DP and UPP of TD论文阅读

- 摘要
  - GPS、应用定位广泛使用使信息收集变的容易
  - 这些位置信息对第三方和政府有用
  - 提出DP-star轨迹数据发布框架
    - 属性：隐私性&实用性&准确性
    - ①：轨迹选取代表点、归一化算法使用MDL，平衡精确性和简洁性
    - ②：构建一个密度感知的网格空间
    - ③：私有行程分布保留轨迹起终点之间的相关性，中间的相关点-马尔可夫移动模型
    - ④：中值长度估计法->合成用户的行程长度
- 介绍部分
  - 收集信息便利
  - 政府&第三方需求
  - 应用
    - 城市规划
    - 旅行模式分析
    - 路线推荐
    - 交通项目管理
    - 挖掘跨市跨县跨区轨迹数据
  - 发布会引出隐私担忧
    - 传统匿名、K-匿名、空间位置隐匿等不能够很好地保护隐私
    - 因为可以连接攻击和背景知识来重建数据库
  - 引出DP
    - 定义一个DP算法：
      - 个人是否存在不影响OUTPUT
    - 由此引出用DP来发布合成轨迹数据SD
      - 合成数据集和原数据集没有一一对应关系（躲避连接攻击）
      - 攻击者不能推断出用户是否在数据集中
  - 之前的工作
    - 解决的是单个位置点的发布问题
    - 假设位置信息只来自小的离散区域
      - 地铁站、公共汽车站
    - 两个最新的相关工作

        - Ngram
        - DPT
- 相关工作
    - 位置匿名
        - 传统的匿名是K-匿名及其扩展的匿名化
        - [18] (k, δ)-匿名 k个不同轨迹在δ半径中
        - [3] 位置泛化执行轨迹K-匿名
        - .....
        - [4], [5]表明个人轨迹 高度独特性&可预测性，仅在空间粗化下辨识只是略有下降
    - 位置差分隐私
        - 当时研究的两个里程碑
            - 统一位置的K-匿名
            - 用户定义的、个性化的K-匿名
        - 提出了基于DP两个保护位置的方法
            - 地理不可辨别性
            - 攻击者推理错误
        - 指出位置保护与轨迹保护不同
            - 位置关注瞬时 单一；轨迹关注连续 时间顺序点列
            - 现有轨迹挖掘希望获得原始轨迹组
            - 目的不在将用户的出行事实隐藏
    - 发布差分隐私数据
        - 研究工作 分两类
            - 发布DP密度统计
            - 发布DP合成数据集
        - 工作集中 计数查询 空间数据图
            - 空间索引 （四叉树、kd树、 Hilbert R-trees）
            - 以上过渡划分 提出单层和双层
            - 提出PrivTree
            - 通话记录等 只可发布密度和计数统计数据 仍不能发布轨迹数据
            - 构成合成轨迹仍然是一个开放性问题.....
            - [15] Ngram [17] DPT
- 设计概述
    - 符号说明
    - 差分隐私定义
        - 邻接数据集
        - ε-差分隐私

        - 敏感度
        - 拉普拉斯机制
        - 指数机制
        - 简单组合属性
            - 串行组合 相加
            - 并行组合 划分 取最大
            - 后处理 看复合函数定义域
- 解决方案
    - 设计目标
        - 采用网格作为索引结构（离散化Ω(D)为一组网格单元)
        - 在SD中保留空间效用
            - 保存轨迹 起 终点以及其中关联性
            - 保存轨迹内的流动性
        - Ngram 中间区域过于密集 其余区域过于稀疏
        - DPT 分层参考 保留轨迹内移动性 但起 终点之间没有相关性
    - DP-star设计
        - 预处理
            - 选择代表性点 （归一化使用MDL）
                - 在假设T˜下 找到最小 $L(T\tilde{})+L(T|T\tilde{})$
            - 配置参数 如ε
                - 将总隐私预算ε 分配到前4个核心环节
                    - 网格构建和线路从长估计可以承受较低隐私预算
                - 自动分配隐私预算ε
                    - 4个误差衡量标准
                        - 空间密度查询平均相对误差（AvRE）
                        - 频繁模式系数 （Kendall-tau）
                        - 行程误差 （ trip error）
                        - 直径误差 （ diameter error）
                    - 爬坡随机重启
                        - 随机一组参数，每次迭代增量为0.02ε
                    - 随机重启技术
                        - 运行若干个爬坡实例 跟踪最优参数
        - 核心步骤
            - 使用一个密度感知的自适应结构来离散Ω(D˜)
                - AG是一个二维网格
                    - 高密度 多细小单元

        - 低密度 粗大单元
    - 将A看成一个划分了很多层的网格的总和，最多网格的那一层个数为 N × N
    - 计数查询中除以|T˜|是必要的，以便于定一个灵敏度上限
    - 输出A这个网格的算法：
        - 先将初始区域均分为N × N个cells 一般为N=7
        - 计算每个cell的计数查询+拉普拉斯噪声
        - 再将每个cell根据公式再分成M × M个更小cells
        - β=（ε−ε1）/80
- 保存起 终点 来凝练行程分布
    - 公式中除以|D˜| 是确保R仍然是一个概率质量函数
    - 这一步我们需要提炼出一个关于不同cells的行程分布Rˆ
    - 公式中我们给每个|D˜Ci~Cj |都加了一个拉普拉斯噪声
- 用马尔可夫模型保存轨迹内移动行程
    - r阶马尔可夫模型 r低越好
    - 下一个位置确定 由 p(n−r+1)p(n−r+2)...pn
    - 由于r1<r2,虽然有拉普拉斯噪声，但是r1的信噪比还是会更好 r2由于点数更少，于是噪声占主导（解释第一条）
    - 马尔可夫模型
        - 一系列cells的状态
        - 一系列由这个cell到其它cells的过渡路径
            - 过渡的值是概率
            - 概率出自一个过渡矩阵中的元素
            - 根据公式获得一个过渡矩阵
- 中值法估计路径长度
    - 用一条轨迹的代表点个数近似表示轨迹长度
    - 按论文表述不是直接就是长度吗，为什么会出来一个中位数长度?
- 生成合成数据集

以上内容整理于 幕布文档