面向轨迹数据发布的个性化差分隐私保护机制

田 丰[1] 吴振强[1] 鲁来凤[2] 刘 海[3] 桂小林[4]

[1](陕西师范大学计算机科学学院 西安 710062)

[2](陕西师范大学数学与信息科学学院 西安 710062)

[3](贵州大学公共大数据国家重点实验室 贵阳 550025)

[4](西安交通大学计算机科学与技术学院 西安 710049)

摘 要 移动互联网和智能手机的普及大大方便了人们的生活,并由此产生了大量的轨迹数据.通过对发布的轨迹数据进行分析,能够有效提高基于位置服务的质量,进而推动智慧城市相关应用的发展,例如智能交通管理、基础设计规划以及道路拥塞预警与检测.然而,由于轨迹数据中包含用户的敏感信息,直接发布原始的轨迹数据会对个人隐私造成严重威胁.差分隐私作为一种具备严格形式化定义、强隐私性保证的安全机制,已经被广泛应用于轨迹数据的发布中.但是,现有的方法假定用户具有相同的隐私偏好,并且为所有用户提供相同级别的隐私保护,这会导致对某些用户提供的隐私保护级别不足,而某些用户则获得过多的隐私保护.为满足不同用户的隐私保护需求,提高数据可用性,本文假设用户具备不同的隐私需求,提出了一种面向轨迹数据的个性化差分隐私发布机制.该机制利用 Hilbert 曲线提取轨迹数据在各个时刻的分布特征,生成位置聚簇,使用抽样机制和指数机制选择各个位置聚簇的代表元,进而利用位置代表元对原始轨迹数据进行泛化,从而生成待发布轨迹数据.在真实轨迹数据集上的实验表明,与基于标准差分隐私的方法相比,本文提出的机制在隐私保护和数据可用性之间提供了更好的平衡.

关键词 个性化差分隐私;Hilbert 曲线;抽样机制;轨迹数据发布

中图法分类号 TP311 **DOI 号** 10.11897/SP.J.1016.2021.00709

A Sample Based Personalized Differential Privacy Mechanism for Trajectory Data Publication

TIAN Feng[1] WU Zhen-Qiang[1] LU Lai-Feng[2] LIU Hai[3] GUI Xiao-Lin[4]

[1](*School of Computer Science, Shaanxi Normal University, Xi'an* 710062)

[2](*School of Mathematics and Information Science, Shaanxi Normal University, Xi'an* 710062)

[3](*State key Laboratory of Public Big Data, Guizhou University, Guiyang* 550025)

[4](*School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an* 710049)

Abstract The widespread of smart phones and mobile internet facilitates people's lives. Meanwhile, a large number of users' trajectory data are collected and analyzed to provide better location-based services. Publishing the trajectory data can benefit the applications such as the intelligent transportation management, infrastructure planning, and road congestion prediction and detection. As the trajectory data contains users' sensitive information, the publication of the original trajectory data may lead to the privacy leakage risks. To solve this problem, researchers have proposed privacy-preserving schemes to obfuscate the original trajectory data. These schemes are mainly on the basis of partition-based privacy models, such as k-anonymity and confidence bounding. Thus,

收稿日期:2019-11-12;在线发布日期:2020-08-15.本课题得到国家自然科学基金项目(61602290,61902229,61672334,61802242,61802241)、陕西省自然科学基础研究计划项目(2017JQ6038,2020JM-288)、贵州省科技重大专项计划项目(2018BDKFJJ004)、中央高校基本科研业务费(GK202103090, GK202103084)资助.**田 丰**,博士,副教授,硕士生导师,中国计算机学会(CCF)会员,主要研究方向为空间数据外包隐私保护、差分隐私、网络安全.E-mail: tianfeng@snnu.edu.cn. **吴振强**(通信作者),博士,教授,博士生导师,中国计算机学会(CCF)高级会员,主要研究领域为隐私保护、网络科学.E-mail: zqiangwu@snnu.edu.cn. **鲁来凤**,博士,副教授,中国计算机学会(CCF)会员,主要研究方向为差分隐私、数据安全. **刘 海**,博士,助理研究员,中国计算机学会(CCF)会员,主要研究方向为差分隐私、基因隐私保护. **桂小林**,博士,教授,博士生导师,中国计算机学会(CCF)高级会员,主要研究领域为云计算、网络安全、物联网.

they cannot resist the inference analysis of attackers with background knowledge. As a de facto standard, differential privacy guarantees the privacy level of the released data set, and privacy leakage risk is not affected by the background knowledge of the attackers. However, the existing differentially private schemes provide the users with the same privacy level, while the users usually have various privacy preferences. These schemes may narrow down the scope of available trajectory data, since some users' privacy preference cannot be guaranteed. In this paper, we propose a sample based personalized differential privacy mechanism for trajectory data publication, which provides users with different privacy budget. Firstly, a location clustering algorithm is designed based on the linear indexes generated by the Hilbert curve. Inspired by the space filling curve, we partition the location set and employ the Hilbert curve to traverse the space regions to generate linear indexes of the location set. Different from traditional two-dimensional data clustering algorithms, this algorithm takes linear indexes as input to generate clusters in one-dimensional space. The algorithm can maintain the distance and distribution characteristics of the locations in the trajectory data set. By linearly scanning the indexes, the location clusters can be effectively obtained. In addition, the algorithm does not need to set the same number of clusters on the location sets at different timestamps, but generates different numbers of clusters according to the different distributions of the location sets. Secondly, a generalization method is proposed to meet personalized differential privacy. This method takes into account the different privacy preferences of individuals, and generates the representative element of each location cluster in a personalized differential privacy way. Specifically, the method determines the selection probability of each location according to its privacy budget, and samples the locations in the clusters at each timestamp. Then, the exponential mechanism is employed to select the representative location of each cluster to ensure that the trajectory generalization process satisfies the personalized differential privacy. The privacy analysis confirms that the proposed mechanism satisfies the definition of personalized differential privacy. The experiments on real trajectory data set show that the proposed mechanism achieves better tradeoff between privacy protection and data utility, compared with the standard differential privacy mechanism. Moreover, the generated representative locations are taken from the original location set, thus it will not lead to the generation of meaningless representative locations, ensuring that the generalized trajectory data set can resist filtering attacks.

Keywords personalized differential privacy; Hilbert curve; sample mechanism; trajectory data publication

1 引 言

智能终端和移动互联网的发展,极大地便利了人们的生活,通过提交个人位置信息,移动用户可以获得大量基于位置的服务(Location-Based Services, LBS),如邻近社交、生活信息获取、签到、导航等.假定每个移动用户平均 15 秒上传一次位置信息,则全球数十亿台智能终端每秒钟提交的位置数据超过十亿条.这些序列化的位置数据即为轨迹数据,具有规模大、生成速度快、蕴含价值高等特点[1-3].通过 LBS 服务商采集的轨迹数据进行发布,能够用于支持政府、商业机构进行设施规划、交通管理、应急处理等智能公共服务,在国民经济资源的优化配置方面有着重要作用.例如,AirSage 公司每天通过处理上百万手机用户的 150 亿条位置信息,为美国超过 100 个城市提供实时交通信息.根据瑞典市场研究公司 Berg Insight 发布的最新报告预测,全球基于位置的服务市场规模将以 22.5% 的复合年增长率从 2014 年的 103 亿欧元,增加至 2020 年的 348 亿欧元,由此带来了轨迹数据的大规模增长.然而,轨迹数据在带给人们巨大收益的同时,也带来了日益严重的安全问题.轨迹数据通常包含敏感的个人信息,简单的共享或发布都可能会导致严重的隐私泄露问题[4].因此,随着轨迹数据的发布和基于位置服务的发展,满足个人隐私要求是亟待解决的问题.

面向 LBS 的隐私保护机制通常采用基于划分的隐私模型[5]，如 k-匿名性[6]和置信限界[7]. 但是，这些模型都无法应对具有背景知识的攻击者，例如前景知识攻击[8]、deFinetti 攻击[9]、组合攻击[5]等. 针对这些缺陷，差分隐私[10]被引入到隐私保护数据发布中，并成为事实上的隐私模型. 差分隐私对于攻击者的背景知识不做假设，并提供可证明的隐私保护. 这样，在发布数据集上的任何计算均不会对单个记录的存在敏感. 因此，无论记录是否包含在原始数据集中，都不会导致记录所有者的隐私泄漏. Chen 等人[4]首先应用差分隐私机制发布轨迹数据，同时提供良好的数据可用性. 在该方案中，轨迹数据由噪声前缀树进行维护，该前缀树基于轨迹数据集自适应地缩小输出域，从而达到提高数据效用的目的. 此外，还引入了前缀树的两组内在约束进行受限推理，使得生成的轨迹满足语义要求. 之后，他们使用可变长度的 n-gram 来保护序列数据的隐私[11]，其通用性更强. 这两种方法假设原始轨迹包含足够多的公共前缀或 n-gram. 然而，由于位置集合和时间集合都非常庞大，在实际的轨迹数据集中无法满足这一假设. 为了解决这些问题，Hua 等人[12]提出了一种差分隐私位置泛化算法，该算法采用指数机制对每个时刻的位置进行概率合并，并利用 K 均值算法对同一时刻的位置进行聚类. 该方案在保证差分隐私的同时，提供了良好的数据可用性. 但是，他们假设所有用户都有相同的隐私需求，并且向所有用户提供相同的隐私保护级别，从而导致对某些用户提供了过多的隐私保护，而其他用户则没有得到足够的隐私保护. 为了对传统的差分隐私进行扩展，Jorgensen 等人[13]提出了个性化差分隐私(PDP)框架，它规定了用户级的隐私需求，而不是所有用户共享相同的全局隐私预算. 该框架在满足所有用户个性化隐私需求的同时，提供了与传统差分隐私相同的隐私保障.

在前期工作[14]中，我们提出了基于隐私预算加权的轨迹数据泛化方法，满足了用户的个性化隐私偏好. 但是，该方法生成的位置代表元不具有语义信息，导致发布的泛化轨迹容易被敌手识别过滤. 本文对该工作进行了扩展，提出了一种面向轨迹数据发布的个性化差分隐私保护机制，在位置聚簇的代表元生成中引入了抽样机制和指数机制，避免隐私预算的加权平均所生成的位置代表元不具有语义信息，进而导致无效轨迹数据的问题，并严格证明了该机制所能满足的差分隐私特性. 本文的主要工作与贡献如下：

(1) 提出了一种基于 Hilbert 线性索引的聚类算法，通过对线性索引的单向扫描生成每个时刻的位置聚簇，避免了多次迭代，从而提高位置聚簇的生成效率.

(2) 设计了基于抽样机制和指数机制的位置代表元生成方法，该方法避免了无语义的位置被设定为代表元，从而能够抵御过滤攻击. 利用生成的各个时刻的位置代表元对原始轨迹数据进行泛化，从而生成待发布轨迹数据集，通过理论分析证明了该方法满足个性化差分隐私.

(3) 在真实数据集上验证了本文方法的轨迹生成效率，并对生成轨迹数据的可用性进行了评估.

本文第 2 节介绍国内外相关工作的研究进展；第 3 节对轨迹数据与个性化差分隐私进行形式化定义；第 4 节提出面向轨迹数据发布的个性化差分隐私保护机制；第 5 节对该机制的隐私性进行分析证明；第 6 节通过实验对本文机制的效率进行验证；第 7 节对全文进行总结，并讨论未来的研究方向.

2　相关工作

随着轨迹数据的广泛挖掘和分析，关于轨迹数据隐私保护的研究逐渐增多. Abul 等人[15]提出 (k,δ)-匿名模型来修改原始轨迹，并保证 k 个不同的轨迹将被半径为 δ 的圆柱体包含. Domingo-Ferrer 等人[16-17]提出 (k,δ)-匿名无法对已发布的轨迹提供足够的保护，并提出位置交换算法以获得更好的隐私保护级别. 该算法首先对原始轨迹进行聚类，然后对每个聚簇的位置进行排列，从而提供匿名性. Yarovoy 等人[18]利用 k 匿名模型对移动对象数据库进行处理，并将其时间戳作为准标识符，他们首先识别匿名群，然后基于准标识符将群泛化到公共区域. 除了 k-匿名模型，混合区域[19]也是隐私保护方法的另一种选择. 如果将一组用户添加到混合区域，那么混合区域将更改其假名. Ying 等人[20]提出一个动态混合区域来提供不同的隐私需求，用户提交服务请求时其真实位置将被忽略. Poulis 等人[21]针对现有数据匿名方法产生的数据效用较低的问题，提出了基于先验原理的匿名算法，保护轨迹数据的位置距离、语义相似度等数据特征. 针对轨迹的语义特性产生的隐私风险，Monreale 等人[22]基于位置泛化定义了一个称为 c-安全性的模型，将敏感位置链接到轨迹的最大概率限定为 c. 类似地，Cicek 等人[23]提出 p-机密性，通过限定用户访问轨迹中敏感位置的概率来确保位置多样性，避免了传统 k 匿名

方法在敏感区域缺乏候选位置的问题.

这些基于扰动和抑制的方法容易受到具有背景知识的敌手的攻击.为了提供可证明的隐私保护数据发布方法,差分隐私被广泛使用在各种应用中[10].通过差分隐私机制处理的数据集可以支持特定的数据分析任务,如计数查询[11]和频繁模式挖掘[24].Andres 等人[25]提出地理不可区分性,以保护 LBS 用户的真实位置信息.地理不可区分性不仅防止了对个人位置的恶意推测,而且限制了攻击者的通过观察所带来的背景知识的增长.Zhang 等人[26]研究了差分隐私在位置推荐系统中的应用,并引入多维空间索引和 n-gram 树来提高位置推荐的准确度.Cormode 等人[27]采用标准的空间索引技术来提供对数据分布特征的隐私描述,并提出基于差分隐私的空间分解技术,它能够生成高精度的查询结果,同时提供有意义的隐私保证.然而,这些方法均基于传统的差分隐私,即由单一全局的隐私预算决定了隐私保护级别.

为了满足用户不同的隐私需求,除了个性化差分隐私[13]之外,Alaggan 等人[28]还提出了异构差分隐私(HDP),该机制也考虑了非统一隐私预算下的差分隐私,以 Laplace 机制为基础,根据相关隐私预算重新调整原始值,但是,这种机制的局限性在于它只支持数值型数据集,无法对其他类型数据集进行处理.Tian 等人[14]提出了用于轨迹数据发布的个性化隐私保护方法,该方法基于隐私预算加权生成位置聚簇的代表元,满足了用户的个性化隐私偏好.但是,该方法生成的位置代表元不具有语义信息,导致发布的泛化轨迹容易被敌手识别过滤.Cao 等人[29]提出了用于轨迹数据流发布的差分隐私模型:l-轨迹隐私,保证所有预定长度的轨迹都满足 ε-差分隐私,并进一步研究了传统差分隐私在连续数据发布中产生的潜在隐私损失,利用马尔科夫模型对发布数据的时序关系进行建模,提出了能够抵御隐私损失的差分隐私转换机制[30].个性化差分隐私[13]作为一种推广的差分隐私概念,需要根据应用场景的不同,设计相应的实现机制.Wang 等人[31]对于众包任务工作者的不同隐私需求,提出了用于移动众包任务分发的个性化隐私保护框架,根据工作者提交的混淆距离和个性化隐私级别,使用概率机制分配众包任务,在保证工作者不同隐私要求的前提下,提高任务完成的成功率.Xu 等人[32]针对不可信环境下的个性化推荐系统,提出了一种多层次的差分隐私方案,该方案使用 Laplace 机制同时保护服务提供者的总体隐私与每一个数据提供者的隐私,减少了发布数据与原始数据之间的误差.针对智能家居中不同的数据源所具有的不同的隐私级别,且数据可能被传输给多个云服务器,Zhang 等人[33]定义了信任距离,并基于此设计了支持不同噪声输出的个性化差分隐私模型,利用马尔科夫链生成一系列噪声,从而打破了噪声之间的关联关系,能够抵御共谋攻击.Qu 等人[34]提出了基于社交距离的个性化差分隐私机制,对于具有不同社交距离的用户设定不同的隐私预算,社交距离较近的具有较多的隐私预算,社交具体较远的则分配较少的隐私预算,从而提高数据的效用.Li 等人[35]针对标准差分隐私机制无法满足社交网络中不同用户的不同隐私偏好的问题,基于个性化差分隐私,提出了用于社交网络中特定统计数据的隐私保护发布方法.这些研究针对不同的数据类型和应用特征,设计了满足用户差异化隐私需求的数据发布机制.为了满足轨迹数据发布的个性化隐私偏好与位置语义需求,本文将概率抽样与指数机制引入到泛化位置代表元的生成中,减少了发布轨迹中无效数据的生成.

3 问题描述

本节对轨迹数据库和个性化差分隐私进行形式化定义.本文使用的符号描述如表 1 所示.

表 1 符号表

符号	含义
T	一条轨迹数据,由 n 个时刻的位置组成的序列
D	若干条轨迹构成的轨迹数据集
D_{-T}	轨迹数据集 D 的邻近数据集,比 D 少一条轨迹 T
$\boldsymbol{D}$	轨迹空间,所有轨迹均取自该集合
$\boldsymbol{\Omega}$	隐私配置,为用户集合到个人隐私偏好的映射
H_d^N	d 维空间中的 N 阶 Hilbert 曲线
LC_i^k	时刻 t_i 的第 k 个位置聚簇
LC_i	时刻 t_i 的位置集合,即 $LC_i=\cup_k LC_i^k$
$\boldsymbol{\Omega}_i^k$	与位置聚簇 LC_i^k 关联的隐私配置信息
$\boldsymbol{\Omega}_i$	与时刻 t_i 的位置集合关联的隐私配置信息
π_l	位置 l 被选中的概率
φ	用于位置抽样的隐私预算阈值
SLC_i^k	基于隐私配置 $\boldsymbol{\Omega}_i^k$ 和阈值 φ 从位置聚簇 LC_i^k 获得的抽样位置集合
L,L_{-l}	邻近的位置数据集,其中 L_{-l} 比 L 少一个位置 l
Z	由位置数据集 L_{-l} 抽样获得的位置集合
Z_{+l}	比位置数据集 Z 多一个位置 l 的位置集合
U_i^k	与隐私配置 $\boldsymbol{\Omega}_i^k$ 关联的用户集合
$U(L)$	与位置数据集 L 关联的用户集合
$\boldsymbol{A}_{m+1}$	从 m 个位置聚簇选择各个聚簇代表元的机制
$\boldsymbol{B}_i$	在时刻 t_i 的位置集合 LC_i 上选择各个聚簇代表元的机制
$\boldsymbol{G}$	在轨迹数据集的 n 个时刻中,选择各时刻各个聚簇代表元的机制

3.1　轨迹数据库

轨迹是一个位置与相应时间戳的有序列表，表示为 $T=(l_1,t_1)\rightarrow(l_2,t_2)\rightarrow\cdots\rightarrow(l_n,t_n)$，其中 n 为该条轨迹的长度，$l_i\in L_i(1\leq i\leq n)$，$L_i$ 为时刻 t_i 的位置集合. 位置 l_i 表示地图上的离散点，使用经度、纬度表示. 时刻 t_i 则是取自时间戳集合 $\{t_1,t_2,\cdots,t_n\}$，$T(t_i)$ 则表示轨迹 T 在时刻 t_i 的位置，即 $T(t_i)=l_i$.

一个规模为 $|D|$ 的轨迹数据库 D 是由若干轨迹组成的一个多重集，即 $D=\{T_1,T_2,\cdots,T_{|D|}\}$，且轨迹数据库 D 取自于轨迹空间 $\boldsymbol{D}$，即满足 $D\subset\boldsymbol{D}$. 本文假定所有轨迹的附属时刻均取自相同的时间戳集合，因此，我们可以对处于相同时刻的轨迹位置进行处理.

定义 1.　邻近轨迹数据集.

如果两个轨迹数据集 $D,D'\subset\boldsymbol{D}$，之间仅相差一条轨迹，即 $D\oplus D'=T$，其中轨迹 $T\in D$ 且 $T\notin D'$，则称 D 和 D' 为邻近轨迹数据集，表示为 $D\sim D'$.

3.2　差分隐私

一个随机机制 $\boldsymbol{A}:D\rightarrow Range(\boldsymbol{A})$，它接收一个数据集作为输入，并从 $Range(\boldsymbol{A})$ 中取出一个元素作为结果输出.

定义 2.　ε-差分隐私.

如果对所有的邻近数据集 $D,D'\subset\boldsymbol{D}$，以及任意的输出 $O\in Range(\boldsymbol{A})$，随机机制 $\boldsymbol{A}:D\rightarrow Range(\boldsymbol{A})$ 均满足以下条件，则称机制 $\boldsymbol{A}$ 满足 ε-差分隐私.

$$Pr[\boldsymbol{A}(D)=O]\leq e^{\varepsilon}\times Pr[\boldsymbol{A}(D')=O].$$

定义 3.　l_1-敏感度.

一个函数 $f:D\rightarrow R^k$ 的 l_1-敏感度定义如下：

$$\Delta f=\max_{D\sim D'}\|f(D)-f(D')\|_1,$$

其中，$\|\cdot\|_1$ 为 L_1 范式.

函数 f 的 l_1-敏感度表示出了单个记录在最差情况下对函数输出值大小的影响. 敏感度给出了函数在输入值为邻近数据集的情况下，函数输出值的差值上界. 敏感度是函数本身的属性，并不依赖于输入数据集.

定理 1.　Laplace 机制.

对任意给定的函数 $f:D\rightarrow R^k$，Laplace 机制 $\boldsymbol{M}_L(D,f(\cdot),\varepsilon)=f(D)+(Y_1,Y_2,\cdots,Y_k)$ 满足 ε-差分隐私，其中 Y_i 是取自于 $Lap(\Delta f/\varepsilon)$ 的独立同分布随机变量.

Laplace 机制通过在函数 f 的输出中添加数值噪声来保护输入数据集中的个体隐私. 对于输出非数值结果的函数，则可以使用指数机制来实现差分隐私. 指数机制使用效用函数 $u(D,r)$ 来评估输出结果 r 的质量，效用分数高的候选输出结果 r 有更高的概率成为最终的输出结果.

定理 2.　指数机制.

若指数机制 $\boldsymbol{M}_E(D,u(\cdot,\cdot),\mathbb{R})$ 选择并输出 $r\in\mathbb{R}$ 作为最终输出结果的概率正比于 $e^{\frac{\varepsilon u(D,r)}{2\Delta u}}$，则该指数机制满足 ε-差分隐私.

其中，Δu 是效用函数 $u(D,r)$ 的 l_1-敏感度，通过下式计算得到.

$$\Delta u=\max_{r\in\mathbb{R},D\sim D'}|u(D,r)-u(D',r)|.$$

3.3　个性化差分隐私

传统的差分隐私机制对所有的轨迹提供了相同的隐私保护水平，可能导致对某些敏感轨迹的隐私保护不足，而对其余轨迹的隐私保护过度. 为解决这一问题，本文假设在轨迹数据库中，每条轨迹都属于一个用户，不同的用户可能有不同的隐私偏好，隐私保护机制应满足所有用户的隐私需求.

定义 4.　隐私配置.

隐私配置是从用户集到个人隐私偏好的映射 $\boldsymbol{\Omega}:U\rightarrow R_+$，个人隐私偏好即差分隐私预算. 本文使用 $\boldsymbol{\Omega}(u)$ 表示用户 $u\in U$ 的隐私偏好.

与传统的差分隐私相比，本文模型中的每条轨迹都有其隐私偏好，与该条轨迹的所有者对应. 该假设的合理性在于现实世界中的个体具有不同的隐私偏好. 若不给隐私保护需求较低的个体提供过高的隐私保护级别，可以实现更好的数据可用性.

定义 5.　个性化差分隐私(PDP)

给定隐私配置 $\boldsymbol{\Omega}$，如果随机化机制 $\boldsymbol{A}:D\rightarrow Range(\boldsymbol{A})$，对所有满足 $D\oplus D'=T$ 的邻近数据集 $D,D'\subset\boldsymbol{D}$，以及所有可能的输出值 $O\in Range(\boldsymbol{A})$，均有下式成立，

$$Pr[\boldsymbol{A}(D)=O]\leq e^{\Omega(T.u)}\times Pr[\boldsymbol{A}(D')=O],$$

则机制 $\boldsymbol{A}$ 满足 $\boldsymbol{\Omega}$-个性化差分隐私($\boldsymbol{\Omega}$-PDP).

其中，$D\oplus D'$ 表示数据集 D 和 D' 之间相差的轨迹集合，$T.u$ 则为轨迹 T 所对应的用户.

显然，如果假设所有用户具有相同的全局隐私偏好，那么 $\boldsymbol{\Omega}$-PDP 提供的隐私保护级别与传统的差分隐私相同. 但是，如果我们允许用户有自己的隐私偏好，那么发布的轨迹数据集就可以提供更好的可用性.

与传统的差分隐私类似，$\boldsymbol{\Omega}$-PDP 也具有组合性质，我们对其正式定义如下.

定理 3.　个性化差分隐私的组合规则.

给定两个分别满足 $\boldsymbol{\Omega}_1$-PDP 和 $\boldsymbol{\Omega}_2$-PDP 的随机

机制 $\boldsymbol{A}_1:\boldsymbol{D}_1\rightarrow Range(\boldsymbol{A}_1)$，$\boldsymbol{A}_2:\boldsymbol{D}_2\rightarrow Range(\boldsymbol{A}_2)$，$U_1$ 和 U_2 分别表示与两个机制的隐私配置所对应的用户集. 令 $\boldsymbol{D}_3=\boldsymbol{D}_1\cup\boldsymbol{D}_2$，则对任意的轨迹数据集 $D\subset\boldsymbol{D}_3$，随机机制 $\boldsymbol{A}_3(D)=f(\boldsymbol{A}_1(D\cap\boldsymbol{D}_1),\boldsymbol{A}_2(D\cap\boldsymbol{D}_2))$ 满足 $\boldsymbol{\Omega}_3$-PDP，其中 $\boldsymbol{\Omega}_3=\{(u,\boldsymbol{\Omega}_1(u)+\boldsymbol{\Omega}_2(u))\mid u\in U_1\cap U_2\}\cup\{(u,\boldsymbol{\Omega}_1(u))\mid u\in U_1\backslash U_2\}\cup\{(u,\boldsymbol{\Omega}_2(u))\mid u\in U_2\backslash U_1\}$，$f$ 是以 $\boldsymbol{A}_1(D\cap\boldsymbol{D}_1)$ 和 $\boldsymbol{A}_2(D\cap\boldsymbol{D}_2)$ 作为输入的映射函数.

4 解决方案

轨迹数据是由若干位置和对应时间戳所组成的序列数据，我们不能直接将传统的位置隐私保护方法应用于轨迹数据. 特别的，如果轨迹所属的用户具有不同的隐私需求，那么情况就变得更加复杂. 通过将隐私偏好需求超过某个阈值的轨迹进行过滤并去除，能够实现对用户隐私的保护. 但是，这种处理方法会导致轨迹的发布数量减少，并且隐私预算阈值的设定也比较困难.

差分隐私是一种有效的保护原始轨迹数据隐私的技术. 通过在原始数据中添加差分隐私机制生成的定制化噪声，使得攻击者即使获得了足够的背景知识信息，也很难从轨迹数据集中区分出不同个体的轨迹. 另一种更为有效的方法是给每条可能轨迹的计数值添加受控的噪声，但是这种方法并未考虑到原始轨迹数据，因而可能在发布的数据中引入无意义的轨迹. Chen 等人[4]首先将差分隐私应用在轨迹数据发布中，通过维护一个前缀树结构来表示原始轨迹数据集，进而将受控噪声注入到前缀树节点，即该节点对应轨迹的计数值. 然而，由于前缀树结构要求在轨迹数据集中具有足够的公共前缀，而真实轨迹数据则很难满足这种需求. Hua 等人[12]提出了一种针对一般轨迹数据的差分隐私保护机制，使用 K 均值聚类和指数机制来泛化每个时刻的位置集合，进而在生成的轨迹数据中添加 Laplace 噪声. 该机制能够提供预先设定的隐私保证，但是该机制给所有的轨迹数据应用了相同的隐私预算，不能满足具有不同隐私保护偏好的个体需求.

为了满足个体不同的隐私保护需求，本文提出一种面向轨迹数据发布的个性化差分隐私保护机制. 该机制首先使用基于 Hilbert 曲线的位置聚类方法，根据位置的分布特征，对每个时刻的位置集合进行泛化. 然后，对每个时刻的位置聚簇进行位置抽样，进而使用指数机制来选择每个位置聚簇的代表元. 最后，对每一条原始轨迹在各个时刻的位置，使用该位置所处聚簇的代表元进行替换，从而完成原始轨迹数据的隐私保护处理. 具体的，本文提出的机制包含以下两个步骤：

(1) 基于 Hilbert 曲线的位置聚类算法. 与传统的二维数据聚类算法不同，该算法以线性索引数据作为输入，在一维空间上生成聚类. 该算法能够保持轨迹数据中位置的距离特性和分布特性. 在空间填充曲线的启发下，我们对位置集合进行区域划分，并利用 Hilbert 曲线串联空间区域，生成每个时间戳位置集合的线性索引. 通过对索引进行线性扫描，能够有效获得各个时刻的位置聚簇. 并且该算法不需要在不同时刻的位置集合上设置相同的聚簇数量，而是根据不同时刻位置集合的不同分布，生成不同数量的聚簇.

(2) 满足个性化差分隐私的轨迹数据泛化处理. 该方法考虑了个体的不同隐私偏好，以个性化差分隐私方式生成每个位置聚簇的代表元. 传统的差分隐私只能对整个数据集应用单一的隐私预算，从而导致部分用户隐私保护不足，而其他用户则隐私保护过度. 为了解决这一问题，该方法根据位置对应的隐私预算来确定其被选中的概率，对每个时刻的位置聚簇进行采样. 然后，采用指数机制来选择每个聚簇的代表元，保证轨迹泛化过程满足个性化差分隐私.

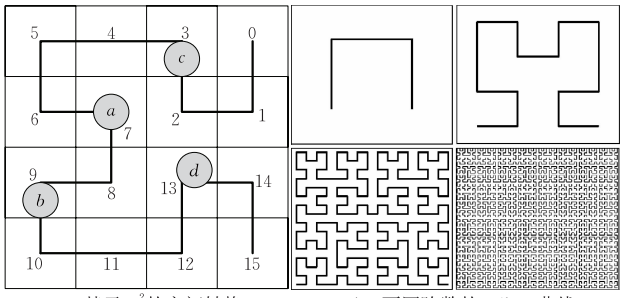
4.1 基于 Hilbert 曲线的位置聚类

基于聚类的算法能够对轨迹数据的位置信息进行泛化，然而，由于聚类算法的局限性，很难对聚类算法的参数进行合理的设置. 以 K 均值算法为例，我们很难在不同的时刻为每个位置集合设置合适的 K 值，这是因为位置集合的分布状况随时间而变化，对于不同时刻的位置集合设置相同的 K 值是不合理的. 为了解决这个问题，本文提出了一种基于 Hilbert 曲线的位置聚类方法，该方法利用了 Hilbert 曲线的距离保持特性.

与 Z 曲线和 Gray 曲线相比，Hilbert 曲线以其优越的聚类和距离保持特性得到了广泛的应用[36]. 与文献[36]类似，本文使用 H_d^N 来表示 d 维空间中的 N 阶 Hilbert 曲线，其中 $N\geq1$，且 $d\geq2$. 这样，d 维整数空间 $[0,2^N-1]^d$ 就可以映射到一维整数集 $[0,2^{Nd}-1]$ 中，即对 d 维空间中的任何位置 l，都有一个函数 f，满足 $V_{\mathrm{H}}=f(l)$，其中，$V_{\mathrm{H}}\in[0,2^{Nd}-1]$.

图 1(a)描述了一个通过 H_2^2 的 Hilbert 曲线将二维位置坐标转换为 Hilbert 值的示例. Hilbert 曲

线依次通过平面空间的每个区域，根据各个区域的访问顺序生成每个位置坐标的 Hilbert 值，该值即可用来建立位置坐标的索引. 图中位置 a、b、c、d 的 Hilbert 值分别为 7、9、3、13. 图 1(b)显示了曲线阶数为 1、2、4、6 的 Hilbert 曲线. 由图可见，曲线阶数越高，则空间分割的粒度就越细，从而可以捕获有关位置数据集分布的更多特征信息.



(a) 基于H_2^2的空间转换　　(b) 不同阶数的Hilbert曲线

图 1　Hilbert 曲线及转换

算法1首先为位置集合中的每个位置生成Hilbert索引值，然后在一维线性空间中对这些索引进行扫描，提取其中的聚簇. 利用这些信息，我们可以直接对每个时刻的位置数据进行泛化，或者为 K 均值聚类提供参数设定的参考指标. 本文不直接使用聚簇的质心作为聚簇的代表元，因为该质心可能没有语义信息，且由此进行的位置泛化过程无法满足用户不同的隐私需求. 算法基于 Hilbert 曲线的一个简单特性，即当两个位置对应的索引足够接近时，则这两个位置属于同一个聚簇. 通过比例因子即可设定两个索引值之间的距离阈值，超过该值则认为两个位置属于不同的聚簇.

算法 1.　线性索引聚类算法 LIC.

输入：位置集合 L，缩放因子 s

输出：位置聚簇 C

1. 对位置集合 L 中的每一个位置 l_i，计算其 Hilbert 索引值 h_i
2. 对索引进行增序排序，得到索引表 H
3. 初始化位置聚簇 $C \leftarrow \varnothing$
4. $start \leftarrow H[0]$
5. $j \leftarrow 1$
6. $C_j \leftarrow \{H[0]\}$
7. FOR $i \leftarrow 1$ to $|H|$
8. 　IF $H[i].index - H[i-1].index \leq s$
9. 　　$C_j \leftarrow C_j \cup H[i]$
10. 　ELSE
11. 　　$C \leftarrow C \cup C_j$
12. 　　$j{+}{+}$
13. 　　$C_j \leftarrow \{H[i]\}$
14. 　END IF
15. END FOR
16. RETURN C

图 2 显示了一个长度为 3 的示例轨迹数据集，其中，在不同时刻的位置分布是有差别的. 由于位置聚类算法被用来分别处理轨迹数据集中每个时刻的位置数据，因此，在所有的时刻均设置相同的聚类参数(如聚簇个数)不符合实际应用场景的需求.

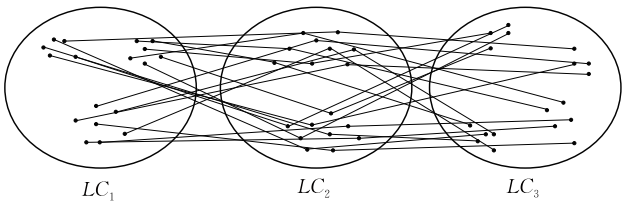


图 2　长度为 3 的轨迹数据集示例

可以发现，图 3 中的位置聚簇在不同的时刻表现出了不同的分布模式，这意味着这些位置数据集应该使用不同的聚簇数量作为聚类算法的配置参数. 例如，位置集合 LC_1、LC_2、LC_3 的聚簇数量分别为 3、2 和 4. 算法 1 根据位置数据集的分布特征，使用一个可伸缩的参数，生成不同时刻的位置聚簇，该方法比传统的聚类方法更适合于位置聚类，能够处理轨迹数据在各个时刻的不同分布状况.

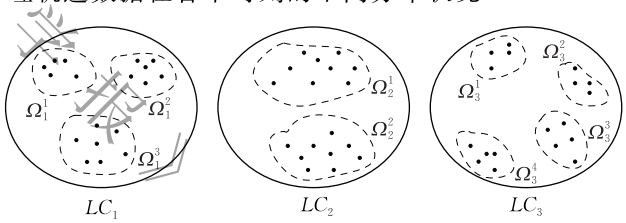


图 3　不同时刻的位置聚簇

图 4 直观地显示了轨迹数据集在不同时刻的 Hilbert 索引分布情况，可见其不同时刻的索引分布并不相同. 为不同时刻的位置数据设置相同的聚簇数量可能会改变原始数据的真实分布特征，从而导致发布的轨迹数据可用性较低. 事实上，在不同时刻采用不同的聚簇数量是可行的，可以根据不同的参

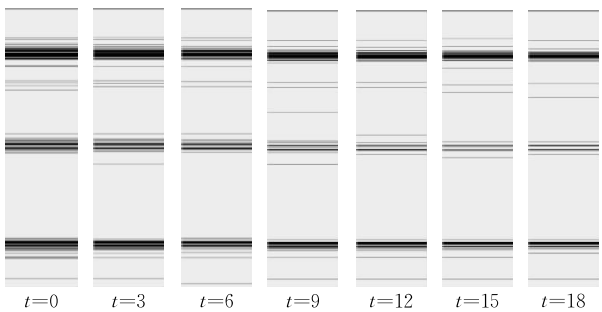


图 4　不同时刻的 Hilbert 索引分布情况

数(聚簇数量、位置分布、位置重要性等)对原始位置进行泛化,从而提高发布的轨迹数据的可用性.

4.2 满足 PDP 的位置代表元生成方法

在 LBS 应用中,位置泛化通常用来保护用户的位置隐私.该方法也可以用来压缩轨迹数据集中每个时刻的位置集合的大小.为了对轨迹数据进行泛化,需要对轨迹数据在不同时刻的位置进行泛化.算法 1 能够将邻近的位置聚集在一起,但是在满足个性化差分隐私的前提下,生成位置聚簇的代表元则比较困难.如果把聚簇的质心作为其代表元,则可能会导致代表元无语义,因为质心可能是我们无法到达的某个位置.若是此种情况,则具有该聚簇所属位置的轨迹将被视为伪轨迹,从而可以很容易地被过滤掉,无法达到隐私保护的目的.此外,在相同时刻,可能在多个聚簇中生成无语义的代表元,并且随着轨迹的延伸,导致无语义代表元出现的概率增加.

本文中假设轨迹所关联的用户具有不同的隐私偏好,这使得代表元的生成更加复杂.如图 3 所示,每个聚簇都有自己的隐私配置 $\boldsymbol{\Omega}_i^k$,即时刻 t_i 的第 k 个聚簇 LC_i^k 的隐私配置信息,其中分配给位置 $l\in LC_i^k$ 的隐私预算为 $\boldsymbol{\Omega}_i^k(l.u)$.轨迹的隐私预算可以通过任意的策略分配给该轨迹各个时刻的位置,例如均匀分配、线性分配、Fibonacci 分配等,本文在第 5 节证明了隐私预算的分配策略不会影响个性化差分隐私保证.

受抽样机制启发,本文使用轨迹的部分隐私预算来计算线性索引聚类算法生成的聚簇中各个位置的选中概率.具体地,位置 $l\in LC_i^k$ 的选中概率由以下等式确定.

$$\pi_l=\begin{cases}\dfrac{e^{\boldsymbol{\Omega}_i^k(l.u)}-1}{e^{\varphi}-1}, & \boldsymbol{\Omega}_i^k(l.u)<\varphi\\ 1, & \boldsymbol{\Omega}_i^k(l.u)\geqslant\varphi\end{cases},$$

其中,φ 表示隐私预算的阈值.由该等式可见,隐私预算较高的位置更有可能被选为抽样位置集合 $SLC_i^k=S(LC_i^k,\boldsymbol{\Omega}_i^k,\varphi)$ 的成员,该抽样位置集合 $SLC_i^k\subseteq LC_i^k$.其中,阈值 φ 的值可以在区间 $[\min\limits_{l.u}\boldsymbol{\Omega}_i^k(l.u),\max\limits_{l.u}\boldsymbol{\Omega}_i^k(l.u)]$ 内抽取.

在生成抽样位置集合 SLC_i^k 之后,本文采用指数机制来选择对应聚簇的代表元.具体地,指数机制 $\varepsilon_\varphi^s(S(LC_i^k,\boldsymbol{\Omega}_i^k,\varphi))$ 选择并输出位置 $l\in SLC_i^k$,其概率与 $e^{\frac{\varphi s(SLC_i^k,l)}{2\Delta s}}$ 成正比,其中,$s(SLC_i^k,l)=\dfrac{\boldsymbol{\Omega}_i^k(l.u)}{\max\limits_{l\in SLC_i^k}\boldsymbol{\Omega}_i^k(l.u)}$ 为效用函数,该函数将抽样位置集合/输出位置映射为效用值.

定理 4. 从轨迹数据集时刻 t_i 的第 k 个聚簇 LC_i^k 选择代表元的机制 ε_φ^s 满足个性化差分隐私 $\boldsymbol{\Omega}_i^k$-PDP.

证明. 根据个性化差分隐私的定义,需要证明,对于任意的邻近位置数据集 $L,L_{-l}\subset LC_i^k$,以及所有的输出集合 $O\in Range(\varepsilon_\varphi^s)$,满足

$$Pr[\varepsilon_\varphi^s(S(L,\boldsymbol{\Omega}_i^k,\varphi))=O]\leqslant e^{\boldsymbol{\Omega}_i^k(l.u)}\times Pr[\varepsilon_\varphi^s(S(L_{-l},\boldsymbol{\Omega}_i^k,\varphi))=O].$$

根据抽样函数 S 和指数机制 ε_φ^s 的定义,可得到以下等式:

$$\begin{aligned}&Pr[\varepsilon_\varphi^s(S(L,\boldsymbol{\Omega}_i^k,\varphi))=O]\\&=\sum_{Z\subseteq L_{-l}}(\pi_l Pr[S(L_{-l},\boldsymbol{\Omega}_i^k,\varphi)=Z])\\&\quad Pr[\varepsilon_\varphi^s(Z_{+l})=O]+\\&\quad\sum_{Z\subseteq L_{-l}}((1-\pi_l)Pr[S(L_{-l},\boldsymbol{\Omega}_i^k,\varphi)=Z])\\&Pr[\varepsilon_\varphi^s(Z)=O]\\&=\sum_{Z\subseteq L_{-l}}(\pi_l Pr[S(L_{-l},\boldsymbol{\Omega}_i^k,\varphi)=Z])\\&\quad Pr[\varepsilon_\varphi^s(Z_{+l})=O]+\\&\quad(1-\pi_l)Pr[\varepsilon_\varphi^s(S(L_{-l},\boldsymbol{\Omega}_i^k,\varphi))=O].\end{aligned}$$

考虑到指数机制 ε_φ^s 的可用性函数为 $s(L,l)=\dfrac{\boldsymbol{\Omega}_i^k(l.u)}{\max\limits_{l\in LC_i^k}\boldsymbol{\Omega}_i^k(l.u)}$,隐私预算为 φ,则根据指数机制的性质,有 $Pr[\varepsilon_\varphi^s(Z_{+l})=O]\leqslant e^{\varphi}Pr[\varepsilon_\varphi^s(Z)=O]$.

因此,等式 $Pr[\varepsilon_\varphi^s(S(L,\boldsymbol{\Omega}_i^k,\varphi))=O]$ 可重写为

$$\begin{aligned}&Pr[\varepsilon_\varphi^s(S(L,\boldsymbol{\Omega}_i^k,\varphi))=O]\\&=\sum_{Z\subseteq L_{-l}}(\pi_l Pr[S(L_{-l},\boldsymbol{\Omega}_i^k,\varphi)=Z])\\&\quad Pr[\varepsilon_\varphi^s(Z_{+l})=O]+\\&\quad(1-\pi_l)Pr[\varepsilon_\varphi^s(S(L_{-l},\boldsymbol{\Omega}_i^k,\varphi))=O]\\&\leqslant\sum_{Z\subseteq L_{-l}}(\pi_l Pr[S(L_{-l},\boldsymbol{\Omega}_i^k,\varphi)=Z])\\&\quad e^{\varphi}Pr[\varepsilon_\varphi^s(Z)=O]+\\&\quad(1-\pi_l)Pr[\varepsilon_\varphi^s(S(L_{-l},\boldsymbol{\Omega}_i^k,\varphi))=O]\\&=e^{\varphi}\pi_l Pr[\varepsilon_\varphi^s(S(L_{-l},\boldsymbol{\Omega}_i^k,\varphi))=O]+\\&\quad(1-\pi_l)Pr[\varepsilon_\varphi^s(S(L_{-l},\boldsymbol{\Omega}_i^k,\varphi))=O]\\&=(e^{\varphi}\pi_l+1-\pi_l)Pr[\varepsilon_\varphi^s(S(L_{-l},\boldsymbol{\Omega}_i^k,\varphi))=O],\end{aligned}$$

其中,位置 l 被抽样机制选中的概率为 π_l.对于上式中的 $e^{\varphi}\pi_l+1-\pi_l$,其计算需要考虑两种情况.第

一种情况，如果位置 l 所分配的隐私预算 $\boldsymbol{\Omega}_i^k(l.u)$ 大于或等于 φ，则有 $e^{\varphi}\pi_l+1-\pi_l=e^{\varphi}\leq e^{\boldsymbol{\Omega}_i^k(l,u)}$. 从而，可以得到不等式 $Pr[\varepsilon_{\varphi}^s(S(L,\boldsymbol{\Omega}_i^k,\varphi))=O]\leq e^{\boldsymbol{\Omega}_i^k(l,u)}Pr[\varepsilon_{\varphi}^s(S(L_{-l},\boldsymbol{\Omega}_i^k,\varphi))=O]$，即满足 $\boldsymbol{\Omega}_i^k$-PDP. 第二种情况，如果位置 l 所分配的隐私预算 $\boldsymbol{\Omega}_i^k(l.u)$ 小于 φ，则有 $e^{\varphi}\pi_l+1-\pi_l=e^{\boldsymbol{\Omega}_i^k(l,u)}$，也满足 $\boldsymbol{\Omega}_i^k$-PDP. 因此，在时刻 t_i 的第 k 个聚簇 LC_i^k 中选择代表元的过程满足 $\boldsymbol{\Omega}_i^k$-PDP.　　证毕.

5　隐私分析

假定原始轨迹数据集的大小为 $|D|$，且所有的轨迹数据的时间标签取自相同的时间域，每一条轨迹包含 n 个位置. 由定理 4 可知，从轨迹数据集时刻 t_i 的第 k 个聚簇 LC_i^k 选择代表元的机制 ε_{φ}^s 满足个性化差分隐私 $\boldsymbol{\Omega}_i^k$-PDP.

为了证明轨迹泛化过程满足个性化差分隐私，接下来，我们证明在时刻 t_i，选择各个聚簇代表元的过程满足 $\boldsymbol{\Omega}_i$-PDP，其中，$\boldsymbol{\Omega}_i=\cup_k\boldsymbol{\Omega}_i^k$.

由轨迹数据库的定义，在时刻 t_i 的任意两个聚簇不存在交集，即对满足 $\boldsymbol{\Omega}_i^k$-PDP 和 $\boldsymbol{\Omega}_i^j$-PDP 的聚簇 LC_i^k,LC_i^j，有 $\boldsymbol{\Omega}_i^k\cap\boldsymbol{\Omega}_i^j=\varnothing$. 那么，对任意邻近位置数据集 $L,L_{-l}\subset LC_i^k\cup LC_i^j$，分两种情况进行讨论. 如果位置 l 所属的用户 $l.u\in U_i^k\backslash U_i^j$，其中 U_i^k 即为与隐私配置 $\boldsymbol{\Omega}_i^k$ 关联的用户集合. 令机制 $\boldsymbol{A}_3(L)=f(\varepsilon_{\varphi_1}^s(S(L\cap LC_i^k,\boldsymbol{\Omega}_i^k,\varphi_1)),\varepsilon_{\varphi_2}^s(S(L\cap LC_i^j,\boldsymbol{\Omega}_i^j,\varphi_2)))$，其中 f 是任意的以指数机制 $\varepsilon_{\varphi_1}^s$ 和 $\varepsilon_{\varphi_2}^s$ 的输出作为其输入的方法. 对所有的集合 $O\in Range(\boldsymbol{A}_3)$，有

$$\begin{aligned}&Pr[\boldsymbol{A}_3(L)=O]\\&=\sum_{f(O_1,O_2)=O}Pr[\varepsilon_{\varphi_1}^s(S(L\cap LC_i^k,\boldsymbol{\Omega}_i^k,\varphi_1))=O_1]\\&\quad Pr[\varepsilon_{\varphi_2}^s(S(L\cap LC_i^j,\boldsymbol{\Omega}_i^j,\varphi_2))=O_2].\end{aligned}$$

由 $l.u\in U_i^k\backslash U_i^j$，可得 $U(L)\cap U_i^j=U(L_{-l})\cap U_i^j$. 因此，上式可以改写如下：

$$\begin{aligned}&Pr[\boldsymbol{A}_3(L)=O]\\&\leq\sum_{f(O_1,O_2)=O}(e^{\boldsymbol{\Omega}_i^k(l.u)}Pr[\varepsilon_{\varphi_1}^s(S(L_{-l}\cap LC_i^k,\boldsymbol{\Omega}_i^k,\varphi_1))=\\&\quad O_1])Pr[\varepsilon_{\varphi_2}^s(S(L_{-l}\cap LC_i^j,\boldsymbol{\Omega}_i^j,\varphi_2))=O_2]\\&=e^{\boldsymbol{\Omega}_i^k(l,u)}Pr[\boldsymbol{A}_3(L_{-l})=O].\end{aligned}$$

对于第二种情况，即 $l.u\in U_i^j\backslash U_i^k$，也可用上述方法进行证明. 因此，对于在时刻 t_i 的任意两个聚簇 LC_i^k,LC_i^j，机制 $\boldsymbol{A}_3(L)$ 满足 $\boldsymbol{\Omega}_{\boldsymbol{A}_3}$-PDP，其中 $\boldsymbol{\Omega}_{\boldsymbol{A}_3}=\{(u,\boldsymbol{\Omega}_i^k(u))\mid u\in U_i^k\backslash U_i^j\}\cup\{(u,\boldsymbol{\Omega}_i^j(u))\mid u\in U_i^j\backslash U_i^k\}$. 由于任意的用户集合均无交集，因此，可以将隐私配置 $\boldsymbol{\Omega}_{\boldsymbol{A}_3}$ 简化为 $\boldsymbol{\Omega}_{\boldsymbol{A}_3}=\{(u,\boldsymbol{\Omega}_i^k(u))\mid u\in U_i^k\}\cup\{(u,\boldsymbol{\Omega}_i^j(u))\mid u\in U_i^j\}$. 假设在时刻 t_i 共有 m 个位置聚簇，我们通过重复上述的组合过程，即可得到 $\boldsymbol{\Omega}_{\boldsymbol{A}_3},\boldsymbol{\Omega}_{\boldsymbol{A}_4},\cdots,\boldsymbol{\Omega}_{\boldsymbol{A}_{m+1}}$，其中，$\boldsymbol{\Omega}_{\boldsymbol{A}_{m+1}}=\{(u,\boldsymbol{\Omega}_i^1(u))\mid u\in U_i^1\}\cup\{(u,\boldsymbol{\Omega}_i^2(u))\mid u\in U_i^2\}\cup\cdots\cup\{(u,\boldsymbol{\Omega}_i^m(u))\mid u\in U_i^m\}$. 类似地，可以证明，机制 $\boldsymbol{A}_{m+1}$ 满足 $\boldsymbol{\Omega}_{\boldsymbol{A}_{m+1}}$-PDP. 因此，在时刻 t_i 选择代表元的过程满足 $\boldsymbol{\Omega}_{\boldsymbol{A}_{m+1}}$-PDP. 为简化叙述，我们使用 LC_i 和 $\boldsymbol{\Omega}_i$ 分别表示时刻 t_i 的位置集合与用户的隐私配置，即该过程满足 $\boldsymbol{\Omega}_i$-PDP.

由于一条轨迹包含 n 个位置以及与其关联的时间戳，接下来我们还需要证明，对于轨迹的泛化过程满足个性化差分隐私. 我们使用 $\boldsymbol{B}_i$ 表示作用在位置集合 LC_i 上的机制，由前面的分析可知，该机制满足 $\boldsymbol{\Omega}_i$-PDP. 对任意的邻近轨迹数据集 $D,D_{-T}\subset\cup_iLC_i$，令机制 $\boldsymbol{G}(D)=g(\boldsymbol{B}_1(D\cap LC_1),\boldsymbol{B}_2(D\cap LC_2),\cdots,\boldsymbol{B}_n(D\cap LC_n))$，其中，$g$ 是任意的以机制 $\boldsymbol{B}_i$ 的输出作为其输入的方法. 则对于所有的集合 $O\in Range(\boldsymbol{G})$，有

$$\begin{aligned}Pr[\boldsymbol{G}(D)=O]=&\sum_{g(O_1,O_2,\cdots,O_n)=O}Pr[\boldsymbol{B}_1(D\cap LC_1)=O_1]\cdots\\&Pr[\boldsymbol{B}_n(D\cap LC_n)=O_n].\end{aligned}$$

由 $\boldsymbol{B}_i$ 满足 $\boldsymbol{\Omega}_i$-PDP，可得 $Pr[\boldsymbol{B}_i(D\cap LC_i)=O_i]\leq e^{\boldsymbol{\Omega}_i(T,u)}Pr[\boldsymbol{B}_i(D_{-T}\cap LC_i)=O_i]$，因此，$Pr[\boldsymbol{G}(D)=O]$ 可以重写为

$$\begin{aligned}&Pr[\boldsymbol{G}(D)=O]\\&\leq\sum_{g(O_1,O_2,\cdots,O_n)=O}e^{\boldsymbol{\Omega}_1(T,u)}Pr[\boldsymbol{B}_1(D_{-T}\cap LC_1)=O_1]\cdots\\&\quad e^{\boldsymbol{\Omega}_n(T,u)}Pr[\boldsymbol{B}_n(D_{-T}\cap LC_n)=O_n]\\&=e^{\boldsymbol{\Omega}_1(T,u)+\cdots+\boldsymbol{\Omega}_n(T,u)}\sum_{g(O_1,O_2,\cdots,O_n)=O}Pr[\boldsymbol{B}_1(D_{-T}\cap LC_1)=O_1]\cdots\\&\quad Pr[\boldsymbol{B}_n(D_{-T}\cap LC_n)=O_n]\\&=e^{\sum_{i=1}^n\boldsymbol{\Omega}_i(T,u)}Pr[\boldsymbol{G}(D_{-T})=O].\end{aligned}$$

因此，机制 $\boldsymbol{G}$ 满足 $\boldsymbol{\Omega}$-PDP，其中 $\boldsymbol{\Omega}=\left\{\left(u,\sum_{i=1}^n\boldsymbol{\Omega}_i(u)\right)\mid u\in\cap_iU_i\right\}$，$U_i$ 表示隐私配置 $\boldsymbol{\Omega}_i$ 所对应的用户集合. 注意到，所有时刻的用户集合都是完全相同的，即 $U_1=U_2=\cdots=U_n$. 从而可以将 $\boldsymbol{\Omega}$ 简化为 $\boldsymbol{\Omega}=\left\{\left(u,\sum_{i=1}^n\boldsymbol{\Omega}_i(u)\right)\mid u\in U_1\right\}$，需要指出的是，用

户的隐私预算可以采用任意的策略分配给不同的位置.

6 实验评估

在本节中,通过实验比较了本文方法与基于标准差分隐私的轨迹发布方法在数据可用性与性能方面的表现.实验在 Intel Xeon E3-1505M 2.8GHz 处理器,16GB 内存计算机上完成.

6.1 实验设置与数据集

实验中使用的真实轨迹数据集来自于 T-Drive,该数据集包含 2008 年 2 月 2 日至 2 月 8 日期间在北京的 10357 辆出租车的 GPS 轨迹[37].轨迹数据集中的每条记录由出租车 ID、时间戳和当前位置(经度、纬度)组成.由于轨迹的时间戳并不统一,我们不能直接使用这个数据集进行实验.因此,我们从 8:30 到 14:30 的时间段中提取轨迹记录,发现有意义的轨迹包含的位置数量在 20 到 37 之间变化.为了获得尽可能多的轨迹,我们从每条轨迹中提取 20 个位置.此外,每条轨迹上相邻位置之间的时间间隔不少于 10 min.经过该处理过程,共获得 6225 条轨迹.我们使用墨卡托投影将各个位置的经纬度信息转换成平面坐标,便于直观显示.图 5 显示了原始轨迹数据,不同灰度的颜色表示不同的轨迹,图 6 则显示了用于发布的泛化轨迹数据.

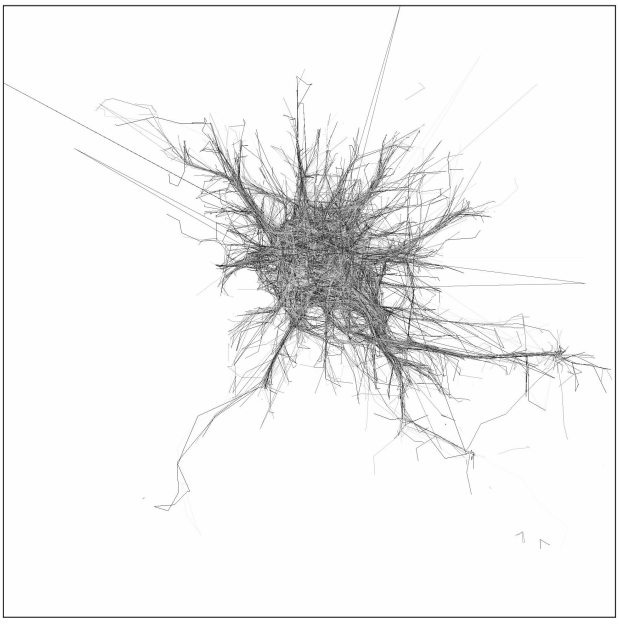


图 5 原始 T-Drive 轨迹数据

本文假定用户具有不同的隐私偏好,用户(即轨迹的所有者)被随机分成 3 组,包括自由派(具有较低的隐私需求)、保守派(具有较高的隐私需求)、和中间派(具有适中的隐私需求).参考文献[13]关于不同类别用户比例的设定,本文用参数 f_M 和 f_C 分别表示中间派用户和保守派用户在所有用户中所占的比例,则自由派用户所占的比例即为 $1-(f_M+f_C)$.基于对用户隐私偏好的调研[38],本文设定 $f_M=0.37$,$f_C=0.54$,并设定对应的隐私预算 $\varepsilon_M=0.2$,$\varepsilon_C=0.01$,$\varepsilon_L=1$.中间派和保守派用户的隐私预算分别在区间$[\varepsilon_M,\varepsilon_L]$、$[\varepsilon_C,\varepsilon_M]$中随机抽取,自由派用户的隐私预算则固定为 ε_L.抽样机制的隐私预算阈值 φ 设定为$\frac{1}{|LC_i^k|}\sum_{l\in LC_i^k}\boldsymbol{\Omega}(l.u)$,即取各个聚簇隐私预算的平均值作为阈值.

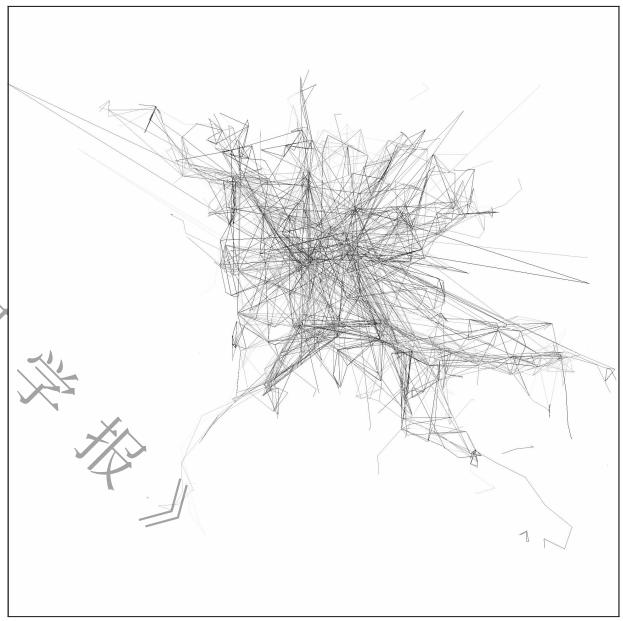


图 6 泛化后的轨迹数据

6.2 可用性评估

本文假定发布的轨迹数据集 D' 包含的轨迹数量与原始轨迹数据集 D 相同,使用轨迹平均距离 $AvgTrjDist(D',D)$ 来评估发布轨迹的质量.

$$AvgTrjDist(D',D)=\frac{1}{n|D|}\sum_{T_i'\in D',T_i\in D}d(T_i',T_i),$$

其中,$d(T_i',T_i)$ 表示泛化轨迹 T_i' 和原始轨迹 T_i 之间的距离,通过计算两条轨迹每个时刻位置之间的欧式距离得到,较小的轨迹平均距离值可以表示更高的数据可用性.

本文比较了基于统一隐私预算的方法 UDP(Uniform Differential Privacy)和本文方法 SPDP(Sample based Personalized Differential Privacy)在真实轨迹数据集上的可用性.UDP 使用的隐私预算

从 0.4 至 1.2 取值. 由于 UDP 和 SPDP 使用不同的方法和参数对每个时刻的位置数据进行聚类，首先，我们研究了聚簇数量和缩放因子之间的关系，其中缩放因子是 SPDP 方法的输入参数，它们之间的关系如图 7 所示，聚簇数量随着缩放因子的增加而逐渐降低. 而且，当 Hilbert 曲线阶数较高时，则需要使用更大的缩放因子，从而获得与 Hilbert 曲线阶数较低的聚类方法所类似的聚簇数量. 这是由于当我们使用较高的曲线阶数时，位置数据所对应的索引范围会呈指数级别的增长，且索引之间的分布间距也会急剧变大，需要我们采用更大的缩放因子来平衡这种差异. 在实验中，为了更精确地刻画位置数据的分布状态，我们设置 Hilbert 曲线阶数为 12.

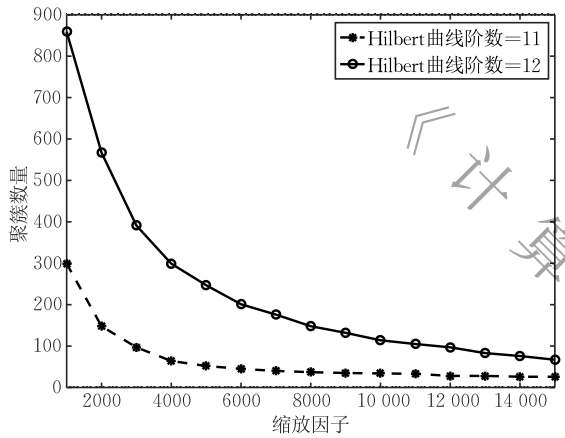


图 7　缩放因子与聚簇数量的关系

图 8 显示了不同方法所生成的发布轨迹数据集与原始轨迹数据集的轨迹平均距离，由图可见，该距离随着聚簇数量以及隐私预算的变化而改变. 直观上说，在一个轨迹数据集中的任何时刻，都应当存在一个最优的聚簇数量，并且其生成的发布数据集的平均轨迹距离应当随着聚簇数量接近这个最优值而

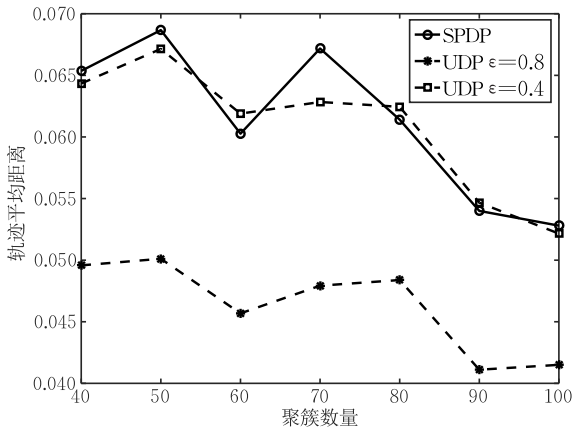


图 8　发布轨迹数据集的轨迹平均距离

不断下降. 我们发现，SPDP 所生成的泛化轨迹数据集的可用性与隐私预算为 0.4 的 UDP 方法接近. 由于在 SPDP 中，用户具有不同的隐私需求，并且保守派用户占了 54% 的比例，会降低保守派用户数据被选为位置代表元的概率，因此 SPDP 生成轨迹数据集的可用性要略低于 UDP. 然而，SPDP 能够给用户提供个性化的隐私保护，使得隐私预算较低用户的数据也能够以受控的概率构成发布的轨迹数据集，这是 UDP 所不能提供的.

图 9 显示了在使用不同的聚簇数量 K 时，对轨迹数据集第一个时刻的位置进行聚类的结果. 由该图可见，随着聚簇数量的增加，位置集合的划分粒度也越细，可以推测，生成的聚簇也会更加接近位置集合的真实分布. 通过对不同时刻的聚簇结果进行分析，发现不同时刻的位置分布有所差异，因此，对轨迹数据集进行位置泛化，需要考虑其不同时刻的分布差异，选择适应于当前时刻的划分参数. 在给定 Hilbert 曲线阶数的情况下，原始的位置坐标被转换为一维索引，索引分布的疏密即可表示位置分布的聚簇状态，从而能够通过使用相同的缩放因子在不同时刻进行线性索引聚类，并在不同时刻生成不同数量的聚簇，降低生成聚簇的平均距离.

为进一步评估发布轨迹数据集 D' 的可用性，本文通过计数查询 $Q(D')=|\{T|T\in D',t\subseteq T\}|$ 来测试生成的泛化轨迹数据集的效用，即给定待查询轨迹 t，计算轨迹数据集 D' 中有多少条轨迹包含轨迹 t. 本文设定查询轨迹长度分别为 4、8、12、16、20，分别随机生成 5000 个计数查询，每个计数查询的位置则随机取自该时刻的原始位置集合，重复运行 20 次取相对误差的平均值作为实验结果. 计数查询 Q 的相对误差由 $\frac{|Q(D')-Q(D)|}{\max\{Q(D),s\}}$ 计算，其中，s 是用于防止查询 $Q(D)$ 的结果集轨迹数量过小而设置的下界，在实验中，其取值设定为数据集中轨迹数量的 0.1%.

计数查询的平均相对误差在不同查询长度和聚簇数量下的变化如图 10 所示. 平均相对误差随着查询长度和聚簇数量的增加而呈下降趋势，这是由于较长的查询轨迹产生的查询结果误差也较小，且聚簇数量的增加能够降低位置泛化的误差. SPDP 的平均相对误差与隐私预算为 0.8 的 UDP 相比，在查询长度大于 12 时非常接近，说明 SPDP 能够在保证保守用户隐私需求的同时，提供较高数据可用性.

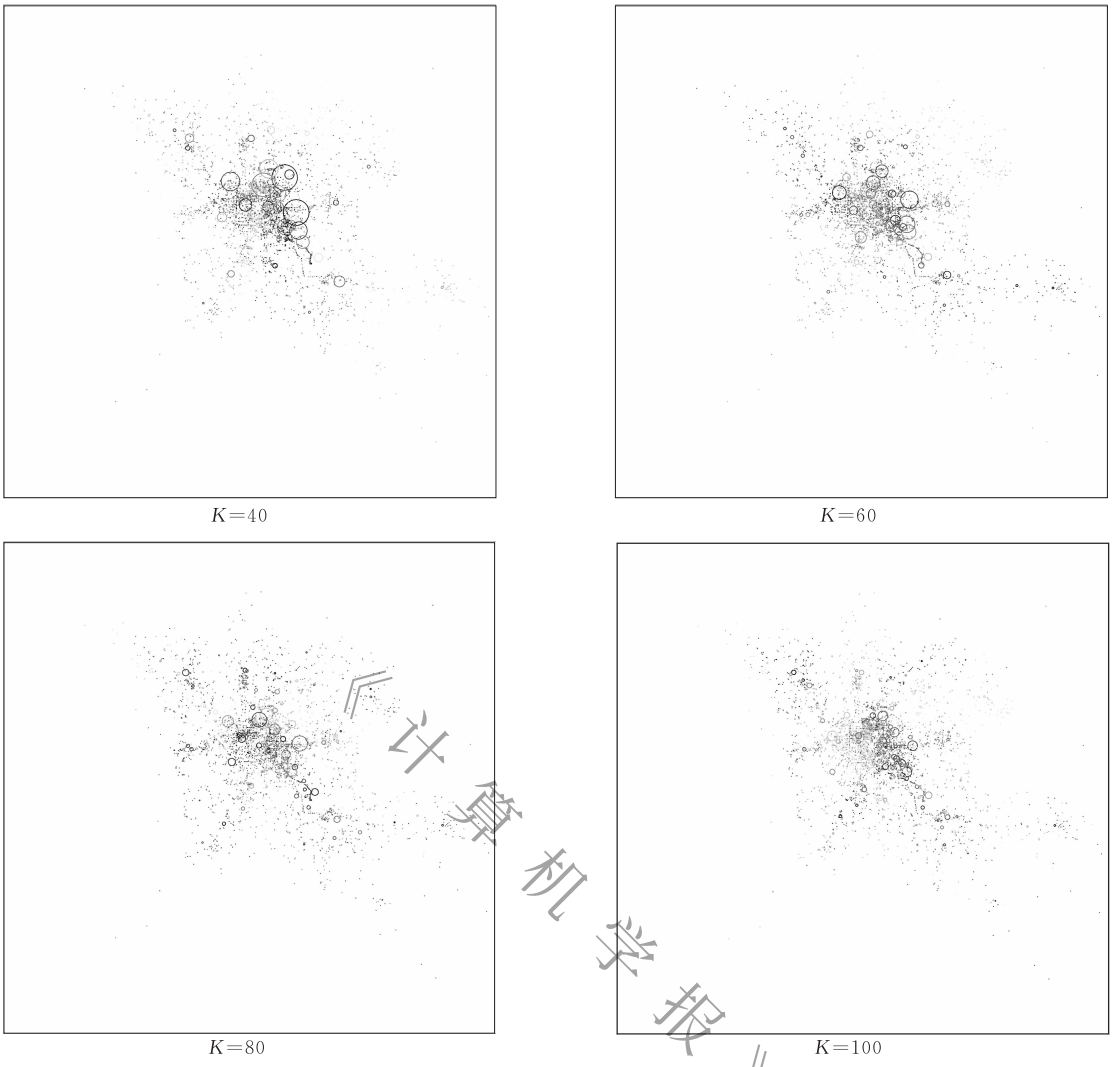


图 9 使用不同聚簇数量所生成的位置聚簇

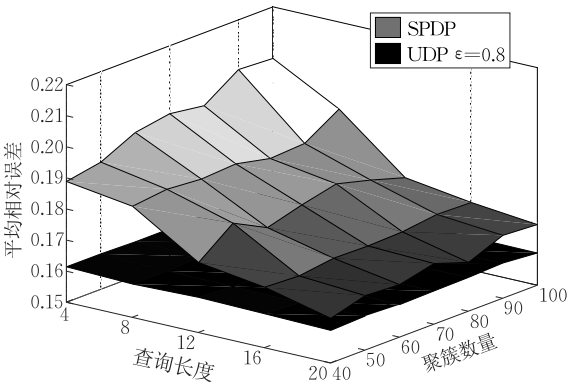


图 10 发布轨迹数据集的计数查询平均相对误差

UDP 和 SPDP 均需要对每个时刻的位置集合进行泛化，并且后续处理过程都基于位置集合泛化的结果. 为进一步研究不同方法生成位置聚簇的可用性，我们使用聚簇平均距离 $MeanDist(t_i)$ 来度量 UDP 和 SPDP 在时刻 t_i 的泛化效果.

$$MeanDist(t_i)=\frac{1}{mn|LC_i^k|}\sum_{k=1}^{m}\sum_{T_j^i\in LC_i^k}d(T_j,\widetilde{T_k}),$$

其中，$\widetilde{T_k}(i=1,2,\cdots,m)$ 表示时刻 t_i 的第 k 个聚簇的平均轨迹.

图 11 显示了线性索引聚类方法 LIC 和 K 均值

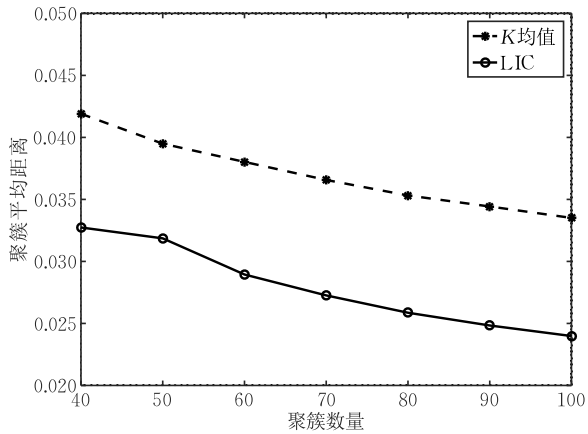


图 11 聚类算法所生成聚簇的平均距离

聚类方法在不同时刻所生成聚簇的平均距离均值.线性聚簇算法 LIC 的聚簇平均距离均值要低于 K 均值算法,并且该均值随着聚簇数量的增加而降低.由聚簇平均距离的定义可知,该值会随着聚簇数量的增加而不断减小.聚簇数量的增加意味着被放入错误聚簇的位置会减少,特别地,如果聚簇数量接近轨迹数据集的规模,会使得每个聚簇仅包含一个位置,从而导致在聚簇平均距离中几乎不会引入误差,此时聚簇的平均距离会趋近于 0.

我们还研究了聚类算法在不同时刻所生成位置聚簇的平均距离,如图 12 所示.在大部分时刻,LIC 算法所生成聚簇的平均距离均显著低于 K 均值算法,其值与 K 均值算法相比,要低 28.4%.说明 LIC 算法具有较好的聚类和距离保持特性,更适合于空间聚类应用.

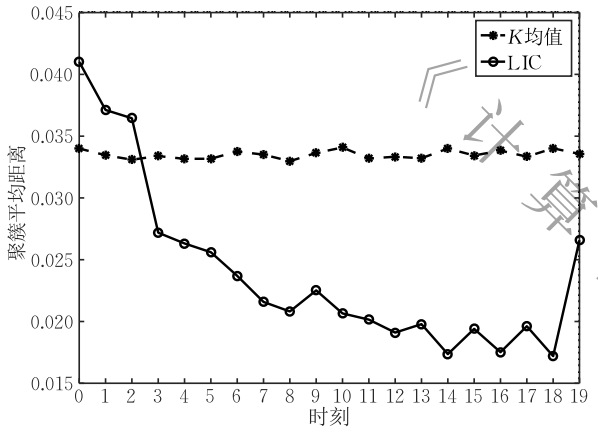


图 12　聚类算法在不同时刻所生成聚簇的平均距离

6.3　性能评估

UDP 在位置泛化过程中使用了 K 均值算法,该算法消耗了轨迹数据生成过程中的大部分计算资源.因此,在性能评估方面,我们对线性聚类算法 LIC 和 K 均值算法在不同聚簇数量时的处理时间进行比较.图 13 显示了 LIC 和 K 均值算法对真实数据集所有时刻的位置进行聚类的平均处理时间.由图可见,K 均值算法消耗了更多的时间来生成最终的位置聚簇,这是由于 K 均值算法需要一定的迭代次数来收敛到设定的中心偏差值,LIC 算法则是通过单向线性扫描来实现对一维索引值的聚类,不需要进行多次迭代.而且,与 K 均值算法相比,LIC 算法在选择不同的缩放因子时,其处理时间差别不大,这使得 LIC 算法的聚类参数选择过程开销较低,适用于规模较大的位置数据集.

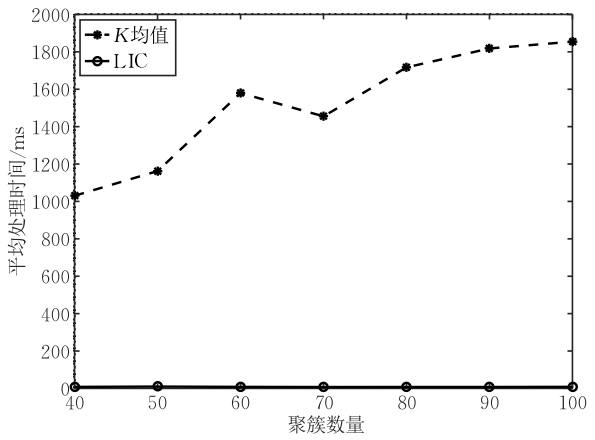


图 13　聚类算法的平均处理时间

7　总　结

本文将轨迹数据发布中,用户所具有的不同隐私需求定义为隐私配置,基于此提出了一种满足个性化差分隐私需求的轨迹数据发布方法.本文对轨迹数据集中不同时刻的位置数据,使用基于 Hilbert 索引的线性扫描聚类算法进行处理,该算法不需要对每个时刻设定相同的聚簇数量,更符合真实轨迹数据在不同时刻的分布特征.在此基础上,利用抽样机制对各个原始聚簇进行抽样,获得抽样位置聚簇,进而使用指数机制选择各个聚簇的位置代表元.最后通过位置代表元对原始轨迹数据进行泛化处理,生成用于发布的轨迹数据集.通过理论分析,证明了该机制能够满足隐私配置所定义的个性化差分隐私.在真实轨迹数据集上的实验表明,与标准差分隐私机制相比,该方案在隐私保护和数据可用性之间取得了很好的平衡,而且该方案所生成的位置代表元由原始位置集合中产生,不会导致无语义的代表元生成,从而保证了泛化轨迹能够抵抗过滤攻击.在未来的工作中,我们将研究处理动态流式轨迹数据的算法,适应未来轨迹数据处理的需求.

参　考　文　献

[1] Wu Yun-Cheng, Chen Hong, Zhao Su-Yun, et al. Differential private trajectory protection based on spatial and temporal correlation. Chinese Journal of Computers, 2018, 41(2): 309-322(in Chinese)
(吴云乘,陈红,赵素云等.一种基于时空相关性的差分隐私轨迹保护机制.计算机学报,2018,41(2):309-322)

[2] Ji Ya-Li, Gui Xiao-Lin, Dai Hui-Jun, Peng Zhen-Long. Construction users' interest regions with two steps for trajectory privacy protection. Chinese Journal of Computers, 2017, 40(12): 2734-3747(in Chinese)
(冀亚丽,桂小林,戴慧珺,彭振龙.支持轨迹隐私保护的两阶段用户兴趣区构建方法.计算机学报,2017,40(12):2734-3747)

[3] Meng Xiao-Feng, Zhang Xiao-Jian. Big data privacy management. Journal of Computer Research and Development, 2015, 52(2): 265-281(in Chinese)
(孟小峰, 张啸剑. 大数据隐私管理. 计算机研究与发展, 2015, 52(2): 265-281)

[4] Chen R, Fung B C M, Desai B C. Differentially private trajectory data publication. CoRR, abs/1112.2020, 2011

[5] Ganta S R, Kasiviswanathan S P, Smith A. Composition attacks and auxiliary information in data privacy//Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Las Vegas, USA, 2008: 265-273

[6] Sweeney L. K-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002, 10(5): 557-570

[7] Terrovitis M, Mamoulis N. Privacy preservation in the publication of trajectories//Proceedings of the 9th International Conference on Mobile Data Management. Beijing, China, 2008: 65-72

[8] Wong R C W, Fu A W C, Wang K, et al. Can the utility of anonymized data be used for privacy breaches? ACM Transactions on Knowledge Discovery from Data, 2011, 5(3), article no. 16

[9] Kifer D. Attacks on privacy and deFinetti's theorem// Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data. Providence, USA, 2009: 127-138

[10] Dwork C, McSherry F, Nissim K, Smith A. Calibrating noise to sensitivity in private data analysis//Proceedings of the 3rd Theory of Cryptography Conference. New York, USA, 2006: 265-284

[11] Chen R, Acs G, Castelluccia C. Differentially private sequential data publication via variable-length n-grams// Proceedings of the ACM Conference on Computer & Communications Security. New York, USA, 2012: 638-649

[12] Hua J Y, Gao Y, Zhong S. Differentially private publication of general time-serial trajectory data//Proceedings of the 2015 IEEE Conference on Computer Communications (INFOCOM). Hong Kong, China, 2015: 549-557

[13] Jorgensen Z, Yu T, Cormode G. Conservative or liberal? personalized differential privacy//Proceedings of the IEEE 31st International Conference on Data Engineering (ICDE). Seoul, South Korea, 2015: 1023-1034

[14] Tian F, Zhang S Y, Lu L F, et al. A novel personalized differential privacy mechanism for trajectory data publication// Proceedings of the International Conference on Networking and Network Applications (NaNA). Kathmandu, Nepal, 2017: 61-68

[15] Abul O, Bonchi F, Nanni M. Never walk alone: Uncertainty for anonymity in moving objects databases//Proceedings of the IEEE 31st International Conference on Data Engineering. Cancun, Mexico, 2008: 376-385

[16] Domingo-Ferrer J, Trujillo-Rasua R. Micro-aggregation and permutation-based anonymization of movement data. Journal of Information Sciences, 2012, 208: 55-80

[17] Trujillo-Rasua R, Domingo-Ferrer J. On the privacy offered by (k,δ)-anonymity. Information Systems, 2013, 38(4): 491-494

[18] Yarovoy R, Bonchi F, Lakshmanan L, Wang W. Anonymizing moving objects: How to hide a mob in a crowd?//Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology. Saint Petersburg, Russia, 2009: 72-83

[19] Palanisamy B, Liu L. Attack-resilient mix-zones over road net-works: Architecture and algorithms. IEEE Transactions on Mobile Computing, 2015, 14(3): 495-508

[20] Ying B, Makrakis D, Mouftah H. Dynamic mix-zone for location privacy in vehicular networks. IEEE Communications Letters, 2013, 17(8): 1524-1527

[21] Poulis G, Skiadopoulos S, Loukides G, Gkoulalas-Divanis A. Apriori-based algorithms for k^m-anonymizing trajectory data. Transactions on Data Privacy, 2014, 7(2): 165-194

[22] Monreale A, Trasarti R, Pedreschi D, et al. C-safety: A framework for the anonymization of semantic trajectories. Transactions on Data Privacy, 2011, 4(2): 73-101

[23] Cicek A, Nergiz M, Saygin Y. Ensuring location diversity in privacy-preserving spatio-temporal data publishing. The VLDB Journal, 2014, 23(4): 609-625

[24] Bonomi L, Xiong L. A two-phase algorithm for mining sequential patterns with differential privacy//Proceedings of the 22nd ACM International Conference on Information and Knowledge Management. San Francisco, USA, 2013: 269-278

[25] Andrés M, Bordenabe N, Chatzikokolakis K, Palamidessi C. Geo-indistinguishability: Differential privacy for location-based systems//Proceedings of the 2013 ACM SIGSAC Conference on Computer and Communications Security. Berlin, Germany, 2013: 901-914

[26] Zhang J, Ghinita G, Chow C. Differentially private location recommendations in geosocial networks//Proceedings of the IEEE 15th International Conference on Mobile Data Management. Brisbane, Australia, 2014: 59-68

[27] Cormode G, Procopiuc M, Shen E, et al. Differentially private spatial decompositions//Proceedings of the IEEE 28th International Conference on Data Engineering. Washington, USA, 2012: 20-31

[28] Alaggan M, Gambs S, Kermarrec A. Heterogeneous differential privacy. Journal of Privacy and Confidentiality, 2017, 7(6): 127-158

[29] Cao Y, Yoshikawa M. Differentially private real-time data publishing over infinite trajectory streams. IEICE Transactions on Information and Systems, 2016, 99(1): 163-175

[30] Cao Y, Yoshikawa M, Xiao Y H, Xiong L. Quantifying differential privacy in continuous data release under temporal correlations. IEEE Transactions on Knowledge and Data Engineering, 2019, 31(7): 1281-1295

[31] Wang Z B, Hu J H, Lv R Z, et al. Personalized privacy-preserving task allocation for mobile crowdsensing. IEEE Transactions on Mobile Computing, 2019, 18(6): 1330-1341

[32] Xu G W, Li H W, Wang W L, et al. Towards practical personalized recommendation with multi-level differential privacy controls//Proceedings of the IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS). Honolulu, USA, 2018: 796-801

[33] Zhang Y P, Qu Y Y, Gao L X, et al. APDP: Attack-proof personalized differential privacy model for a smart home. IEEE Access, 2019, 7: 166593-166605

[34] Qu Y Y, Cui L, Yu S, et al. Improving data utility through game theory in personalized differential privacy//Proceedings of the IEEE International Conference on Communications (ICC). Kansas City, USA, 2018: 1-6

[35] Li Y, Liu S, Li D, Wang J. Release connection fingerprints in social networks using personalized differential privacy. Chinese Journal of Electronics, 2018, 27(5): 1104-1110

[36] Moon B, Jagadish H, Faloutsos C, Saltz J. Analysis of the clustering properties of the Hilbert space-filling curve. IEEE Transactions on Knowledge and Data Engineering, 2001, 13(1): 124-141

[37] Yuan J, Zheng Y, Xie X, Sun G. Driving with knowledge from the physical world//Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA, 2011: 316-324

[38] Acquisti A, Grossklags J. Privacy and rationality in individual decision making. IEEE Security and Privacy, 2005, 3(1): 26-33

TIAN Feng, Ph. D., associate professor, M. S. supervisor. His research interests include privacy preserving of spatial data outsourcing, differential privacy, and network security.

WU Zhen-Qiang, Ph. D., professor, Ph. D. supervisor. His research interests include privacy preserving, network science.

LU Lai-Feng, Ph. D., associate professor. Her research interests include differential privacy and data security.

LIU Hai, Ph. D., assistant researcher. His research interests include differential privacy and gene privacy preserving.

GUI Xiao-Lin, Ph. D., professor, Ph. D. supervisor. His research interests include cloud computing, network security and Internet of Things.

Background

With the development of smart city and increasing users' privacy demand, in order to provide better location-based services, more attention have been paid on analyzing people's trajectory data. Although it helps a lot, there still exists threatens to individuals while directly publishing the original trajectory data. Differential privacy has been applied in the trajectory data publication because of its powerful framework for providing formal and strong privacy guarantees. The existing approaches treated each individual equally, regarding individuals have the same privacy preference, and the same level of privacy protection has been provided for all individuals. By doing so, insufficient privacy guarantees would be provided for some individuals, while the other individuals who need little protections will receive excess privacy protection.

To solve these problems, this paper assumes that each individual requires different level of privacy and proposes a personalized differential privacy publication mechanism for trajectory data. Hilbert curve has been applied to extract the distribution characteristics of the trajectory data at each time and a personalized differential privacy generalization algorithm for trajectories with different privacy preferences has been proposed. Personalized differential privacy protects the users' various privacy preferences, and achieves better data utility than conventional differential privacy schemes.

This research is supported in part by the National Natural Science Foundation of China under Grant No. 61602290, No. 61902229, No. 61672334, No. 61802242, and No. 61802241, the Natural Science Basic Research Plan in Shaanxi Province of China under Grant No. 2017JQ6038 and No. 2020JM-288, the Foundation of Guizhou Provincial Key Laboratory of Public Big Data under Grant No. 2018BDKFJJ004, and the Fundamental Research Funds for the Central Universities under Grant No. GK202103090 and No. GK202103084. These projects aim to provide secure spatial data outsourcing and publication. Our group has focused on differential privacy and privacy-preserving data publication for many years, we proposed schemes that support secure spatial data retrieving, lots of papers have been published, some are indexed by SCI or EI, and several patents have been applied.