doi:10.11959/j.issn.1000-436x.2017119

CLM：面向轨迹发布的差分隐私保护方法

王豪[1,2]，徐正全[1,2]，熊礼治[3]，王涛[1,2]

(1. 武汉大学测绘遥感信息工程国家重点实验室，湖北 武汉 430079；
2. 武汉大学地球空间信息技术协同创新中心，湖北 武汉 430079；
3. 南京信息工程大学计算机与软件学院，江苏 南京 210044)

摘 要：针对现有轨迹差分隐私保护发布方法面临的独立噪声容易被滤除的问题，提出一种轨迹差分隐私发布方法——CLM。CLM 提出一种相关性拉普拉斯机制，利用高斯噪声通过特定的滤波器，产生与原始轨迹序列自相关函数一致的相关性噪声序列，叠加到原始轨迹中并发布。实验结果表明，与现有的轨迹差分隐私保护发布方法相比，CLM 能够达到更高的隐私保护强度并能保证较好的数据可用性。

关键词：轨迹发布；隐私保护；差分隐私；相关性拉普拉斯
中图分类号：TP309.2 **文献标识码**：A

CLM: differential privacy protection method for trajectory publishing

WANG Hao[1,2], XU Zheng-quan[1,2], XIONG Li-zhi[3], WANG Tao[1,2]

(1. State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China;
2. Collaborative Innovation Center for Geospatial Technology, Wuhan University, Wuhan 430079, China;
3. School of Computer and Software, Nanjing University of Information Science and Technology, Nanjing 210044, China)

Abstract: In order to solve the problem existing in differential privacy preserving publishing methods that the independent noise was easy to be filtered out, a differential privacy publishing method for trajectory data (CLM), was proposed. A correlated Laplace mechanism was presented by CLM, which let Gauss noises pass through a specific filter to produce noise whose auto-correlation function was similar with original trajectory series. Then the correlated noise was added to the original track and the perturbed track was released. The experimental results show that the proposed method can achieve higher privacy protection and guarantee better data utility compared with existing differential privacy preserving publishing methods for trajectory data.

Key words: trajectory publishing, privacy preserving, differential privacy, correlated Laplace

1 引言

随着社会信息化的发展和移动终端设备的日益普及，产生了大量的轨迹数据，对于信息咨询组织、商业机构以及政府决策部门来说，为了对轨迹数据进行分析挖掘以获得有价值的信息，要求用户将采集到的轨迹数据进行发布和共享[1~3]。但轨迹数据中可能包含用户的敏感信息，用户出于对隐私泄露的担忧不愿发布自身的轨迹数据[4~6]，因此，如何在保护个人位置信息不被泄露的同时，使经过隐私保护处理后发布的轨迹数据仍然能够支持轨迹挖掘类应用，如轨迹频繁、聚集和伴随等模式挖掘，是轨迹发布亟待解决的问题[7]。

当前的轨迹隐私保护发布方法主要有匿名[8]、随机扰动[9]和加密[10]。在这 3 种方法中，由于随机扰动方法能够保证较好的数据可用性，对分析挖掘

收稿日期：2017-02-06；**修回日期**：2017-04-12
通信作者：徐正全，xuzq@whu.edu.cn
基金项目：国家自然科学基金资助项目（No.41671443）；武汉市应用基础研究计划基金资助项目（No.2016010101010024）；中美计算机科学研究中心开放基金资助项目（No.KJR16228）；南京信息工程大学人才引进基金资助项目（No.2016r055）
Foundation Items: The National Natural Science Foundation of China (No.41671443), Applied Basic Research Program of Wuhan (No.2016010101010024), Open Funding of NUIST and PAPD (No.KJR16228), Introducing Talenet of NUIST Program (No.2016r055)

结果的影响较小，是目前重要的保护手段。在基于随机扰动的方法中，差分隐私机制[11]由于具有严格的数学公理化模型，并对攻击者的背景知识没有限制，是一种从数学上严格定义保护强度和数据可用性的隐私保护手段。由于差分隐私机制是根据数据挖掘的结果计算敏感度函数，进而生成噪声加入到数据集中的单个数据中，因此，加入的噪声对数据集的挖掘结果影响很小，同时又能保护单个数据的真实值，近年来已成为隐私保护研究的热门领域。

轨迹数据作为一种来源及应用都非常广泛的数据类型，其差分隐私发布方法也被研究者关注，目前的轨迹差分隐私发布方法主要分为2类：基于相关性建模和序列变换。基于相关性建模的方式通过建立相关性模型重构敏感度函数，主要有马尔可夫[12]、贝叶斯[13~15]等概率模型和系数矩阵模型[16]；基于序列变换的方式是将相关性的轨迹序列变换到另一个域的独立序列进行处理，包括离散傅里叶变换（DFT）[17]及其改进算法的离散小波变换（DWT）[18,19]以及主成分分析（PCA）[20]等。

虽然差分隐私机制作为一种随机扰动方法被广泛研究和应用，但其最初是为了解决由相互独立的数据构成的静态数据集的隐私泄露问题而被提出的，通过对待保护数据加入独立的噪声来保护隐私[21]。因此，目前，基于差分隐私机制的轨迹发布方法也都是将轨迹数据作为独立序列来处理，通过对待发布轨迹序列叠加独立的噪声序列来保护隐私。然而，当轨迹数据非独立（具有相关性）时，独立的噪声方式仍然会泄露隐私，现有的轨迹差分隐私保护方法仍然面临以下问题。

1) 隐私保护程度不足。由于轨迹数据具有相关性而现有方法中加入的噪声是独立的，攻击者可以利用一些滤除噪声的方法（如滤波）过滤加入的噪声，从而提高攻击成功的概率。

2) 数据可用性较低。由于轨迹数据是相关的，导致差分隐私的敏感度函数的权重增大。为了达到相同的隐私保护强度，需要加入更多的噪声，而更多的噪声意味着数据可用性的降低。

针对现有轨迹差分隐私发布方法中的缺陷，本文提出了一种相关性拉普拉斯机制的轨迹差分隐私发布方法，通过生成与要保护轨迹序列自相关函数相同的噪声序列，叠加到原始轨迹中并发布，与现有方法相比，可以防止滤波等攻击方式的攻击，可以为单个位置点提供更高的保护强度；由于加入噪声的相关性与原始序列相关性一致，可以更好地支持轨迹规律性的发现，对轨迹挖掘的结果影响较小。本文贡献主要有以下4点。

1) 从信号处理角度给出了针对现有轨迹差分隐私发布方法的攻击模型，阐述了现有轨迹差分隐私发布方法所面临的问题。

2) 提出了序列不可区分性的概念，在轨迹相关性已知的情况下，序列不可区分性使噪声和待发布轨迹序列相关性一致，在满足差分隐私需求的同时，能够保证较好的数据可用性。

3) 提出了一种高效且易实现的相关性拉普拉斯机制，将高斯白噪声通过特定冲激响应的滤波器，生成与待发布轨迹相关性一致的拉普拉斯噪声，以实施序列不可区分性。

4) 本文在真实轨迹数据集上对CLM方法与现有的轨迹隐私保护方法进行对比实验。实验结果表明CLM具有更加严格的隐私保护强度，在相同隐私保护效果的情况下，CLM具有更好的数据可用性。

2 相关工作

现有的轨迹差分隐私发布机制可以分为2种类型。一类是基于建模的方法，通过建立轨迹序列的相关性模型，将模型中的系数作为敏感度函数的权重重新计算应加入的噪声大小。另一类是基于变换的方法，由于原始差分隐私机制要求数据集的独立性，基于变换的方法将相关性的轨迹序列变换为独立的形式进行处理。

1) 相关性建模。该类方法利用相关性模型来描述轨迹数据的相关性，轨迹相关性可以用各种方式进行建模。Gehrke等[21,22]专注于相关数据的隐私，从密码学的概念出发，提出了一种零背景知识的隐私保护模型，但其中的聚合函数需要仔细选择。作为改进方案，Shen等[12]考虑了攻击者的背景知识，提出了一个基于采样和概率转移的马尔可夫链蒙特卡罗（MCMC）框架，该框架可以在频繁图模式挖掘时保证用户数据的差分隐私效果。文献[23]定义一个通用框架——Pufferfish，来对独立和非独立数据进行统一建模。另外，受到Pufferfish思想的启发，文献[24]提出了Blowfish机制，它给出了Pufferfish框架增加数据可用性所需的必要约束条件。以Pufferfish为基础的另一个隐私定义是由Yang等[13]提出的贝叶斯差分隐私机制。为了更加精确地描述轨迹数据之间的相关性，他们采用高斯模型来描述数据之间的相关性，

同时，利用高斯模型评估其他算法所能达到的隐私保护强度。Zhu 等[16]利用相关系数矩阵来描述轨迹数据之间的相关性，将相关系数作为差分隐私敏感度函数的权重，重新计算噪声大小。

此外，在未考虑隐私保护的相关性建模方面，也有很多研究。Cao 等[25]利用内耦合和外耦合行为函数建模相关性信息，且相关性程度用行为函数来表示。Zhou 等[26]将相关性数据映射为无向图，提出了多示例学习算法。文献[27]假设网络结构的查询结果未被公布，在这种情况下，攻击者无法知晓关于网络结构的背景知识。与文献[27]不同的是，本文不对攻击者的背景知识进行任何假设，即轨迹的相关性已经完全被攻击者知晓。

随着越来越多的建模方法的出现，相关性的建模方法会越来越完善。在处理轨迹隐私发布问题上，这类方法的关键是如何建立更加精确的相关性模型。

2) 序列变换。该类方法的思想是将轨迹相关性序列转化为独立的序列进行处理，同时，尽量保留其主要特征。Rastogi 等[17]提出了一种傅里叶扰动算法（FPA）来解决这个问题。FPA 中的离散傅里叶变换将相关性数据变换到独立的傅里叶变换域，然后对 DFT 变换系数添加噪声扰动，最后利用傅里叶反变换重建原始序列。此外，为了克服 FPA 无法处理短期非平稳序列的缺点，文献[18,19]引入了离散小波变换（DWT）。DWT 扩展了 FPA 的适用范围并能保持更多的序列特征。文献[20]利用主成分分析方法提取数据集的主要特征，并变换到另一个独立的维度中进行处理，从而确保差分隐私的效果，并利用这种方法测试了一些常用的统计学习应用。

尽管在数据可用性方面，基于序列变换的方法表现较好，但这种方法破坏了轨迹序列的相关特性，导致分析挖掘结果不准确。

由于现有方法加入的噪声是独立的，且现有的建模方式不够完善，因此，现有研究方法不能确保实际中有效的隐私保护强度达到设定的保护强度值，容易造成用户隐私信息的泄露，同时引入了额外的噪声，导致挖掘分析结果的可用性较差。本文的目标是提供一个实用的解决方案，发布相关性的轨迹数据，同时，确保发布的轨迹数据满足差分隐私的需求。更具体地说，本文试图解决以下问题。

1) 如何在相关性背景知识公开给攻击者的情况下实现差分隐私。

2) 如何解决当前方法中由于独立噪声所导致的隐私保护程度不足的问题。

3) 是否可能实现不添加额外的噪声，即可保证与原始差分隐私机制相同的隐私保护程度。

3　理论基础及问题定义

3.1　差分隐私

差分隐私的主要思想是在数据集 D 中的每个记录上加入噪声，使数据泄露的概率控制在一定的范围。差分隐私的正式定义如下。

定义 1　ε-差分隐私[11]。设 $f(\cdot)$是查询函数，若有随机算法 K 以及 K 所有可能输出的集合 S，对于给定数据集 D 以及与其最多相差一个记录的任意邻近数据集 D'，若算法 K 满足

$$\Pr\left[K(D)\in S\right]\leqslant \mathrm{e}^{\varepsilon}\Pr\left[K(D')\in S\right] \tag{1}$$

则算法 K 提供 ε-差分隐私保护。

如图 1 所示，算法 K 对输出结果提供 ε-差分隐私保护，通过隐私保护强度 ε 保证对于数据集 D 与其最多相差一个记录的邻近数据集 D' 的查询结果在一定概率上不可区分。

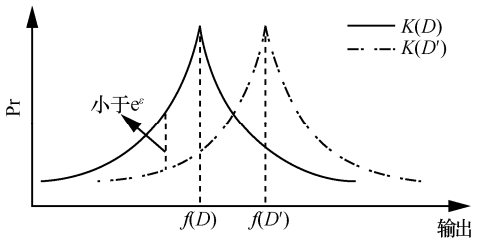


图 1　随机算法在邻近数据集上的输出概率

由于差分隐私机制本质上是一种噪声扰动机制，当前，普遍使用 Laplace 机制在原始数据集中加入噪声使其满足 ε-差分隐私。

定义 2　Laplace 机制[11]。对于查询函数 $f:D\to R^d$，则式(2)所示的随机算法提供 ε-差分隐私保护。

$$K(D)=f(D)+Lap(\lambda) \tag{2}$$

其中，$Lap(\lambda)$ 为服从 Laplace 分布的噪声，λ 的计算式为

$$\lambda=\frac{\Delta f}{\varepsilon} \tag{3}$$

其中，Δf 表示全局敏感度，ε 表示隐私保护强度。ε 越小，数据的隐私保护强度就越大。

全局敏感度 Δf 衡量了从数据集 D 中移除一个记录后输出 S 的最大变化，其定义如下。

定义 3 全局敏感度[28]。对于查询函数 $f: D \to R^d$，f 的全局敏感度为

$$\Delta f = \max_{D,D'} \| f(D) - f(D') \|_p \tag{4}$$

其中，R 表示映射的实数空间，d 表示函数 f 的查询维度，p 表示度量 Δf 使用的范数距离。

原始差分隐私机制主要针对独立数据集的隐私保护，而在相关性轨迹序列中，由于相关性会导致差分隐私全局敏感度的增加，原始差分隐私机制不能适应相关性轨迹序列的发布。

3.2 自相关函数

自相关函数是描述随机信号 $X(t)$ 在任意 2 个不同时刻 t_1 和 t_2 的取值之间的相关程度。

定义 4 自相关函数[29]。实随机信号 $X(t)$ 任意 2 个时刻 t_1 和 t_2 的自相关函数 $R_{XX}(t_1,t_2)$ 定义为

$$\begin{aligned} R_{XX}(t_1,t_2) &= \mathrm{E}\left[X(t_1)X(t_2)\right] \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_1 x_2 \zeta(x_1,x_2;t_1,t_2)\mathrm{d}x_1\mathrm{d}x_2 \end{aligned} \tag{5}$$

其中，$x_1, x_2 \in X$，$\zeta(\cdot)$ 表示概率密度函数。

若 $X(t)$ 是平稳随机信号，那么 $X(t)$ 的统计特性与时间的起点无关，令 $t_2 = t_1 + \tau$，则有 $\zeta(x_1,x_2;t_1,t_2) = \zeta(x_1,x_2;\tau)$。所以，平稳随机信号的自相关函数是时间间隔 τ 的函数，可以记为 $R_{XX}(\tau)$。平稳随机信号的自相关函数 $R_{XX}(\tau)$ 有以下性质。

1) 当平稳随机信号是实函数时，其相关函数是偶函数。

$$R_{XX}(\tau) = R_{XX}(-\tau)$$

2) $\tau = 0$ 时的自相关函数取最大值。

$$R_{XX}(0) \geqslant \left| R_{XX}(\tau) \right|$$

3) 若 $X(t)$ 与 $X(t+\tau)$ 相互独立，则 $R_{XX}(t, t+\tau) = 0$，即 2 个独立时刻的自相关函数为 0。

由于轨迹序列主要由时间、经度和纬度 3 个部分组成，且轨迹序列在短时间内前后位置点之间的相关性很强，因此，轨迹序列可以当成短时平稳过程进行处理，而短时平稳过程的相关性可以用自相关函数表示。

3.3 问题陈述

由于时间序列的相关性会导致差分隐私全局敏感度的增加，现有的方法从降低相关性时间序列全局敏感度的角度提出了各种差分隐私保护模型，本节具体阐述现有方法在基于滤波的攻击模型下所面临的问题。

如图 2 所示，假设时间序列 X 经过差分隐私机制处理加入噪声序列 N 后，得到带噪序列 X'。对原始时间序列 X、带噪序列 X' 以及原始时间序列的邻近序列 X'' 进行查询后，分别得出查询结果的概率密度分布曲线 $K(X)$、$K(X')$ 和 $K(X'')$。由于原始时间序列 X 具有相关性，且通过差分隐私机制加入的噪声 N 是独立同分布的，因此，基于滤波的攻击模型可以滤除部分噪声并得到序列 $\tilde{X}$。相比于 $K(X')$，滤波后查询结果的概率密度分布曲线 $K(\tilde{X})$ 将更接近原始时间序列的概率密度分布曲线，若记设定的隐私保护强度为 ε，实际滤波后的隐私保护强度为 ε'，则 $\varepsilon < \varepsilon'$，即隐私保护强度减小。

与轨迹差分隐私发布最为接近的工作是贝叶斯差分隐私[13]，通过攻击者掌握相关性背景知识的多少来定义攻击者的强弱程度。实际上，为了达到无条件的安全性，对于攻击者来说，相关性背景知识应当假设为已知的，即攻击者已知全部轨迹相关性信息。

轨迹差分隐私发布的目标是设计一个扰动机

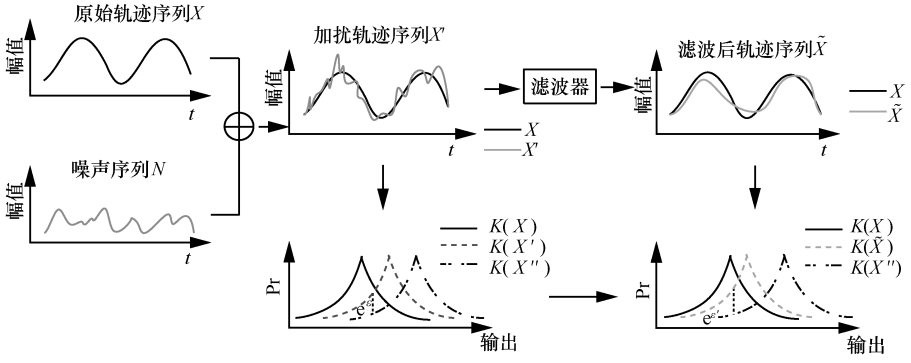


图 2 相关性对隐私保护强度的影响

制，使在轨迹相关性背景知识公开的情况下，用户的隐私仍然不被泄露。假设 CLM 代表隐私保护机制，那么用户隐私泄露的风险可以表示为

$$CLM(M):=\sup_{x_i,x_i',R_{XX}(\tau),S}\ln\frac{\Pr\left[M(X)\in S\mid x_i,R_{XX}(\tau)\right]}{\Pr\left[M(X')\in S\mid x_i',R_{XX}(\tau)\right]} \tag{6}$$

其中，$x_i\in X$、$x_i'\in X'$、$M(X)$ 和 $M(X')$ 分别表示原始和发布序列的统计结果。

如果隐私泄露的上限是 ε，即

$$CLM(M)\leqslant\varepsilon \tag{7}$$

$$\sup_{x_i,x_i',R_{XX}(\tau),S}\ln\frac{\Pr\left[M(X)\in S\mid x_i,R_{XX}(\tau)\right]}{\Pr\left[M(X')\in S\mid x_i',R_{XX}(\tau)\right]}\leqslant\varepsilon \tag{8}$$

那么 CLM 可以达到 ε 水平的隐私保护效果，即满足 ε-差分隐私。轨迹差分隐私发布的关键就是设计一种机制，能够满足上述定义中的目标，同时保持较高的数据可用性。

4　算法原理

本节首先阐述本文提出的满足轨迹差分隐私发布问题定义的概念——序列不可区分性，然后给出了实现序列不可区分性的方法，最后分析了算法的安全性。

4.1　序列不可区分性

在现代密码学中，如果加密机制能够保证攻击者无法从加密的 2 个密文当中区分出正确的明文，那么这个加密机制叫作“不可区分性”，不可区分性可以达到无条件的安全性。事实上，差分隐私机制的思想同样来自于不可区分性。在 3.3 节中的攻击模型中，由于发布序列中的噪声是独立的，知晓相关性背景知识的攻击者就可以利用原始序列的自相关函数过滤噪声。相反，如果加入发布轨迹序列中的噪声是相关的，并且与原始序列的自相关函数相同，那么原始序列和发布序列对于攻击者来说是无法区分的，可以达到无条件的安全性。下面，定义这种不可区分性并在定理 1 中证明它的安全性。

定义 5　序列不可区分性。如果原始轨迹和发布轨迹序列的自相关函数 $R_{XX}(\tau)$ 和 $R_{XX'}(\tau)$ 满足

$$R_{XX'}(\tau)=R_{XX}(\tau) \tag{9}$$

那么发布轨迹和原始轨迹对于攻击者来说是不可区分的，攻击者在已知原始序列相关性的条件下无法发起攻击。下面，通过定理 1 证明序列不可区分性能够满足式(8)。

定理 1　如果发布序列 X' 和原始序列 X 的自相关函数满足定义 5 中的不可区分性，那么发布的序列满足 ε-差分隐私。

证明　根据条件概率公式，将式(6)展开，可以得到

$$\begin{aligned}CLM(M)&:=\sup_{x_i,x_i',R_{XX}(\tau),S}\ln\frac{\Pr\left[M(X)\in S\mid x_i,R_{XX}(\tau)\right]}{\Pr\left[M(X')\in S\mid x_i',R_{XX}(\tau)\right]}\\&=\sup_{x_i,x_i',R_{XX}(\tau),S}\ln\frac{\Pr\left[M(X)\in S\mid X'\right]\Pr\left[X'\mid x_i,R_{XX}(\tau)\right]}{\Pr\left[M(X')\in S\mid X'\right]\Pr\left[X'\mid x_i',R_{XX}(\tau)\right]}\\&=\sup_{x_i,x_i',R_{XX}(\tau),S}\ln\frac{\Pr\left[M(X)\in S\mid X'\right]\Pr\left[X'\mid R_{XX}(\tau)\right]}{\Pr\left[M(X')\in S\mid X'\right]\Pr\left[X'\mid R_{XX'}(\tau)\right]}\end{aligned}$$

由于 $R_{XX'}(\tau)=R_{XX}(\tau)$，那么

$$CLM(M):=\sup_{x_i,x_i',R_{XX}(\tau),S}\ln\frac{\Pr\left[M(X)\in S\mid X'\right]}{\Pr\left[M(X')\in S\mid X'\right]}$$

更进一步，如果噪声大小和原始差分隐私中加入的独立噪声的大小保持一致，那么

$$\begin{aligned}CLM(M)&:=\sup_{x_i,x_i',R_{XX}(\tau),S}\ln\frac{\Pr\left[M(X)\in S\mid X'\right]}{\Pr\left[M(X')\in S\mid X'\right]}\\&=\sup_{x_i,x_i',R_{XX}(\tau),S}\ln\frac{\Pr\left[M(X)\in S\right]}{\Pr\left[M(X')\in S\right]}\end{aligned}$$

在原始差分隐私中，有

$$\sup_{x_i,x_i',R_{XX}(\tau),S}\ln\frac{\Pr\left[M(X)\in S\right]}{\Pr\left[M(X')\in S\right]}\leqslant\varepsilon$$

因此，

$$CLM(M)\leqslant\varepsilon$$

定义 5 阐述了相关性轨迹差分隐私发布的原则，并且在定理 1 中给出了证明。通过上述分析可以得出结论，序列不可区分性在保证和原始差分隐私相同的隐私保护强度下，并没有提高噪声水平。

尽管定理 1 阐述了相关性轨迹发布的设计原则，但是，目前尚未有实用高效的办法产生满足序列不可区分性的噪声序列。

4.2　CLM

实现相关性轨迹差分隐私发布的条件已经在序列不可区分性中推导得出。本节提出了一个简单高效的方案来实现序列不可区分性。根据信号处理的有关知识，滤波器主要由加法器和延时器组成，因此，如果一个独立信号输入滤波器，那么在输出端输出的信号将会是非独立的，CLM[30]就是根据这

个思想来进行设计的。此外，文献[17]讨论了分布式系统中由高斯噪声生成拉普拉斯噪声的方法，受文献[17]工作的启发，CLM 将 4 组高斯白噪声序列通过特定冲激响应的线性系统来产生相关性的拉普拉斯噪声，CLM 框架如图 3 所示。

图 3 描述了 CLM 的框架机制。首先，产生 4 组高斯白噪声，高斯白噪声的参数由原始轨迹序列和隐私保护强度决定；然后，将高斯白噪声通过滤波器（一种经典的线性系统），可以得到相关性的高斯噪声序列；最后，根据高斯噪声生成拉普拉斯噪声序列的原则生成相关性的拉普拉斯噪声。

在这个过程当中，主要需要解决滤波器的设计问题。假设 $R_G(\tau)$ 表示高斯白噪声的自相关函数，有 $R_G(\tau)=R_G(t,t+\tau)$。$R_{G'}(\tau)$ 表示高斯噪声序列 G_i' 中每组高斯噪声的自相关函数，δ 为脉冲函数。在设计滤波器前，首先要考虑相关性的高斯噪声序列 G_i' 的自相关函数 $R_{G'}(\tau)$ 应该满足的条件。由于文献[17]阐述了高斯噪声生成拉普拉斯噪声的方法，这里只需要推导出 $R_{G'}(\tau)$ 应满足的条件，如定理 2 所示。

定理 2 如果 G_i' 的自相关函数 $R_{G'}(\tau)$ 满足

$$R_{G'}(\tau)=\sqrt{\frac{R_{XX}(\tau)}{8}} \tag{10}$$

那么，由 $Z=G_1'^2+G_2'^2-G_3'^2-G_4'^2$ 计算得到的噪声序列 Z 的自相关函数满足 $R_Z(\tau)=R_{XX}(\tau)$。

证明 因为 $Z=G_1'^2+G_2'^2-G_3'^2-G_4'^2$，那么，

$$\begin{aligned}R_Z(\tau)=&\mathrm{E}\Big[\big(g_1'^2(t)+g_2'^2(t)-g_3'^2(t)-g_4'^2(t)\big)\cdot\\&\big(g_1'^2(t+\tau)+g_2'^2(t+\tau)-g_3'^2(t+\tau)-g_4'^2(t+\tau)\big)\Big]\\=&4\Big[R^2_{G_1'}(\tau)+R^2_{G_2'}(\tau)-R^2_{G_3'}(\tau)-R^2_{G_4'}(\tau)\Big]\\=&8R^2_{G'}(\tau)\end{aligned}$$

如果 $R_{G'}(\tau)=\sqrt{\dfrac{R_{XX}(\tau)}{8}}$，那么

$$R_Z(\tau)=R_{XX}(\tau) \tag{11}$$

定理 2 给出了 $R_{G'}(\tau)$ 应该满足的条件，那么滤波器的参数（冲激响应）就可以由此条件得到。滤波器的冲激响应应满足的条件在定理 3 给出。

定理 3 如果滤波器的脉冲响应满足

$$h(\tau)=\sqrt{\frac{R_{XX}(\tau)}{16\pi N_0}} \tag{12}$$

其中，N_0 表示高斯白噪声的功率谱密度。那么 4 组通过此滤波器的高斯白噪声序列的自相关函数是 $R_{G'}(\tau)$。

证明 由于滤波器是一个线性系统，输出序列的自相关函数 $R_{G'}(\tau)$ 满足

$$R_{G'}(\tau)=R_G(\tau)*\big[h(\tau)*h(-\tau)\big] \tag{13}$$

其中，$R_G(\tau)$ 表示高斯白噪声的自相关函数，符号 * 表示卷积。

由于自相关函数和功率谱密度是一对傅里叶变换，对式(13)进行傅里叶变换，可以得到相关性高斯噪声序列的功率谱密度

$$P_{G'}(\omega)=P_G(\omega)\big|H(\omega)\big|^2 \tag{14}$$

由于 G_i 是高斯白噪声序列，那么 G_i 的功率谱密度为

$$P_G(\omega)=N_0 \tag{15}$$

另外，根据定理 2，G_i' 的自相关函数 $R_{G'}(\tau)$ 应该满足条件

$$R_{G'}(\tau)=\sqrt{\frac{R_{XX}(\tau)}{8}} \tag{16}$$

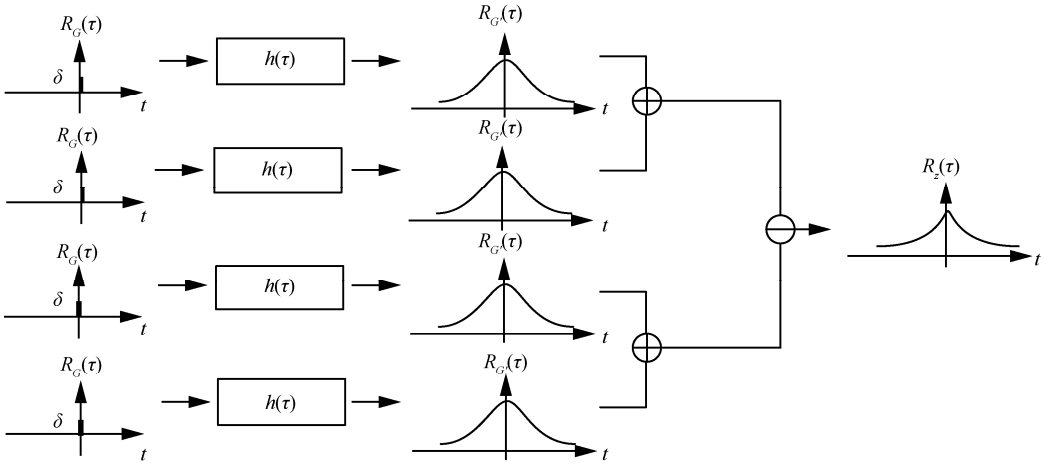


图 3 CLM 框架

对式(16)进行傅里叶变换，可以得到相关性高斯序列的功率谱密度

$$P_{G'}(\omega)=2\pi\sqrt{\frac{R_{XX}(\tau)}{8}}\delta(\omega) \tag{17}$$

结合式(14)、式(15)和式(17)，可以得到滤波器的系统函数

$$\left|H(\omega)\right|^2=\frac{1}{N_0}2\pi\sqrt{\frac{R_{XX}(\tau)}{8}}\delta(\omega) \tag{18}$$

将式(18)进行傅里叶逆变换，可以得到滤波器的冲激响应

$$h(\tau)=\sqrt{\frac{R_{XX}(\tau)}{16\pi N_0}} \tag{19}$$

定理 3 给出了滤波器的设计准则。由信号处理的知识可知，平稳时间序列的相关性可以用自相关函数表示。而轨迹数据由于前后位置点之间的相关性很强，因此，轨迹序列可以当成短时平稳过程进行处理，那么，轨迹序列的相关性就可以用自相关函数表示。

由于轨迹数据对于实施隐私保护的一方来说是已知的，因此，可以分别在经度和纬度上根据时间戳计算轨迹的自相关函数，然后按照本文方法加入噪声，最后再确定其空间位置点，CLM 方案的详细步骤如算法 1 所示。

算法 1　轨迹序列差分隐私发布算法

输入　原始轨迹序列 X

输出　发布轨迹序列 X'

1) 用户计算待发布序列 X 的自相关函数 $R_{XX}(\tau)$。

2) 用户产生与待发布序列 X 长度相同的 4 个高斯白噪声序列 G_1、G_2、G_3、G_4。其中，$G_i\sim N\left(0,\sqrt{2\lambda}\right)$，$i\in\{1,2,3,4\}$。

3) 将 G_1、G_2、G_3、G_4 通过冲激响应 $h(\tau)=\sqrt{\frac{R_{XX}(\tau)}{16\pi N_0}}$ 的滤波器，得到 4 个自相关函数为 $R_{G'}(\tau)=\sqrt{\frac{R_{XX}(\tau)}{8}}$ 的相关性高斯噪声序列 G_1'、G_2'、G_3'、G_4'。

4) 计算噪声序列 $Z=G_1'^2+G_2'^2-G_3'^2-G_4'^2$。

5) 将噪声序列 Z 叠加到 X 上得到加扰的轨迹序列 $X'=X+Z$。

6) 返回 X'

4.3　隐私分析

4.2 节给出了 CLM 方案的设计原则，本节分析 CLM 的安全性，如定理 4 所示。

定理 4　如果噪声和原始轨迹序列的自相关函数 $R_Z(\tau)$ 和 $R_{XX}(\tau)$ 满足

$$\left|\frac{R_Z(\tau)}{R_{XX}(\tau)}\right|\leqslant\mu \tag{20}$$

那么 $X'=X+Z$ 满足 $(1+\zeta)\ \varepsilon$-差分隐私，其中，$\zeta=\frac{\mu}{\varepsilon}$，$\mu$ 是容忍参数。

证明　从定理 1 可以得到

$$\begin{aligned}
&CLM(M):=\\
&\sup_{x_i,x_i',R_{XX}(\tau),S}\ln\frac{\Pr\left[M(X)\in S\mid X'\right]\Pr\left[X'\mid R_{XX}(\tau)\right]}{\Pr\left[M(X')\in S\mid X'\right]\Pr\left[X'\mid R_{X'X'}(\tau)\right]}\\
&=\sup_{x_i,x_i',S}\ln\frac{\Pr\left[M(X)\in S\mid X'\right]}{\Pr\left[M(X')\in S\mid X'\right]}+\\
&\sup_{x_i,x_i',R_{XX}(\tau),R_{X'X'}(\tau)}\ln\frac{\Pr\left[X'\mid R_{XX}(\tau)\right]}{\Pr\left[X'\mid R_{X'X'}(\tau)\right]}
\end{aligned}$$

关注等式的右半部分

$$\begin{aligned}
&\sup_{x_i,x_i',R_{XX}(\tau),R_{X'X'}(\tau)}\ln\frac{\Pr\left[X'\mid R_{XX}(\tau)\right]}{\Pr\left[X'\mid R_{X'X'}(\tau)\right]}\\
&=\sup_{x_i,x_i',R_{XX}(\tau),R_{X'X'}(\tau)}\ln\frac{\Pr\left[X'\mid R_{XX}(\tau)\right]}{\Pr\left[X'\mid R_{XX}(\tau)\right]\Pr\left[R_{XX}(\tau)\mid R_{X'X'}(\tau)\right]}\\
&=\sup_{x_i,x_i',R_{XX}(\tau),R_{X'X'}(\tau)}\ln\frac{1}{\Pr\left[R_{XX}(\tau)\mid R_{X'X'}(\tau)\right]}
\end{aligned}$$

如果 $\left|\frac{R_Z(\tau)}{R_{XX}(\tau)}\right|\leqslant\mu$，$R_{XX}(\tau)$ 和 $R_{X'X'}(\tau)$ 满足

$$\frac{R_{XX}(\tau)}{R_{X'X'}(\tau)}\geqslant\frac{1}{1\pm\mu}$$

因此，有

$$\begin{aligned}
&\sup_{x_i,x_i',R_{XX}(\tau),R_{X'X'}(\tau)}\ln\frac{1}{\Pr\left[R_{XX}(\tau)\mid R_{X'X'}(\tau)\right]}\\
&\leqslant\sup_{x_i,x_i',R_{XX}(\tau),R_{X'X'}(\tau)}\ln\frac{1}{\frac{1}{1\pm\mu}\Pr\left[R_{X'X'}(\tau)\mid R_{X'X'}(\tau)\right]}\\
&=\sup_{x_i,x_i',R_{XX}(\tau),R_{X'X'}(\tau)}\ln(1\pm\mu)\\
&\approx\sup_{x_i,x_i',R_{XX}(\tau),R_{X'X'}(\tau)}\ln e^{\mu}\\
&=\mu
\end{aligned}$$

那么有

$$CLM(M) := \sup_{x_i, x_i', S} \ln \frac{\Pr[M(X) \in S \mid X']}{\Pr[M(X') \in S \mid X']} + \mu$$

另外，

$$\sup_{x_i, x_i', S} \ln \frac{\Pr[M(X) \in S \mid X']}{\Pr[M(X') \in S \mid X']} \leqslant \varepsilon$$

$CLM(M)$ 满足 $(1+\zeta)\varepsilon$ -差分隐私，其中，$\zeta = \frac{\mu}{\varepsilon}$。

5 实验评估

本节评估了所提出的相关性轨迹序列差分隐私发布方法 CLM 的性能，并与当前的主要算法进行了对比分析，主要包括安全性、可用性和计算复杂度评估。

5.1 数据集和配置

本文实验环境是 Intel Core 2 Quad 3.06 GHz Windows 7 系统，8 GB 内存。每个实验运行 100 次。本文在真实的数据集中对提出的算法进行测试，主要包括以下 3 个不同的轨迹数据集。

1) Geolife。Geolife 项目[31]的轨迹数据集搜集来自 182 个志愿者 5 年（2007 年 4 月~2012 年 8 月）的轨迹数据，由微软亚洲研究院提供。每条 GPS 轨迹由时间戳的序列构成，包括经纬度、海拔和时间等信息。此数据包含 17 621 条轨迹数据，总长度 1 292 951 km，总时长 50 176 h。

2) T-Drive Taxi[32]。此数据集描述了 2009 年 5 月中国北京 8 602 辆出租车的 GPS 轨迹数据。轨迹区域覆盖了经纬度（39.788°N, 116.148°W）和（40.093°N, 116.612°W）之间的矩形区域，面积接近 34 km×40 km。数据集中轨迹的采样频率从 30~300 s 不等，包含大约 4 300 000 次的乘客记录，每条记录是由大约 30 s 间隔的插值序列构成。

3) Check-in[33]。数据集包含由美国纽约超过 49 000 名和洛杉矶 31 000 名社交用户的签到位置点信息。每个签到信息包括用户 ID、时间戳、地点 ID 和地点类型。

在这 3 个数据集中，Geolife 数据集较其他 2 个数据集覆盖了更长的时间戳，轨迹数据的相关性也更强。对于每个数据集来说，对其进行了 1 000 次随机查询，查询结果数据集表示为 Q。查询结果的精确程度可以用平均绝对误差（MAE）来衡量，其定义为

$$MAE = \frac{1}{L} \sum_{x_i \in X} |x_i' - x_i| \tag{21}$$

其中，L 表示轨迹序列的长度，MAE 越小代表数据可用性越高。

5.2 隐私性能评估

本文通过与现有的 5 种方法进行对比来验证 CLM 的性能，实验在所有数据集上进行测试，所对比方法的具体参数设置按照其论文中的建议进行设定。本文设置隐私强度 ε 的变化范围为 0.1~0.9，间隔是 0.2，计算得到查询结果的概率密度函数，隐私评估结果如图 4 所示。

图 4 展示了所有数据集上的隐私保护强度对比结果。从图 4 可以看出，CLM 的隐私保护强度值 ε' 比设定值要小，意味着 CLM 具有更高的保护强度。在图 4(a)中，当 $\varepsilon = 0.1$ 时，CLM 的 ε' 值是 0.081，而 CIM[16]机制的 ε' 值是 0.086，CLM 机制比 CIM 机制提高了 5.8%；当 $\varepsilon = 0.9$ 时，CLM 的 ε' 值是 0.873，而 CIM 机制的 ε' 值是 0.895，提高了 2.5%。从图 4 可以看出，CLM 和 CIM 机制的隐私保护强度小于设定值，而其他算法的隐私保护强度均大于这 2 种算法，证明了本文算法的有效性。

另外，相比于图 4(a)，图 4(b)和图 4(c)中不同方法的隐私保护强度更为接近，引起这种现象的原因是，图 4(a)中的轨迹数据集的相关性较强。在图 4 中，ε' 的最大值出现在图 4(a)中，当 $\varepsilon = 0.9$ 时，CLM 的性能较马尔可夫方法提高 30.9%。实验结果表明，轨迹的相关性确实对隐私保护强度有影响。由于不同方法的相关性建模方式不同，不同方法加入的噪声大小也不一样，从而导致隐私保护强度不同。

隐私性能评估表明，CLM 方法在相关性轨迹差分隐私发布中是有效的，比现阶段最优的 CIM 方法的性能提高了 2.9%。实验结果同时表明，轨迹数据的相关性对于隐私保护强度确实是有影响的，现有方法的隐私性能取决于它们自身的建模原则。

5.3 可用性评估

为了评估本文方法在数据可用性上的表现，评估了 CLM 机制在 3 种不同轨迹挖掘类应用中的 MAE，主要包括轨迹聚类、轨迹频繁和轨迹伴随 3 种应用，并与现有方法进行对比。隐私强度 ε 的变

化范围设置为 0.1~0.9，间隔范围是 0.2。对比结果如图 5~图 7 所示。

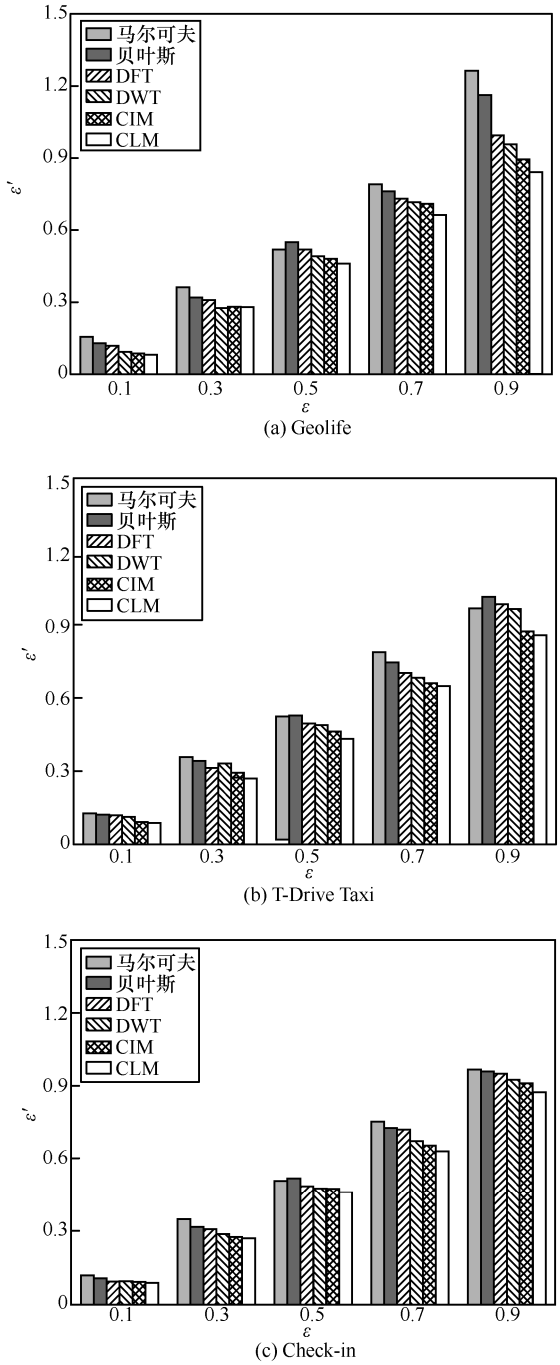


图 4　隐私保护程度对比结果

从图 5~图 7 可以看出，CLM 较其他方法具有更小的 *MAE*。另外，由于 Geolife 轨迹数据集具有更强的相关性，图 5(a)的 *MAE* 比另外 2 个数据集要大。在图 5(a)中，当 $\varepsilon = 0.1$ 时，CLM 的 *MAE* 是 12.1，而现有的最优方法 CIM 机制的 *MAE* 是 21.6，数据可用性提高（*MAE* 降低）了 44.0%；当 $\varepsilon = 0.9$ 时，CLM 的 *MAE* 是 4.1 而 CIM 机制的 *MAE* 是 6.0，可用性提高（*MAE* 降低）了 31.7%。这种现象在图 5(b)和图 5(c)中同样可以观察到。实验结果表明，对于相关性轨迹数据发布，CLM 在数据可用性的表现上优于其他算法，DFT 和 DWT 的数据可用性虽然优于马尔可夫和贝叶斯方法，但它们的可用性受保留的变换系数个数的影响较大。这是因为 CLM 是一种相关性噪声机制，相比于其他方法中的独立噪声机制，在设定相同的隐私保

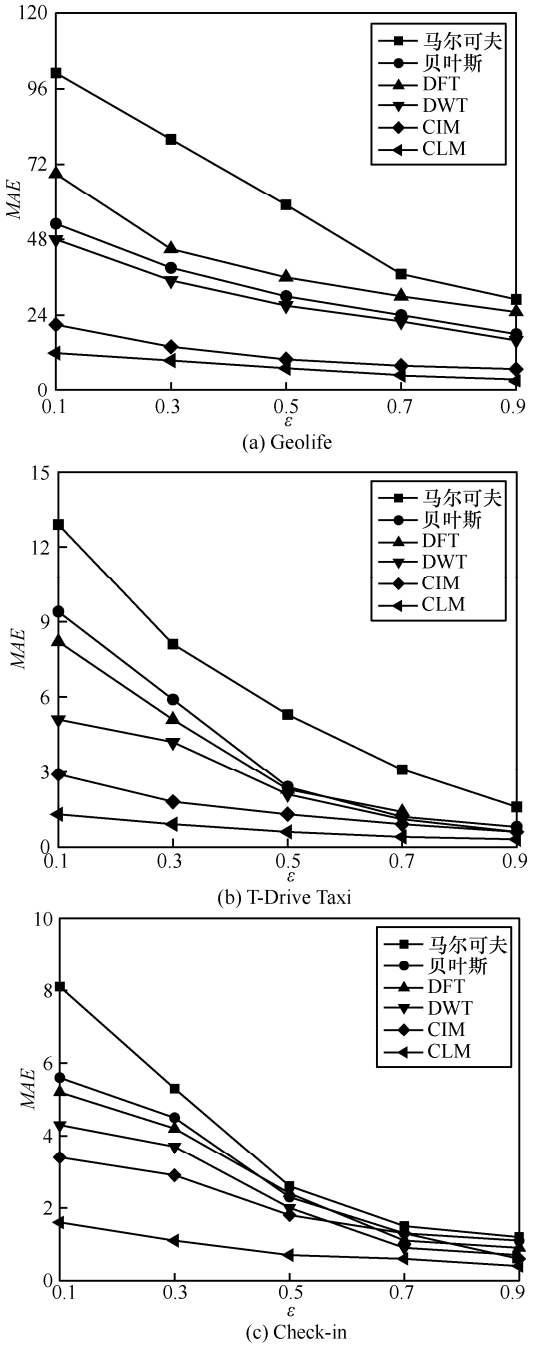


图 5　轨迹聚类可用性对比结果

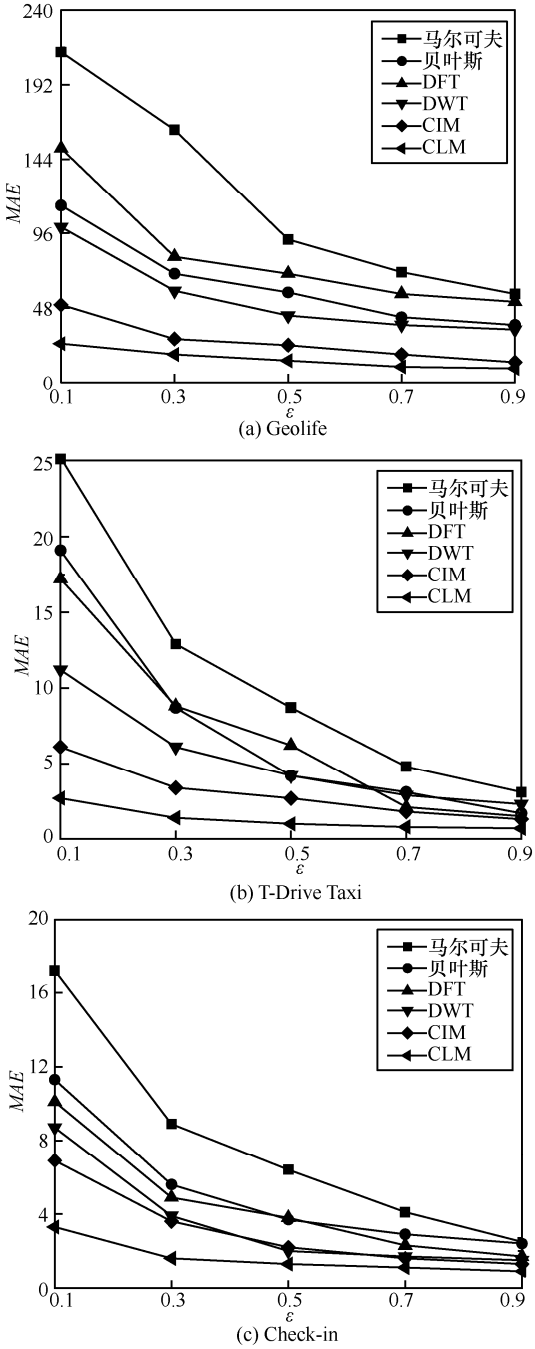


图 6　轨迹频繁可用性对比结果

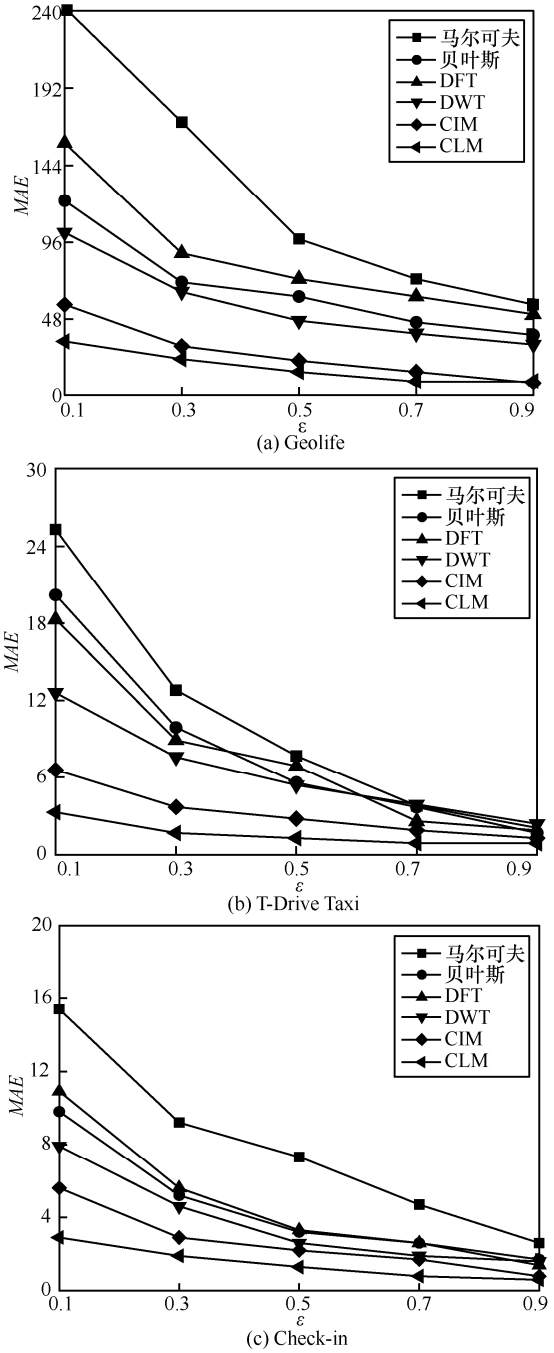


图 7　轨迹伴随可用性对比结果

护强度 ε 下，需要更小的噪声就可以达到相同的隐私保护强度。因而，CLM 具有更高的数据可用性。

在差分隐私中 ε 作为一个关键参数，用以确定隐私强度。根据 Dwork[28]的研究，当 ε=1 或更小时，数据可用性可以达到较为合适的水平。因此，本文评估了 ε 为 0.1~0.9、步长为 0.2 的不同隐私保护强度下的数据可用性，可以看出随着 ε 的增加，*MAE* 越来越小，这意味着较低的隐私保护水平具有更好的数据可用性。图 5(a)中，当 ε=0.1 时，CLM 的 *MAE* 为 12.1，而当 ε=0.9，*MAE* 下降到 4.1，在其他数据集中可以观察到同样的趋势。

此外，从图 5~图 7 可以看出，当 ε 从 0.1 上升到 0.4 时，本文中所有方法 *MAE* 下降得更快，这表明，要想达到更高的隐私级别（ε=0.1），需要牺牲更多的数据可用性。从图 5 还可以看出，当 $\varepsilon \geqslant 0.7$ 时，CLM 的性能较为稳定，这表明 CLM 在保证较高的数据可用性时仍然能够满足适当

的隐私保护需求。

图 5~图 7 的对比结果说明 CLM 机制在以下几个方面的有效性：1）与其他机制相比，CLM 可以保留较高的数据可用性；2）随着隐私强度的增加，CLM 的数据可用性表现显著增强。因此，可以设定一个合适的隐私强度值，以实现数据可用性和隐私保护强度更好的权衡；3）当有充足的隐私保护强度时，数据可用性的损失很小。

5.4　计算复杂度评估

图 8 给出了不同算法在不同查询数目下的计算复杂度。从图中可以看出，基于马尔可夫的算法性能最佳，因为该模型在实际的数据集中实施较为简单，而 CLM 机制的计算复杂度次优，大约是基于马尔可夫方法的 2 倍。CLM 机制每个查询的平均运行时间为 1.88 s，仍能保证较快的查询速度。此外，随着查询量的增大，基于 DFT 和 DWT 的方法的查询时间是线性增加的，而其他方法的计算复杂度几乎保持不变。这是因为每次当新的查询到来时，基于 DFT 和 DWT 变换的机制都需要对查询的序列进行变换。

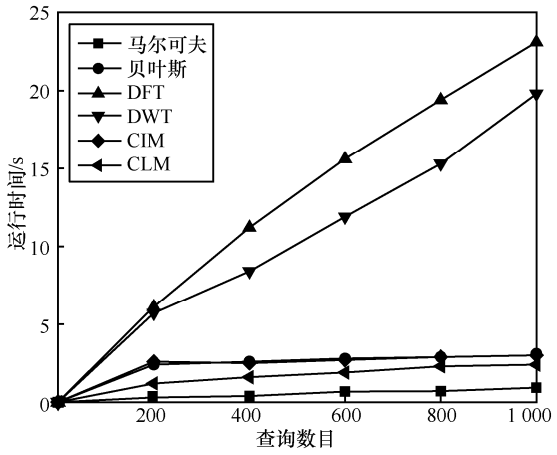


图 8　不同算法在不同查询数目下的计算复杂度

从图 8 可以得出，基于相关性建模的方法的计算复杂度低于基于变换的方式，此外，当查询的数据量越来越大时，CLM 机制仍然可以保证较高的噪声生成速度。

6　结束语

为了解决现有轨迹差分隐私发布方法中存在的独立噪声导致的隐私保护程度不足，并引入额外的噪声的问题，本文提出了一种相关性拉普拉斯机制的噪声添加机制，生成与原始待发布轨迹序列相关性一致的噪声序列，叠加到原始序列中发布。实验结果表明，与现有方法相比，对于轨迹数据来说，本文机制在保证差分隐私的前提下，仍能保证较高的数据可用性，且计算复杂度没有明显增加。

由于统计相关特性的方法只适用于相关性较强的平稳序列（如轨迹序列）中，对于非平稳序列来说，自相关函数的统计方式不太适用，接下来的工作将会研究本文机制在非平稳序列中的适用性及可扩展性；同时，也会尝试寻求高维数据的差分隐私发布方法，以期得到更好的分析挖掘结果。

参考文献：

[1] ZHENG Y. Trajectory data mining: an overview[J]. ACM Transactions on Intelligent Systems & Technology,2015,6(3): 1-41.

[2] 张扶桑, 金蓓弘, 汪兆洋, 等. 基于轨迹挖掘的公交车自组织网络路由机制[J]. 计算机学报, 2015,38(3):648-662.
ZHANG F S, JIN B H, WANG Z Y, et al. Routing mechanism of ad hoc network for bus based on trajectory mining[J]. Chinese Journal of Computers, 2015,38(3):648-662.

[3] 宋雪涛, 蒲英霞, 刘大伟, 等. 利用行人轨迹挖掘城市区域功能属性[J]. 测绘学报, 2015, 44(s1): 82-88.
SONG X T, PU Y X, LIU D W, et al. Mining urban regional functional attributes by pedestrian trajectory[J]. Acta Geodaetica et Cartographica Sinica, 2015, 44(s1): 82-88.

[4] HUA J, GAO Y, ZHONG S. Differentially private publication of general time-serial trajectory data[C]//Computer Communications. 2015: 549-557.

[5] CAO Y. Differentially private real-time data release over infinite trajectory streams[J]. IEICE Transactions on Information & Systems,2016,E99-D(No.1): 68-73.

[6] CHEN R, FUNG B C M, MOHAMMED N, et al. Privacy-preserving trajectory data publishing by local suppression[J]. Information Sciences, 2013, 231(1): 83-97.

[7] CHOW C Y, MOKBEL M F. Privacy of spatial trajectories[M]// Computing with Spatial Trajectories, 2011:109-141.

[8] GEDIK B, LIU L. Protecting location privacy with personalized *k*-anonymity: architecture and algorithms[J]. IEEE Transactions on Mobile Computing,2008,7(1): 1-18.

[9] 王璐, 孟小峰. 位置大数据隐私保护研究综述[J]. 软件学报,2014,25(4):693-712.
WANG L, MENG X F. A survey of the research on privacy preserving for location data[J]. Journal of Software,2014,25(4):693-712.

[10] MAGLOGIANNIS I, KAZATZOPOULOS L, DELAKOURIDIS K, et al. Enabling location privacy and medical data encryption in patient telemonitoring systems[J]. IEEE Transactions on Information Technology in Biomedicine a Publication of the IEEE Engineering in Medicine & Biology Society,2009,13(6): 946-954.

[11] DWORK C. Differential privacy[M]//Automata, Languages and Programming. Springer Berlin Heidelberg,2006: 1-12.

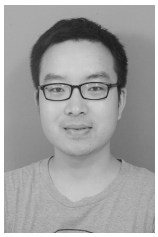
[12] SHEN E, YU T. Mining frequent graph patterns with differential privacy[C]//The 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2013: 545-553.

[13] YANG B, SATO I, NAKAGWA H. Bayesian differential privacy on correlated data[C]//SIGMOD/PODS. 2015: 747-762.
[14] XIAO X. Differentially private data release: improving utility with wavelets and Bayesian networks[M]//Springer International Publishing, 2014: 25-35.
[15] SHORKRI R. Privacy games: optimal protection mechanism design for Bayesian and differential privacy[J]. IEEE Transactions on Dependable & Secure Computing, 2014, 11(3): 266-279.
[16] ZHU T, XIONG P, LI G, et al. Correlated differential privacy: hiding information in non-IID data set[J]. IEEE Transactions on Information Forensics & Security, 2015,10(2):229-242.
[17] RASTOGI V, NATH S. Differentially private aggregation of distributed time series with transformation and encryption[C]//The ACM SIGMOD International Conference on Management of Data. 2010: 735-746.
[18] XIAO X. Differentially private data release: improving utility with wavelets and Bayesian networks[C]//Springer International Publishing. 2014:25-35.
[19] XIAO X, et al. Differential privacy via wavelet transforms[J]. IEEE Transactions on Knowledge & Data Engineering, 2009, 23(8): 23-30.
[20] JIANG W, XIE C, ZHANG Z. Wishart mechanism for differentially private principal components analysis[J]. Computer Science, 2015, 9285(2): 458-473.
[21] GEHRKE J, HAY M, LUI E, et al. Crowd-blending privacy[C]// Cryptology Conference on Advances in Cryptology. Springer-Verlag New York. 2012: 479-496.
[22] GEHRKE J, LUI E, PASS R. Towards privacy for social networks: a zero-knowledge based definition of privacy[C]//Theory of Cryptography - 8th Theory of Cryptography Conference. 2011: 432-449.
[23] KIFER D, MACHANAVAJJHALA A. Pufferfish: a framework for mathematical privacy definitions[J]. ACM Transactions on Database Systems, 2014, 39(1):1-36.
[24] HE X, MACHANAVAJJHALA A, DING B. Blowfish privacy: tuning privacy-utility trade-offs using policies[C]//The 2014 ACM SIGMOD International Conference on Management of Data. 2014: 1447-1458.
[25] CAO L, OU Y, YU P S. Coupled behavior analysis with applications[J]. IEEE Transactions on Knowledge & Data Engineering, 2012, 24(8):1378-1392.
[26] ZHOU Z H, SUN Y Y, LI Y F. Multi-instance learning by treating instances as non-IID samples[C]//The 26th Annual International Conference on Machine Learning. 2009: 1249-1256.
[27] RASTOGI V, HAY M, MILLAU G, et al. Relationship privacy: output perturbation for queries with joins[C]//The 28th ACM Sigmod-Sigact-sigart Symposium on Principles of Database Systems. 2009: 107-116.
[28] DWORK C, ROTH A. The algorithmic foundations of differential privacy[J]. Foundations and Trends in Theoretical Computer Science, 2014, 9(3-4):211-407.
[29] BOJAN, BASRAK. The sample autocorrelation function of non-linear time series[M]//University of Groningen, 2000.
[30] WANG H, XU Z Q. CTS-DP: publishing correlated time-series data via differential privacy[J]. Knowledge-Based Systems, 2017, 122: 167-179.
[31] ZHENG Y, XIE X, MA W Y. GeoLife: a collaborative social networking service among user, location and trajectory[J]. Bulletin of the Technical Committee on Data Engineering, 2010, 33(2):32-39.
[32] YUAN J, ZHENG Y, ZHANG C, et al. T-drive: driving directions based on taxi trajectories[C]//ACM SIGSPATIAL International Symposium on Advances in Geographic Information Systems. 2010:99-108.
[33] WANG H, TERROVITIS M, MAMOULIS N. Location recommendation in location-based social networks using user check-in data[C]//ACM Sigspatial International Conference on Advances in Geographic Information Systems. 2013:374-383.

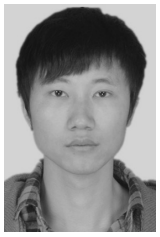
作者简介：

王豪（1990-），男，河南驻马店人，武汉大学博士生，主要研究方向为数据挖掘、隐私保护。

徐正全（1962-），男，湖北黄冈人，博士，武汉大学教授，主要研究方向为数据挖掘、隐私保护、图像处理等。



熊礼治（1988-），男，湖北荆州人，博士，南京信息工程大学讲师，主要研究方向为图像处理、信息安全。



王涛（1985-），男，河南开封人，博士，武汉大学副研究员，主要研究方向为并行与分布式系统、轨迹挖掘。