基于信息论方法的多等级位置隐私度量与保护

张文静，刘樵，朱辉
（西安电子科技大学网络与信息安全学院，陕西 西安 710071）

摘 要：针对位置数据拥有者对数据使用者具有不同的信任程度时，会对使用者进行不同等级的划分并向其发布不同扰动程度的位置数据这一场景中的隐私泄露问题，提出了基于信息论中的互信息的隐私度量方法与保护问题。此外，基于互信息提出了度量攻击者获取不同等级的扰动数据而对真实位置数据进行更精确的推断分析所造成的隐私泄露的方法。借鉴用于求解率失真函数的 Blahut-Arimoto 算法提出了多等级位置隐私保护机制。实验结果表明，在上述 2 种问题场景中，所提位置隐私保护机制与基于差分隐私的位置隐私保护方法相比具有更低的隐私泄露，且当真实位置数据具有显著不同的受欢迎程度时，优势更明显。

关键词：位置隐私度量；多级位置隐私保护；信息论方法；率失真理论
中图分类号：TP393
文献标识码：A
doi: 10.11959/j.issn.1000−436x.2019235

Evaluation and protection of multi-level location privacy based on an information theoretic approach

ZHANG Wenjing, LIU Qiao, ZHU Hui
School of Cyber Engineering, Xidian University, Xi’an 710071, China

Abstract: A privacy metric based on mutual information was proposed to measure the privacy leakage occurred when location data owner trust data users at different levels and need to publish the distorted location data to each user according to her trust level, based on which an location privacy protection mechanism (LPPM)was generated to protect user’s location privacy. In addition, based on mutual information, a metric was proposed to measure the privacy leakage caused by attackers obtaining different levels of distorted location data and then performing inference attack on the original location data more accurately. Another privacy metric was also proposed to quantify the information leakage occurred in the scenario based on mutual information. In particular, the proposed privacy mechanism was designed by modifying Blahut-Arimoto algorithm in rate-distortion theory. Experimental results show the superiority of the proposed LPPM over an existing LPPM in terms of location privacyutility tradeoff in both scenarios, which is more conspicuous when there are highly popular locations.

Key words: location privacy metric, multi-level location privacy protection, information theoretic approach, rate- distortion theory

收稿日期：2019–07–08；**修回日期**：2019–10–08
通信作者：刘樵，qiaoliu@xidian.edu.cn
基金项目：国家重点研发计划基金资助项目（No.2017YFB0802200）；国家自然科学基金资助项目（No.61932015, No.61672411, No.61902297）；陕西省自然科学基金资助项目（No.2019ZDLGY12-02）；陕西省科技创新团队基金资助项目（No.2018TD-007）
Foundation Items: The National Key Research and Development Program of China (No.2017YFB0802200), The National Natural Science Foundation of China (No.61932015, No.61672411, No.61902297), The Natural Science Foundation of Shaanxi Province (No.2019ZDLGY12-02), Shaanxi Innovation Team Project (No.2018TD-007)

1 引言

随着移动设备、无线网络的高速发展以及先进的感知和定位技术的出现，产生了大量的基于位置的服务（LBS, location-based services），如谷歌地图、打车软件 Uber、Foursquare、Yelp，以及用于广告推广的应用等。发布位置数据具有很高的应用价值，为了保护用户的位置隐私，学者们已在研究中提出大量的位置隐私保护技术。然而，对外发布的位置数据的使用者可能具有不同的使用需求、使用权限或信任等级，因此，有必要对位置数据进行分级发布。使用权限高或信任程度高的数据使用者被认为是高等级的使用者。高等级的数据使用者（或数据挖掘者）可以访问数据可用性更高（即失真更低或扰动程度更低）的位置数据，而低等级的数据使用者被允许访问的位置数据的扰动程度会更高。然而，当前大多数的位置隐私保护机制（LPPM, location privacy protection mechanism）都将数据使用者考虑为同一等级，并未区分不同数据使用者的级别。在数据使用者都属于同一等级的假设下，位置数据发布者通过LPPM仅生成一种具有固定隐私泄露的扰动数据，这种假设不再适用于数据使用者具有不同等级的应用场景。一个等级分为2个级别的例子如下。某一位置服务提供商的内部员工可能需要使用收集到的位置数据进行数据分析，同时这些位置数据也可以被发布给外部人员使用。由于内部员工信任等级比外部人员高，因此这些位置数据在发布给内部人员和外部人员时，需要被提前划分好等级。此外，与只拥有单一级别权限用户的场景相比，具有多等级权限数据使用者的场景中存在某一数据使用者可能会通过恶意截取、共谋等方式获取多个不同等级（即扰动程度不同）的位置数据。此时的数据使用者被视为攻击者。攻击者通过分析这几种数据间的差异，能够更加精确地推测数据拥有者的真实位置数据。

本文针对数据使用者对发布的扰动位置数据具有不同的访问权限，在问题描述中将数据使用者分级，分析不同级别情况下的隐私泄露，设计基于信息论方法用于发布给不同权限（等级）的数据使用者位置数据的 LPPM。选择信息论中的互信息作为隐私度量方法是因为需要一种能够从本质上考虑到位置数据中先验知识的度量方法，而互信息则能够非常清晰地描述这种信息。此外，本文分析了攻击者拥有不同等级发布数据，并且能够利用这些不同等级的发布数据来更精确地推测真实位置数据这一场景下的隐私泄露，并提出一种可能用于最小化该场景下隐私泄露的优化方案。

2 相关工作

近年来，位置隐私的研究已成为非常活跃的研究领域。大部分位置隐私保护机制都是基于位置数据的扰动来实现的。这种扰动技术包括假位置[1]、加密[2]、空间位置隐身[3-8]、基于差分隐私的位置扰动[9-10]等。使用假位置来代替真实位置可以保护位置隐私，但隐私保护程度和服务质量却依赖于真实位置和假位置之间的距离。基于加密技术的位置隐私保护提供非常强的隐私保护程度，然而，被加密后的位置数据可以被使用的范围仅限于具有解密能力的使用者，因此应用场景十分受限。*k*-匿名技术最初被用于保护数据库隐私[11]，然后被应用到保护位置隐私的场景[12]。此后，基于空间位置隐身的位置隐私保护技术被广泛研究[5,8,13]。然而，这些方案都没有考虑位置隐私泄露的最小化问题。在数据分级发布方面，文献[14]中研究了支持多级位置隐私保护的位置隐身方法，当用户的访问权限更高时，允许用户访问的扰动位置数据的精度更高。然而，基于位置隐身技术的方法并不能从信息论的角度来最小化位置隐私泄露。

3 问题描述

本文使用随机变量L和V_k来分别表示用户的真实位置和发布给等级为 k 的数据使用者的扰动位置，l 和 v_k 分别表示这 2 个随机变量的可能取值。假设数据使用者是不可信的，即其会利用扰动后的位置数据来推测用户的真实位置信息。不同等级的数据使用者被允许访问的位置数据扰动程度不同，等级越高的数据使用者获得的位置数据扰动程度越小，即扰动数据更接近真实位置数据，可用性更高，反之亦然。下文中将交替使用数据拥有者和数据发布者。

定义 1 隐私保护等级为 k 时的位置隐私度量方法。当位置数据拥有者使用隐私保护等级为 k 的LPPM 生成扰动位置数据 V_k，并发布给等级为 k 的数据使用者时，由扰动位置 V_k 导致的隐私泄露定义为 $I(L;V_k)$。其中，$I(L;V_k)$是数据拥有者的真实位置和扰动位置之间的互信息；k 取自正整数集，k 越小

表示等级越高。使用 $I(L;V_k)$作为发布扰动位置数据 V_k 时的隐私泄露的度量方法。

定义 2　扰动位置的可用性度量方法。给定用户的真实位置 L 和要发布给等级为 k 的数据使用者的扰动位置 V_k，将扰动位置的可用性度量方法定义为 $D(L;V_k)=\sum_{l,v_k} p(l)q(v_k|l)d(l,v_k)$，其中，$p(l)$ 是真实位置 L 的先验概率分布；$q(v_k|l)$ 为条件概率，即 LPPM；$d(l,v_k)$ 是真实位置与扰动位置间的失真函数（如汉明距离或欧几里得距离）。

命题 1　发布扰动位置数据 V_k 时的隐私−可用性折中问题。给定用户在某一时刻的真实位置 L，该时刻要发布给等级为 k 的不可信数据使用者的扰动位置 V_k 和可用性约束 D_k，一个 LPPM $q(v_k|l)$ 在给定的可用性约束 D_k 的条件下达到了位置隐私的最小泄露时，这个 LPPM 是如下优化问题的解

$$\text{Leakage}_k^*(D_k)=\min_{q(v_k|l):D(L;v_k)\geqslant D_k} I(L;V_k)$$

其中，$I(L;V_k)$ 是位置隐私的度量方法。

然而，在存在着多个等级数据使用者的场景下，若某一数据使用者成为具有恶意目的的攻击者，他可能会通过恶意截取或与其他等级数据使用者共谋等方式，来获取原本要发布给其他等级数据使用者的扰动位置数据，然后该攻击者即可通过对多个具有不同隐私保护等级的发布数据进行联合分析，进而更精确地推测真实位置数据 L。将这类攻击定义如下。

定义 3　多样性攻击。数据发布者对数据使用者信任程度不同的场景下，数据发布者会依据不同的信任程度来对数据使用者进行等级划分。在这种场景中会存在以下一种攻击。设数据发布者的真实位置数据是 l，其发布给不同等级数据使用者的位置数据分别为 $v_1,v_2,\cdots,v_m,\cdots,v_M$，其中，$m$ 为等级序号，m 值越大表示数据使用者等级越高。在这种场景中，当等级为 m 的数据使用者成为恶意攻击者时，其可通过恶意截取等方式获取发布给其他等级数据使者的数据集为 $V\setminus M$，其中，$V\setminus M$ 表示 $v_1,v_2,\cdots,v_m,\cdots,v_M$ 中除了 v_m 以外任意发布数据组成的集合。例如，攻击者可获得等级为 2 和等级为 M 的发布数据 v_2 和 v_M，有 $V\setminus M=v_2v_M$，然后将其根据自身权限获取的发布数据 v_m 与 $V\setminus M$ 进行联合数据分析，攻击者可以更精确地推断出数据发布者的真实位置数据 l。将这类攻击定义为多样性攻击。

当多等级隐私保护的位置数据发布中存在多样性攻击时，会对数据拥有者的真实位置数据 l 造成更多的隐私泄露。为了衡量这种隐私泄露的多少，本文提出一种基于信息论方法的用于衡量多样性攻击造成的隐私泄露的度量方法。

定义 4　多样性攻击隐私泄露的度量方法。设等级为 m 的数据使用者为恶意攻击者，当其获取了多个发布给不同等级数据使用者的数据集 $V\setminus M$ 后，能够将 v_m 与 $V\setminus M$ 进行联合数据分析来推测数据拥有者的真实位置数据 L，将这种攻击对数据拥有者的真实位置数据 L 造成的隐私泄露定义为 $I(L;v_m\cup V\setminus M))$，其中，$I(L;v_m\cup V\setminus M)$ 是真实位置 L 与数据集 $v_m\cup V\setminus M$ 之间的互信息。

定义 4 提供了一种通用的、用于度量隐私保护机制在受到多样性攻击情况下所产生的隐私泄露。该度量方法可用于衡量任何能够计算出互信息 $I(L;v_m\cup V\setminus M)$ 的隐私保护机制的多样性攻击隐私泄露。第 6 节将详细介绍多样性攻击隐私泄露的计算过程。

4　预备知识

本节介绍率失真函数的定义以及计算率失真函数的算法。率失真函数问题最初被用在有损压缩的研究中，有损压缩的目的是在一定失真约束条件下最小化压缩率。注意到，命题 1 中的隐私−可用性折中问题与率失真问题有紧密关联。实际上 Sankar 等[15]、Calmon 等[16]已分别研究过这种关联。特别地，当分析隐私−可用性折中问题时，他们把信息速率和失真分别类比为折中问题中的信息（隐私）泄露和可用性。然而，这些工作是将率失真与隐私−可用性折中问题的关联关系用于数据库的场景中。此外，Oya 等[17]将这种关联关系用于研究单一时刻位置的隐私−可用性折中问题，并设计了相应的位置隐私保护机制，但其并未考虑数据使用者分为多等级的情况。本节简要描述率失真问题和该问题的计算，以及其与命题 1 中的隐私−可用性折中问题的关联关系。

定义 5　率失真函数[18]。假设编码器的输入是 X，相应的输出是 X'，设信源的失真度量为 $d(x,x')$，分布服从 $X\sim p(x)$，该信源的率失真函数定义为

$$R(D)=\min_{p(X'|X):\sum_{xx'} p(x)p(x'|x)d(x,x')\leqslant D} I(X;X')$$

其中，最小值取自使联合分布 $p(x'|x)=p(X)p(x'|x)$ 满足期望失真限制的所有条件分布 $q(x'|x)$，并且 $I(X;X')=\sum_{x,x'}p(x)p(x'|x)\text{lb}\frac{p(x'|x)}{p(x')}$。

为了便于介绍用于计算率失真函数的算法，首先简要描述一个用于寻找 2 个凸集之间最小距离的算法。这个算法可被用于求解率失真函数中的最优化问题。

寻找 2 个凸集之间最小距离的算法[18-19]。已知 $d(a,b)$ 表示元素 a 和 b 之间的欧几里得距离，给定 2 个凸集 A 和 B，它们之间的最小距离 $d_{\min}=\min_{a\in A}\min_{b\in A}d(a,b)$ 可通过以下的步骤来寻找。首先，在集合 A 中任取一点 $x\in A$，在集合 B 中找出与 $x\in A$ 距离最近的一点 $y\in B$。然后，固定点 $y\in B$，找出集合 A 中与 $y\in B$ 最近的点。重复上述过程，很明显，该距离会随着重复次数的增加而减小。文献[19]中提出如果 2 个集合都是凸集，并且距离度量满足一定的条件，那么这个交替最小化算法最终会收敛到距离的最小值。特别地，若 2 个集合是概率分布的集合且距离度量是相对熵时，该算法的结果会收敛到 2 个概率分布集合之间的最小相对熵。

下面简要介绍使用上述算法中基本思想的用于计算率失真函数的 Blahut-Arimoto 算法。

Blahut-Arimoto 算法[18,20]是一种最终会收敛到率失真函数中凸优化问题最优解的迭代算法。首先，为 $r(x')$ 选择一个初始分布（如均匀分布），使用 $r(x')$ 和 $q(x'|x)=\frac{r(x')\text{e}^{-\lambda d(x,x')}}{\sum_{x'}r(x')\text{e}^{-\lambda d(x,x')}}$ 计算 $q(x'|x)$。在获得 $q(x'|x)$ 后，通过等式 $r(x')=\sum_{x}p(x)q(x'|x)$ 更新 $r(x')$。然后，使用新的 $r(x')$ 和等式 $q(x'|x)=\frac{r(x')\text{e}^{-\lambda d(x,x')}}{\sum_{x'}r(x')\text{e}^{-\lambda d(x,x')}}$ 来更新 $q(x'|x)$。重复上述步骤直到算法收敛，即可获得率失真函数中凸优化问题的最优解 $q(x'|x)$。

本质上来说，Blahut-Arimoto 算法可以被用于求解命题 1 中的最优 LPPM，即 $q(v_k|l)$。

5 多等级隐私保护位置发布机制

基于第 3 节的预备知识，本节提出多等级隐私保护位置发布机制。具体地，提出了基于互信息的位置数据分级发布机制，该发布机制可保证在一定的可用性约束条件下，每一级别的扰动位置数据对真实位置数据具有最小的隐私泄露。需要特别强调的是，多级隐私保护位置数据发布机制中的级别 k 由该算法中的输入参数 λ（即拉格朗日乘子）决定，即数据发布者根据 λ 来定义等级 k。例如：当 λ 的值取自集合{0.01, 2, 5} 时，可以定义 $\lambda=$ 5, 2, 0.01 时对应的等级分别为 $k=1,2,3$。

为了生成基于 $\text{Leakage}_k^*(D_k)$ 的最优 LPPM $q(v_k|l)$，使用 l 和 v_k 分别替代 Blahut-Arimoto 算法中的 X 和 X'，具体形式为

$$q(v_k|l)=\frac{r(v_k)\text{e}^{-\lambda d(l,v_k)}}{\sum_{v_k}r(v_k)\text{e}^{-\lambda d(l,v_k)}} \tag{1}$$

$$r(v_k)=\sum_{v_k}p(l)q(v_k|l) \tag{2}$$

反复迭代 $q(v_k|l)$ 和 $r(v_k)$ 直至算法收敛，即可获得最优 LPPM $q(v_k|l)$。此时，根据互信息的定义可以计算出发布隐私保护等级为 k 的扰动数据对真实位置造成的隐私泄露，如式(3)所示。

$$I(L;v_k)=\sum_{l,v_k}p(l)p(v_k|l)\text{lb}\frac{p(v_k|l)}{q(v_k)} \tag{3}$$

在算法 1 中详细介绍数据发布者如何通过控制输入参数来获取用于生成发布给多个不同级别数据使用者扰动位置数据的 LPPM（即 $q(v_i|l)$）。获取 $q(v_i|l)$ 后，按照概率分布 $q(v_i|l)$ 进行采样来发布扰动位置 v_i。

算法 1 多隐私保护等级的位置数据发布机制

输入 拉格朗日乘子（即等级控制因子）λ，数据使用者的等级 i，真实位置的概率分布 $p(l)$，真实位置数据与发布的扰动位置数据间的失真函数 $d(l,v_i)$，算法收敛设置的阈值 δ

输出 发布扰动位置数据给等级为 i 的数据使用者时所使用的 LPPM $q(v_i|l)$，将 v_i 发布给等级为 i 的数据使用者所导致的最小隐私泄露 I_i^*，对应于 I_i^* 的失真 D_i，扰动位置 v_i 的边缘分布 $p(v_i)$

1) 初始化 $r_0(v_i)$ 为均匀分布
2) 用 $p(v_i)$ 和式(1)计算出 $q_0(v_i|l)$
3) 用 $q_0(v_i|l)$ 和式(2)计算出 $r(v_i)$
4) 用 $r_0(v_i)$，$q_0(v_i|l)$，$p(l)$ 和式(3)计算出

$I_i^* = I(L;V_i)$

5) while true do
6)　用 $r(v_i)$ 和式(1)计算 $q(v_i \mid l)$
7)　用 $r(v_i)$，$q(v_i \mid l)$ 和式(3)计算 $I_i = I(L;V_i)$
8)　if ($I_i^0 - I_i^0 \leqslant \delta$) then
9)　　$I_i^0 \leftarrow I_i$
10)　　$D_i = \sum_{i,v_i} p(l)p(v_i \mid l)d(l,v_i)$
11)　　return $q_0(v_i \mid l)$，I_i^*，D_i
12)　else
13)　　用 $q(v_i \mid l)$，$p(l)$和式(2)计算 $r(v_i)$
14)　end if
15) end while

本文通过给出算法 1 中每一步迭代的计算复杂度的表达式，来分析该算法的计算复杂度。在每次迭代中，计算复杂度是由计算 $q(v_i \mid l)$和 $r(v,k)$主导的。计算式(1)中 $q(v_k \mid l)$ 的复杂度分析如下。针对变量 l 的每个取值，对于一个特定的 v_k，在分母上需要进行 $|v_k|$次乘法。考虑到对每个 v_k 都使用这个分母，因此共需要 $O(|v_k|)$次计算操作。考虑到变量 l 的所有取值，计算 $q(v_k \mid l)$ 的复杂度为 $O(|V_k||L|)$。计算式(2)中 $r(v_k)$ 的复杂度分析如下。类似地，对于一个特定的 v_k 需要$|L|$次乘法。考虑到所有的 v_k，计算 $r(v_k)$ 时的复杂度为 $O(|V_k||L|)$。因此，算法 1 中的每次迭代需要 $O(|V_k||L|)$次计算。

6　多样性攻击隐私泄露的计算方法

定义 3 中指出，攻击者可能截取到的发布给其他等级数据使用者的数据集 $V \setminus m$ 包括了除 v_m 以外的任意发布数据。以数据使用者分为 3 个等级为例，详细介绍如何使用本节提出的多样性攻击隐私泄露的度量方法来衡量当攻击者获得其他 2 个等级的发布数据时导致的隐私泄露。设该场景中数据拥有者的真实位置数据为 L，通过算法 1 生成了用于发布给 3 个不同等级的数据使用者的扰动位置数据 V_1,V_2,V_3。设攻击者的等级为 2，当他通过截获等方式获得其他 2 个等级用户的数据 V_1 和 V_3 时，该攻击者可通过将发布数据 V_1,V_2,V_3 进行联合分析，进而更好地推断真实位置数据 L 的值。根据定义 4 中提出的度量方法，这种场景下的多样性隐私泄露为 $I(L;V_1,V_2,V_3)$。

为了清楚地描述出如何计算不同隐私保护机制在受到多样性攻击时导致的隐私泄露，下面将对互信息 $I(L;V_1,V_2,V_3)$进行展开计算。根据互信息的定义有

$$I(L;V_1,V_2,V_3) = H(V_1,V_2,V_3)H(V_1,V_2,V_3;L)$$

根据信息熵的定义有

$$H(V_1,V_2,V_3) = \sum_{v_1,v_2,v_3} p(v_1,v_2,v_3)\mathrm{lb}p(v_1,v_2,v_3)$$

由于 V_1,V_2,V_3 分别为真实位置 L 经由 3 种不同隐私保护等级处理后的发布数据，因此 V_1,V_2,V_3 之间相互独立。因此有

$$p(v_1,v_2,v_3) = p(v_1)p(v_2)p(v_3)$$

其中，v_1,v_2,v_3 的边缘分布为

$$p(v_m) = \sum_l p(l)\mathrm{lb}p(v_m \mid l), m = 1,2,3 \tag{4}$$

根据条件熵的定义，有

$$H(V_1,V_2,V_3 \mid L) = -\sum_{v_1,v_2,v_3,l} p(v_1,v_2,v_3)\mathrm{lb}p(v_1,v_2,v_3 \mid l)$$

其中，有

$$p(v_1,v_2,v_3,l) = p(v_1 \mid v_2,v_3,l)p(v_2 \mid v_3,l)p(v_3 \mid l)p(l) = p(v_1 \mid l)p(v_2 \mid l)p(v_3 \mid l)p(l)$$

该式中第一个等号是基于概率论中的贝叶斯公式，第二个等号是由于考虑单一时刻的位置发布，因此当给定真实位置 L 时，V_l 完全由 L 决定，而与其他变量无关，因此有

$$p(v_2 \mid v_3,l) = p(v_2,l)$$

同样地，有

$$p(v_1,v_2,v_3 \mid l) = p(v_1 \mid l)p(v_2 \mid l)p(v_3 \mid l)$$

由此可以看出，计算互信息 $I(L;V_1,V_2,V_3)$的关键是需要知道发布数据 V_1,V_2,V_3 的边缘分布，条件概率分布（即 LPPM）$p(v_1 \mid l)$，$p(v_2 \mid l)$，$p(v_3 \mid l)$，以及真实位置 L 的概率分布 $p(l)$。

7　实验与性能

本节在模拟数据集上评估算法 1 中提出的多级隐私保护的位置发布机制，并与文献[21]中提出的 LPPM 进行对比。文献[21]中提出一种基于差分隐私的 LPPM—— Geo-indistinguishability，该 LPPM 保证真实位置 x 会被以很高的概率映射到一个邻近的位置，而不是映射到较远的位置。

Geo-indistinguishability 扩展了差分隐私的定义，以达到将用户的真实位置保护在一定直径范围内的目的。为了合理地比较 2 种 LPPM 的隐私泄露，本文也用互信息来计算文献[21]中 LPPM 的隐私泄露。

选择模拟数据集的原因是可以通过改变模拟数据集中真实位置 L 的先验概率分布，来理解不同的先验概率分布对于位置隐私泄露的影响。位置 L 的概率分布不同表明数据集中的位置具有不同的受欢迎程度（即用户位于某一位置的概率明显高于其他位置）。在这个模拟数据集中，假设地图中有 6 个位置。具体地，分别在模拟数据集上计算并比较 2 种 LPPM 在发布给单一级别数据使用者时的隐私泄露（即数据使用者无法获取其他隐私保护级别的发布数据）和受到多样性攻击时的隐私泄露。所有的实验都是在一台配备了 2.3 GHz Intel i5 处理器和 8 GB 内存的笔记本电脑上完成的。

7.1 单一级别的位置隐私泄露

本节在模拟数据集上进行仿真，分析了算法 1 中提出的 LPPM 在数据使用者仅拥有符合自己相应权限的发布数据情况下的隐私泄露，并与文献[21]中的 LPPM 产生的隐私泄露进行对比。为了分析真实位置 L 的先验概率分布的变化对隐私泄露的影响，在模拟数据集中将真实位置 L 的先验概率分布分别设置为 $p_1(L)=\left\{\frac{1}{6},\frac{1}{6},\frac{1}{6},\frac{1}{6},\frac{1}{6},\frac{1}{6}\right\}$，$p_2(L)$= {0.3, 0.1, 0.2, 0.25, 0.05, 0.1}，$p_3(L)$ = {0.8, 0.04, 0.04, 0.04, 0.04, 0.04}和 $p_4(L)$= {0.04, 0.04, 0.04, 0.04, 0.04, 0.8}。注意到模拟数据集中真实位置 L 的 4 个先验概率分布的变化有一定的特点，即在每个概率分布中，位置受欢迎程度的差别逐渐增大。换句话说，当真实位置 L 的概率分布为 $p_1(L)=\left\{\frac{1}{6},\frac{1}{6},\frac{1}{6},\frac{1}{6},\frac{1}{6},\frac{1}{6}\right\}$，即均匀分布时，每个位置受欢迎程度的大小是同等的；而当真实位置 L 的概率分布为 $p_1(L)$= {0.8, 0.04, 0.04, 0.04, 0.04, 0.04}时，每个位置受欢迎程度的差别最大。本文将通过实验来了解当位置受欢迎程度区别较大或较小时，对单一时刻隐私泄露的影响。

失真使用欧几里得距离来计算。文献[21]中的 LPPM 是通过使用与模拟数据集中相同的初始位置概率分布和失真生成的。此外，为了描绘出平滑的隐私−可用性曲线，在 0.01~10 这个范围内逐渐递增地选择 λ。λ 越小，失真越大。将算法 1 中结果收敛时的阈值设置为 1×10^{-8}，分别在 $p_1(L)$、$p_2(L)$、$p_3(L)$ 和 $p_4(L)$上进行仿真，结果如图 1~图 4 所示。图中曲线上的每一个点对应于一个相应的 λ 取值。每一个 λ 对应一个数据使用者的级别，λ 越大，对应的数据使用者的级别越高。

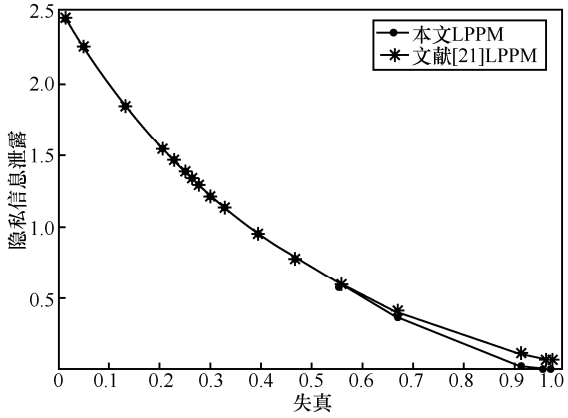


图 1 真实位置的先验概率分布为 $p_1(L)$时的隐私泄露

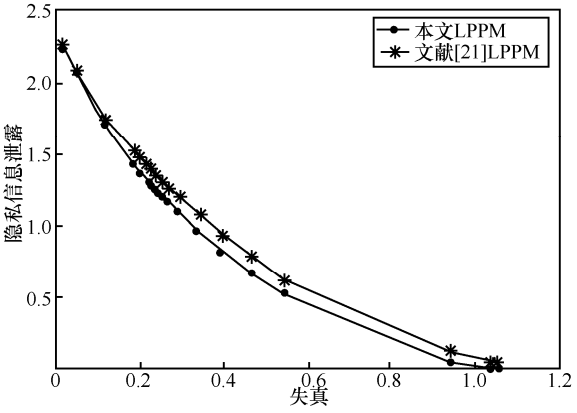


图 2 真实位置的先验概率分布为 $p_2(L)$时的隐私泄露

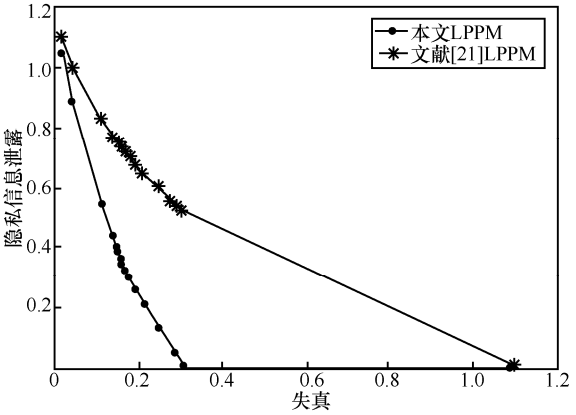


图 3 真实位置的先验概率分布为 $p_3(L)$时的隐私泄露

比较图 1～图 4 中的实验结果可以看出，本文提出的 LPPM 的隐私泄露均低于文献[21]中 LPPM

的隐私泄露。这是因为本文的 LPPM 是通过求解最小化隐私泄露的优化问题而得到的。此外，还观察到当真实位置 L 的概率分布越趋向于集中在某些位置时（即用户位于某些位置的概率远高于其他位置，如概率分布为 $p_3(L)$= {0.8, 0.04, 0.04, 0.04, 0.04, 0.04}时，用户真实位置是第一个位置的概率为 0.8，远高于其他位置的概率），本文所提 LPPM 在隐私泄露方面的优势就越明显。这是由于攻击者具有关于真实位置 L 的先验概率 $p(L)$ 的知识，因此当某一位置非常受用户欢迎时，真实位置的先验概率分布本身已经泄露了很多信息，为了保证可用性，生成的 LPPM 几乎不会再通过降低可用性来减小隐私泄露了；相比之下，当用户真实位置的先验概率为均匀分布时，先验概率分布本身泄露的隐私很少，因此当攻击者一旦观测到了扰动位置数据后，会对真实位置造成较多的隐私泄露。

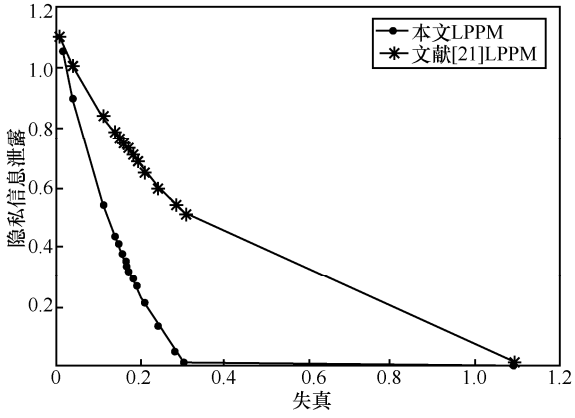


图 4　真实位置的先验概率分布为 $p_4(L)$时的隐私泄露

7.2　存在多样性攻击时的隐私泄露

与第 6 节中的举例一致，本节在分析多样性攻击隐私泄露的实验部分也将数据使用者分为 3 个等级。首先，需要保证数据使用者等级的选取具有一般性。考虑到本文所提方案中的数据使用者等级由 λ 决定，因此使用程序在[0.01;0.01;10]这个区间范围内随机选出 3 个 λ 值，对比分析 2 种方案在受到多样性攻击时的隐私泄露。在模拟数据集中仍然假设真实位置 L 的先验概率分布为 $p_1(L)=\left\{\frac{1}{6},\frac{1}{6},\frac{1}{6},\frac{1}{6},\frac{1}{6},\frac{1}{6}\right\}$，$p_2(L)$ = {0.3, 0.1, 0.2, 0.25, 0.05, 0.1}，$p_3(L)$ = {0.8, 0.04, 0.04, 0.04, 0.04, 0.04}和 $p_4(L)$={0.04, 0.04, 0.04, 0.04, 0.04, 0.8}。实验结果分别如表 1～表 4 所示。

表 1～表 4 中的实验结果表明，本文提出的 LPPM 在存在多样性攻击的场景中仍然有隐私泄露低于文献[21]中 LPPM 的隐私泄露的优势。这是因为，发布给不同等级数据使用者的扰动数据之间相互独立，因此，每个扰动数据都是通过求解最小化隐私泄露的优化问题得到的，多个扰动数据隐私泄露的求和也是最小的。此外，为了分析攻击者所拥有的发布数据的等级差距大小与多样性攻击隐私泄露多少的关联关系，本文固定 3 个等级中的 2 个等级，改变第 3 个等级，进而分析这种关联关系。相比只有 2 个等级扰动数据时的多样性隐私泄露，当攻击者可以额外获得一个等级的发布数据时，隐私泄露更多。实验数据表明，文献[21]中 LPPM 在受到多样性攻击时产生的隐私泄露也有类似的结论。此外，注意到 $p_3(L)$ 和 $p_4(L)$ 的概率分布中，仅是受欢迎的位置不同，而位置的受欢迎程度是相同的。由表 3 和表 4 中的实验数据可以看出，具体哪一个位置受欢迎对隐私泄露的多少并无影响，影响隐私泄露的是位置的受欢迎程度。最后，与 7.1 节

表 1　　**真实位置的先验概率分布为 $p_1(L)$时的多样性隐私泄露**

方案	λ					
	0.01, 0.05	0.01, 0.05, 2	0.01, 0.05, 5	2, 3	2, 3, 5	2, 3, 10
本文 LPPM	2.1×10^{-4}	1.127 8	2.437 6	2.964 4	5.401 9	5.547 7
文献[21]LPPM	1.213 3	2.803 9	3.658 0	3.509 2	5.953 9	6.080 5

表 2　　**真实位置的先验概率分布为 $p_2(L)$时的多样性隐私泄露**

方案	λ					
	0.01, 0.05	0.01, 0.05, 2	0.01, 0.05, 5	2, 3	2, 3, 5	2, 3, 10
本文 LPPM	3.4×10^{-4}	1.089 8	2.231 2	2.778 7	5.009 6	5.143 1
文献[21]LPPM	1.163 3	2.641 9	3.399 2	3.245 1	5.481 1	5.598 0

表 3 真实位置的先验概率分布为 $p_3(L)$时的多样性隐私泄露

方案	λ					
	0.01, 0.05	0.01, 0.05, 2	0.01, 0.05, 5	2, 3	2, 3, 5	2, 3, 10
本文 LPPM	1.5×10^{-5}	0.267 6	1.056 0	0.812 9	1.868 8	1.997 7
文献[21]LPPM	0.524 8	1.204 1	1.621 8	1.507 7	2.604 7	2.683 9

表 4 真实位置的先验概率分布为 $p_4(L)$时的多样性隐私泄露

方案	λ					
	0.01, 0.05	0.01, 0.05, 2	0.01, 0.05, 5	2, 3	2, 3, 5	2, 3, 10
本文 LPPM	1.5×10^{-5}	0.267 6	1.056 0	0.812 9	1.868 8	1.997 7
文献[21]LPPM	0.524 8	1.204 1	1.621 8	1.507 7	2.604 7	2.683 9

的结果类似，当真实位置的概率分布中不同位置的受欢迎程度区别越大时，隐私泄露越少。

8 结束语

本文提出了基于信息论方法、独立于任何攻击、用于单一时刻的位置隐私度量方法。特别地，考虑了数据使用者由于可信度不同而被分为多个等级的场景。在一定的可用性约束条件下，依据本文提出的位置隐私度量方法建立最小化位置隐私泄露的优化问题，即隐私−可用性折中问题。通过在该优化问题中设置不同的输入参数来生成不同隐私保护等级的扰动位置数据，并发布给相应等级的数据使用者。此外，还提出了一种用于衡量当攻击者掌握多个不同等级的扰动数据时对真实位置数据造成的隐私泄露的度量方法，并将此类攻击定义为多样性攻击。实验结果表明，在没有多样性攻击和有多样性攻击的 2 种场景中，本文所提 LPPM 在隐私−可用性折中方面相比于基于差分隐私的 LPPM 具有显著优势，尤其是当真实位置的先验概率分布存在特别受欢迎的一些位置时，这种优势更加明显。未来的研究工作是如何获取可最小化多样性攻击场景下隐私泄露的位置隐私保护机制。

参考文献：

[1] KIDO H, YANAGISAWA Y, SATOH T. Protection of location privacy using dummies for location based services[C]// 21st International Conference on Data Engineering Workshops. 2005 : 1248-1248.

[2] AHAMED S I, HAQUE M M, HASAN C S. A novel location privacy framework without trusted third party based on location anonymity prediction[J]. ACM SIGAPP Applied Computing Review, 2012, 12(1): 24-34.

[3] BAMBA B, LIU L, PESTI P, et al. Supporting anonymous location queries in mobile environments with privacygrid[C]//The 17th International Conference on World Wide Web. 2008 :237-246.

[4] CHENG R, ZHANG Y, BERTINO E, et al. Preserving user location privacy in mobile data management infrastructures[C]// International Workshop on Privacy Enhancing Technologies. 2006: 393-412.

[5] GEDIK B, LIU L. A customizable *k*-anonymity model for protecting location privacy[R]. Georgia Institute of Technology, 2004.

[6] GHINITA G, KALNIS P, SKIADOPOULOS S. PRIVE: anonymous location-based queries in distributed mobile systems[C]//The 16th International Conference on World Wide Web. 2007: 371-380.

[7] KALNIS P, GHINITA G, MOURATIDIS K, et al. Preventing location-based identity inference in anonymous spatial queries[J]. IEEE Transactions on Knowledge and Data Engineering, 2007, 19(12): 1719-1733.

[8] MOKBEL M F, CHOW C Y, AREF W G. The new casper: query processing for location services without compromising privacy[C]// The 32nd International Conference on Very Large Data Bases. 2006 : 763-774.

[9] ANDRÉS M E, BORDENABE N E, CHATZIKOKOLAKIS K, et al. Geo-indistinguishability: differential privacy for location-based systems[C]//ACM Conference on Computer and Communications Security. ACM, 2013.

[10] BORDENABE N E, CHATZIKOKOLAKIS K, PALAMIDESSI C. Optimal geo-indistinguishable mechanisms for location privacy[C]// The 2014 ACM SIGSAC Conference on Computer and Communications Security. 2014 : 251-262.

[11] SWEENEY L. k-anonymity: a model for protecting privacy[J]. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002, 10(5): 557-570.

[12] GRUTESER M, GRUNWALD D. Anonymous usage of location-based services through spatial and temporal cloaking[C]//The 1st International Conference on Mobile systems, Applications and Ser-

vices. 2003: 31-42.

[13] XIAO Z, MENG X, XU J. Quality aware privacy protection for location-based services[C]//International Conference on Database Systems for Advanced Applications. 2007: 434-446.

[14] LI C, PALANISAMY B. ReverseCloak: Protecting multi-level location privacy over road networks[C]//The 24th ACM International on Conference on Information and Knowledge Management. 2015: 673-682.

[15] SANKAR L, RAJAGOPALAN S R, POOR H V. Utility-privacy tradeoffs in databases: an information-theoretic approach[J]. IEEE Transactions on Information Forensics and Security, 2013, 8(6): 838-852.

[16] CALMON F D P, FAWAZ N. Privacy against statistical inference[C]// The 50th Annual Allerton Conference on Communication, Control, and Computing. 2012: 1401-1408.

[17] OYA S, TRONCOSO C, PÉREZ-GONZÁLEZ F. Back to the drawing board: revisiting the design of optimal location privacy-preserving mechanisms[C]//The 2017 ACM SIGSAC Conference on Computer and Communications Security. 2017: 1959-1972.

[18] COVER T M, THOMAS J A. Elements of Information Theory[M]. New Jersey: John Wiley & Sons, 2012.

[19] CSISZ I, TUSNÁDY G. Information geometry and alternating minimization procedures[J]. Statistics and Decisions, 1984(1): 205-237.

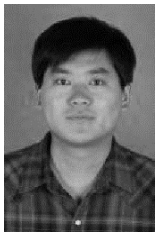
[20] BLAHUT R. Computation of channel capacity and rate-distortion functions[J]. IEEE Transactions on Information Theory, 1972, 18(4) : 460-473.

[21] ANDRÉS M E, BORDENABE N E, CHATZIKOKOLAKIS K, et al. Geo-indistinguishability: differential privacy for location-based systems[C]//The 20th ACM conference on Computer and Communications Security. 2013 : 901-914.

[作者简介]

张文静（1988− ），女，黑龙江绥化人，西安电子科技大学博士生，主要研究方向为数据隐私和隐私度量。

刘樵（1989− ），男，陕西咸阳人，博士，西安电子科技大学讲师，主要研究方向为物理层安全。



朱辉（1981− ），男，河南周口人，博士，西安电子科技大学教授、博士生导师，主要研究方向为数据安全与隐私保护、安全方案及协议设计、网络及应用安全。