doi:10.11959/j.issn.1000-436x.2015305

面向时序数据发布的隐私保护方法研究

于东，康海燕
（北京信息科技大学 信息管理学院，北京 100192）

摘 要：针对动态数据（时序数据）提出一种抽样过滤技术的差分隐私保护模型及评价机制。首先，利用固定抽样法对原始时序数据进行抽样，非抽样数据直接发布；其次，对抽样数据采取差分隐私保护机制进行加噪；然后，运用 Kalman 过滤技术对保护后的抽样数据进行预测修正；最后，通过互信息评价机制对不同抽样间隔下的数据进行评价。通过实验证明抽样过滤机制在安全性和实用性上达到最优的平衡性。
关键词：差分隐私；时序数据；数据发布；抽样；Kalman 过滤；互信息
中图分类号：TP309 **文献标识码**：A

Privacy protection method on time-series data publication

YU Dong, KANG Hai-yan
(School of Information Management，Beijing Information Science and Technology University，Beijing 100192，China)

Abstract: A differential privacy model was proposed based on the sampling filtering and the mechanism of evaluation. Firstly, fixed sampling method was used to sample the original data and the non-sampling data be published directly. Secondly, for the sampling date, utilize the differential privacy mechanism to add the noise. Then, use Kalman to correct the sampling date. Finally, use the mutual information to evaluate data under different sampling intervals. Through the experiment, it is proved that the mechanism can achieve a good balance between the practicality and protective.
Key words: differential privacy; time-series data; data publication; sample; Kalman filter; mutual information

1 引言

随着网络的迅速发展，各种信息数据不断的扩大延伸,使信息技术的应用在当今社会下已经随处可见。以电子商务为例，电子商务作为信息技术运用频繁且发展迅速的行业，数据的共享不仅可以让公众获得和学习更多关于真实世界的知识,也为研究机构和政府对于市场消费水平和经济的走势提供了大量研究价值。数据的共享和发布给生活提供了许多的有利因素，但是如果被不法机构利用，将会给个人带来很大的危险。因此，对于拥有大量用户数据的行业机构来说，数据发布的安全性变得尤为重要。

数据共享的一个普遍的方式，一个被信任的机构从个体用户中收集大量的相关数据，然而这些被收集的数据不仅会被政府部门用于研究也可能被非法机构或个体以各种形式和目的进行持续的分享和收集。被信任的机构有责任和义务去保护用户的利益，必须确保发布的数据不会泄露，提供相关数据的任何个体隐私。本文的目标是保证信任机构持续不断（时间序列）的分享用户数据集时，个体用户隐私不被泄露且数据集具有特定的统计特性。

本文的主要贡献是针对时序数据提出了一种基于抽样过滤技术的差分隐私保护模型以及互信息评价机制，该模型可以使加噪后的数据更加接近

收稿日期：2015-10-23
基金项目：北京市社会科学基金资助项目（15JGB099）；北京市优秀人才培养基金资助项目（2013E005007000001）；国家自然科学基金资助项目(61370139)；教育部人文社会科学青年基金资助项目（11YJC870011）
Foundation Items: The Social Science Foundation of Beijing (15JGB099); The Excellent Talents Program of Beijing (2013E005007000001); The National Natural Science Foundation of China (61370139); Humanity and Social Science Youth Foundation of Ministry of Education (11YJC870011)

原始值，详细如图 1 所示。

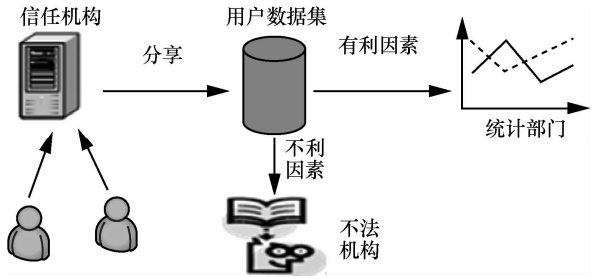


图 1　数据发布示意

1) 采用 Kalman 过滤技术对加噪后的数据进行预测再修正使发布数据更加接近原始值，使数据的实用性更高。

2) 采用抽样技术，提高系统效率。对抽样点的数据进行加噪及过滤，降低了整个系统的运行时间。

3) 首次将互信息的概念运用在差分隐私中，利用互信息选择最优的抽样间隔。

2　相关工作

研究发现针对静态数据发布的差分隐私机制无法满足时序数据发布，为了更好地完善差分隐私保护机制，国内外的学者对此作了突破性的研究。

Papadimitrios 等[1]提出以快速傅里叶变换（FFT）和离散的小波变换（DWT）为基础的 2 种算法去干扰时序数据的频率，从而达到噪音干扰与时间序列压缩属性之间的平衡。但是 2 种算法噪音不能保证满足差分隐私机制，对于掌握大量背景知识的攻击者来说，敏感信息保护是非常脆弱的。随后，在此基础上，Rastogi 等[2]提出一种建立在算法基础上的离散傅里叶变换（DFT），通过干扰离散的傅里叶系数 k 和从完全相反的离散的傅里叶变换中重新构建一个发布类型，进来保证满足差分隐私。但是不能应用在实时数据的发布中。Ny 等[3]运用 Kalman 过滤技术，对添加噪音后的时序数据进行修正，从而减少噪音的干扰,但是发布数据的实用性不高。Fan 等[4~6]研究中发现，Laplace 分布与高斯分布非常相似，文献将 Laplace 噪音模拟成近似的高斯噪音，将差分隐私噪音机制近似成系统测量模型，从而使 Kalman 过滤技术运用到差分隐私保护中。但是全部数据进行过滤处理会导致系统运作效率低下。

本文针对动态数据中的一类数据（即时序数据）进行研究。对 Kalman 过滤技术进行了优化改进，改进后的 Kalman 过滤技术通过对加噪后的数据进行优化修正，提高了数据的实用性。与抽样技术进行融合，提高了系统运行的效率，是一种基于抽样过滤技术的差分隐私模型。

3　差分隐私理论基础

定义 1　差分隐私[7]。设随机算法 K，Pr[z]表示事件 z 的披露风险。对于任意差别至多为一个记录的 2 个数据集 D_1 和 D_2，若算法 K 在数据集 D_1 和 D_2 上任意输出结果 D 满足

$$\Pr[K(D_1)=D] \leqslant \mathrm{e}^{\varepsilon} \Pr[K(D_2)=D] \tag{1}$$

则称算法 M 满足 ε-差分隐私保护。ε 为隐私预算，e 为自然底数。

定义 2　全局敏感度[8]。对任意一个函数 f：$D_1 \to R^d$ 函数 f 的全局敏感度为

$$\Delta f = \max \left\| f(D_1) - f(D_2) \right\|_p \tag{2}$$

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。

噪音机制

典型的差分隐私机制是 laplace 噪音机制和指数机制。形式定义如下所示。

1) Laplace 噪音

定义 3　Laplace 机制[8]。对于任意一个函数 f：$D_1 \to R^d$，算法 Y 满足下列等式，则 Y 满足 ε-差分隐私。

$$Y(D) = f(D) + Lap\left(\frac{\Delta f}{\varepsilon}\right) \tag{3}$$

其中，函数 $Lap(\ )$表示 Laplace 密度函数。

2) 指数机制

定义 4　指数机制[9]。给定一个打分函数 q：（$D\times O$）$\to R$，算法 K 满足下列等式，则 K 满足 ε-差分隐私。

$$\boldsymbol{K}(D,u) = \{r : |\Pr[r \in O] \propto \exp(\frac{\varepsilon q(D,r)}{2\Delta q})\} \tag{4}$$

若算法 K 以正比于 exp()的概率输出 r，那么 $K(D, u)$满足 ε-差分隐私。

4　基于差分隐私的时序数据发布算法

4.1　问题的描述

简单来说，时序数据就是按时间顺序记录的数据。时序数据的形式定义如下。

定义 5　时序数据。设有一个离散型单变量序列集 $X=\{x_1, x_2, \cdots, x_k\}$，$x_k$ 代表在离散时间点 k 上的统计值，$0 \leqslant k \leqslant T$，$T$ 是时间序列的长度，那么 X 所代表的是时序数据集。

本文研究的问题是时序数据是动态数据，随着时间的变化，数据不断更新。更新后数据发布必定包含前一个时间点更新发布的噪音，那么随着时间的推移，噪音的累积会使发布数据与原始数据有着巨大的差异，降低数据的实用性。因此，在时序数据的发布中，既能满足用户隐私安全又具有很高实用性的发布机制是迫切需要的。本文把 X 当成是一个计数序列，提出了一种基于抽样过滤技术的时序数据发布机制，对原始的时序数据集 X 进行隐私保护，保护后的发布数据集 $R=\{r_k\}$ 满足差分隐私保护且实用性好。

4.2　时序数据发布机制

基于抽样过滤技术的时序数据发布运行机制，如图 2 所示。

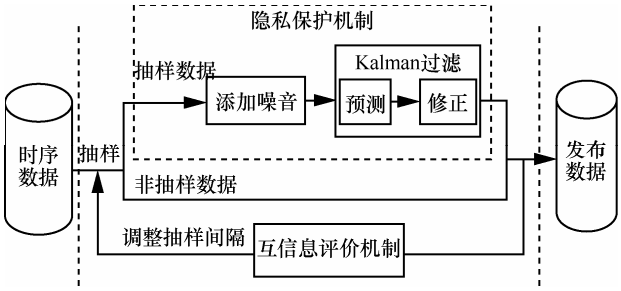


图 2　时序数据发布机制

抽样过滤机制具体步骤如下。

1) 首先确定抽样间隔 t，对原始的时序数据进行抽样，将数据分为抽样数据和非抽样数据。

2) 对抽样点的时序数据添加 Laplace 噪音。由于 Laplace 噪音添加值为随机值可能会造成发布数据与原始数据误差较大，影响实用性。因此，本文采用 Kalman 过滤系统进行预测修正,得出抽样点添加噪音后的最优估计值（即后验估计值）进而发布出去，且该机制满足差分隐私保护。

3) 对于非抽样点的时序数据直接发布。

4) 采用 Kalman 过滤技术，过滤机制的运作是将前一个时间点的发布数据值当作抽样点数据的先验估计值，然后通过抽样点添加噪音后的观测值与先验估计值之间的公式关系得出抽样点的后验估计值。

5) 进行多次不同时间间隔的固定抽样率抽样，通过互信息的评价机制，比较原始数据和发布数据之间的关系，确定最优的抽样间隔，进而确定最终发布数据。

基于差分隐私的时序数据发布算法，如算法 1 所示。算法 1 概括了时序数据发布机制的实现算法。运用抽样技术对 Kalman 过滤进行改进，可以很大程度地提高系统的效率和数据的实用性。实验将证明抽样过滤机制无论在实用性和系统效率上都是优于原始的 Kalman 过滤机制。

算法 1　时序数据发布算法

输入：时序数据集 X，隐私预算 ε，k 为任意时间点

输出：发布数据集 R

1) 自定义抽样间隔 t，for　t=1 to n do

2) 数据按每秒导入，　　if　$k \% t==0$

3) $z_k \leftarrow x_k + Lap(0, \frac{r}{a})$

4) Kalman 过滤，①预测：$\hat{x}_k = \hat{x}_{k-1}$，②修正：$\hat{x}_k = \hat{x}_{\bar{k}} + K_k(z_k - \hat{x}_{\bar{k}})$

5) 发布数据 $r_k \leftarrow \hat{x}_k$

6) else

7) 发布数据 $r_k \leftarrow x_k$

8) k++

9) end

10) 求出互信息值 $I(X,R)$，评价 t

4.3　过滤机制

差分隐私保护中，过滤机制的应用极其的稀少。通过对过滤技术的研究，本文最终选择了 Kalman 滤波[10]。Kalman 最先提出基于状态空间的递推滤波算法，即 Kalman 滤波算法。本文对 Kalman 滤波进行优化，使差分隐私保护后的发布数据更加精确且系统运行效率较高。

系统过程模型（线性方程模型）如下

$$X_{k+1} = X_k + U_k + \omega \tag{5}$$

其中，X_k 是离散时间点 k 上的系统状态，即 k 时刻原始数据的值，U_k 是 k 时刻对系统的控制量，本文中 U_k=0。ω 为高斯白噪音，它服从期望为 0,方差为 Q 高斯分布

$$\omega \sim N(0,Q)$$

系统测量模型如下

$$Z_k = X_k + v \tag{6}$$

其中，Z_k 表示系统的观测值，即原始数据值加上系

统的噪音值，v 为高斯白噪音（测量噪音），它是服从期望为 0，方差为 R 的高斯分布

$$v \sim N(0,R)$$

系统测量模型与 laplace 噪音模型十分相似，从式(6)和式(3)不难看出两者区别在于噪音服从分布上。通过高斯噪音模拟 Laplace 噪音，Kalman 过滤便可应用于差分隐私保护中（R 的取值）。Fan 等[4]结合 Kalman 过滤的相关公式，给出了高斯噪音模拟 Laplace 噪音时，方差 R 的最优取值方程。

定义 6 假设在每个时间点添加的噪音分布服从 $Lap\left(0,\frac{T}{\varepsilon}\right)$，然后将服从 $N(0，R)$分布的高斯噪音入后验误差方程 $\mathrm{var}(x_k-\hat{x}_k)$ 中，得到

$$\mathrm{var}(x_k-\hat{x}_k)=\frac{R^2[\mathrm{var}(x_{k-1}-\hat{x}_{k-1})+Q]}{(P_{\bar{k}}+R)}+\frac{2P_{\bar{k}}^2T^2}{(P_{\bar{k}}+R)^2\varepsilon^2} \tag{7}$$

显然，要得到最优的后验估计值，必须满足 $R=\frac{T^2}{\varepsilon^2}$。

上述解决了 Kalman 过滤在差分隐私中的使用性问题。设时间点 k 上的先验估计值为 $\hat{x}_{\bar{k}}$；后验估计值为 $\hat{x}_k$；先验误差方差为 $p_{\bar{k}}$；后验误差方差为 p_k；抽样点数据加噪值 z_k；k 为抽样点。通过式(4)可以得出时间点 k 上的先验估计值 $\hat{x}_{\bar{k}}$ 等于上一个时间点的后验估计值

$$\hat{x}_{\bar{k}}=\hat{x}_{k-1} \tag{8}$$

式(8)是对抽样点数据值进行预测，得出抽样点数据的先验估计值。关于先验误差方差和上一个时间点的后验误差方差的关系，Kalman 又给出了一个关系式(Q 为方差）

$$p_{\bar{k}}=p_{k-1}+Q \tag{9}$$

式(8)和式(9)为 Kalman 滤波的时间更新方程，它是对原始的时序数据值进行预测，得出下一个时间状态的先验估计值。Kalman 过滤最终的目的是修正加噪后抽样点数据值，即得出抽样点数据的后验估计值

$$\hat{x}_k=\hat{x}_{\bar{k}}+K_k(z_k-\hat{x}_{\bar{k}}) \tag{10}$$

式(10)可以求抽样点数据后验估计值。其中，K_k 表示 Kalman 增益，K_k 不是个固定的值，它在每一个时刻都会进行调整，从而使后验误差方差尽可能的最小化。Kalman 增益的方程式（R 为方差）

$$K_k=p_{\bar{k}}(p_{\bar{k}}+R)^{-1} \tag{11}$$

其中，$p_{\bar{k}}$ 为先验误差方差，本文数据为计数数据，因此研究的内容是在一阶矩阵中，先验误差方差具体求解（E(x)表示 x 的期望值）为

$$p_k=\mathrm{E}[(x_k-\hat{x}_{\bar{k}})^2] \tag{12}$$

通过上述公式，得出抽样点 k 上数据的后验估计值，即抽样点数据的发布值。通过得出的 Kalman 增益，可以得出同一时间点先验误差方差和后验误差方差的关系

$$p_k=(1-K_k)p_k \tag{13}$$

本文所使用 Kalman 过滤算法是通过时间预测方程和测量更新方程，不断地对抽样点数据进行预测和修正。后续实验将证明使用 Kalman 过滤技术对数据发布实用性上的优势。

4.4 抽样技术

本文发现对每个时间点的数据都采用 Kalman 过滤技术预测和修正，虽然提高了精确性，但会使系统的效率降低。采用抽样技术，对抽样点的数据进行过滤，不仅可以减少隐私预算的消耗，提高系统整体效率，而且还可以大大降低噪音给整体数据带来的干扰，提高数据的可用性。下面介绍 2 种抽样方法即随机抽样法与固定抽样法，并比较在差分隐私运用中的优劣之处。

随机抽样方法。简单随机抽样是指从总体直接抽取个体。预先确定好抽样点的个数，随机的从数据中抽取数据，然而这种抽样方法存在着一定的随机性（不确定性）。随机抽样方法存在明显的缺陷：数据发布保护不确定性。随机抽样方法中抽样点选择的差异，对发布数据的保护性和实用性有着很大的影响，如抽样点密集在数据的一部分区域内，那么这一部分区域的保护性较强，反之亦然。

固定抽样率方法。本文采用的是固定抽样率方法，给定一个自定义的抽样间隔 t，按照这种抽样间隔定期的抽取时序数据，然后将每个抽样点数据的后验估计值发布出去，在 2 个临近抽样点之间的数据（非抽样点数据）直接发布原始值。隐私预算将被分配给每个抽样点，根据“序列组合性”原则，进而保证全部的数据满足 ε-差分隐私。

抽样点的选择直接影响着发布数据的实用性和安全性。抽样点选择过多，噪音量越大，数据实

用性不高。同样的，抽样点选择过少，发布数据可以更好体现原始数据的“原貌”，但数据安全性也随之降低。因此，怎样决定最优的抽样间隔 t 是该机制的挑战之一。

4.5 互信息评价机制

本文引进信息熵和互信息的观点，通过将不同抽样点下的发布数据集与原始数据集进行比较得出互信息值，然后利用平均值法，将最接近均值的抽样点，作为实用性和安全性达到较优的抽样点。

4.5.1 信息熵

Shannon[11]首次提出“信息熵”的概念，解决了信息的度量问题。离散型随机变量 X 的信息熵

$$H(x)=-\sum p(x)\log_b^{p(x)} \tag{14}$$

其中，$p(x)$为发生事件 x 的概率。本文采用 b=2 的比特计量。

4.5.2 互信息

互信息[12](mutual information)表示 2 个变量或多个变量之间共享的信息量。互信息越大，变量之间的相关性越强，即原始数据集与发布数据集相关性越强。互信息定义如下

$$I(X,Y)=H(X)+H(Y)-H(X,Y) \tag{15}$$

互信息越大，数据实用性越强。用互信息去评价发布数据与原始数据之间的关系，是新颖可行的。

4.5.3 时序数据的互信息计算

时序数据是离散型函数，使用最普遍的方法是插入法，其中包括等概率法和等间距法。

等概率法[13]将数据值从小到大进行重排列,然后按落入每组的数据值相等来对变量进行等概率划分。对时序数据而言，计算较为繁琐并且费时,给应用带来不便。

等间距法[14]按数据值的取值范围，等间隔划分为若干组,然后分别统计每组中的数据数量。等间距法在计算上较为简单,易于理解。本文选用等间距法进行离散信息熵的计算。

时序数据是一系列的随机数据，没有固定的分布特征。确定分组数的经验公式为

$$m=1.87(n-1)^{\frac{2}{5}} \tag{16}$$

其中，n 为数据量。利用一维和二维等间距法，求出原始数据集和发布数据集的互信息，在运用平均值法将最接近均值的抽样点，作为实用性和安全性达到较优的抽样点。

5 实验与分析

5.1 实验数据与环境

实验数据集来自数据堂[15]某电子商务平台上某卖家 2012 年 6 月交易明细，包含 24 104 条交易记录。数据格式为“订单号/交易日期/销售单价/购买数/成交金额/用户 ID/性别/”。实验中删去销售单价和购买数 2 个属性，只把成交金额看作敏感属性进行差分隐私保护。抽样过滤算法使用 Java 语言实现，编程环境为 Eclipse 8.5，实验环境为 Windows7 2.60 GHz，4.0 GB，数据挖掘工具 SQL Server 2008R2。

5.2 过滤机制中参数的选择

过滤机制中，参数 Q、R 的选择直接影响到整个机制的精确性（不涉及抽样）。Q 值的选择与数据源本身的内部状态有关，是通过对数据源本身状态的观测或查看历史数据状态来确定的。R 值的选择以式(7)的取值为指导。实验通过对隐私预算 ε=1 时 R 值的选择对比，验证了 R 的取值方程，如图 3 所示。图中横坐标代表数据量，设定数据每秒钟更新一次。明显地看出 R=10^6 时发布数据的相对误差越小且越平稳。隐私预算平均分配给每个抽样点，即 $\left(\frac{\varepsilon}{T}\right)$，那么近似于 $Lap\left(0,\frac{T}{\varepsilon}\right)$ 噪音的高斯噪音为 $N\left(0,\frac{T^2}{\varepsilon^2}\right)$，因此，$R$=$\left(\frac{T^2}{\varepsilon^2}\right)$。

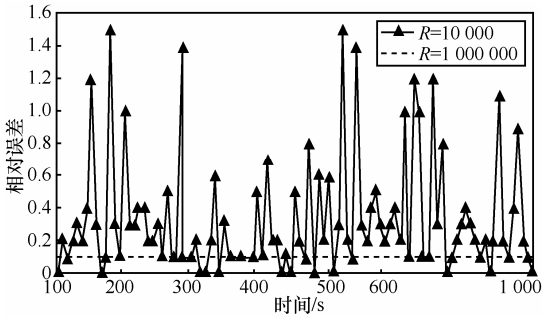


图 3　Kalman 过滤参数 R 的选择

5.3 抽样间隔的确定

利用等间距法，对 100 个抽样间隔下的数据进行计算，得出不同抽样点下原始数据与发布数据的互信息值，从而得到最优的抽样间隔。图 4 展示了不同抽样间隔下互信息值的变化趋势，可以看出：1）随着抽样点间隔的增大，互信息值越大；2）互

信息的增长量变化集中在[1,12]抽样间隔下。利用平均值法，对 100 个抽样间隔下的互信息值取均值，得出互信息 $I(X,Y)$=2.536。再与不同抽样间隔下的互信息值比较，最终确定 t=6 时，I（X，Y）=2.528 与均值最为接近。t=6 为最优抽样间隔。

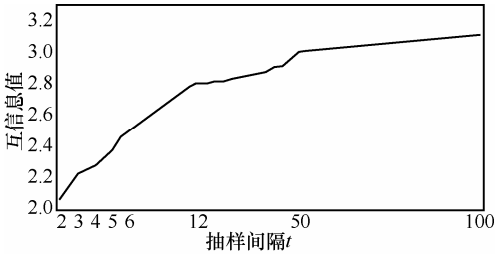


图 4 不同抽样间隔下互信息变化趋势

5.4 性能测试与分析

5.4.1 系统效率测试与分析

为了验证改进后的 Kalman 过滤在系统效率上的优越性，让本文提出的时序数据发布机制与原始的 Kalman 过滤机制以及 Laplace 噪音机制进行比较。为了简化实验，本文选择 1 000 条数据进行实验。实验过程中，时序数据发布机制最优抽样间隔 t 的选择上述实验已经证明（t=6），3 种差分隐私机制的系统效率比较与隐私预算无关，选择隐私预算参数为 ε=1。图 5 展示了 3 种差分隐私机制下系统效率的比较。

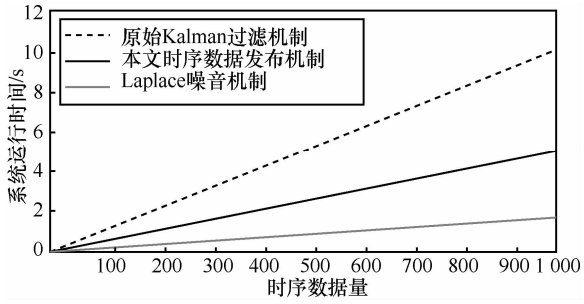


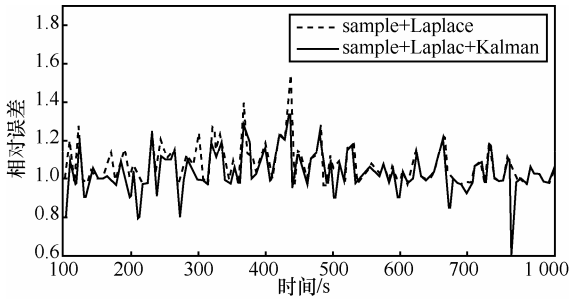
图 5 3 种差分隐私保护机制的系统效率比较

通过改进 Kalman 过滤技术后提出的时序数据发布机制（sample+Laplace+Kalman）在系统运行时间上明显的低于原先的 Kalman 过滤机制。时序数据发布机制由于只是对抽样点数据进行加噪过滤，系统的运行时间也低于对全部数据加噪的 Laplace 噪音机制。与传统的差分隐私保护机制进行比较，抽样过滤机制也有着显著的优势。

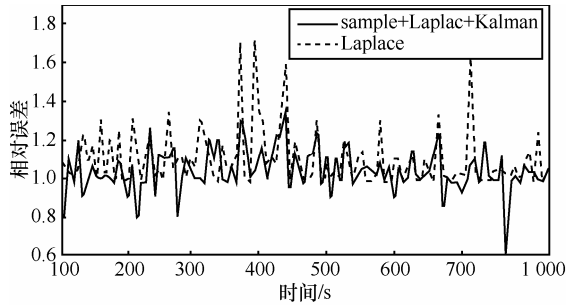
5.4.2 数据实用性测试与分析

为了验证本文提出的时序数据发布机制（sample+Laplace+Kalman）的数据实用性，本文与传统的两种差分隐私技术（添加噪音机制（Laplace）与抽样加噪机制（sample+Laplace））进行比较实验。在数据安全性方面，3 种方法均使用差分隐私进行保护，数据安全性相当。在数据实用性测试方面，由于 3 种机制采用不同的机制算法，本文将通过实验进行对比，得出 3 种机制中数据实用性和安全性达到最优平衡的差分隐私保护机制。

参考以往文献，分别选取隐私预算参数 ε=1,0.1,0.01 的 3 种情况进行实验比较，时序数据集设定为每秒更新一次。实验通过得出 3 种保护机制之间的相对误差大小，从而判断发布数据实用性的优劣(当 ε=0.1，ε =0.01，ε =1 时，3 种隐私预算下的数据实用性比较趋势相当且由于篇幅原因，本文只给出 ε=1 时的比较)，如图 6 所示。



（a）时序数据发布机制与抽样加噪机制比较



（b）时序数据发布机制与加噪机制比较

图 6 时序数据发布实用性比较

本文提出的时序数据发布机制的数据实用性明显高于添加噪音机制（Laplace）与抽样加噪机制（sample+Laplace），图 6 中时序数据发布机制（sample+Laplace+Kalman）与抽样加噪机制（sample+Laplace）有部分区域近似于重叠，相对误差区分不明显，因为添加的噪音值自身过小导致过滤后的发布值优化不明显导致。

本文提出的时序数据发布机制（sample+Laplace+Kalman）是实用性和保护性达到最优平衡的数据发布机

制且该机制在系统效率上也是十分优越的。

6　结束语

本文针对电子商务数据提出了基于差分隐私的数据发布算法。首先，对差分隐私的基本原理和性质进行了阐述，其次，对现有的动态时序数据发布的文献进行了总结，并指出了不足之处。最后，提出了一种基于抽样过滤技术的差分隐私保护模型及新型的互信息评价机制。通过实验，将抽样过滤机制与常见的两种差分隐私保护机制进行比较，结果显示，抽样过滤算法在保护性和实用性上达到最优的平衡性。

参考文献：

[1] PAPADIMITRIOU S, LI F, KOLLIOS G., *et al*. Time series compressibility and privacy[A]. ser VLDB '07[C]. VLDB Endowment, 2007. 459-470

[2] RASTOGI V, NATH S. Differentially private aggregation of distributed time series with transformation and encryption[A]. SIGMOD[C]. 2010. 735-746.

[3] NY J L, PAPPAS G J. Differentially private Kalman filtering[A]. Proceedings of the Annual Allerzon Conference on Communication, Control and Computing[C]. 2012. 1618-1625.

[4] FAN L, XIONG L. Adaptively sharing series time with differential privacy[J].CoRR, 2012, 2012.

[5] FAN L, XIONG L. Real-time aggregate monitoring with differential privacy[J]. IEEE Transactions on, Knowledge & Data Engineering 2012,26(9):2169-2173.

[6] FAN L, XIONG L. An adaptive approach to real-time aggregate monitoring with dif-ferential privacy[J]. IEEE Transactions on Knowledge & Data Engineering, 2013, 26(9):1.

[7] DWORK C. Differential privacy [A]. Proceedings of the 33rd International Colloquium on Automata, Languages and Programming (ICALP)[C]. Venice, Italy, 2006. 1-12.

[8] DWORK C，Calibrating noise to sensitivity in private data analysis[A]. Proceedings of the 3th Theory of Crypto-graphy Conference (TCC)[C]. New York, USA, 2006. 363-385.

[9] Mcsherry F, TALWAR K. Mechanism design via differential privacy[A]. Proceedings of the48th Annual IEEE Symposium on Foundations of Computer Science (FOCS)[C].Providence RI,USA, 2007. 94-103.

[10] KALMAN R E, A new approach to linear filtering and prediction problems[J]. Basic Eng 1960, 82, 35-45,doi:10.1115/1.3662552.

[11] SHANNON C A. Mathematical Theory of Communication[J]. Bell System Technical Journal, 1948. 27:379-423.

[12] Cover. Elements of Information Theory[M]. The United States: Wiley-Black well. 2006.

[13] 龙海辉，张佃中. 基于等概率符号分析方法计算互信息确定延迟时间[J]. 计算物理， 2010, (3): 155-161.
LONG H H, ZHANG D Z. Calculate the mutual information to determine the delay time based on the equal probability symbols analysis method [J]. Chinese Journal of Computational Physics, 2010, (3): 155-161

[14] 丁晶, 王文圣, 赵永龙. 以互信息为基础的广义相关系数[J]. 四川大学学报(工程科学版), 2002,34(3):1-5.
DING J, WANG S W, ZHAO Y L. General correlation coefficient between variables based on mutual information [J]. Journal of Sichuan University (Engineering Science Edition), 2002, 34(3):1-5.

[15] http://datatang.com/Datatang, http://datatang.com/[EB/OL].2012.

作者简介：

于东（1989-），男，安徽蚌埠人，北京信息科技大学硕士生，主要研究方向为物流信息安全。

康海燕[通信作者]（1971-），男，河北石家庄人，北京信息科技大学教授、博士生导师，主要研究方向为信息系统安全和网络隐私保护。