说明：笔记中的“当前”是以论文发表的时间为基准，例如5中的当前指的是2014年。

笔记中的“作者”指的是论文原作者，“笔者”指的是我自己

# 5．位置大数据隐私保护研究综述

## 背景

出版时间：2014年

作者：王璐，孟小峰

## 问题

1. 经典位置隐私保护方法不能解决的主要问题是：攻击者可以从多种途径获得各个角度关于用户的位置数据或非位置数据，这些数据可以直接或者间接地重构出用户希望保护的位置隐私．
2. 如何防止攻击者利用收集到的各方面数据推测用户的隐私信息

## 面临的挑战

(1) 如何度量用户的敏感信息的泄露程度；

(2) 如何实现对位置大数据隐私全面的保护；

(3) 如何兼顾隐私保护的程度和基于位置服务的可用性．

## 现状

1. 针对不同程度的隐私需求，权衡隐私保护效果和服务可用性
2. 将位置大数据隐私保护技术分为 3 类：
3. 基于启发式隐私度量的位置大数据隐私保护技术．
4. 缺点：只考虑当前时刻的位置信息是否会暴露用户的敏感位置信息
5. 手段：通过发布不精确的位置数据和抑制发布位置数据达到对位置大数据隐私保护的目的
6. 相应的方法：经典的基于单点或轨迹的位置隐私保护技术，如针对位置或轨迹的k 匿名、，多样性、t紧密性、P 敏感性、m 不变性或空问匿名框等方法
7. 原理：让用户提交不真实的位置数据来避免攻击者获得用户的真实位置数据，
8. 主要技术：

随机化（加入噪声）、

空间模糊化（降低发布位置数据的精度）、

时间模糊化技术（增加位置数据的时间域的不确定性，以减少位置数据的精度）

1. 基于概率推测的位置大数据隐私保护技术．
2. 原理：在攻击者具有完全的背景知识的情况下，在统一的位置大数据攻击模型下，定量地保护用户的位置隐私
3. 手段：通过发布不精确的位置数据和抑制发布位置数据达到对位置大数据隐私保护的目的（同上述基于启发式隐私度量）
4. 满足隐私定义 1 的一种平凡的方法是抑制所有位置数据的发布，但是：(1) 这样做无法获得基于位置的服务；(2) 当用户放松隐私保护需求时，这样做在服务的性能上没有收益
5. 方法：

基于隐马尔可夫模型进行概率计算（受到重视，当前时刻所处的位置和当前时刻前发布的历史数据对是否发布位置数据直接影响了基于位置服务的可用性）、

基于图模型进行概率计算（图模型是对隐马尔可夫模型的一般化，它允许用户在某时刻处于某位置与历史位置数据有关，可阅读文献[75,76]的高效算法）

1. 基于隐私信息检索（PIR）的位置大数据隐私保护技术．
2. 说明：可以在任何情况下完全地保护移动用户的隐私，与加密方法类似
3. 分类：

基于信息论的PIR 方法（保证攻击者无论拥有怎样的计算能力，都不能区分用户对不同数据项的访问）、

基于计算能力的 PIR 方法（假设攻击者不具有计算求解某个难题的能力而保证攻击者不能区分用户对不同数据项的访问，当前通常采用这类方法）

1. 常用方法：

基于一个难解的二次剩余假设问题、

基于伪随机函数的可实现性

1. 关键：查询计划如何使不同位置的查询不可区分

## 研究热点

1. 位置大数据与非位置大数据结合产生的隐私问题．其中，移动社交网络由于天然地将位置数据与非位置数据结合起来，将成为未来的研究热点之一
2. 差分隐私应用到位置大数据隐私保护中
3. 背景知识攻击
4. 隐私保护与可用性的权衡：如何避免向攻击者发布移动对象某一时刻的精确位置，同时获得基于位置大数据的服务，

## 未来展望

1. 位置大数据与非位置大数据结合：

应研究用户位置与行为模式之间的映射关系，以设法降低攻击者根据从匿名用户位置推测 出的个性向量与攻击者收集到的不同用户之间的个性向量的匹配程度：

1. 匿名用户的位置数据增多后，其映射成的偏好程度向量会变得准确；反之，偏好程度向量会变得模糊；
2. 匿名用户偏好程度的向量变得准确，有利于与攻击者收集到的个性向量进行匹配，且映射后不可区分的个性向量会变少：反之，不可区分的匿名数据数量会增多．
3. 移动社交网络：
   1. 背景：用户的位置信息与用户的身份信息显式地结合
   2. 方法：是通过对数据进行隐私等级的分类，然后使用访问控制策略防止攻击者访问敏感位置数据
   3. 最新研究：试图将单一位置大数据隐私保护方法应用到移动社交网络中，提出了 k 匿名的轨迹隐私保护方法[92]以及概率推测的方法[93]
   4. 作者感悟：考虑移动社交网络中用户相关的位置数据与非位置数据之间的关系，并防止攻击者利用该关系推测用户的敏感信息，是未来位置大数据的隐私保护技术的研究方向．
   5. 相关思考：

关于该类的最新研究（相对于2014年）有必要追踪，将隐私等级分类，和上述的基于概率推测中的p值的收敛是不是有联系呢？

1. 背景知识攻击（使用差分隐私技术）
2. 技术：差分隐私（当前最有效的技术，DP主要针对离线数据）
3. 说明：差分隐私的研究主要集中在如何添加假数据干扰分析结果，主要的技术包括针对数据类型是整数的Laplace机制、针对数据类型是浮点数的exponential机制[96]以及它们在性能上的改进geometric机制

## 阅读思考

1. 在隐马尔可夫模型中，信息量阈值怎么设置？计算量很大的问题怎么处理？其中，文献[71-73]关于p的收敛值得学习（攻击者可以利用贝叶斯公式计算用户的敏感位置，反过来保护机制也可以通过计算攻击者可能获得的位置数据，计算用户属于敏感位置的概率（后验概率））
2. 笔者阅读的Paper27就有涉及到马尔可夫模型
3. 文献[74]将“考虑未来所有可能的位置发布”这样一项时间复杂度极高的任务，简化成多项式时间内可完成的任务，值得关注

# 26．面对数据发布和数据分析中的差分隐私保护

## 背景

出版时间：2014年4月

作者：张啸剑、孟小峰

刊物：计算机学报

## 问题

1. 交互式框架只允许数据分析者通过查询接口提交查询，查询数目决定着该框架的误差和性能，若提交查询的数目超过某个上界，隐私预算 ε 会被耗尽，该框架则不能满足差分隐私。
2. 在直方图发布中的采用聚类划分桶和采用傅里叶变化两个方法都面临如何选择均衡参数的问题。（聚类重复划分方法能够找出比较合适的均衡参数，但是原始直方图的桶个数未得到压缩 ；通过傅里叶变换的采样技术虽然能够压缩原始直方图，但是找出的均衡参数未必最佳）
3. 在直方图发布的基于划分的发布方法中，在自顶向下分割数据空间时，如何确定分割停止条件至关重要。分割过细会引入过多的噪音数据；反之范围计数查询的响应精度会非常低

## 面临的挑战

1. 动态环境下的数据发布：
2. 基于数据流的直方图发布（挑战：（1）在滑动窗向前滑动过程中，如何防止直方图自身结构披露隐私；（2）如何利用动态规划技术设计高效的流式直方图发布方法）；
3. 增量更新环境下的集值型数据发布（在设计算法时，记得给出一个合理的更新边界；挑战：如何合理地分配隐私预算与有效地控制累积误差）
4. 差分隐私下图数据的发布：利用双聚类技术挖掘邻接矩阵中的所有满足阈值条件的密集区域（挑战：如何迭代地挖掘邻接矩阵中的密集区域）
5. 分布式差分隐私保护（挑战：在数据提供者不可信时，如何利用差分隐私和加密技术防止聚集信息泄露；实现满足差分隐私保护的数据挖掘）
6. 差分隐私下的大数据分析（作者认为根据大数据的流式特点，利用采样技术、直方图以及概要技术，结合差分隐私可以做大数据的回归分析、模式挖掘、个性化推荐以及概要数据发布等研究）
7. 其他研究点：

如何设置一个合适的预算ε值，来保护敏感数据并最大化所产生数据的可用性；

如何扩展差分隐私保护框架，使其能够保护具有关联性的数据集

## 现状

1. 研究方向：
2. 基于DP的数据发布（常用发布技术：直方图、划分、采样-过滤）；
3. 面向数据挖掘和学习的DP保护技术（主要解决高层隐私需求带来的问题，如何设计满足差分隐私的挖掘和学习算法是其主要目的）；
4. 基于DP的查询处理（主要解决如何以较小的隐私预算与较低的误差来响应查询）；
5. 基于DP的应用系统（提供几种在各类应用环境中可以通用的系统）
6. 保护框架：
7. 交互式：分析者提交的查询通常包含一定的语义约束，数据拥有者常采用后置处理技术对噪音结果进行求精处理；
8. 非交互式：如何设计高效的发布算法，该类算法既满足差分隐私，又具有高的可用性；当前技术：数据拥有者采用数据压缩 、 数据转换与采样过滤等技术，还有，数据发布过程中，合理的隐私预算分配策略
9. 数据发布：
10. 策略：
11. 先对原始数据或者原始数据的统计信息添加噪音，然后对加过噪音后的数据采用规划策略 （ 例如，二次规划、凸规划等 ） 进行优化，最后发布优化结果。
12. 先转换或者压缩原始数据，再对转换后的数据添加噪音。
13. 技术：
14. 以直方图为发布标准的方法：

* 基于上述策略a的直方图发布方法：

原理：通常直接为每个桶的计数添加拉普拉斯噪音，进而达到扰动真实计数的效果

适用范围：支持单位长度的范围查询和较长的范围查询

缺陷：长范围查询时，噪音的累加会导致很大的查询误差；多维范围查询偏差大，也会导致发布误差大

研究方向：实际精度比较好的、误差比较低的、并且能够支持单位和长范围计数查询的多维直方图发布；如何设计高效的后置处理算法

* 基于上述策略b的直方图发布方法：

原理：为各个桶计数添加噪音前，先对直方图的自身结构进行重新组织，再对重组之后的结构添加噪音

分类：

（1）按照层次输结构重组直方图（其中Privelet与Bootst2方法优点在于能够比较精确地响应较长范围的计数查询，缺点在于查询敏感性比较高，实际的性能比较差）；

（2）采用聚类重新划分直方图的各个桶（思路：减少查询敏感性、合理地分配隐私预算；缺陷：只采用简单的启发式规则确定合并桶的个数k， 而这种方法没有顾及到重构误差和噪音误差之间的平衡；其次，对于一些桶个数比较多的直方图，该方法效率非常低）；

（3）采用傅里叶变换有损压缩直方图

1. 基于划分的发布方法

* 思路：考虑如何设计支持数据划分的索引结构，并依据索引结构发布隐私数据
* 是否在原始基础数据上划分分类：数据依赖的划分（在基础数据上划分，可能会导致划分结构自身泄漏数据隐私）；数据独立的划分（在查询空间上的划分，并不涉及基础数据）
* 常用的索引划分结构：

（1）树结构

挑战：如何设计分割条件

分类：

数据独立的树划分：（代表方法：Quad-Post；原理：采用最小二乘法无偏估计对最终噪音响应结果进行后置处理；优点：能够合理分配隐私预算，噪音误差低；缺点：仅适应于二维空间数据，均匀假设误差比较高）；

数据依赖的树划分：（缺点：该类划分通常存在树自身的结构会披露基础数据隐私的问题；代表方法：kd-standard、DiffPart（原理：采用自顶向下的方式随机地分割基于泛化技术的分类树；缺点：仅支持计数查询，并且在泛化集值数据时，没有考虑不同项之间的语义关联）；挑战：如何根据数据空间的中值数确定分割线是划分的关键）

（2）网格结构

关键：划分粒度大小的选择

挑战：如何利用启发式规则指导自适应地划分网络粒度

缺点：局限于二维空间数据；在设计均衡策略方面，为考虑到如何利用启发式规则来自适应地设置均衡参数

* 是否在原始基础数据上划分分类：数据依赖的划分（在基础数据上划分，可能会导致划分结构自身泄漏数据隐私）；数据独立的划分（在查询空间上的划分，并不涉及基础数据）

（注：无论是树结构还是网格结构都需要考虑是否在原始基础数据上划分，上述两类分类是不一样的）

1. 数据分析方法
2. 基于数据挖掘的分析方法：

分类：

基于差分隐私的模式挖掘技术（目的：目的是找出频繁出现在数据集中的模式）；

基于差分隐私的分类技术（目的：找出描述和区分数据类或概念的模型）；

基于差分隐私的聚类方法

1. 基于机器学习的分析方法
2. 目的：采用统计学习方法对已知敏感数据进行分类，或者对未知新敏感数据进行预测和分析
3. 代表技术：

* 基于差分隐私的回归分析：

目的：确定输入数据集中两种或两种以上属性间相互依赖的定量关系

分类：基于拉普拉斯机制的回归分析方法（噪音误差较高；回归分类精度较低）；基于扰动机制的回归分析（仅适用于特定的目标函数）；基于函数机制的回归分析（仅适用于线性表示的目标函数）

* 基于差分隐私的支持向量机

目的：处理输入空间为非线性的分类问题

分类：基于拉普拉斯机制的向量机分类（精度低、噪音大）；基于扰动目标函数的支持向量机（仅适用于特定的目标函数）

1. 查询处理方法：
2. 基于矩阵机制的批量查询处理：
3. 优点：理论角度优于拉普拉斯机制
4. 缺点：适用范围小（仅适合小规模的数据集和查询负载）；产生次优化的查询计划，其返回结果的准确性远不及直接在D上添加噪音产生的结果
5. 基于低秩机制的批量查询处理：

缺点：只考虑了负载矩阵的关联性，而没有顾及数据本身的关联性

1. 相关应用系统：

基于交互式的数据分析系统（微软的PINQ）；

基于非交互式的数据分析系统（美国伯克利大学的GUPT）；

基于MapReduce的聚集分析系统（德克萨斯大学的Airavat）

1. 差分隐私原型系统可能存在的攻击：

状态攻击：指攻击者可能会修改系统中的静态变量；

隐私预算攻击：指预算被耗尽；

定时攻击：指攻击者无限请求同一个分析任务

## 研究热点

1. 如何发布具有代表性的数据，而不披露数据的隐私
2. 实际精度比较好的、误差比较低的、并且能够支持单位和长范围计数查询的多维直方图发布
3. 如何设计高效的后置处理算法

## 未来展望

未来研究方向：

6.1数据发布：

1. 以直方图为发布标准的方法：
2. 基于层次树结构重组：发布多维直方图；如何设定合理的扇出m
3. 基于聚类重新划分：如何发布携带离群点的直方图；如何设计自适应的排序分割方法
4. 采用傅里叶变换发布直方图：如何支持多维直方图的发布以及支持携带离群点的直方图发布；
5. 基于划分的发布方法中：
6. 如何均衡重构误差与噪音误差
7. 如何设计分割条件（例如，基于区域密度的分割条件）
8. 如何利用启发式规则指导自适应地划分网络粒度
9. 如何设计支持高维的、支持数据依赖的树划分方法
10. 如何设计具有启发式的均衡参数设置方法
11. 如何设计支持动态数据的划分发布方法

6.2基于数据挖掘的分析方法：

如何设计同时兼顾Top-k模式的隐私性与可用性的方法

6.3查询处理方法

如何从数据本身的实际相关性出发，设计出通用的批量处理机制

## 阅读思考

1. 在数据发布的两类方法中更倾向于先清洗数据，后添加噪声，但在中心化和本地化DP中肯定更倾向于本地化DP（其中如果是高维数据，先进行降维、再进行保护）。
2. 关于直方图发布的缺陷其实可以考虑限定范围，不追求全部适用，找到合适的适用范围就可以了，但问题是怎么找，用什么手段去证明是要继续的思考。
3. 在基于划分的发布方法中，大概是由于看了网格划分的手段，所以更倾向于网格划分，至于是否在原始基础数据上划分，不是很明白，但直观更倾向于数据依赖划分，确实仅从本文看来数据独立的划分要好
4. 基于树结构的划分发布方法中均匀假设误差是什么意思（参考文献33：一般假设数据在均匀分布的情况下响应查询，而数据实际分布与均匀分布存在偏差，该偏差带来的误差为均匀假设误差）
5. 在kd-standard方法中的隐私预算被分为两部分：一用来确定中值数，另一部分用于为每个分割空间中的数据计数添加噪音
6. 什么是启发式规则？
7. 研究方向要不要暂定为基于差分隐私在查询空间上的网格划分发布方法？还是基于差分隐私在数据独立的树结构划分发布方法？基于关联数据的本地化差分隐私发布方法研究？

# 24．面对数据直方图发布的差分隐私保护综述

## 背景

出版时间：2017年6月

作者：王红，葛丽娜，王丽颖，张静，张翼鹏

刊物：计算机应用研究

说明：暂时不做提炼

## 问题

## 面临的挑战

## 现状

## 研究热点

## 未来展望

## 阅读思考

1. 基于傅里叶变换直方图发布的DP算法开销大，这方面的研究还是不要了，轻量级的更值得研究。
2. 可以考虑推荐系统下的DP保护——保护好关联性数据敏感信息
3. 位置DP相对轨迹DP的研究更简单一点

# 14．本地差分隐私研究综述

## 背景

出版时间：2017年7月（2018年过审）

作者：叶青青、孟小峰、朱敏杰、霍峥

刊物：软件学报

## 问题

1. 中心化差分隐私中：在实际应用中,即使第三方数据收集者宣称不会窃取和泄露用户的敏感信息,用户的隐私依旧得不到保障.
2. 本地化差分隐私中多值频数统计中：（1）需要根据变量的数量分割隐私预算，当变量较多时，数据可用性急剧降低；（2）忽略了变量之间的关联关系 ，损失了部分信息

（注意：有相关研究者解决了部分问题）

## 面临的挑战

1. 中心化差分隐私：如何保证数据收集者不会从中窃取用户的隐私信息
2. 本地化差分隐私：由于噪声的随机性 , 要保证统计结果的无偏性 , 需要海量的数据集来实现满足数据可用性的统计精度
3. 扰动机制（随机响应）中：

离散型数据的随机响应的挑战：属性候选值的编码和匹配策略

连续型数据的随机响应挑战：（1）如何合理设置离散化的两个数值 ; （2）如何保证统计结果的无偏性

1. 本地差分隐私的挑战具体可见未来展望部分

## 现状

1. 中心化差分隐私：
2. 定义：将原始数据集中到一个数据中心,然后发布满足差分隐私的相关统计信息（即先收集原始数据，后进行隐私扰动）
3. 面临的问题：在实际应用中,即使第三方数据收集者宣称不会窃取和泄露用户的敏感信息,用户的隐私依旧得不到保障.
4. 数据保护框架：交互式和非交互式框架（注：与本地DP的框架意义上是不同的）
5. 研究方面：数据发布、数据分析、查询处理
6. 本地化差分隐私：
7. 定义：将数据的隐私化处理过程转移到每个用户上,使得用户能够单独地处理和保护个人敏感信息（即先进行本地数据扰动，后发送给收集者）
8. 原理：对个体数据进行正向和负向的扰动 , 最终通过聚合大量的扰动结果来抵消添加在其中的正负向噪声 , 从而得到有效的统计结果
9. 特点：
10. 充分考虑任意攻击者的背景知识,并对隐私保护程度进行量化;
11. 本地化扰动数据,抵御来自不可信第三方数据收集者的隐私攻击.
12. 应用场景：

众包数据采集、敏感图像特征提取

1. 主流扰动机制：随机响应W-RR
2. 离散型数据的随机响应：

思路：（1）对变量的不同取值进行编码和转化,使其满足W-RR 技术对二值变量的要求；（2）改进 W-RR 技术 , 使其能够直接适用于超过两种取值的变量

1. 连续型数据的随机响应

思想：将连续型数据离散化 , 然后利用离散型数据下的随机响应方法 , 对数据进行扰动

1. 数据保护框架：交互式和非交互式框架（区别：输出结果之间的关联性，前者适用于最终输出结果与前i个输出有依赖关系的情形，后者适用于无依赖关系的情形）
2. 本地化差分隐私技术研究方向：
3. 扰动机制的研究：
4. 随机响应：

针对离散型数据的频数发布和针对连续型数据的均值发布（频数发布形式：列联表、直方图；按变量的数量不同分类：单值频数发布、多值频数发布）；

优点：扰动框架简洁、直观，其扰动程度可以直接量化；

1. 信息压缩和扭曲
2. 统计数据的发布：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.
3. 注：无论是频数统计还是均值统计 , 当给定相同隐私预算时 , 用于统计的数据量大小决定了数据可用性高低 , 数据量大则统计结果的可用性高 , 数据量小则统计结果的可用性低

## 研究热点

1. 如何设计满足Ɛ -本地化差分隐私的数据扰动算法,以保护其中的敏感信息（即理论上 , 设计满足本地化差分隐私的保护机制 ）
2. 数据收集者如何对查询结果进行求精处理,以提高统计结果的可用性
3. 方法上 , 对频数和均值两种统计结果进行保护
4. 基于随机响应技术的扰动机制（当前研究热点）：

分类：针对离散型数据的频数统计查询和针对连续型数据的均值统计查询.

## 未来展望

三个方面：

1. 复杂数据类型的本地化差分隐私保护

挑战：, 数据收集者如何根据扰动后的数据构建可用性高的图结构 , 即如何保证原始数据之间的关联性

1. 不同查询和分析任务的本地化差分隐私保护

问题：

1. 提供对除计数查询和均值查询以外的多种查询方式的支持 ;
2. 数据扰动方式与查询类型的解绑 , 使得扰动后的数据能够同时支持多种查询 ;
3. 提高数据分析结果的可用性
4. 基于本地化差分隐私的高维数据发布
5. 挑战：数据规模变大、信噪比降低、通信代价大（通信代价根据不同的扰动机制随数据维度的增加呈线性增长或指数增长）
6. 思路：

如何在一定隐私预算内衡量属性之间的关联性 , 从而进行降维处理；

如何设计推理模型 , 最小化边缘分布到联合分布的近似误差 ,

提高数据可用性；如何控制高维数据在用户和数据收集者之间的通信代价

1. 现状：利用属性划分的思想

## 阅读思考

位置差分隐私和本地差分隐私是同一个概念，才看到翻译Local Differential Privacy，是自己翻译错误了。本地化差分隐私是相对于中心差分隐私而言的，关键之处在与去中心化，没有一个可信的第三方数据中心，是基于位置服务的差分隐私保护。

# 19．本地差分隐私保护及其应用

## 背景

出版时间：2018年6月

作者：高志强，崔俺龙，周沙，袁琛

刊物：计算机工程与科学

## 问题

基于随机应答中的RAPPOR技术中：

（1）使用RAPPOR的数据收集者只能孤立地了解单一变量的分布；

（2）数据收集者必须事先知道潜在字符串字典、安装软件的报告、名称、hash值，然而这些是不可能作为先验知识的。

## 面临的挑战

流式频繁项集挖掘中：（1）解决top-k频繁项集任务；（2）大量的通信消耗、隐私预算的损耗与可用性的平衡

基于LDP的众包机器学习中：分布式条件下的数据同步与集中统计分析

## 现状

1. 应用领域：集值型流式频繁项集挖掘的Heavy Hitters估计、众包模式下字符串边缘频率估计和联合概率估计（众包模式下的智能终端数据收集与统计分析）、针对智能设备的机器学习
2. 主要思想：用户在本地通过差分隐私技术来置乱原始数据，然后再把加噪数据发送给收集者（即保证收集者不能收集或拥有任何个人的精确信息；并且可以推断出用户群体的泛化统计信息）
3. 现有的RR技术的LDP机制的局限性：只适用于用户数据类型为数值型或范围型
4. 研究方向：
5. 基于随机应答与BloomFiher的编解码方式研究、
6. 针对流式频繁项集挖掘Heavy Hitters挖掘、
7. 针对智能终端的收集与机器学习：着重关注苹果（局部抽样、Hash加密、噪声扰动）

## 研究热点

1. RR技术的LDP机制：在收集群体层面的统计数据而不泄露个体数据方面具有优越性能

## 未来展望

1. 基于LDP的众包机器学习：在支持用户隐私保护的数据收集、支持隐私保护的机器学习、统计分析等领域具有很大的研究前景。
2. 抵抗新型攻击的能力方面：例如具有隐私保护数据生成能力的生成式对抗网络GAN(Generative Adversarial Networks)
3. 差分隐私下的大数据分析：基于RAPPOR技术的本地隐私保护算法在大规模数据字典的构建与计算效能间的优化

## 阅读思考

值得关注：

2014年谷歌将基于随机应答和BloomFilt的RAPPOR(Randomized Aggregatable Privacy—Preserving Ordinal Response)技术成功应用于Google Chrome中。

2017年6月，圣何塞McEnery会议中心的苹果开发者大会(WWDC2017)Ez0]发布了面向开发者的机器学习API——CoreML

为什么在收集用户数据前，先加入随机噪声，只要被注入的噪音抽样是正态分布的，那么整体来看，这些噪音最终将相互抵消？

# 相关内容

1.（链接：<https://zhuanlan.zhihu.com/p/72189648>）

基于差分隐私的LBS用户地理位置保护（知乎）：Geo-indistinguishability在经典的差分隐私模型上进行扩展，针对单用户的位置发布情况，在用户的真实位置上添加噪声，使得最终发布出的用户位置在真实位置的一个圆形范围内，并且离真实位置越近的位置点会有越高的概率被作为加噪后的位置点进行发布，最终使得服务提供者无法从收集到的位置信息得到用户的真实地理位置。

总体悲观：基于差分隐私的方法离理想的保护技术任然相去甚远；服务提供者掌握了太多的先验知识，保护在很多时候没有意义

保护总比不保护好：陌生地点保护有效（陌生地点通常也是敏感地点）；过度追踪可以避免

未来在哪里：密码学技术的突破

# 笔者感悟

我对大数据的定义还没有理解懂，以至于我在开题定题目的时候很艰难

## Paper 5（2014）

5.2基于隐私信息检索方法没看懂

## Paper 11（2011）

Q1：算法5的第4条语句——dir是方向的意思？w.r.t是什么意思？（with regard to关于的意思吗）

Q2：该文的真实数据Athens卡车数据集是卡车上的人使用手机查询的数据，还是车本身导航/系统查询的数据呢（是自动追踪还是被动追踪？好吧，有点没说清楚）？

注意snapshot LBS指的就是单时刻的LBS，还有一个叫做连续的LBS

## Paper 12（2018）

当关联数据很多时，组差分隐私会失效，达不到保护隐私的效果。

关于相关数据对敏感性的影响有待研究。

DP本质上属于数据扰动技术

该文对理解Laplace机制有帮助，容易理解

该文主要针对的是外部相关数据的影响

Ⅳ部分有必要再多看几次

要知道对于大数据而言，相关数据占比是很小的，该情况下，分组差分隐私就可以实现隐私保护了；对于单个数据实现DP太耗费资源，可以尝试分块（即划分），在每个块上实现不同隐私

如何使用MIC来判断不同ID子数据集是否相关

MIC的概念都不理解，应该查找相关的论文去学习，参考文献【27-30】

## Paper 14（2017）

Part 5的具体方法部分没有仔细看

该文的参考文献[47,48]（高维数据发布问题）,[26]（线性回归、SVM分类、Logistic回归）,[64]（大数据隐私风险）值得学习

## Paper 19（2018）

去了解对抗神经网络、众包模式、统计分析

学习相关数学知识：最大似然估计、无偏离估计、哈希函数、最小二乘求解

苹果资源短期内是不可能开源的

## Paper 20（2018）

隐私预算Ɛ的设置有什么依据呢？

针对于k-匿名、l-多样性的思想没能理解，

在攻击模型中的f的敏感度不知道怎么计算

很认同老师说的那句话“数学是种素养”。看到LPBDP算法，从头懵逼到最后，最后得出一个p-差分可识别是LPBDP的特例时，只想说句666。

UCI的Adult数据库是什么

## Paper 21（2006）

任何安全问题都会涉及到的一个问题：什么是保护隐私的失败？该问题在1970’s已经有了很明确的定义。

经典文献要好好读，多读几遍总会有新的体会的

本文采用给等宽的直方图添加噪声的思想LP，但LP存在长区间查询、噪声累加导致查询误差较大的情况。

## Paper 22（2018）

## Paper 23（2019）

## Paper 24（2017）

## Paper 25（2018）

## Paper 26（2014）

1. 在数据发布的两类方法中更倾向于先清洗数据，后添加噪声，但在中心化和本地化DP中肯定更倾向于本地化DP（其中如果是高维数据，先进行降维、再进行保护）。
2. 关于直方图发布的缺陷其实可以考虑限定范围，不追求全部适用，找到合适的适用范围就可以了，但问题是怎么找，用什么手段去证明是要继续的思考。
3. 在基于划分的发布方法中，大概是由于看了网格划分的手段，所以更倾向于网格划分，至于是否在原始基础数据上划分，不是很明白，但直观更倾向于数据依赖划分，确实仅从本文看来数据独立的划分要好
4. 基于树结构的划分发布方法中均匀假设误差是什么意思（参考文献33：一般假设数据在均匀分布的情况下响应查询，而数据实际分布与均匀分布存在偏差，该偏差带来的误差为均匀假设误差）
5. 在kd-standard方法中的隐私预算被分为两部分：一用来确定中值数，另一部分用于为每个分割空间中的数据计数添加噪音
6. 什么是启发式规则？
7. 研究方向要不要暂定为基于差分隐私在查询空间上的网格划分发布方法？还是基于差分隐私在数据独立的树结构划分发布方法？基于关联数据的位置差分隐私发布方法研究

## Paper 27（2018）

将大数据划分成块，对每个块进行差分隐私保护，由于并行性特征，最终可实现对整个大数据保护

是利用MIC（互信息相关分析）还是利用距离计算相关分析来判断不同ID子数据集是否相关呢？论文12用的是MIC，然后通过机器学习得到相关模型。

## Paper 28（2018）

# 相关词汇

基于隐私信息检索（PIR）

基于位置的服务（LBS：location based services）

本地差分隐私（LDP：Local differential privacy）