Bivariate Correlation-Based Attack: A Challenge for Privacy Preservation of Location Sequence Data

Lihui Mao [1], Zhengquan Xu [1,2]

[1] State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China - maolihui@whu.edu.cn

[2] Collaborative Innovation Center for Geospatial Technology, Wuhan 430079, China - xuzq@whu.edu.cn

Keywords: Location Privacy Preservation, Differential Privacy, Geo-indistinguishability, Series-indistinguishability, Correlated Laplace Mechanism, Correlation-based Attack.

Abstract

Data correlation is a critical issue for location sequence data privacy protection. Series-indistinguishability provides a theoretical basis for the differential privacy protection of temporally correlated location data and is implemented by the correlated Laplace mechanism (CLM), which has become a novel privacy protection method. However, location sequence data is essentially a multivariate time series with data correlation not only within each dimension but also between different dimensions. The CLM-based location privacy scheme only adapts to the auto-correlation in each dimension and ignores cross-correlation between different dimensions, which compromises its privacy performance. To evaluate its actual privacy protection performance, we propose a bivariate correlation-based attack (BCA) utilizing a filtering method and theoretically derive the optimal filter parameters. Based on simulations and real data experiments, the results show that the privacy performance of the CLM-based privacy scheme is significantly reduced under BCA, confirming that this scheme cannot achieve complete series-indistinguishability for bivariate sequences. Furthermore, the results suggest that BCA is an effective tool for privacy performance evaluation.

1. Introduction

Location data privacy protection is a significant issue in the big data era, and several works (Wernke et al., 2014; Chatzikokolakis et al., 2017; Liu et al., 2018; Jiang et al., 2021) provide systematic reviews of the research in this area. Among the existing privacy definitions, differential privacy (Dwork, 2006; Dwork et al., 2006) is based on the idea of "computational indistinguishability" in cryptology and defines privacy protection using a probabilistic statistical model, which guarantees that the actual privacy strength is not affected by the attacker's background knowledge, thus becoming one of the preferred privacy protection techniques. Furthermore, geo-indistinguishability (Andrés et al., 2013) extends the idea of indistinguishability to the two-dimensional continuous space, providing a theoretical basis for the differential privacy preservation of single-point locations.

However, the privacy problem of location sequence data is more serious because of temporal correlation. Some studies (Shokri et al., 2011; Yang et al., 2015; Cao et al., 2017; Wang et al., 2021) have shown that this makes it difficult for many differential privacy mechanisms to achieve the desired effect. For this problem, many works take a modeling approach to describe the correlation between locations, e.g. kinematics (Jiang et al., 2013), linear prediction (Chatzikokolakis et al., 2014; Al-Dhubhani and Cazalas, 2018), Markov (Xiao and Xiong, 2015; Xiong et al., 2019), etc., and consequently adjust the privacy preserving approach to make the perturbed location sequences more consistent with actual patterns. These methods achieve privacy protection for location sequence data in specific application scenarios. For time series data, series-indistinguishability (Wang and Xu, 2017) has been proposed from the perspective of resisting the correlation-based attack (Wang et al., 2021). It requires that the time series data before and after differential privacy perturbation are consistent in terms of data correlation, which can be achieved by the correlated Laplace mechanism (CLM), providing a systematic approach to differential privacy preservation for correlated time series data. However, there are still some problems in applying it to the privacy protection of location sequence data.

Location data are typically multivariate variables represented by 2D or 3D coordinate values. Without loss of generality, this paper considers the location data to be represented as a bivariate variable in a planar Cartesian coordinate system, and accordingly, location sequence data are bivariate time series. Compared to univariate time series, the correlation of location data is much more complex, as it exists not only within the same dimension, but also to some extent between different dimensions. However, existing studies typically address the data correlation problem by assuming that the original data is a univariate sequence, thereby simplifying the problem. For high-dimensional location sequence data, existing studies usually adopt certain simplified methods to transform them into univariate sequences, such as spatial gridding (Xiao and Xiong, 2015; Xiong et al., 2019), separate treatment for each dimension (Chatzikokolakis et al., 2014), etc. Similarly, although series-indistinguishability and CLM are also ideologically applicable to multivariate sequences, they still treat univariate sequences as the object in practical applications. For location sequence data, CLM-based privacy schemes apply CLM in different dimensions separately. However, from the perspective of series-indistinguishability, this scheme is incomplete because it completely ignores the correlation between different dimensions. In fact, there is non-negligible cross-correlation between different dimensions in location sequence data, which allows the attacker to launch an attack by simultaneously exploiting the auto-correlation in each dimension and the cross-correlation between different dimensions, referred to in this paper as the bivariate correlation-based attack (BCA). Incomplete series-indistinguishability can make it difficult for privacy-preserving schemes to resist BCA, resulting in less effective privacy preservation.

Correlation-based attacks on bivariate time series, especially for location sequence data, have been poorly studied. There are two main issues: 1) whether effective BCA can be achieved; and 2) the performance of the CLM-based location sequence privacy scheme under BCA. To address these problems, this paper proposes a filter-based implementation of BCA based on the results of correlation characteristic analysis of real location sequence data and verifies the privacy performance loss of CLM-based privacy schemes under BCA through simulation and real data experiments. Our main contributions can be summarized as follows:

1. The correlation characteristics of real location sequence data are analyzed, revealing significant correlation between different dimensions, which enable the bivariate correlation-based attack.

2. A filter-based implementation of the bivariate correlation-based attack is proposed, and the optimal filter parameters are theoretically derived, providing an effective tool for privacy performance evaluation.

3. Based on simulations and real data experiments, the effectiveness of the bivariate correlation-based attack is verified, and the results confirm that the CLM-based location sequence privacy scheme is incomplete in terms of series-indistinguishability.

The remainder of this paper is organized as follows: Section 2 introduces the basic model of location sequence data privacy protection and its privacy requirements; Section 3 details the CLM-based privacy protection scheme; Section 4 presents the filter-based bivariate correlation-based attack and the theoretical optimal filter parameter computation method; We report the experimental evaluation in Section 5 and give the conclusion in Section 6.

2. Preliminary

In this paper, let $\mathbf{L}(I)=[\mathbf{l}(0),\cdots,\mathbf{l}(i),\cdots,\mathbf{l}(I)]^T$ denote the location sequence data from successive observations, where T indicates the matrix transposition, $\mathbf{l}(i)=[x(i),y(i)]^T$ denotes the i-th observed location data represented by two-dimensional coordinate values in the planar Cartesian coordinate system XOY. For privacy protection, $\mathbf{l}(i)$ should be perturbed by the privacy mechanism $\mathcal{M}$,

$$\tilde{\mathbf{l}}(i)=\mathcal{M}[\mathbf{l}(i)]=\mathbf{l}(i)+\mathbf{n}(i) \tag{1}$$

where $\mathbf{n}(i)=[n_x(i),n_y(i)]^T$ is the noise introduced by $\mathcal{M}$, $\tilde{\mathbf{l}}(i)=[\tilde{x}(i),\tilde{y}(i)]^T$ is the perturbed location data used for release, the perturbed location sequence data is denoted as $\tilde{\mathbf{L}}(I)=[\tilde{\mathbf{l}}(0),\cdots,\tilde{\mathbf{l}}(i),\cdots,\tilde{\mathbf{l}}(I)]^T$.

To protect privacy, it should be guaranteed that the attacker cannot accurately infer the true location $\mathbf{l}(i)$ based on the published data $\tilde{\mathbf{L}}(I)$, which poses two basic requirements for the privacy mechanism $\mathcal{M}$,

1. The attacker cannot accurately infer $\mathbf{l}(i)$ from $\tilde{\mathbf{l}}(i)$ at each moment. Based on geo-indistinguishability, if $\mathcal{M}$ satisfies ε-geo-indistinguishability,

$$\frac{Pr\{\mathcal{M}[\mathbf{l}(i)]=\tilde{\mathbf{l}}(i)\}}{Pr\{\mathcal{M}[\mathbf{l}'(i)]=\tilde{\mathbf{l}}(i)\}}\le exp\{\varepsilon\cdot d_E[\mathbf{l}(i),\mathbf{l}'(i)]\} \tag{2}$$

it is difficult for an attacker to accurately distinguish $\mathbf{l}(i)$ from its surrounding location $\mathbf{l}'(i)$ according to $\tilde{\mathbf{l}}(i)$, thus providing privacy protection for single-point locations, where $Pr(\cdot)$ denotes probability, $\mathbf{l}'(i)=[x'(i),y'(i)]^T$ is a location near $\mathbf{l}(i)$ with a Euclidean distance $d_E[\mathbf{l}(i),\mathbf{l}'(i)]$.

2. The attacker cannot accurately infer $\mathbf{l}(i)$ from $\tilde{\mathbf{L}}(I)$ by exploring data correlation. Based on series-indistinguishability, if the auto-correlation functions of $\mathbf{L}(I),\tilde{\mathbf{L}}(I)$, denoted as $\mathbf{r}_{\mathbf{LL}}(i,i-\tau),\mathbf{r}_{\tilde{\mathbf{L}}\tilde{\mathbf{L}}}(i,i-\tau)$, are equal after normalization,

$$\mathbf{r}_{\tilde{\mathbf{L}}\tilde{\mathbf{L}}}(i,i-\tau)\sim\mathbf{r}_{\mathbf{LL}}(i,i-\tau) \tag{3}$$

the attacker can hardly infer $\mathbf{l}(i)$ by using correlation-based attacks, where τ denotes the lag of the auto-correlation function, $\sim$ indicates that the two functions are equal after normalization.

It should be noted that for a univariate sequence, the normalized equality of the correlation function is well defined, i.e., both sides of the symbol $\sim$ are equal after normalization according to the result at $\tau=0$, for which the correlated Laplace mechanism (CLM) provides an implementation. However, $\mathbf{L}(I)$ is a bivariate sequence, and its auto-correlation function $\mathbf{r}_{\mathbf{LL}}(i,i-\tau)$ is a 2×2 matrix,

$$\mathbf{r}_{\mathbf{LL}}(i,i-\tau)=E[\mathbf{l}(i)\mathbf{l}^T(i-\tau)]=\begin{bmatrix} r_{xx}(i,i-\tau) & r_{xy}(i,i-\tau)\\ r_{yx}(i,i-\tau) & r_{yy}(i,i-\tau)\end{bmatrix} \tag{4}$$

where $r_{xx}(i,i-\tau),r_{yy}(i,i-\tau)$ denote the auto-correlation functions of sequence data in the X and Y directions, respectively, $r_{xy}(i,i-\tau),r_{yx}(i,i-\tau)$ denote the cross-correlation functions between X and Y directions. This applies similarly to the auto-correlation functions of $\tilde{\mathbf{L}}(I),\mathbf{N}(I)$, denoted as $\mathbf{r}_{\tilde{\mathbf{L}}\tilde{\mathbf{L}}}(i,i-\tau),\mathbf{r}_{\mathbf{NN}}(i,i-\tau)$, respectively. The correlation problem of bivariate sequences is much more complex than that of univariate sequences. Although the idea of series-indistinguishability is also applicable to multivariate sequences, there is no clear conclusion on what requirements the multivariate correlation function should satisfy in practical implementations. Additionally, the corresponding implementation mechanism is also lacking, which will be the focus of our future research.

3. The CLM-based Privacy Protection Scheme for Location Sequence Data

For the bivariate sequence $\mathbf{L}(I)$, a simple approach is to separate it into two univariate sequences: $\mathbf{X}(I)=[x(0),\cdots,x(I)]^T$ and $\mathbf{Y}(I)=[y(0),\cdots,y(I)]^T$. Then, CLM is applied individually to generate noise sequences, $\mathbf{N}_\mathbf{X}(I)=[n_x(0),\cdots,n_x(I)]^T$, $\mathbf{N}_\mathbf{Y}(I)=[n_y(0),\cdots,n_y(I)]^T$, resulting in the perturbed sequence data $\tilde{\mathbf{X}}(I)=[\tilde{x}(0),\cdots,\tilde{x}(I)]^T$, $\tilde{\mathbf{Y}}(I)=[\tilde{y}(0),\cdots,\tilde{y}(I)]^T$

from which the perturbed location sequence is constructed, $\tilde{\mathbf{L}}(I)=[\tilde{\mathbf{X}}(I),\tilde{\mathbf{Y}}(I)]$.

In this scheme, $\mathbf{X}(I)$ and $\tilde{\mathbf{X}}(I)$, $\mathbf{Y}(I)$ and $\tilde{\mathbf{Y}}(I)$ satisfies series-indistinguishability,

$$\begin{cases} r_{\tilde{x}\tilde{x}}(i,i-\tau)\sim r_{xx}(i,i-\tau) \\ r_{\tilde{y}\tilde{y}}(i,i-\tau)\sim r_{yy}(i,i-\tau) \end{cases} \tag{5}$$

Moreover, the scheme completely ignores the correlation between $\mathbf{X}(I),\mathbf{Y}(I)$, and the generated noise sequences $\mathbf{N}_{\mathbf{X}}(I),\mathbf{N}_{\mathbf{Y}}(I)$ are independent of each other, thus we can get the following result,

$$\begin{cases} r_{\tilde{x}\tilde{y}}(i,i-\tau)=r_{xy}(i,i-\tau) \\ r_{\tilde{y}\tilde{x}}(i,i-\tau)=r_{yx}(i,i-\tau) \end{cases} \tag{6}$$

That is, the scheme makes the four corresponding components in the correlation function $\mathbf{r}_{\tilde{\mathbf{L}}\tilde{\mathbf{L}}}(i,i-\tau)$, $\mathbf{r}_{\mathbf{LL}}(i,i-\tau)$ are equal after normalization.

Essentially, the CLM-based scheme first performs dimensionality reduction on the multivariate sequence, and then applies CLM to each univariate sequences obtained. The final effect is to make the corresponding components in the correlation function before and after the perturbation satisfy the normalized equivalence, thereby extending CLM to multivariate sequence privacy preservation. However, this method ignores the correlation between different dimensions, failing to fully exploit the data correlation. Therefore, its actual privacy performance requires further evaluation.

4. Bivariate Correlation-based Attack

By analyzing the correlation characteristics of actual location sequence data, we found that the cross-correlation between X and Y is significant and cannot be ignored (details are given in Section 5.3). This allows the attacker to launch an attack on the privacy scheme by simultaneously exploiting $r_{xx}(i,i-\tau)$, $r_{yy}(i,i-\tau)$ and $r_{xy}(i,i-\tau), r_{yx}(i,i-\tau)$. This paper refers to this attack mode as the Bivariate Correlation-based Attack (BCA).

Filtering is a commonly used analysis method in the field of time series. In this paper, we consider the bivariate correlation-based attack achieved by filtering. The basic idea is that let $\tilde{\mathbf{L}}(I)$ pass through a specific filter to obtain the best estimate about $\mathbf{L}(I)$ under the minimum mean square error criterion, thereby eliminating the privacy protection effect as much as possible. Let $\hat{\mathbf{L}}(I)=[\hat{\mathbf{l}}(0),\cdots,\hat{\mathbf{l}}(i),\cdots,\hat{\mathbf{l}}(I)]^T$ denote the estimated result, where $\hat{\mathbf{l}}(i)=[\hat{x}(i),\hat{y}(i)]^T$ is the estimate about $\mathbf{l}(i)$. The filtering process can be expressed as follows,

$$\hat{\mathbf{l}}(i)=\sum_{m=0}^{M}\mathbf{h}_m\tilde{\mathbf{l}}(i-m) \tag{7}$$

where M is the filter's order, the parameter $\mathbf{h}_m$ is 2×2 real matrix, $m=0,1,\cdots,M$. Then, the estimate error $\mathbf{e}(i)=[e_x(i), e_y(i)]^T$ can be calculated as follows,

$$\mathbf{e}(i)=\mathbf{l}(i)-\hat{\mathbf{l}}(i)=\mathbf{l}(i)-\sum_{m=0}^{M}\mathbf{h}_m\tilde{\mathbf{l}}(i-m) \tag{8}$$

Under the minimum mean square error criterion, the problem of calculating the optimal filter parameters can be expressed as

$$\underset{\mathbf{h}_m\in\mathbb{R}^{2\times 2},m=0,\cdots,M}{\arg\min} E[\mathbf{e}^T(i)\mathbf{e}(i)] \tag{9}$$

For this problem, we use the knowledge of matrix differentiation to compute the partial derivatives of $E[\mathbf{e}^T(i)\mathbf{e}(i)]$ with respect to the parameter $\mathbf{h}_m, m=0,\cdots,M$,

$$\frac{\partial E[\mathbf{e}^T(i)\mathbf{e}(i)]}{\partial \mathbf{h}_m}=2E[\mathbf{e}(i)\tilde{\mathbf{l}}^T(i-m)] \tag{10}$$

By substituting Equation (8) into the above equation and setting it equal to the 2×2 zero matrix $\mathbf{0}$, we can get

$$\mathbf{r}_{\mathbf{L}\tilde{\mathbf{L}}}(m)=\sum_{k=0}^{M}\mathbf{h}_k\mathbf{r}_{\tilde{\mathbf{L}}\tilde{\mathbf{L}}}(m-k) \qquad m=0,1,\cdots,M \tag{11}$$

Here, we assume that the sequence data is stationary and the correlation function is only dependent on the time delay, $\mathbf{r}_{\tilde{\mathbf{L}}\tilde{\mathbf{L}}}(m-k)=\mathbf{r}_{\tilde{\mathbf{L}}\tilde{\mathbf{L}}}(i-k,i-m)$, and $\mathbf{r}_{\mathbf{L}\tilde{\mathbf{L}}}(m)=E[\mathbf{l}(i)\tilde{\mathbf{l}}^T(i-m)]$ is the cross-correlation function between $\mathbf{L}(I)$, $\tilde{\mathbf{L}}(I)$. The Equation (11) is formally similar to the Wiener-Hopf equation, which can be further expressed in matrix form,

$$\mathbf{\Gamma}_{\tilde{\mathbf{L}}\tilde{\mathbf{L}}}(M)\mathbf{H}^T(M)=\mathbf{R}^T_{\mathbf{L}\tilde{\mathbf{L}}}(M) \tag{12}$$

From this, the optimal filter parameter $\mathbf{H}_{\text{opt}}(M)$ can be obtained,

$$\mathbf{H}_{\mathbf{opt}}(M)=\mathbf{R}_{\mathbf{L}\tilde{\mathbf{L}}}(M)[\mathbf{\Gamma}^{-1}_{\tilde{\mathbf{L}}\tilde{\mathbf{L}}}(M)]^T \tag{13}$$

where $\mathbf{\Gamma}_{\tilde{\mathbf{L}}\tilde{\mathbf{L}}}(M)$, $\mathbf{H}(M)$, $\mathbf{R}_{\mathbf{L}\tilde{\mathbf{L}}}(M)$ are defined as follows,

$$\mathbf{\Gamma}_{\tilde{\mathbf{L}}\tilde{\mathbf{L}}}(M)=\begin{bmatrix} \mathbf{r}_{\tilde{\mathbf{L}}\tilde{\mathbf{L}}}(0) & \mathbf{r}_{\tilde{\mathbf{L}}\tilde{\mathbf{L}}}(1) & \cdots & \mathbf{r}_{\tilde{\mathbf{L}}\tilde{\mathbf{L}}}(M) \\ \mathbf{r}^T_{\tilde{\mathbf{L}}\tilde{\mathbf{L}}}(1) & \mathbf{r}_{\tilde{\mathbf{L}}\tilde{\mathbf{L}}}(0) & \cdots & \mathbf{r}_{\tilde{\mathbf{L}}\tilde{\mathbf{L}}}(M-1) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{r}^T_{\tilde{\mathbf{L}}\tilde{\mathbf{L}}}(M) & \mathbf{r}^T_{\tilde{\mathbf{L}}\tilde{\mathbf{L}}}(M-1) & \cdots & \mathbf{r}_{\tilde{\mathbf{L}}\tilde{\mathbf{L}}}(0) \end{bmatrix} \tag{14}$$

$$\mathbf{H}(M)=[\mathbf{h}_0 \quad \mathbf{h}_1 \quad \cdots \quad \mathbf{h}_M] \tag{15}$$

$$\mathbf{R}_{\mathbf{L}\tilde{\mathbf{L}}}(M)=[\mathbf{r}_{\mathbf{L}\tilde{\mathbf{L}}}(0) \quad \mathbf{r}_{\mathbf{L}\tilde{\mathbf{L}}}(1) \quad \cdots \quad \mathbf{r}_{\mathbf{L}\tilde{\mathbf{L}}}(M)] \tag{16}$$

In the attack process described in the Equation (7), the bivariate sequence is no longer separated into different univariate sequences, but is treated as a whole to exploit its correlation for

the attack. From the above derivation, it is clear that there exists an optimal filter that achieves the best estimation under the minimum mean square error criterion, and that the optimal filter parameters are related to both auto-correlation in the same dimension and cross-correlation between different dimensions. The CLM-based scheme completely ignores this cross-correlation, which makes it difficult to achieve the desired privacy protection under BCA.

5. Experiment Evaluation

We have theoretically analyzed that the CLM-based location privacy protection scheme is incomplete in terms of series-indistinguishability. To evaluate the impact of this incompleteness on privacy protection performance, we first analyze the correlation characteristics of real-world location sequence data and verify the effectiveness of the bivariate correlation-based attack, and then evaluate the actual privacy protection effect of the CLM-based privacy scheme. All experiments were performed using MATLAB 2023a on a computer with an Intel(R) Xeon(R) CPU E3-1240 v5 @ 3.50GHz and 16GB of memory.

5.1 Experimental Datasets

The real-world location datasets used for the experiments include GeoLife (Zheng et al., 2010, 2009, 2008), T-Drive (Yuan et al., 2011, 2010), and OpenStreetMap[1]. In the data preprocessing stage, location data were transformed from the geographic coordinate system to the Cartesian coordinate system XOY, and then the sequences with a constant time interval $\Delta t \in [1,10]$ were extracted from the dataset with linear interpolation. Based on the dataset obtained, we analyzed the correlation between different dimensions of the location sequence data. In addition, both CLM and the correlation-based attack are implemented based on filters, which makes them inapplicable to the sequence data with drastically varying correlation, and thus the sequences with approximately stationary incremental parts were filtered for subsequent experiments on privacy performance evaluation. See the work (Mao and Xu, 2024) for specific details about the experimental dataset, the coordinate transformation, and the definition of approximate stationarity.

5.2 Evaluation Metrics

In the experiments, two privacy-preserving schemes are compared: the classical Laplace mechanism (denoted as IID) and the extension of CLM in the time-varying environment, QCLM-Lowpass (denoted as CLM). We focuse on the privacy performance loss of CLM-based privacy schemes under bivariate correlation-based attacks. For a single location $\mathbf{l} = [x, y]$, we evaluated the ε_l -geo-indistinguishability achieved by the privacy scheme in the limited region $Area = \{\mathbf{l}' \mid d_E(\mathbf{l}, \mathbf{l}') \le r_{eff}\}$, where r_{eff} is the radius of the focus area. That is, $\forall \tilde{\mathbf{l}}, \tilde{\mathbf{l}}' \in Area$, the privacy mechanism $\mathcal{M}$ satisfies,

$$\frac{Pr[\mathcal{M}(\mathbf{l}) = \tilde{\mathbf{l}}]}{Pr[\mathcal{M}(\mathbf{l}) = \tilde{\mathbf{l}}']} \le \exp[\varepsilon_l \cdot d_E(\tilde{\mathbf{l}}, \tilde{\mathbf{l}}')] \tag{17}$$

For the location dataset $\mathbf{L}_{Set}$, we evaluated the overall privacy strength $\mathcal{E}_\phi, \phi \in [0,1]$ by using the privacy strength satisfied by most locations in it. That is, for $\forall \mathbf{l} \in \mathbf{L}_{Set}$,

$$Pr(\varepsilon_l \le \mathcal{E}_\phi) \ge 1 - \phi \tag{18}$$

In this way, the impact of individual statistical results on the overall evaluation results can be reduced.

In this paper, the correlation-based attack is used to evaluate the actual privacy performance, and for bivariate sequence data, the following two attacks are considered,

1. Univariate correlation-based attack (UCA). In this scheme, the filtering attack realized by a second-order all-pole single-input single-output (SISO) filter was performed on the sequence data in the X and Y directions, respectively.

2. Bivariate correlation-based attack (BCA). Similarly, a second-order all-pole multiple-input multiple-output (MIMO) filter was used to realize the attack, which can be expressed as,

$$\hat{\mathbf{l}}(i) = \sum_{m=1}^{M} \mathbf{a}_m \hat{\mathbf{l}}(i-m) + \mathbf{b}_0 \tilde{\mathbf{l}}(i) \tag{19}$$

where $\mathbf{b}_0, \mathbf{a}_m$ are 2×2 real matrix, which can be solved by vector auto-regressive model theory, the schematic diagram is shown in Figure 1.

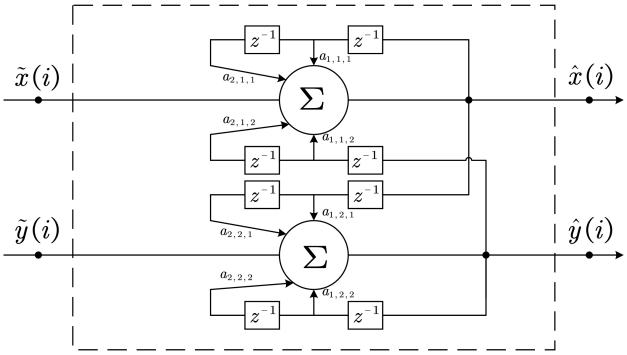

Figure 1. Schematic diagram of the second-order all-pole filter in BCA.

All experiments were repeated 500,000 times to count the distribution of perturbation noise at each location, from which the privacy intensity was calculated.

5.3 Correlation Characterization of Location Sequence Data

Whether incomplete series-indistinguishability has a practical impact on the privacy scheme depends mainly on the correlation characteristics of the location sequence data. For this reason, we analyzed the correlation characteristics between the data in the X and Y directions. In the experiment, we set the sliding window size to $W_L = \lfloor 60 / \Delta t \rfloor$, and calculated the absolute values of Pearson's coefficients and Spearman's coefficients between the data in the X and Y directions at different delay times $\tau \Delta t$, denoted as $|\rho_p(\tau \Delta t)|, |\rho_s(\tau \Delta t)|$. These results were counted and are shown in Figure 2.

[1] https://www.openstreetmap.org/traces

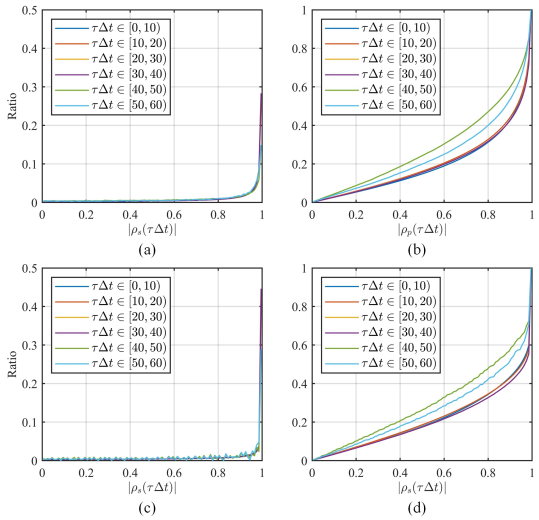

Figure 2. Statistical results of correlation coefficients. (a) and (b) are the density and cumulative distributions of the Pearson's coefficients, respectively. (c) and (d) are the density and cumulative distributions of the Spearman's coefficients, respectively.

The results of the two correlation evaluation metrics are very close to each other. Although the correlation gradually decreases as the delay time $\tau\Delta t$ increases, which is reflected in the fact that their cumulative distribution curves are located at the top of the other curves, but the overall distribution is still concentrated around 1. The result reveals that there is a negligible correlation between the X and Y directions in the actual location sequence data, and BCA that fully consider this cross-correlation is practical challenges for privacy protection.

5.4 Validity Analysis of Bivariate Correlation-based Attack

Based on simulations and real data experiments, we verified the effectiveness of the bivariate correlation-based attack. Figure 3 demonstrates the results of one experiment. Figure 3(a) shows the experimental location sequence data, from which we selected four location points to evaluate the privacy strength achieved by the CLM-based privacy scheme, where the privacy budget in geo-indistinguishability, ε, was set from 0.05 to 0.15 at intervals of 0.025. The results of the actual privacy strength of the privacy scheme under attacks are shown in Table 1.

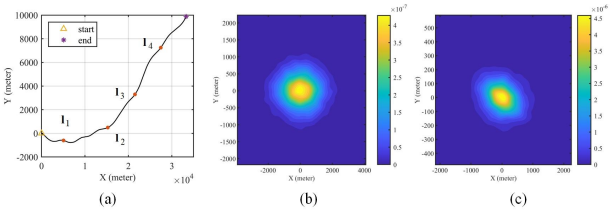

Figure 3. The results under correlation-based attacks. (a) is the real location sequence data. (b) and (c) are the distribution of perturbation noise at $\mathbf{l}_1$ after UCA and BCA, respectively, where $\varepsilon = 0.1$.

It can be seen that the privacy strength of CLM-based privacy scheme under UCA is close to the expectation, which indicates that the scheme achieves better series-indistinguishability in the X and Y directions, but in the majority of scenarios, the result of BCA is larger than that of UCA, by calculating the relative percentage change (RPC) of BCA with respect to UCA, with 5% used as the significance criterion, the result shows that the privacy protection performance of the CLM-based privacy scheme is significantly reduced under BCA, which suggests that the scheme is incomplete in series-indistinguishability for bivariate sequence.

	ε	0.05	0.075	0.1	0.125	0.15
$\mathbf{l}_1$	UCA	0.0465	0.0776	0.1069	0.1396	0.1758
	BCA	0.052	0.0811	0.1047	0.1505	0.1856
	RPC(%)	11.83	4.51	-2.06	7.81	5.57
$\mathbf{l}_2$	UCA	0.0526	0.0775	0.0993	0.139	0.1725
	BCA	0.0782	0.1417	0.2038	0.2362	0.3248
	RPC(%)	48.67	82.84	105.24	69.93	88.29
$\mathbf{l}_3$	UCA	0.0497	0.0777	0.1096	0.149	0.1781
	BCA	0.0964	0.1662	0.2563	0.3216	0.3721
	RPC(%)	93.96	113.9	133.85	115.84	108.93
$\mathbf{l}_4$	UCA	0.0499	0.0766	0.107	0.1379	0.1724
	BCA	0.1054	0.1938	0.2377	0.3499	0.3544
	RPC(%)	111.22	153	122.15	153.73	105.57

Table 1. The privacy strength achieved by the CLM-based privacy scheme under correlation-based attacks.

Figure 3(b) and (c) show the distribution of the perturbation noise of the privacy scheme under different correlation-based attacks, and the distribution is more concentrated and more unevenly distributed in space under BCA. This is because BCA fully exploits the correlation between different dimensions and can further filter out the perturbation noise that is inconsistent with the data in terms of correlation. The same result has been observed in other experiments as well. With this experiment, we have verified that BCA is effective, which implies that it can be used as a tool for testing the completeness of series-indistinguishability.

5.5 Performance Evaluation

To evaluate the actual privacy-preserving performance of CLM-based privacy schemes on real location datasets, we randomly selected 10 location sequences from the dataset with time intervals Δt of 1 and 5 seconds, respectively, for evaluating the overall privacy protection strength after attacks, denoted as $\mathcal{E}'_{0.05}$, where the privacy budget in geo-indistinguishability was set from 0.05 to 0.15 at intervals of 0.025. The results are shown in Figure 4. "IID-UCA" and "IID-BCA" denote the results of the scheme IID after UCA and BCA, respectively, and "CLM-UCA" and "CLM-BCA" are defined in the same way.

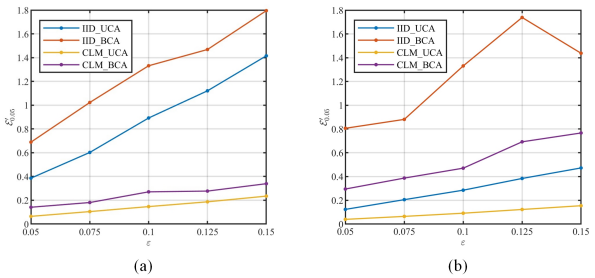

Figure 4. Overall privacy strength under correlation-based attacks. (a) $\Delta t = 1$. (b) $\Delta t = 5$.

It can be seen that the results of BCA are overall higher than those of UCA in both IID and CLM. Specifically, on the dataset with $\Delta t = 1$, the average change in BCA compared to UCA are 51.1% for IID and 73.79% for CLM, while on the dataset with $\Delta t = 5$, the corresponding changes increase to 361.69% for IID

and 485.8% for CLM. This may be because, as the time interval Δt increases, the data correlation decreases, and correspondingly, the filtering effect of UCA also decreases, but according to the result in Section 5.3, there is still a strong correlation between different dimensions of location sequence, which makes the BCA scheme still has a strong ability to filter out the noise, and leads to the gap between the two attacking schemes becomes larger. Therefore, BCA has stronger attack performance and can filter out more perturbation noise.

6. Conclusions

In this paper, the practical problems faced by existing CLM-based location sequence privacy schemes were analyzed. To analyze whether it is reasonable for this scheme to apply CLM independently on each dimension, the correlation characteristics between different dimensions in the actual location sequence data was analyzed, and the statistical results of both correlation coefficients revealed that there is a non-negligible correlation between different dimensions, which provides an opportunity for the attacker to launch an attack on the privacy protection by simultaneously exploiting the auto-correlation of the same dimensions and the cross-correlation between different dimensions, i.e., bivariate correlation-based attack (BCA). Further, a filter-based method was proposed to realize BCA and the optimal filter parameters under the minimum mean square error criterion was derived, which proves in principle that there is such an optimal filter for correlation-based attack, and after that, the effectiveness of the scheme was verified by experiments. Finally, BCA was used to evaluate the actual privacy performance of the CLM-based location sequence privacy preservation scheme. As the theoretical analysis, the experimental results show that the CLM-based scheme can effectively resist correlation attacks in the same dimension, but it is difficult to resist the bivariate correlation-based attack, which indicates that the series-indistinguishability achieved by the scheme is not complete.

The filter-based implementation scheme of BCA proposed in this paper not only provides a quantitative evaluation tool for privacy performance on bivariate sequences, but also lays the groundwork for subsequent research to construct a completely series-indistinguishable privacy protection method. Future work includes investigating the specific requirements for series-indistinguishability on multivariate sequences and the corresponding implementation methods.

References

Al-Dhubhani, R., Cazalas, J.M., 2018. An adaptive geo-indistinguishability mechanism for continuous LBS queries. *Wireless Networks*, 24, 3221–3239.

Andrés, M.E., Bordenabe, N.E., Chatzikokolakis, K., Palamidessi, C., 2013. Geo-indistinguishability: differential privacy for location-based systems. *Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security*, Association for Computing Machinery, 901–914.

Cao, Y., Yoshikawa, M., Xiao, Y., Xiong, L., 2017. Quantifying Differential Privacy under Temporal Correlations. *2017 IEEE 33rd International Conference on Data Engineering (ICDE)*, IEEE, 821–832.

Chatzikokolakis, K., ElSalamouny, E., Palamidessi, C., Anna, P., 2017. Methods for location privacy: A comparative overview. *Foundations and Trends® in Privacy and Security*, 1(4), 199–257.

Chatzikokolakis, K., Palamidessi, C., Stronati, M., 2014. A Predictive Differentially-Private Mechanism for Mobility Traces. *Privacy Enhancing Technologies: 14th International Symposium*, Springer International Publishing, 21–41.

Dwork, C., 2006. Differential Privacy. *International colloquium on automata, languages, and programming*, Springer Berlin Heidelberg, 1–12.

Dwork, C., McSherry, F., Nissim, K., Smith, A., 2006. Calibrating Noise to Sensitivity in Private Data Analysis. *Theory of Cryptography: Third Theory of Cryptography Conference*, Springer, Berlin, Heidelberg, 265–284.

Jiang, H., Li, J., Zhao, P., Zeng, F., Xiao, Z., Iyengar, A., 2021. Location Privacy-preserving Mechanisms in Location-based Services: A Comprehensive Survey. *ACM Computing Surveys*, 54(1), 1–36.

Jiang, K., Shao, D., Bressan, S., Kister, T., Tan, K.-L., 2013. Publishing trajectories with differential privacy guarantees. *Proceedings of the 25th International conference on scientific and statistical database management*, ACM Press, 1-12.

Liu, B., Zhou, W., Zhu, T., Gao, L., Xiang, Y., 2018. Location Privacy and Its Applications: A Systematic Study. *IEEE Access*, 6, 17606–17624.

Mao, L., Xu, Z., 2024. Differential Privacy Preservation for Continuous Release of Real-Time Location Data. *Entropy*, 26(2), 138.

Shokri, R., Theodorakopoulos, G., Le Boudec, J.-Y., Hubaux, J.-P., 2011. Quantifying Location Privacy. *2011 IEEE symposium on security and privacy*, IEEE, 247–262.

Wang, H., Xu, Z., 2017. CTS-DP: Publishing correlated time-series data via differential privacy. *Knowledge-Based Systems*, 122, 167–179.

Wang, H., Xu, Z., Jia, S., Xia, Y., Zhang, X., 2021. Why current differential privacy schemes are inapplicable for correlated data publishing? *World Wide Web*, 24, 1–23.

Wernke, M., Skvortsov, P., Frank Duerr, Rothermel, K., 2014. A classification of location privacy attacks and approaches. *Personal and ubiquitous computing*, 18, 163–175.

Xiao, Y., Xiong, L., 2015. Protecting Locations with Differential Privacy under Temporal Correlations. *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, ACM Press, 1298–1309.

Xiong, X., Liu, S., Li, D., Wang, J., Niu, X., 2019. Locally differentially private continuous location sharing with randomized response. *International Journal of Distributed Sensor Networks*, 15(8), 155014771987037.

Yang, B., Sato, I., Nakagawa, H., 2015. Bayesian Differential Privacy on Correlated Data. *Proceedings of the 2015 ACM SIGMOD international conference on Management of Data*, ACM Press, 747–762.

Yuan, J., Zheng, Y., Xie, X., Sun, G., 2011. Driving with knowledge from the physical world. *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM Press, 316-324.

Yuan, J., Zheng, Y., Zhang, C., Xie, W., Xie, X., Sun, G., Huang, Y., 2010. T-drive: driving directions based on taxi trajectories. *Proceedings of the 18th SIGSPATIAL International conference on advances in geographic information systems*, ACM Press, 99-108.

Zheng, Y., Li, Q., Chen, Y., Xie, X., Ma, W.-Y., 2008. Understanding mobility based on GPS data. *Proceedings of the 10th international conference on Ubiquitous computing*, Association for Computing Machinery, 312–321.

Zheng, Y., Xie, X., Ma, W.-Y., 2010. GeoLife: A collaborative social networking service among user, location and trajectory. *IEEE Data Eng Bull.*, 33(2), 32–39.

Zheng, Y., Zhang, L., Xie, X., Ma, W.-Y., 2009. Mining interesting locations and travel sequences from GPS trajectories. *Proceedings of the 18th international conference on World wide web*, ACM Press, 791-800.