每一位以的概率翻转。因此，通过将与随机位比较来计算传统概率。（3）联合条件概率可以通过结合单个属性计算出来。（4）给定一个特定位串组合的所有条件分布，其相应的后验概率可由贝叶斯定理计算：



其中，是维联合概率的第次迭代。（5）在识别每个用户的后验概率之后，计算来自大量用户的后验概率的均值来更新先验概率。在下一次迭代中使用先验概率来更新后验概率。迭代地执行上述EM类过程直到收敛，即两个估计之间的最大差异小于指定阈值。

Lasso-based分布估计主要包括以下五个主要步骤：（1）在收到了所有噪声Bloom过滤后，对于每一个属性中的每一位，服务器计算1的数量。（2）根据应用RR到真实计数，每一位的真实计数总和可以被估计。这些所有位的计数总和形成长度为的向量。（3）每个维度上的Bloom过滤器由服务器使用的相同散列函数构造。假设上所有不同的Bloom过滤器是，候选的Bloom过滤器集合为。（4）将Lasso回归模型拟合到计数器向量和候选矩阵，然后选择非零系数作为每个候选字符串的相应频率。通过将系数向量整形为维矩阵，我们可以推导出维联合分布估计 。

如果有足够的样本，EM\_JD可以表现出良好的收敛性，但也会产生很高的复杂性。另一方面，与基于EM的算法相比，Lasso\_JD可以非常有效，具有轻微的精度降低。为了在准确性和效率之间取得平衡，作者提出了一种混合算法Lasso + EM\_JD，它首先消除冗余候选并用Lasso\_JD估计初始值，然后使用EM\_JD细化收敛。因此，Lasso\_JD将非常有效地选择稀疏的候选人，EM\_JD可以只计算那些稀疏候选者而不是所有候选者的条件概率，从而显着降低时间和空间复杂度。同时，Lasso\_JD可以生成联合分布的良好初始估计。与使用具有均匀随机分配的初始值相比，使用由Lasso\_JD生成的初始值可以进一步加速EM\_JD的收敛，EM\_JD对初始值敏感，尤其是当候选空间稀疏时。

## 7.8 小结

本章首先描述几种复杂类型数据，包括图数据、连续数据、高维数据、地理位置数据、社交网络数据等，其次针对不同类型数据分析过程存在的隐私泄露问题，分析其差分隐私结合方法，并着重针对社交网络数据进行拓展，讲述动态网络、本地化差分隐私的应用等。

## 7.9 参考文献

1. Backstrom L, Dwork C, Kleinberg J. Wherefore art thou r3579x?: anonymized social networks, hidden patterns, and structural steganography[C]//Proceedings of the 16th international conference on World Wide Web. ACM, 2007: 181-190.
2. Liu P, Li X. An improved privacy preserving algorithm for publishing social network data[C]//2013 IEEE 10th International Conference on High Performance Computing and Communications & 2013 IEEE International Conference on Embedded and Ubiquitous Computing. IEEE, 2013: 888-895.
3. Liu K, Terzi E. Towards identity anonymization on graphs[C]//Proceedings of the 2008 ACM SIGMOD international conference on Management of data. ACM, 2008: 93-106.
4. Narayanan A, Shmatikov V. De-anonymizing social networks[J]. arXiv preprint arXiv:0903.3276, 2009.
5. Zhou B, Pei J. Preserving Privacy in Social Networks Against Neighborhood Attacks[C]//ICDE. 2008, 8: 506-515.
6. Bhagat S, Cormode G, Krishnamurthy B, et al. Class-based graph anonymization for social network data[J]. Proceedings of the VLDB Endowment, 2009, 2(1): 766-777.
7. Ying X, Wu X. On link privacy in randomizing social networks[J]. Knowledge and information systems, 2011, 28(3): 645-663.
8. Korolova A, Motwani R, Nabar S U, et al. Link privacy in social networks[C]//Proceedings of the 17th ACM conference on Information and knowledge management. ACM, 2008: 289-298.
9. Zafarani R, Liu H. Social computing data repository at ASU[J]. 2009.
10. Zhang L, Zhang W. Edge anonymity in social network graphs[C]//2009 International Conference on Computational Science and Engineering. IEEE, 2009, 4: 1-8.
11. Masoumzadeh A, Joshi J. Preserving structural properties in edge-perturbing anonymization techniques for social networks[J]. IEEE Transactions on Dependable and Secure Computing, 2012, 9(6): 877-889.
12. Zheleva E, Getoor L. Privacy in social networks: A survey[M]//Social network data analytics. Springer, Boston, MA, 2011: 277-306.
13. Das S, Egecioglu O, El Abbadi A. Anónimos: An lp-based approach for anonymizing weighted social network graphs[J]. IEEE Transactions on Knowledge and Data Engineering, 2010, 24(4): 590-604.
14. Liu L, Wang J, Liu J, et al. Privacy preservation in social networks with sensitive edge weights[C]//proceedings of the 2009 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics, 2009: 954-965.
15. Wang B, Yang J. Personalized (α, k)-anonymity algorithm based on entropy classification[J]. Journal of Computational Information Systems, 2012, 8(1): 259-266.
16. Yuan M, Chen L, Yu P S. Personalized privacy protection in social networks[J]. Proceedings of the VLDB Endowment, 2010, 4(2): 141-150.
17. Liu L, Wang J, Liu J, et al. Privacy preserving in social networks against sensitive edge disclosure[R]. Technical Report Technical Report CMIDA-HiPSCCS 006-08, Department of Computer Science, University of Kentucky, KY, 2008.
18. Ying X, Wu X. Randomizing social networks: a spectrum preserving approach[C]//proceedings of the 2008 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics, 2008: 739-750.
19. Das S, Eğecioğlu Ö, El Abbadi A. Anonymizing weighted social network graphs[C]//2010 IEEE 26th International Conference on Data Engineering (ICDE 2010). IEEE, 2010: 904-907.
20. Liu X, Yang X. Protecting sensitive relationships against inference attacks in social networks[C]//International Conference on Database Systems for Advanced Applications. Springer, Berlin, Heidelberg, 2012: 335-350.
21. Hay M, Miklau G, Jensen D, et al. Resisting structural re-identification in anonymized social networks[J]. Proceedings of the VLDB Endowment, 2008, 1(1): 102-114.
22. Ying X, Wu X. Randomizing social networks: a spectrum preserving approach[C]//proceedings of the 2008 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics, 2008: 739-750.
23. Hay M, Miklau G, Jensen D, et al. Anonymizing social networks[J]. Computer science department faculty publication series, 2007: 180.
24. Li Y, Li Y, Yan Q, et al. Privacy leakage analysis in online social networks[J]. Computers & Security, 2015, 49: 239-254.
25. Campan A, Truta T M. Data and structural k-anonymity in social networks[C]//International Workshop on Privacy, Security, and Trust in KDD. Springer, Berlin, Heidelberg, 2008: 33-54.
26. Nguyen H H, Imine A, Rusinowitch M. Anonymizing social graphs via uncertainty semantics[C]//Proceedings of the 10th ACM Symposium on Information, Computer and Communications Security. ACM, 2015: 495-506.
27. Boldi P, Bonchi F, Gionis A, et al. Injecting uncertainty in graphs for identity obfuscation[J]. Proceedings of the VLDB Endowment, 2012, 5(11): 1376-1387.
28. Watts D J, Strogatz S H. Collective dynamics of ‘small-world’networks[J]. nature, 1998, 393(6684): 440.
29. Li X, Yang J, Sun Z, et al. Differential privacy for edge weights in social networks[J]. Security and Communication Networks, 2017, 2017.
30. Hay M, Li C, Miklau G, et al. Accurate estimation of the degree distribution of private networks[C]//2009 Ninth IEEE International Conference on Data Mining. IEEE, 2009: 169-178.
31. Raskhodnikova S, Smith A. Efficient lipschitz extensions for high-dimensional graph statistics and node private degree distributions[J]. arXiv preprint arXiv:1504.07912, 2015.
32. Borgs C, Chayes J, Smith A. Private graphon estimation for sparse graphs[C]//Advances in Neural Information Processing Systems. 2015: 1369-1377.
33. Dwork C, McSherry F, Nissim K, et al. Calibrating noise to sensitivity in private data analysis[C]//Theory of cryptography conference. Springer, Berlin, Heidelberg, 2006: 265-284.
34. Nissim K, Raskhodnikova S, Smith A. Smooth sensitivity and sampling in private data analysis[C]//Proceedings of the thirty-ninth annual ACM symposium on Theory of computing. ACM, 2007: 75-84.
35. Blocki J, Blum A, Datta A, et al. Differentially private data analysis of social networks via restricted sensitivity[C]//Proceedings of the 4th conference on Innovations in Theoretical Computer Science. ACM, 2013: 87-96.
36. Blocki J, Blum A, Datta A, et al. The johnson-lindenstrauss transform itself preserves differential privacy[C]//2012 IEEE 53rd Annual Symposium on Foundations of Computer Science. IEEE, 2012: 410-419.
37. Rastogi V, Hay M, Miklau G, et al. Relationship privacy: output perturbation for queries with joins[C]//Proceedings of the twenty-eighth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems. ACM, 2009: 107-116.
38. 白川平, 杨志翀. 基于加权滑动窗口的数据流频繁项集挖掘算法[J]. 宁夏师范学院学报, 2017(6):49-55.
39. C. Dwork, M. Naor, T. Pitassi, and G. N. Rothblum. Differential privacy under continual observation. In STOC, 2010.
40. C. Dwork. Differential privacy in new settings. In SODA, 2010.
41. D. Mir, S. Muthukrishnan, A. Nikolov, and R. N. Wright. Pan-private algorithms via statistics on sketches. In PODS, 2011.
42. T. Chan, E. Shi, and D. Song. Private and continual release ofstatistics. TISSEC, 14(3):26, 2011.
43. T.-H. H. Chan, M. Li, E. Shi, and W. Xu. Differentially private continual monitoring of heavy hitters from distributed streams. In PETS, 2012.
44. L. Fan and L. Xiong. Real-time aggregate monitoring with differential privacy. In CIKM, 2012.2008.
45. Cheng M, Sun Y, Zhao B, et al. An Event Grouping Approach for Infinite Stream with Differential Privacy[M]// Advances in Services Computing. Springer International Publishing, 2016:106-116.
46. 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.
47. Cao L, Ou Y, Philip S Y. Coupled behavior analysis with applications[J]. IEEE Transactions on Knowledge and Data Engineering, 2012, 24(8): 1378-1392.
48. Yang B, Sato I, Nakagawa H. Bayesian differential privacy on correlated data[C]//Proceedings of the 2015 ACM SIGMOD international conference on Management of Data. ACM, 2015: 747-762.
49. Zhu T, Xiong P, Li G, et al. Correlated differential privacy: hiding information in non-IID data set[J]. IEEE Transactions on Information Forensics and Security, 2015, 10(2): 229-242.
50. Rastogi V, Nath S. Differentially private aggregation of distributed time-series with transformation and encryption[C]//Proceedings of the 2010 ACM SIGMOD International Conference on Management of data. ACM, 2010: 735-746.
51. Xiao X, Wang G, Gehrke J. Differential privacy via wavelet transforms[J]. IEEE Transactions on Knowledge and Data Engineering, 2011, 23(8): 1200-1214.
52. Xiao X. Differentially private data release: Improving utility with wavelets and bayesian networks[C]//Asia-Pacific Web Conference. Springer, Cham, 2014: 25-35.
53. Jiang W, Xie C, Zhang Z. Wishart Mechanism for Differentially Private Principal Components Analysis[C]//AAAI. 2016: 1730-1736.
54. 熊文君,徐正全,王豪.基于滤波原理的时间序列差分隐私保护强度评估[J].通信学报,2017,38(05):172-181.
55. 张啸剑, 孟小峰. 基于差分隐私的流式直方图发布方法[J]. 软件学报, 2016, 27(2):381-393.
56. 张啸剑,邵超,孟小峰.差分隐私下一种精确直方图发布方法[J].计算机研究与发展,2016,53(05):1106-1117.
57. Li Y, Li S. Research on Differential Private Streaming Histogram Publication Algorithm[C]//2018 5th IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS). IEEE, 2019: 598-603.
58. Hardt M, Roth A. Beyond worst-case analysis in private singular vector computation[C]//Proceedings of the forty-fifth annual ACM symposium on Theory of computing. ACM, 2013: 331-340.
59. Kapralov M, Talwar K. On differentially private low rank approximation[C]//Proceedings of the twenty-fourth annual ACM-SIAM symposium on Discrete algorithms. Society for Industrial and Applied Mathematics, 2013: 1395-1414.
60. Chaudhuri K, Sarwate A, Sinha K. Near-optimal differentially private principal components[C]//Advances in Neural Information Processing Systems. 2012: 989-997.
61. Matoušek J. On variants of the Johnson–Lindenstrauss lemma[J]. Random Structures & Algorithms, 2008, 33(2): 142-156.
62. Frankl P, Maehara H. The Johnson-Lindenstrauss lemma and the sphericity of some graphs[J]. Journal of Combinatorial Theory, Series B, 1988, 44(3): 355-362.
63. Ahmed F, Jin R, Liu A X. A random matrix approach to differential privacy and structure preserved social network graph publishing[J]. arXiv preprint arXiv:1307.0475, 2013.
64. 戚名钰, 黄刘生, 陆潇榕,等. 采用成分分析的差分隐私数据发布算法[J]. 小型微型计算机系统, 2017, 38(3):437-443.
65. Jiang X, Ji Z, Wang S, et al. Differential-private data publishing through component analysis[J]. Transactions on data privacy, 2013, 6(1): 19.
66. Kasiviswanathan S P, Nissim K, Raskhodnikova S, et al. Analyzing graphs with node differential privacy[C]//Theory of Cryptography Conference. Springer, Berlin, Heidelberg, 2013: 457-476.
67. Qin Z, Yu T, Yang Y, et al. Generating synthetic decentralized social graphs with local differential privacy[C]//Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. ACM, 2017: 425-438.
68. Li C, Palanisamy B, Joshi J. Differentially Private Trajectory Analysis for Points-of-Interest Recommendation[C]// IEEE International Congress on Big Data. IEEE, 2017:49-56.
69. Wang S, Sinnott R, Nepal S. Protecting the location privacy of mobile social media users[C]// IEEE International Conference on Big Data. IEEE, 2017:1143-1150.
70. Samarati P, Sweeney L. Generalizing data to provide anonymity when disclosing information[C]//PODS. 1998, 98: 188.
71. Sweeney L. k-anonymity: A model for protecting privacy[J]. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002, 10(05): 557-570.
72. Backstrom L, Dwork C, Kleinberg J. Wherefore art thou r3579x?: anonymized social networks, hidden patterns, and structural steganography[C]//Proceedings of the 16th international conference on World Wide Web. ACM, 2007: 181-190.
73. Zheleva E, Getoor L. Preserving the privacy of sensitive relationships in graph data[M]//Privacy, security, and trust in KDD. Springer, Berlin, Heidelberg, 2008: 153-171.
74. Korolova A, Motwani R, Nabar S U, et al. Link privacy in social networks[C]//Proceedings of the 17th ACM conference on Information and knowledge management. ACM, 2008: 289-298.
75. De Cristofaro E, Soriente C, Tsudik G, et al. Hummingbird: Privacy at the time of twitter[C]//Security and Privacy (SP), 2012 IEEE Symposium on. IEEE, 2012: 285-299.
76. Beresford A R, Rice A, Skehin N, et al. Mockdroid: trading privacy for application functionality on smartphones[C]//Proceedings of the 12th workshop on mobile computing systems and applications. ACM, 2011: 49-54
77. Wang L, MENG X F. Location privacy preservation in big data era: a survey[J]. Journal of Software, 2014, 25(4): 693-712.
78. Xiao-Feng H U O Z M. A Survey of Trajectory Privacy-Preserving Techniques [J][J]. Chinese Journal of Computers, 2011, 10: 008.
79. Bamba B, Liu L, Pesti P, et al. Supporting anonymous location queries in mobile environments with privacygrid[C]//Proceedings of the 17th international conference on World Wide Web. ACM, 2008: 237-246.
80. Liu L. From data privacy to location privacy: models and algorithms[C]//Proceedings of the 33rd international conference on Very large data bases. VLDB Endowment, 2007: 1429-1430.
81. Liu F, Hua K A, Cai Y. Query l-diversity in location-based services[C]//Mobile Data Management: Systems, Services and Middleware, 2009. MDM'09. Tenth International Conference on. IEEE, 2009: 436-442.
82. Li Y D, Zhang Z, Winslett M, et al. Compressive mechanism: Utilizing sparse representation in differential privacy[C]//Proceedings of the 10th annual ACM workshop on Privacy in the electronic society. ACM, 2011: 177-182.
83. Jia O, Jian Y, Shaopeng L, et al. An Effective Differential Privacy Transaction Data Publication Strategy[J]. Journal of Computer Research and Development, 2014, 10: 007.
84. Zhang X, Wang M, Meng X. An accurate method for mining top-k frequent pattern under differential privacy[J]. Journal of Computer Research and Development, 2014, 51(1): 104-114.
85. 杨理皓, 谷科, 李威. 基于差分隐私机制的位置数据隐私保护策略[J]. 计算机应用研究, 2018(3).
86. Nguyen, Hiep H., Abdessamad Imine, and Michaël Rusinowitch. "Detecting communities under differential privacy." Proceedings of the 2016 ACM on Workshop on Privacy in the Electronic Society. ACM, 2016.
87. Mülle, Yvonne, Chris Clifton, and Klemens Böhm. "Privacy-Integrated Graph Clustering Through Differential Privacy." EDBT/ICDT Workshops. 2015.
88. Dong, Kang, et al. "Differentially private big data publication via structural inference and community detection." 2017 14th International Symposium on Pervasive Systems, Algorithms and Networks & 2017 11th International Conference on Frontier of Computer Science and Technology & 2017 Third International Symposium of Creative Computing (ISPAN-FCST-ISCC). IEEE, 2017.
89. Jyothi, Vadisala, and V. Valli Kumari. "Privacy preserving in dynamic social networks." Proceedings of the International Conference on Informatics and Analytics. ACM, 2016.
90. Liu, Changchang, and Prateek Mittal. "LinkMirage: Enabling Privacy-preserving Analytics on Social Relationships." NDSS. 2016.
91. Yan F, Zhang X, Li C, et al. Differentially private histogram publishing through Fractal dimension for dynamic datasets[C]//2018 13th IEEE Conference on Industrial Electronics and Applications (ICIEA). IEEE, 2018: 1542-1546.
92. Liu Z, Dong Y, Zhao X, et al. A Dynamic Social Network Data Publishing Algorithm Based on Differential Privacy[J]. Journal of Information Security, 2017, 8(04): 328.
93. Zhu J, Xu C, Guan J, et al. Differentially private distributed online algorithms over time-varying directed networks[J]. IEEE Transactions on Signal and Information Processing over Networks, 2018, 4(1): 4-17.
94. Song S, Little S, Mehta S, et al. Differentially Private Continual Release of Graph Statistics[J]. arXiv preprint arXiv:1809.02575, 2018.
95. Wang Q, Zhang Y, Lu X, et al. Real-time and spatio-temporal crowd-sourced social network data publishing with differential privacy[J]. IEEE Transactions on Dependable and Secure Computing, 2018, 15(4): 591-606.
96. Zhang Y, Wei J, Zhang X, et al. A Two-Phase Algorithm for Generating Synthetic Graph Under Local Differential Privacy[C]//Proceedings of the 8th International Conference on Communication and Network Security. ACM, 2018: 84-89.
97. Ren X , Yu C M , Yu W , et al. LoPub: High-Dimensional Crowdsourced Data Publication with Local Differential Privacy[J]. IEEE Transactions on Information Forensics and Security, 2018:1-1.
98. Erlingsson Ú, Pihur V, Korolova A. Rappor: Randomized aggregatable privacy-preserving ordinal response[C]//Proceedings of the 2014 ACM SIGSAC conference on computer and communications security. ACM, 2014: 1054-1067.
99. Wang Z , Pang X , Chen Y , et al. Privacy-preserving Crowd-sourced Statistical Data Publishing with An Untrusted Server[J]. IEEE Transactions on Mobile Computing, 2018:1-1.
100. Qin Z, Yang Y, Yu T, et al. Heavy hitter estimation over set-valued data with local differential privacy[C]//Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. ACM, 2016: 192-203.