一种基于时空相关性的差分隐私轨迹保护机制

吴云乘 陈 红 赵素云 梁文娟 吴 垚 李翠平 张晓莹

(中国人民大学数据工程与知识工程教育部重点实验室 北京 100872)

(中国人民大学信息学院 北京 100872)

摘 要 近年来,基于位置的服务(LBS)越来越成为人们生活中一种重要的查询方式,具有广阔的应用前景和价值.然而,在连续地使用LBS时会暴露用户的位置甚至轨迹,用户对这种位置或轨迹隐私泄露的顾虑一方面阻碍了LBS的应用,另一方面降低了用户得到的服务质量.目前,轨迹隐私保护技术已成为研究热点,但是现有的技术极少考虑到地理空间的限制以及时间序列上位置的相关性,使得攻击者仍有较大可能推断出用户的真实敏感位置和轨迹.该文针对轨迹隐私保护问题,首先根据地理空间的拓扑关系,提出了CPL算法计算地图上各区域的隐私级别,并定义了一种结合隐私级别与差分隐私预算的隐私模型.然后,该文基于马尔可夫概率转移矩阵,分析了发布位置对当前真实位置和之前真实位置的影响,提出了一种差分隐私位置发布机制DPLRM,以保护用户的位置和轨迹隐私.最后,在真实数据集上的实验验证了该文提出的隐私模型和差分隐私位置发布机制的准确性和有效性.

关键词 轨迹隐私;差分隐私;时空相关性;位置隐私

中图法分类号 TP391 **DOI号** 10.11897/SP.J.1016.2018.00309

Differentially Private Trajectory Protection Based on Spatial and Temporal Correlation

WU Yun-Cheng CHEN Hong ZHAO Su-Yun LIANG Wen-Juan

WU Yao LI Cui-Ping ZHANG Xiao-Ying

(*Key Laboratory of Data Engineering and Knowledge Engineering of Ministry of Education, Renmin University of China, Beijing* 100872)

(*School of Information, Renmin University of China, Beijing* 100872)

Abstract In recent years, location based services (LBS) is becoming one of the most important ways for information retrieval in our daily life, and it has broad application prospects and great value. However, people's locations or trajectory may be disclosed when they continuously use LBS to retrieve point of interests. This privacy disclosure problem not only restricts the development of LBS, but also reduces the quality of service the users obtained. Recently, trajectory privacy protection has attracted more and more attention, such as cloaking based technique, perturbation based technique, and so on. However, existing techniques seldom consider the geo-spatial and temporal correlation of the locations between several timestamps, which might degrade the location privacy of users. In this paper, aiming at dealing with the trajectory privacy problem, we explore a popular paradigm for providing privacy with strong theoretical guarantees, differential privacy, which has recently gained significant attention for its robustness to known attacks, and

收稿日期:2016-06-23;在线出版日期:2017-03-28.本课题得到国家自然科学基金(61532021,61702522,61772537,61772536)、国家"八六三"高技术研究发展计划项目(2014AA015204)资助.**吴云乘**,男,1989年生,博士研究生,主要研究方向为隐私数据分析、物联网数据管理.E-mail:yunchengwu@ruc.edu.cn.**陈 红**(通信作者),女,1965年生,博士,教授,博士生导师,中国计算机学会(CCF)高级会员,主要研究领域为数据库、数据仓库、物联网、隐私保护.E-mail:chong@ruc.edu.cn.**赵素云**,女,1979年生,博士,副教授,主要研究方向为模糊集、粗糙集、不确定信息处理.**梁文娟**,女,1980年生,博士研究生,主要研究方向为隐私保护、数据库.**吴 垚**,男,1990年生,博士研究生,主要研究方向为群智感知、大数据管理.**李翠平**,女,1971年生,教授,博士生导师,主要研究领域为社会网络分析、数据挖掘、推荐系统.**张晓莹**,女,1987年生,博士,讲师,主要研究方向为无线传感器网络、隐私保护.

define a new privacy model based on differential privacy for trajectory protection. Specifically, we firstly propose an algorithm (CPL) to calculate the privacy level of each location on the map according to geo-spatial correlation. This algorithm transforms the topology of map into an undirected weighted graph. Based on the initial sensitive locations and the corresponding pre-defined privacy levels that provided by users, CPL algorithm iteratively allocates the privacy level of a location to its adjacent locations by the edge weights, and computes the aggregated privacy levels for all other locations that are not in the set of initial sensitive locations. Secondly, we present a privacy model, called γ-trajectory privacy, that combines the privacy level and differential privacy budget. Fundamentally, for any location in a trajectory, this privacy model requires that the multiplication of privacy level that computed from CPL algorithm and differential privacy budget of this location should equal to γ. In other words, the higher the privacy level is, the lower the differential privacy budget should be. Therefore, we can use this privacy model as a guideline to determine the differential privacy budget for every location. Thirdly, we figure out that if we release locations of a trajectory according to a differentially private location perturbation algorithm independently (which is widely used in existing work), malicious adversaries may also compromise the location privacy by the temporal correlation between perturbed locations. Thus, we propose a differentially private location release mechanism (DPLRM) that considers the temporal correlation to protect the trajectory privacy of users. Specifically, we model the temporal correlation between user's true locations by Markov chain transition matrix, and define the DPLRM as an optimization problem by minimizing an objective function based on the total distance between the true locations and possible released locations. We also give a mathematical deduction to calculate the constraints for this optimization problem. Finally, we conduct extensive experiments on two real world datasets, and show that it is computational efficient for CPL algorithm to compute the privacy levels, and the performance of DPLRM algorithm is close to an optimal approach and better than an existing mechanism.

Keywords Trajectory privacy; differential privacy; spatial and temporal correlation; location privacy

《计算机学报》

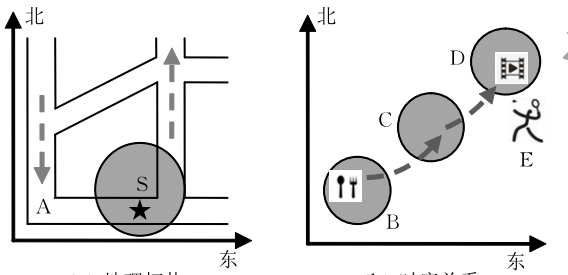
1 引 言

随着定位技术的快速发展,如 GPS、Wi-Fi、蜂窝网络、射频识别(RFID)等,基于位置的服务(Location-Based Services,LBS)在人们的生活中日益普及. LBS 主要可分为两大类:单时刻 LBS 和连续 LBS. 单时刻 LBS 需要用户不定时地提供其当前位置给服务商来获取所需的信息(如查询当前周边的 PM2.5 浓度),而连续 LBS 需要用户周期性地提供其位置给服务商来获取服务(如行驶过程中不断地提出查询附近的餐馆). 尽管 LBS 为用户提供了许多便捷的服务,但是将敏感的位置信息上传给不可信的服务商存在一定的隐私威胁,如用户访问过的敏感位置(医院、家等)不愿被外界知晓. 从 2003 年开始,研究者们针对位置隐私保护技术展开了研究,并取得了丰富的研究成果[1-2]. 然而,多数研究仅考虑单时刻 LBS 的情形,对连续 LBS 所产生的用户轨迹隐私关注较少. 近年来,研究者们发现即使用户每个时刻的位置得到保护,攻击者仍能用数据挖掘、关联分析等手段获知用户的兴趣爱好、行为模式、生活习惯等信息. 因此,如何在保证服务质量的同时,保护用户的轨迹隐私是当前亟待解决的问题.

当前 LBS 中轨迹隐私保护方案大致包括以下四类[3]:泛化(将轨迹上每个时刻的真实位置泛化到一个区域)[4-8]、混合区(在车联网中应用较多,通过在路口变换车辆的假名)[9-11]、抑制(根据真实位置所在区域的敏感程度有选择地抑制发布)[12-14]以及扰乱(对每个时刻的真实位置添加随机噪声生成扰乱位置)[15-19]. 现有的解决方案大多数都是从单时刻 LBS 位置隐私保护方案演化而来,并存在以下两点问题:

(1)某些个性化的轨迹隐私保护机制(即用户对不同位置自定义隐私程度,如抑制策略)仅考虑单个位置的隐私级别,而忽略了地理拓扑对位置隐私级别的影响.如图 1(a)所示,设敏感位置为 S(星号位置),并将 S 泛化到灰色区域中(即灰色区域为敏感区域),且当前按左侧灰色箭头移动;在抑制策略中,当用户移动至位置 A 时,会如实上报位置(因为 A 不敏感);但是若过一段时间,用户移动至右侧的箭头处,即使在灰色区域范围内不上报用户的位置,攻击者根据 A 与 S 的地理位置关系,也极易猜测出用户经过或停留在了 S 处,从而泄露了用户的位置隐私甚至是轨迹隐私.

(2)连续时刻间的位置可能存在某种相关性,即使单个时刻的敏感位置都得到保护,连续的轨迹仍会有很大可能暴露用户的真实位置.如图 1(b)所示,即使用户每个时刻的真实位置被泛化保护到灰色区域中,根据用户的行为模式(如晚上九点后通过这条轨迹的用户 80% 都是吃完饭后去看电影的人)、地理约束等其它信息,攻击者有很大可能推测出用户此时在电影院,从而造成隐私的泄露.



(a) 地理拓扑　　(b) 时序关系

图 1　地理拓扑和时序关系对隐私泄露的影响

针对上述两点问题,本文首先根据地理位置间的拓扑关系提出了一种个性化隐私级别界定方案 CPL 算法,CPL 算法将地理拓扑关系转换成带权重的无向图,依据用户自定义的敏感位置集合及相应的隐私级别,利用图性质计算其它位置的隐私级别;然后提出了一种融合位置隐私级别和差分隐私预算 γ-隐私模型定义;最后提出了保证轨迹可用性的差分隐私位置扰乱机制 DPLRM,该机制在考虑各个时刻时序相关性的情况下,不仅能够保护当前的位置隐私,还能保护之前时刻的位置隐私不会因为当前发布的位置而泄露.真实数据集上的实验证明本文所提出的方法能够较好地保护用户的轨迹隐私并获得较高的轨迹质量.

本文第 2 节介绍轨迹隐私保护技术和差分隐私的相关工作;第 3 节介绍预备知识;第 4 节介绍基于地理拓扑关系的 γ-隐私模型;第 5 节提出能保护轨迹隐私的差分隐私位置发布机制 DPLRM;第 6 节阐述实验评估的结果;最后总结全文并展望未来研究工作.

2　相关工作

如前文所述,目前 LBS 中轨迹隐私保护技术主要包括泛化、混合区、抑制和扰乱四类.泛化的思想是将真实轨迹上的位置表示成一个区域,以达到隐私保护的目的.文献[4]在每个时刻选择泛化区域时需要满足该区域中有其他 $k-1$ 个用户存在,以达到 k-匿名保护.文献[5]在文献[4]的基础上,考虑了用户的移动方向和速度来最小化泛化区域.混合区技术的思想是通过在路口变换为用户分配的假名,以切断用户轨迹在时序上的联系.文献[10]提出了一种优化方法来限制混合区的数目,以减少混合区对用户位置发布的影响,同时达到 k-匿名.文献[11]将路网中的时空限制条件考虑在内,并提出一种自适应大小和形状的混合区隐私保护方案.抑制策略的思想是在敏感的位置选择不发布当前位置.文献[12]提出了一种局部抑制的隐私保护方案,可以实现定制化的轨迹隐私保护.扰乱的方法在近年来有较多的研究,它的思想是将各个时刻的真实位置通过一定的变换,从而生成假位置,以保护真实位置的隐私.文献[15]基于可信第三方,提出一种方案将用户查询的起点和目的点隐藏在一系列其它用户的查询中,从而保护用户的隐私.文献[20]中提出一种基于旋转的方法将用户的真实轨迹变换成假轨迹.

差分隐私模型[21-23]因其严格的数学形式和定制化的隐私保护程度受到国内外学者的青睐.其概念由 Dwork 等人[21-22]在 2006 年提出,主要思想是对原始查询结果添加随机噪声以保护隐私,以使得在数据集中添加或删除某一条数据对查询结果的影响很小,从而让攻击者很难通过多次的查询结果来反推数据集中的某条真实数据.

近几年,轨迹数据的差分隐私保护逐渐成为研究热点,并主要分为两个方面:历史轨迹数据的发布[24-29]以及实时位置数据的发布[17-18,30].在差分隐私历史轨迹数据发布的问题上,Chen 等人[24-25]在层次结构的基础上提出了差分隐私合成轨迹的发布.文献[24]采用相同的前缀对轨迹进行分组,并构造前缀树,但是由于模式具有较高的唯一性,叶节点上的计数操作非常稀疏.文献[25]中,在构建基于

马尔可夫假设的树时考虑了子字符串的影响，使得叶节点的计数较高并具有较好的可用性. Shao 等人在文献[26]中提出采样和差值策略来合成历史轨迹，但是仅实现了一个相对较弱的差分隐私定义. 文献[27]提出了 SSD 算法，它对每个时刻的方向和距离进行采样，以此为基础发布下一时刻的数据，形成轨迹. 但是其可用性相对较弱. 文献[28]通过离散化原始轨迹生成多粒度的层次化引用系统，基于该系统构建前缀树，最后采用权重方向提高可用性. Hua 等人在文献[29]中指出现有的 n-gram 和统一前缀树假设可能在一些情况下不成立，并提出了一种更通用的基于轨迹距离的差分隐私历史轨迹发布方案.

当前关于实时差分隐私位置或轨迹发布的研究成果还相对较少[17-18,30]. 主要难点在于：(1) 在历史轨迹的差分隐私保护中，可以采用一些全局优化策略来提高合成轨迹的隐私保护强度或可用性，而实时位置或轨迹发布无法达到；(2) 实时位置或轨迹发布需要实时处理各个时刻的位置发布，对算法的要求较高，而历史轨迹合成则可在线下处理. Andrés 等人在文献[17]中基于差分隐私的思想提出了 Geo-Indistinguishability 的位置隐私模型，并设计了一种极坐标下的拉普拉斯机制来实现该模型的隐私保护. 文献[18]在文献[17]的基础上，将差分隐私保护的发布机制看成最优化问题，提出了一种基于 δ-spanner 的近似解决方案. 文献[30]用马尔可夫链表示位置上的时序关系，并重新定义相邻数据集的概念，提出了一种基于凸包的差分隐私位置发布机制 PIM. 然而现有的这些差分隐私实时发布机制仅仅考虑了单个时刻上的位置发布，忽略了在连续发布过程中，已发布的当前位置对之前真实位置所产生的隐私泄露的影响.

3 预备知识

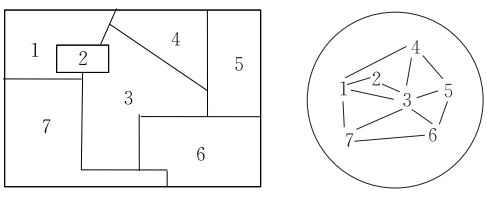
文中粗体小写字母表示向量(如 $\boldsymbol{x}$)，粗体大写字母表示矩阵(如 $\boldsymbol{X}$)，$\boldsymbol{x}[i]$ 指向量 $\boldsymbol{x}$ 中的第 i 个元素. $\|\cdot\|_p$ 表示 L_p 范数. 表 1 总结了文中的常用符号.

表 1 常用符号

符号表示	符号含义	符号表示	符号含义
SL^{initial}	初始敏感区域集合	$\Pr(z_i=a\mid o_j=b)$	当 $o_j=b$ 时 $z_i=a$ 的后验概率
PL^{initial}	初始敏感区域对应的隐私级别	$\Pr(o_t\mid z_t)$	已知 z_t 时发布区域为 o_t 的概率
$\boldsymbol{q}^{(t)}$	t 时刻真实位置的可能区域集合	$\boldsymbol{R}^{(t)}_{N_{t1}\times N_{t2}}$	t 时刻真实区域→发布区域概率矩阵
$\boldsymbol{h}^{(t)}$	t 时刻发布位置的可能区域集合	$\boldsymbol{M}$	时刻间真实区域转移概率矩阵
N_{t1}	t 时刻真实位置的可能区域个数	m_{ij}	当 $z_{t-1}=i$ 时 $z_t=j$ 的概率
N_{t2}	t 时刻发布位置的可能区域个数	ϵ_t	t 时刻真实位置区域的隐私预算
z_t	t 时刻的真实位置所在区域	w	时间窗口长度
o_t	t 时刻的发布位置所在区域	$\boldsymbol{X}^{t(t-1)}$	从 t 时刻真实位置到 $t-1$ 时刻真实位置的后验概率转移矩阵

3.1 隐私级别计算时的地图定义

假设给定一个城市的知识地图，我们首先对地图划分，将其表示成图 2(a)所示的区域分割图并依次编号，每块区域代表一个语义(如小区、商场等)的最小单位. 区域划分的粒度可以调节. 在获得区域划分图后，可以将其转换为带权的无向图 $G=\langle V,E\rangle$，如图 2(b)所示. 图中的节点 V 即为划分后的各区域；E 表示区域间的边，若两节点 v_i 和 v_j 间有边相连，则表示两区域可直接到达，边的权重表示两区域间的直接距离.



(a) 区域划分图 (b) 无向图表示

图 2 隐私级别计算时地图的区域划分和无向图表示

由于用户预先仅可能指定一些敏感区域，而对于地理拓扑上的支配关系无法一一得知，因而会遗漏一些应当敏感的区域(如第 1 节图 1 (a)所示). 假设初始敏感区域以及它们相应的隐私级别分别用集合 SL^{initial}，PL^{initial} 表示. 根据无向图 G 的拓扑关系，并利用第 4 节提出的 CPL 算法我们可以计算出其它区域的隐私级别.

3.2 位置发布时的地图定义

在获得各个区域的隐私级别后，根据第 4 节定义的 γ-隐私模型，可以获得每个区域相应的差分隐私预算. 此时，我们将地图划分为网格形式(如图 3 所示). 若该网络划分的粒度可调，则第 3.1 节中划分的一个区域可能包含多个网格，这些网格的隐私级别与其对应的区域相同.

值得注意的是，一个网格有可能会处于多个区

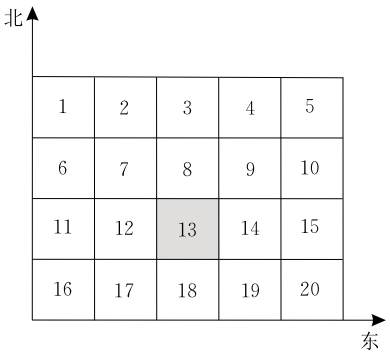


图 3　位置发布时的地图定义

域中(如边界上的网格),此时将各区域隐私级别最高的作为该网格的隐私级别.在图 3 中,如果当前的真实位置所在区域是 13,则我们将采用第 5 节中的差分隐私位置发布机制生成一个扰乱位置,以保护真实位置.

3.3　差分隐私及其在本文中的定义

差分隐私的思想是对原始数据或者原始数据上的函数、查询结果添加随机噪声,以使在数据集中插入或删除某一条记录不会影响输出结果,从而实现隐私保护.差分隐私算法通常是在相邻数据集概念的基础上构造的.相邻数据集指的是如果两个数据集 D 和 D' 的结构相同(即具有相同的维度和属性),并且它们之间有且仅有一条数据不相同,其它所有数据都相同,那么这两个数据集则称作相邻数据集.根据相邻数据集的概念,差分隐私的形式化定义如下.

定义 1.　差分隐私[21-22].给定相邻数据集 D 和 D' 及在 D 和 D' 上的一个算法 A,若算法 A 在 D 和 D' 上的任意输出结果 O 满足不等式

$$\Pr[A(D)=O]\leq \Pr[A(D')=O]\times e^{\varepsilon} \quad (1)$$

则称算法 A 满足差分隐私.

其中非负参数 ε 称为差分隐私预算,表示隐私保护程度,且 ε 越小隐私保护程度越高.可以看出,如果某算法 A 满足差分隐私,当 ε 趋近于 0 时,则 A 在相邻数据集 D 和 D' 上的输出结果趋于相同.这表明算法 A 不会泄露数据集中任何记录的敏感信息.

然而,对于位置隐私和轨迹隐私来说,并不存在传统相邻数据集的概念,因为用户的所有位置都是敏感的.本文采用文献[17,30]中的位置差分隐私模型,即如果在得知 t 时刻发布位置 o_t 后,根据已发布的位置推断 t 时刻真实位置的后验概率 $\Pr(z_t\mid o_t)$ 与 t 时刻真实位置的先验概率 $\Pr(z_t)$ 的比值满足差分隐私定义.

3.4　可用性定义

假设某时刻发布的位置是 o_t,其真实位置是 z_t,本文采用 z_t 和 o_t 之间的距离作为误差评价.即

$$dis(z_t,o_t)=\|z_t-o_t\|_2.$$

特别地,对于长度为 W 的轨迹,同样以距离误差[27,31]为基础定义位置可用性 $RMSE$,如式(2)所示,$RMSE$ 等于轨迹上处于敏感区域内的真实位置与其发布位置之间的均方根误差之和.

$$RMSE=\frac{1}{W}\sum_{t=1}^{W}\mathbb{I}(z_t)\cdot dis(z_t,o_t) \quad (2)$$

其中 $\mathbb{I}(z_t)$ 为指示函数,当 z_t 为敏感位置时,该值等于 1,否则该值等于 0.

4　γ-隐私模型

对于差分隐私位置扰乱的算法来说,在每一点上进行等程度的扰乱并不是很合理.原因有两点:一是用户可能在某些地方对隐私保护的需求很低,对这些位置进行大幅度地扰动会降低用户得到的服务质量;二是过度扰乱所产生的轨迹精确度较低,对于服务提供商进行隐私数据挖掘价值不大.因此,我们借用抑制策略的思想,为不同位置设定不同的隐私级别.如前文所述,预先设定的敏感位置可能会由于地理拓扑等因素有所遗漏.本节在 3.1 节的基础上提出了一种隐私级别计算算法 CPL,并据此提出了 γ-隐私模型的定义.

设用户事先的初始敏感位置集合 $SL^{\text{initial}}=\{sl_1,\cdots,sl_{|SL|}\}$ 及对应的隐私级别集合 $PL^{\text{initial}}=\{pl_1,\cdots,pl_{|SL|}\}$.$SL^{\text{initial}}$ 集合中的元素是区域划分后的区域编号,PL^{initial} 集合中的元素范围为(0,1](值越大表示该区域隐私等级越高,1 表示该区域隐私级别最高).根据区域划分的结果和初始敏感位置集合 SL^{initial},可以将整个地图划分为 $M=\{SL^{\text{initial}},NSL^{\text{initial}},NA\}$,其中 NSL^{initial} 为初始非敏感位置集合,NA 为地理上无法到达的位置(如湖泊).

前面我们提到,如果仅仅对敏感位置采取抑制或扰乱策略,攻击者仍可能根据位置间的地理关系来推断敏感位置.直观地考虑,敏感位置附近的区域也应当具有一定的隐私级别.一种策略是对与敏感位置相连的节点随机分配一定的隐私级别,这种方式安全性较高,然而可能会造成位置可用性的降低(即多数节点的隐私级别高)和部分敏感位置仍会被泄露(当图 1(a)中节点 A 的隐私级别非常小时).因此,我们将节点间的关联关系考虑进来,根据距离和

度将敏感位置的隐私级别分配给相邻节点.尽管在计算发布位置的时候(见第 5 节),攻击者会得知各个时刻用户所在真实位置所对应的隐私级别,但是由于只有用户自己知道所设敏感位置及附近位置在地图上的具体区域和相应的隐私级别(见算法 1),所以攻击者并不能将所知的隐私级别与地图上具体的敏感位置一一对应起来(除非该位置本身的隐私级别非常低,不需要保护).

假设敏感节点 v 的隐私级别为 pl,v 的邻接点集合为 neighborSet,其大小为图 G 中节点 v 的度.则对于 neighborSet 中的任一节点 g,其分配的隐私级别如式(3)所示.

$$g.pl=\frac{\lceil 1/(g.dis)\rceil\times(v.pl)}{\sum_{g'\in \text{neighborSet}} 1/(g'.dis)} \tag{3}$$

其中 $g.pl$ 表示节点 g 分配到的隐私级别,$g.dis$ 表示节点 g 与节点 v 之间的距离.显而易见,若节点 v 的度为 1,则邻接点的隐私级别与节点 v 相同.若节点 v 的度大于 1,则距离节点 v 越近的邻接点所分配的隐私级别越多.如果节点 v 的邻接点当前同时也在 SL 中,则取其本身的隐私级别和分配的隐私级别中最大的作为新的隐私级别.根据这个思想我们提出了 CPL 算法,伪代码如算法 1 所示.其中算法第 4 行的 $findNeighbors(v)$ 函数获得当前节点的邻接点集合,第 7 行的 $allocPrivLevel(g)$ 是根据式(3)进行隐私级别的分配.

算法 1. 计算各区域的隐私级别.

输入:地图的图表示 $G=\langle V,E\rangle$,地图的区域划分 $M=\{SL^{\text{initial}},NSL^{\text{initial}},NA\}$,初始敏感区域集合的隐私级别集合 PL^{initial},隐私级别阈值 $delta$

输出:敏感区域集合和对应的隐私级别

1. $SL=SL^{\text{initial}}$;
2. $v=SL.head()$;
3. WHILE v!=NULL:
4. $neighborSet=findNeighbors(v)$;
5. FOR all $g\in neighborSet$:
6. IF $g\in NA$ THEN: CONTINUE;
7. $newpl=allocPrivLevel(g)$; //式(3)
8. IF$newpl<delta$ THEN: CONTINUE;
9. IF $g\in SL$ THEN:
10. $g.pl=\max(g.pl,newpl)$;
11. ELSE:
12. $g.pl=newpl$;
13. $SL.append(g)$;
14. END FOR
15. $v=v.next()$;
16. END WHILE
17. RETURN SL;

在 CPL 算法中,阈值参数 δ(即隐私级别$\geq\delta$ 的区域需要执行隐私保护算法)可以起到一定的调节作用.当 $\delta=0$ 时,即要地图内的所有位置执行隐私保护算法,这种情况与基于位置扰乱的隐私保护算法类似,但是每个区域的隐私级别不同.当 $\delta=1$ 时,即只有位置在最敏感的区域内时才需要执行隐私保护算法,这种情况与传统的抑制策略类似,只是本文采用的是基于位置扰乱的差分隐私保护方法而不是完全禁止发布.另外,分析可知当图 G 中的每一个节点都与其它节点相连时,CPL 算法的计算量最大,即 CPL 的最坏时间复杂度为 $O(|V||V-1|)$.

在得到各个区域的隐私级别后,我们采用第 3.2 节的网格地图表示,并将隐私级别与差分隐私预算结合,提出了 γ-隐私的概念.

定义 2. 位置 γ-隐私.一个发布的位置满足 γ-隐私当且仅当该点的隐私级别 pl 与分配给该点的差分隐私保护预算 ε 满足 $\varepsilon\times pl=\gamma$.

由定义 2 可知,给定 γ 时,隐私级别 pl 越高,分配的隐私保护预算 ε 越小,隐私保护强度也越大,当 $pl=1$ 时,$\varepsilon=\gamma$.相反,隐私级别 pl 越低时,ε 越大,隐私保护强度也越低,当 $pl\to 0$ 时,$\varepsilon\to\infty$.

值得说明的是,本文根据隐私级别计算得到每个节点的差分隐私保护预算后,除非重新指定初始的敏感位置集合以及其相应的隐私级别,该节点的隐私保护预算即会固定不变.另外,对于一条轨迹来说,轨迹上的每个点都满足 $\varepsilon_i(i\in[1,w])$-差分隐私保护,根据差分隐私的序列组合性[20],整条轨迹必然也满足$(\sum_{i=1}^{w}\varepsilon_i)$-差分隐私.然而,这种结果在实际应用中并没有意义,因为有些节点的隐私级别非常低,其差分隐私保护预算可能会趋于无穷大,从这个角度来说,整条轨迹的隐私性也必然非常低.本文的出发点是保护轨迹上敏感位置及其附近位置的隐私,我们采用 w-滑动窗口的策略,假设当前时刻的位置对 $w-1$ 个时刻之前的真实位置的影响可以忽略不计,并在位置 γ-隐私定义基础上,直接定义轨迹 γ-隐私,如定义 3 所示(与文献[32]不同,在这种定义下我们不需要在轨迹上执行序列组合性然后对每个节点进行差分隐私预算分配).

定义 3. 轨迹 γ-隐私.设轨迹 T 长度为 W,若 $\forall i\in[1,W]$,T_i 满足 γ-隐私,则称 T 满足 γ-隐私.

显而易见,如果 T 上的每个点都满足位置 γ-隐私,则由定义 2 知,T 上任意连续的 w 个点都满足

γ-隐私. 又因为每个点对其 $w-1$ 个时刻之前的影响可以忽略，则轨迹 T 必然也满足 γ-隐私. 现在的问题在于，轨迹上各个点的位置可能存在一定的时序相关性，即 t 时刻发布的位置可能会在某种程度上泄露之前时刻（如 $t-1,t-2,\cdots$）的真实位置，从而对前面时刻的位置发布产生影响，使得其不满足位置 γ-隐私定义. 因此，接下来我们提出一种差分隐私位置发布机制，使得 t 时刻的位置发布不会破坏之前时刻的位置 γ-隐私定义.

5　差分隐私位置发布机制

5.1　时序相关的差分隐私位置发布问题

在每个时刻，如果用户所处的位置是敏感的，则这个真实位置只有自己可见；然后根据差分隐私位置发布机制发布扰乱位置，而外界（包括潜在的攻击者）都可以获得该发布位置. 本文采用马尔可夫链模拟用户真实位置间的时序相关性[33-34]. 状态转移概率矩阵 $\boldsymbol{M}$ 表示真实位置在区域间的转移可能性，如 $\boldsymbol{M}$ 中元素 m_{ij} 表示用户从第 i 个区域移动到第 j 个区域的概率. 这里假设 $\boldsymbol{M}$ 可事先在历史记录上计算得到.

假设用户在每个时刻产生的真实位置区域为 $\boldsymbol{z}=\{z_1,z_2,\cdots,z_t\}$，经过隐私保护算法处理后发布的扰乱位置区域为 $\boldsymbol{o}=\{o_1,o_2,\cdots,o_t\}$. 设在 t 时刻前，可以根据隐私保护算法的位置发布机制和状态转移概率矩阵 $\boldsymbol{M}$ 推测出 t 时刻的位置在集合 $\boldsymbol{q}^{(t)}=\{q_1^{(t)},\cdots,q_{N_{t1}}^{(t)}\}$ 中（其中 N_{t1} 表示集合中元素的个数）以及在每个区域的概率值

$$\boldsymbol{p}^{(t)-}=\{p_1^{(t)-},\cdots,p_{N_{t1}}^{(t)-}\}$$

这里 $\boldsymbol{p}^{(t)-}$ 相当于 t 时刻发布扰乱位置前的先验概率分布. 当 t 时刻发布扰乱位置 o_t 后，可以根据扰乱位置推测真实位置 z_t，即后验概率 $\Pr(z_t\mid o_t)$，由此我们可以得到在集合 $\boldsymbol{q}^{(t)}$ 上的后验概率分布

$$\boldsymbol{p}^{(t)+}=\{p_1^{(t)+},\cdots,p_{N_{t1}}^{(t)+}\}.$$

对于单时刻的差分隐私位置保护[17-18,30]，期望后验概率与先验概率的比值满足差分隐私定义，即

$$\forall i\in[1,N_{t1}],\ p_i^{(t)+}\le e^{\varepsilon_t}\cdot p_i^{(t)-},$$

其中 ε_t 是 t 时刻真实位置所在区域的差分隐私预算. 然而这几种方案仅仅考虑了已发布的扰乱位置对即将发布位置的时序影响，并未考虑当前发布的扰乱位置对之前时刻真实位置的隐私泄露情形. 本文将同时考虑当前时刻发布的扰乱位置（如 o_t）对之前的 $w-1$ 个时刻真实位置（$z_{t-w+1},\cdots,z_{t-1}$）的后验概率的影响，并提出一种基于可用性的优化算法来求解差分隐私发布机制.

如表 2 所示，在第 t 时刻，根据发布的扰乱位置 o_t 可以求得后验概率分布 $\boldsymbol{p}^{(i)+}(o_t)$，若位置发布机制满足差分隐私，则必须满足

$$\boldsymbol{p}^{(t)+}(o_t)\le e^{\varepsilon_t}\cdot\boldsymbol{p}^{(t)-},\ \forall o_t\in\boldsymbol{h}^{(t)}\tag{4}$$

表 2　每个时刻的概率求解图

时刻 1	时刻 2	时刻 3	…	时刻 $t-1$	时刻 t
$\boldsymbol{p}^{(1)+}$	$\boldsymbol{p}^{(1)+}$	$\boldsymbol{p}^{(1)+}$	…		
↓	↑	↑			
$\boldsymbol{p}^{(2)-}$	$\boldsymbol{p}^{(2)+}$	$\boldsymbol{p}^{(2)+}$	…	$\boldsymbol{p}^{(t-w)+}$	
	↓	↑		↑	
	$\boldsymbol{p}^{(3)-}$	$\boldsymbol{p}^{(3)+}$	…	$\boldsymbol{p}^{(t-w+1)+}$	$\boldsymbol{p}^{(t-w+1)+}$
		↓		↑	↑
		$\boldsymbol{p}^{(4)-}$	…	…	…
			…	$\boldsymbol{p}^{(t-1)+}$	$\boldsymbol{p}^{(t-1)+}$
				↓	↑
			…	$\boldsymbol{p}^{(t)-}$	$\boldsymbol{p}^{(t)+}$

其中 $\boldsymbol{p}^{(t)-}=\boldsymbol{p}^{(t-1)+}\boldsymbol{M}$，$\boldsymbol{h}^{(t)}$ 表示发布位置可能在的区域集合. 同时，可根据 $\boldsymbol{p}^{(t)+}$ 求得已知 o_t 时，前 $w-1$ 时刻真实位置的后验概率分布 $\boldsymbol{p}^{(t-1)+},\cdots,\boldsymbol{p}^{(t-w-1)+}$；并将 $t-1$ 时刻所得到的对应后验概率分布作为先验概率分布，若发布的位置所在区域满足差分隐私，则还须满足：

$$\boldsymbol{p}^{(j)+}(o_t)\le e^{\varepsilon_j}\cdot\boldsymbol{p}^{(j)-},$$
$$\forall j\in[t-w+1,t-1]\text{ and }\forall o_t\in\boldsymbol{h}^{(t)}\tag{5}$$

综合式(4)和式(5)，可知在考虑当前发布位置对之前真实位置的影响下，t 时刻的差分隐私位置发布机制须满足：

$$\boldsymbol{p}^{(j)+}(o_t)\le e^{\varepsilon_j}\cdot\boldsymbol{p}^{(j)-},$$
$$\forall j\in[t-w+1,t]\text{ and }\forall o_t\in\boldsymbol{h}^{(t)}\tag{6}$$

即在获知 t 时刻的发布位置后，所得到前 w 个时刻真实位置的后验概率分布与 $t-1$ 时刻所得到的后验概率分布（在 t 时刻可看成先验概率分布）比值小于等于 e^{ε_j}，ε_j 为 j 时刻真实位置所在区域的差分隐私预算.

现在关键问题在于如何求出差分隐私位置发布概率矩阵 $\boldsymbol{R}^{(t)}$. 设 t 时刻发布位置的可能区域集合为 $\boldsymbol{h}^{(t)}=\{h_1^{(t)},\cdots,h_{N_{t2}}^{(t)}\}$. 机制 $\boldsymbol{R}^{(t)}$ 应当是维度为 $N_{t1}\times N_{t2}$ 的概率矩阵（通常 N_{t2} 大于 N_{t1}），其中元素 r_{ij}^t 表示 t 时刻真实区域为 $q_i^{(t)}$ 时发布的扰乱区域为 $h_j^{(t)}$ 的概率. 与文献[18]类似，本文将 $\boldsymbol{R}^{(t)}$ 的求解看成一个优化问题，但本文采用的是后验概率分布与先验概率分布的比作为约束条件，并且考虑了当前发布位置对之前时刻真实位置的影响. 目标函数如下：

$$\min_{\boldsymbol{R}} \sum_{i=1}^{N_{t1}} \sum_{j=1}^{N_{t2}} p_i^{(t)-} \cdot r_{ij}^t \cdot dis(q_i^{(t)}, h_j^{(t)}) \tag{7}$$

约束条件为式(6)及每个元素的值不小于 0 且每行的和为 1. 其中 $dis(\cdot)$ 函数是两个区域中心的欧几里得距离. 这样我们就将 t 时刻的差分隐私位置发布机制转换成了上式的基于欧几里得距离(与第 3.4 节的可用性对应)的优化问题. 上述目标函数的解 $\boldsymbol{R}^{(t)^*}$ 即不仅使得 t 时刻的真实位置发布满足其所在区域的差分隐私强度,还能使得之前的 $w-1$ 个时刻的真实位置满足各自区域的差分隐私强度.

5.2 差分隐私位置发布机制 DPLRM

为了求解上述目标函数,首先应当求解当 t 时刻发布位置 o_t 已知时,t 时刻真实位置 z_t 的后验概率分布. 根据贝叶斯公式有

$$\begin{aligned}
&\Pr(z_t = q_i^{(t)} \mid o_t = h_j^{(t)}) \\
&= \frac{\Pr(o_t = h_j^{(t)} \mid z_t = q_i^{(t)}) \Pr(z_t = q_i^{(t)})}{\sum_{a=1}^{N_{t1}} \Pr(o_t = h_j^{(t)} \mid z_t = q_a^{(t)}) \Pr(z_t = q_a^{(t)})} \\
&= \frac{r_{ij} p_i^{(t)-}}{\sum_{a=1}^{N_{t1}} r_{aj} p_a^{(t)-}},
\end{aligned}$$

用向量形式表示 z_t,当给定发布位置 o_t 时,t 时刻真实位置的后验概率分布向量为

$$\begin{aligned}
\boldsymbol{p}^{(t)+}(o_t = h_j^{(t)}) &= \Pr(z_t \in \boldsymbol{q}^{(t)} \mid o_t = h_j^{(t)}) \\
&= \frac{\boldsymbol{R}_j^{\mathrm{T}} \cdot \boldsymbol{p}^{(t)-}}{\boldsymbol{R}_j^{\mathrm{T}} \times (\boldsymbol{p}^{(t)-})^{\mathrm{T}}}
\end{aligned} \tag{8}$$

其中 $\boldsymbol{R}_j^{\mathrm{T}}$ 是原始矩阵第 j 列的转置,分子表示元素乘,分母是向量内积. 对于所有的 $o_t \in \boldsymbol{h}^{(t)}$,$t$ 时刻的后验概率分布可以写成矩阵形式

$$\boldsymbol{P}^{(t)+} = (\boldsymbol{p}^{(t)+}(o_t = h_1^{(t)}), \cdots, \boldsymbol{p}^{(t)+}(o_t = h_{N_{t2}}^{(t)}))^{\mathrm{T}} \tag{9}$$

现在求当 t 时刻发布位置 o_t 已知时,$t-1$ 时刻真实位置的后验概率分布. 设 $t-1$ 时刻真实位置区域的集合为 $\boldsymbol{q}^{(t-1)}$,则根据全概率公式有

$$\begin{aligned}
&\Pr(z_{t-1} = q_i^{(t-1)} \mid o_t) = \\
&\sum_{a=1}^{N_{t1}} \Pr(z_{t-1} = q_i^{(t-1)} \mid z_t = q_a^{(t)}) \Pr(z_t = q_a^{(t)} \mid o_t),
\end{aligned}$$

其中 $\Pr(z_{t-1} \mid z_t)$ 表示当已知 t 时刻真实位置 z_t 时,$t-1$ 时刻真实位置为 z_{t-1} 的概率. 因为真实位置的转移概率服从马尔可夫链模型,可由贝叶斯公式得该后验概率分布.

$$\begin{aligned}
&\Pr(z_{t-1} = q_i^{(t-1)} \mid z_t = q_a^{(t)}) = \\
&\frac{\Pr(z_t = q_a^{(t)} \mid z_{t-1} = q_i^{(t-1)}) \Pr(z_{t-1} = q_i^{(t-1)} \mid o_{t-1})}{\sum_{b=1}^{N_{(t-1)1}} \Pr(z_t = q_a^{(t)} \mid z_{t-1} = q_b^{(t-1)}) \Pr(z_{t-1} = q_b^{(t-1)} \mid o_{t-1})}
\end{aligned}$$

值得注意的是,这里用的是 $t-1$ 时刻的后验概率分布 $\Pr(z_{t-1} \mid o_{t-1})$ 作为当前的先验概率分布. 特别地,可以得到从 t 时刻真实位置到 $t-1$ 时刻真实位置的后验概率转移矩阵 $\boldsymbol{X}^{t(t-1)}$,矩阵的每列代表 t 时刻真实位置为 $z_t = q_a^{(t)}$,相应 $t-1$ 时刻真实位置的概率转移向量. 因此我们可以将已知 o_t 时,$t-1$ 时刻的后验概率分布写成矩阵相乘的形式:

$$\boldsymbol{P}^{(t-1)+} = \boldsymbol{P}^{(t)+} \boldsymbol{X}^{t(t-1)} \tag{10}$$

依次类推,可得 $t-2, \cdots, t-w+1$ 时刻的后验概率分布的矩阵形式如下所示:

$$\begin{cases}
\boldsymbol{P}^{(t-2)+} = \boldsymbol{P}^{(t)+} \boldsymbol{X}^{t(t-1)} \boldsymbol{X}^{(t-1)(t-2)} \\
\cdots \\
\boldsymbol{P}^{(t-w+1)+} = \boldsymbol{P}^{(t)+} \boldsymbol{X}^{t(t-1)} \cdots \boldsymbol{X}^{(t-w+2)(t-w+1)}
\end{cases} \tag{11}$$

根据上述推导,在 $t-1$ 时刻,计算得出相应的差分隐私位置发布机制 $\boldsymbol{R}^{(t-1)}$ 后,可以得到一个发布位置 o_{t-1}. 根据 o_{t-1} 及 $\boldsymbol{R}^{(t-1)}$ 可以得到 $t-w$ 到 $t-1$ 时刻的真实位置后验概率分布 $\boldsymbol{p}^{(t-1)+}, \cdots, \boldsymbol{p}^{(t-w)+}$ 和 t 时刻真实位置的先验概率分布 $\boldsymbol{p}^{(t)-}$. 值得注意的是 $t-1$ 时求得的真实位置间的概率转移矩阵 $\boldsymbol{X}^{(t-1)(t-2)}$ 与 t 时刻所求的不同. 在 t 时刻,我们希望找到一个差分隐私位置发布机制 $\boldsymbol{R}^{(t)}$,使得根据该机制得到的任意发布位置 o_t 都满足目标函数的条件,且损失最小.

综上所述,我们可以计算出目标函数式(7)的约束条件(式(6))具体值. 特别地,因为 t 时刻之前的 $w-1$ 个时刻的约束条件都可以通过式(10)和式(11)转换成与 t 时刻约束条件类似的不等式,所以目标函数约束条件的数量可以从 wN^2 减少为 N^2,其中 $N \geq \max\{N_{i1}, N_{i2}\}, \forall i \in [t-w+1, t]$.

在获得各线性不等式约束条件后,采用优化算法对其进行迭代求解[35-36]. 下面给出 t 时刻时序相关的差分隐私位置发布机制 DPLRM 的主要流程(伪代码见算法 2).

(1) 初始化矩阵 $\boldsymbol{R}^{(t)}$;

(2) 由式(9)计算 o_t 已知时 z_t 的后验概率分布 $\boldsymbol{P}^{(t)+}$;

(3) 由式(10)和式(11)计算时刻间的概率转移矩阵 $\boldsymbol{X}$ 和 o_t 已知时 $z_{t-1}, \cdots, z_{t-w+1}$ 的后验概率分布 $\boldsymbol{P}^{(t-1)+}, \cdots, \boldsymbol{P}^{(t-w+1)+}$;

(4) 计算各个约束条件的值,并根据各个时刻的后验概率矩阵 $\boldsymbol{P}^{(t)+}, \boldsymbol{P}^{(t-1)+}, \cdots, \boldsymbol{P}^{(t-w+1)+}$ 更新矩阵 $\boldsymbol{R}^{(t)}$ 的参数;

(5) 重复(2)~(4)步,直到达到迭代结束条件;

《计算机学报》

得到差分隐私位置发布机制 $\boldsymbol{R}^{(t)^*}$；

(6)根据当前所在的真实位置和 $\boldsymbol{R}^{(t)^*}$ 相应的行，依概率随机在 $\boldsymbol{h}^{(t)}$ 中选择一个位置发布.

算法 2.　差分隐私位置发布(DPLRM).

输入：真实位置的区域转移概率矩阵 $\boldsymbol{M}$，时间窗口长度 w，各个时刻的隐私保护预算向量 $\boldsymbol{\varepsilon}=\{\varepsilon_1,\cdots,\varepsilon_t\}$，各个时刻真实位置的可能集合 $\boldsymbol{Q}=\{\boldsymbol{q}^{(1)},\cdots,\boldsymbol{q}^{(t-1)}\}$，$t$ 时刻的先验概率分布向量 $\boldsymbol{p}^{(t)-}$，前 $t-1$ 时刻的后验概率分布矩阵集合 $\hat{\boldsymbol{P}}=\{\mathbf{P}^{(1)+},\cdots,\boldsymbol{P}^{(t-1)+}\}$，$t$ 时刻的真实位置 z_t，终止条件 $cond$

输出：t 时刻的发布位置.

1. q_vec, $h_vec=chooseLoc(\boldsymbol{p}^{(t)-})$;
2. $N1=q_vec.length()$, $N2=h_vec.length()$;
3. $R=$initialize();
4. IF $t==1$ THEN:
5. $Q_used=\varnothing$; $P_used=\varnothing$;
6. ELSE IF $t>1$ and $t<w$ THEN:
7. $Q_used=\boldsymbol{Q}$; $P_used=\hat{\boldsymbol{P}}$
8. ELSE:
9. $Q_used=selectQ(\boldsymbol{Q})$; $P_used=selectP(\hat{\boldsymbol{P}})$;
10. WHILE $cond$!=TRUE:
11. $P_plus=calPosterior(\boldsymbol{R},\boldsymbol{p}^{(t)-},q_vec,h_vec)$;
12. $X_used=\varnothing$;
    //时刻间的真实位置概率转移矩阵集合
13. FOR all $j\in Q_used.length()$:
14. $X_used.append(compute(Q_used,P_used,P_plus))$;
15. $P_used[t-j]=calPosterior2(P_used,Q_used,P_plus)$;
    //根据式(10)、(11)计算前面时刻的后验概率矩阵
16. END FOR
17. $Conditions=calConds(P_used,P_plus,\varepsilon,X_used)$;
18. $updateR(\boldsymbol{R},Conditions,N1,N2)$;
19. END WHILE
20. $o_t=RandomChoose(\boldsymbol{R},z_t)$;
21. RETURN o_t;

5.3　算法分析

关于时间复杂度，在 DPLRM 算法的每次迭代中，最耗时的部分在于对 w 个时刻计算其后验概率矩阵 $\boldsymbol{P}^{(t)+},\cdots,\boldsymbol{P}^{(t-w+1)+}$(根据式(9)～(11))，因此每次迭代的时间复杂度为 $O(wN^3)$，其中 N 为 N_{i1}，$N_{i2}(\forall i\in[t-w+1,t])$的最大值.值得注意的是，尽管 DPLRM 算法的计算复杂度相对较高，但 N 的值不会很大，因为下一时刻的位置不会偏离上一时刻的位置太远.另外，也可以将算法中 while 循环这部分计算量(即差分隐私位置发布矩阵 $\boldsymbol{R}$ 的计算)分配到服务器或进行并行计算，而在移动端仅仅通过当前的真实位置计算最终的发布位置(即使服务器知道每个时刻的发布位置和 $\varepsilon_i(i\in[1,t])$，它也很难将 ε_i 与真实位置对应起来，因为这种对应关系只有移动端自身知道).

关于隐私性，根据 t 时刻差分隐私位置发布机制 DPLRM，可知 t 时刻的发布位置 o_t 既使得当前时刻的真实位置 z_t 满足 ε_t 差分隐私，还使得之前 $w-1$ 个时刻的所有真实位置 $z_{t-w+1},\cdots,z_{t-1}$ 分别满足 $\varepsilon_{t-w+1},\cdots,\varepsilon_{t-1}$ 差分隐私.因此，由位置 γ-隐私模型可知，在长度为 w 轨迹上的每个点都满足位置 γ-隐私，即满足轨迹 γ-隐私(见定义 3).

关于可用性，由于 DPLRM 算法主要是通过对式(7)进行求解来生成扰乱的发布位置，而式(7)的目标就是最小化所有情况下真实位置和发布位置间的距离总和，因此在保证之前 $w-1$ 时刻真实位置满足 γ-隐私的前提下，DPLRM 算法发布位置的整体可用性是最优的.当 w 增大时，需要纳入考虑的时刻增多，目标函数(式(7))中的约束条件可能会变得更加严格，从而使发布位置偏离真实位置的程度和概率变大，一定程度上降低可用性.

6　实验与分析

6.1　实验设置

CPL 算法和 DPLRM 机制均采用 Python 实现，在 3.60GHz CPU、8.00GB RAM 的 Windows 7 平台上运行.本文采用的数据集是 Geolife[①] 和 Gowalla[②] 真实数据集.数据集 Geolife 采集了 182 个用户从 2007 年 4 月到 2012 年 8 月在北京活动的真实数据，数据集共包含 17 621 条轨迹.Geolife 数据集中包括用户编号、时间戳、经度、纬度、海拔等属性，我们抽取五环以内轨迹的前 4 个属性作为新的数据集.数据集 Gowalla 采集了 2009 年 2 月到 2010 年 10 月共 15 116 个用户在移动社交网站上(加州范围内)的签到数据.同 Geolife 一样，我们抽取洛杉矶范围内的用户编号、时间戳、经度、纬度作为新的数据集.

对于 CPL 算法，我们将地图转换成无向图表示，并将每个用户停留时间最长或访问次数最多的 $k(k\in[5,10,15,20,25])$ 个区域作为初始敏感位置集合，且设其隐私级别等于 1，我们衡量阈值参数 δ $(\delta\in[0.1,0.2,0.3,0.4,0.5])$对算法运行时间的影

① http://research.microsoft.com/en-us/downloads/
② http://snap.stanford.edu/data/loc-gowalla.html

《计算机学报》

响.对于 DPLRM 机制,与文献[30]类似,我们将北京地图划分为大小为 0.34 km×0.34 km 的网格,将洛杉矶地图划分为 0.89 km×0.89 km 的网格,并在此基础上获得相对应的马尔可夫概率转移矩阵.我们衡量 500 个时间戳内 γ($\gamma\in[0.2,0.4,0.6,0.8,1.0]$)对位置可用性的影响、CPL 算法阈值参数 δ($\delta\in[0.1,0.2,0.3,0.4,0.5]$)和初始敏感集合大小 k ($k\in[5,10,15,20,25]$)对位置可用性的影响及时间窗口 w($w\in[1,3,5,7,9]$)对算法运行时间和位置可用性的影响.其中,位置可用性的定义为时间戳内敏感区域上真实位置和发布位置之间的距离总误差的均值 *RMSE*(见第 3.4 节式(2)),*RMSE* 越小,位置可用性越高.

6.2 实验结果和分析

6.2.1 CPL 算法表现

我们首先分析阈值参数 δ 对 CPL 算法运行时间的影响,结果如图 4 所示.在这一系列实验中,初始敏感位置集合 SL^{initial} 大小 k 的默认值是 10.由图 4 可以看出,当 δ 逐渐增加时,CPL 算法在两个数据集上的运行时间逐渐降低.这是因为,当 δ 变大时,CPL 算法的剪枝操作很有效(见算法 1 的第 8 行),因此算法的运行效率会有很大的提高.同时,我们可以看出 CPL 算法在 Geolife 数据集上的运行时间略大于在 Gowalla 数据集上的运行时间,这主要是因为北京在地理上的语义集合略大于洛杉矶的,同样的遍历,Geolife 数据集所需要的时间稍多.

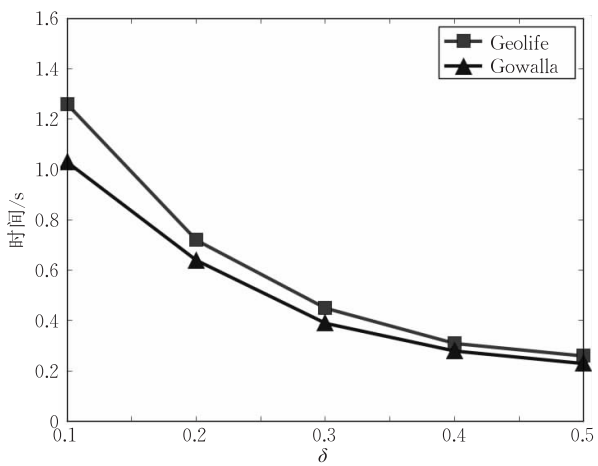


图 4 阈值 δ 对 CPL 算法运行时间的影响

然后,我们分析初始敏感位置集合 SL^{initial} 大小 k 对 CPL 算法运行时间的影响,结果如图 5 所示.在这一系列实验中,δ 的默认值设为 0.2.由图 5 可以看出,当 k 不断增加时,CPL 算法在两个数据集上的运行时间也相应增加.这是因为 k 很大时,算法所遍历的地理空间增加,进而所需时间也增加.

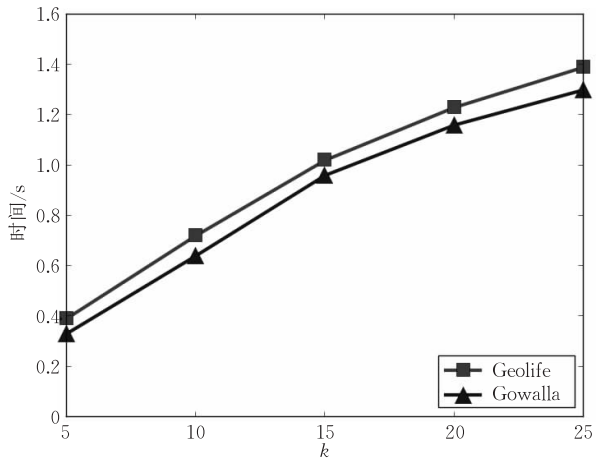


图 5 k 对 CPL 算法运行时间的影响

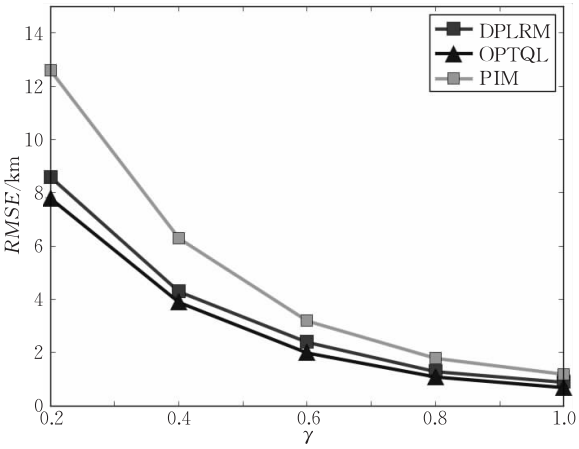
6.2.2 DPLRM 算法表现

对于 DPLRM 算法,我们主要关注隐私模型参数 γ、CPL 算法阈值 δ 和时间窗口长度 w 的影响.在衡量 γ 和 δ 对位置可用性的影响时,我们将 DPLRM 与 OPTQL 算法[18] 和 PIM 算法[30] 对比.OPTQL 考虑单个时刻的最优差分隐私位置发布,PIM 是一种基于敏感度包的差分隐私位置发布机制.

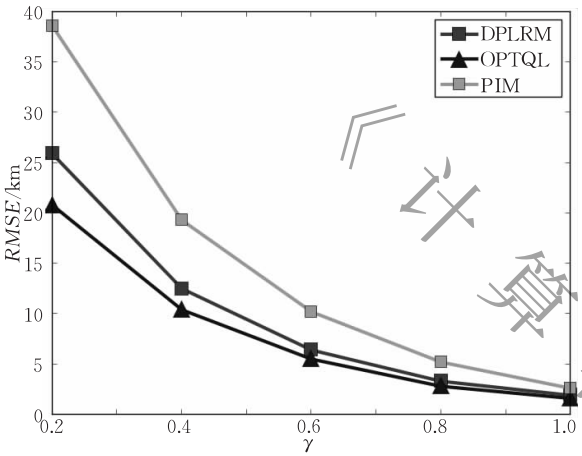
我们首先分析 γ 对位置可用性的影响,结果如图 6 所示.在这一系列实验中,为方便衡量 γ 的影响,我们假设时间窗口长度 $w=3$,且阈值 $\delta=1.0$,也就是说只有初始设定的敏感集合需要采用三种算法.此时的 γ 与传统差分隐私中的隐私保护预算 ε 等价.由图 6(a)可以看出,PIM 算法的位置可用性相对最差,这是因为 PIM 算法仅提出了一种满足差分隐私的位置发布机制,在执行该发布机制时并未将位置的可用性考虑在内;OPTQL 算法的位置可用性最好,这是因为该算法的目标就是在满足差分隐私的同时最小化发布位置的误差;DPLRM 算法介于两个算法之间,并接近于 OPTQL 算法的表现.这是因为 DPLRM 不仅考虑了当前发布位置对当前时刻的隐私影响,还考虑了当前发布位置对之前发布位置的隐私影响,所以位置可用性略弱于 OPTQL 算法.图 6(b)在 Gowalla 数据集上显示了类似的表现.另外,三种算法在 Geolife 数据集上的表现优于在 Gowalla 数据集上的表现,原因是在 Geolife 上划分的网格面积小于在 Gowalla 上的划分,所以网格更密集、结果更精确.

然后,我们分析阈值 δ 对三种算法位置可用性的影响,结果如图 7 所示.在这一系列实验中,我们设定 $\gamma=0.2$,$w=3$;并且三种算法采用相同的 CPL 算法结果,即在敏感区域集合上的差分隐私预算根

《计算机学报》

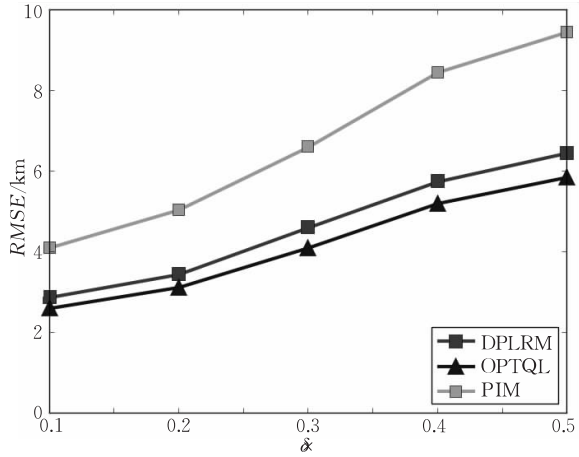


(a) Geolife数据集

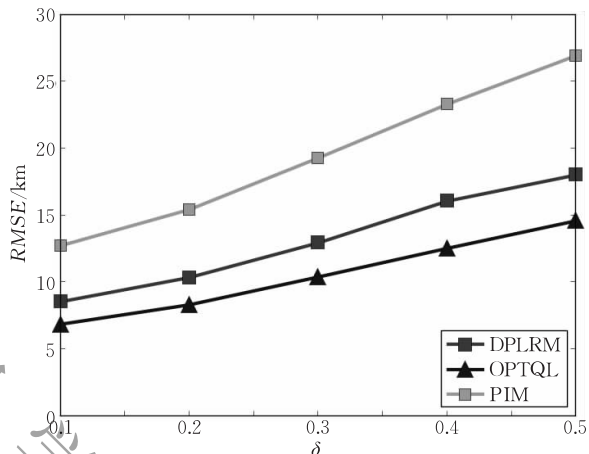


(b) Gowalla数据集

图 6　γ 对 *RMSE* 的影响



(a) Geolife数据集



(b) Gowalla数据集

图 7　阈值 δ 对 *RMSE* 的影响

据定义 2 获得. 由图 7(a)可以看出，δ 越小时，三种算法的 *RMSE* 越小. 这是因为当 δ 小时，CPL 算法所得结果中会包含较多的隐私级别 pl 较低的区域；又由定义 2 可知，各区域的差分隐私预算 ε 与其隐私级别 pl 成反比，因而在这种情况下，会有较多的区域拥有较大的差分隐私预算 ε，最终使得平均的位置可用性提高(即 *RMSE* 降低). 图 7(b)在 Gowalla 数据集上显示了相似的结果.

最后，我们分析时间窗口长度 w 对位置可用性的影响，结果如图 8 所示. 在这一系列实验中，我们假设 $\gamma=0.2$、$\delta=0.2$. 由图 8 可知，当 w 逐渐增大时，*RMSE* 也相应增加，这是因为当 w 很大时，DPLRM 算法求得的解可能要更偏离真实位置才能满足当前发布位置在 w 个时刻上都满足差分隐私，因而位置误差相对较大. 图 9 显示了时间窗口长度 w 对 DPLRM 算法运行时间的影响. 如图所示，当 w 增大时，算法所需的运行时间也越长，因为 DPLRM 算法要逆向在 w 时间窗口内的每个时刻求解当前时刻发布位置对之前真实位置的后验概率，所以所需时间也越长. 值得注意的是，w 的大小对最优化问题的求解影响相对较小，如第 5.2 节所述，最优化问题的条件最终可以约减到 N^2，因此，这部分的运行时间与 w 的值关系较小.

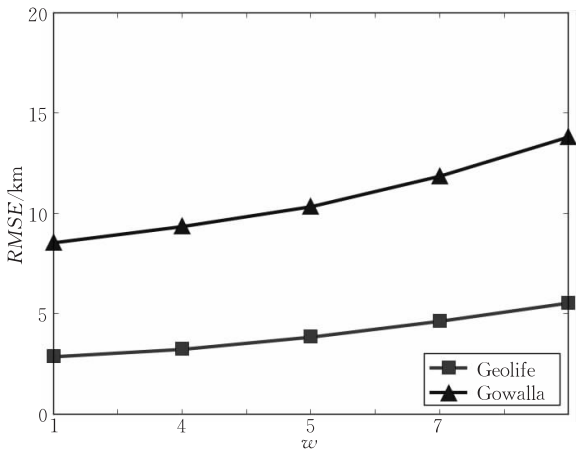


图 8　时间窗口长度 w 对 *RMSE* 的影响

《计算机学报》

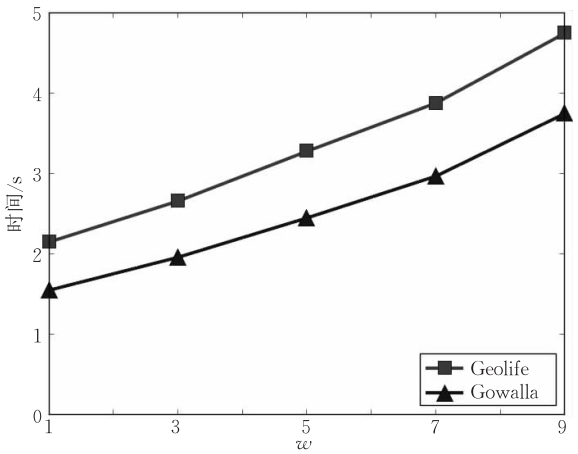


图 9 时间窗口长度 w 对 DPLRM 运行时间的影响

7 总 结

本文针对基于位置的服务中的轨迹隐私保护问题,提出了一种基于地理空间拓扑关系的区域隐私级别计算方法 CPL 以及结合区域隐私级别与差分隐私预算的 γ-隐私模型. CPL 算法利用无向图表示地理间的拓扑关系,并根据节点的度和节点间的距离将预先设定的敏感区域的隐私级别分配给相邻节点,以使得攻击者无法根据地理限制推测用户的隐私. 然后,分析了当前发布位置对轨迹上真实位置的隐私影响,提出了一个基于可用性的差分隐私位置发布机制的最优化问题,以及相应的算法 DPLRM. 实验结果表明,DPLRM 在达到轨迹差分隐私效果的同时具有较好的可用性. 今后的研究将考虑如下两个方面:(1)在根据地理拓扑关系计算初始敏感位置附近位置的隐私级别时,将用户的移动模式(如交通方式)以及时间因素(如用户可以自定义初始敏感位置的隐私级别随时间变化的曲线)考虑在内,以提供更细致的隐私级别计算;(2)研究如何对 DPLRM 算法进行继续优化,以提高扰乱位置发布的可用性并减少算法的运行时间,更好地扩展到实时的位置服务中.

参 考 文 献

[1] Gruteser M, Grunwald D. Anonymous usage of location-based services through spatial and temporal cloaking// Proceedings of the 1st International Conference on Mobile Systems, Applications, and Services (MobiSys2003). San Francisco, USA, 2003: 31-42

[2] Mokbel M F, Chow C Y, Aref W G. The new casper: Query processing for location services without compromising privacy //Proceedings of the 32nd Conference of Very Large Data Bases (PVLDB2006). Seoul, South Korea, 2006: 763-77

[3] Huo Zheng, Meng Xiao-Feng. A survey of trajectory privacy preserving techniques. Chinese Journal of Computers, 2011, 34(10): 1820-1830(in Chinese)
(霍峥, 孟小峰. 轨迹隐私保护技术研究. 计算机学报, 2011, 34(10): 1820-1830)

[4] Chow C Y, Mokbel M F. Enabling private continuous queries for revealed user locations//Proceedings of the 10th International Symposium on Spatial and Temporal Databases (SSTD 2007). Boston, USA, 2007: 258-275

[5] Pan X, Meng X, Xu J. Distortion-based anonymity for continuous queries in location-based mobile services// Proceedings of the 17th ACM SIGSPATIAL International Symposium on Advances in Geographic Information Systems (ACM-GIS 2009). Seattle, USA, 2009: 256-265

[6] Bamba B, Liu L, Pesti P, Wang T. Supporting anonymous location queries in mobile environments with PrivacyGrid// Proceedings of the 17th International Conference on World Wide Web (WWW 2008). Beijing, China, 2008: 237-246

[7] Ghinita G, Kalnis P, Skiadopoulos S. PRIVE: Anonymous location-based queries in distributed mobile systems// Proceedings of the 16th International Conference on World Wide Web (WWW 2007). Banff, Canada, 2007: 371-380

[8] Huo Zheng, Meng Xiao-Feng, Huang Yi. PrivateCheckIn: Trajectory privacy-preserving for check-in services in MSNS. Chinese Journal of Computers, 2013, 36(4): 716-726(in Chinese)
(霍峥, 孟小峰, 黄毅. PrivateCheckIn: 一种移动社交网络中的轨迹隐私保护方法. 计算机学报, 2013, 36(4): 716-726)

[9] Freudiger J, Raya M, Felegyhazi M, et al. Mix-zones for location privacy in vehicular networks//Proceedings of the 1st International Workshop on Wireless Networking for Intelligent Transportation Systems (WiN-ITS 2007). Vancouver, Canada, 2007: 1-7

[10] Freudiger J, Shokri R, Hubaux J P. On the optimal placement of mix zones//Proceedings of the 9th International Symposium on Privacy Enhancing Technologies (PETS 2009). Seattle, USA, 2009: 216-234

[11] Palanisamy B, Liu L. MobiMix: Protecting location privacy with mix zones over road networks//Proceedings of the 27th International Conference on Data Engineering (ICDE 2011). Hannover, Germany, 2011:494-505

[12] Chen R, Fung B C M, Mohammed N, et al. Privacy-preserving trajectory data publishing by local suppression. Information Science, 2013, 231: 83-97

[13] Gruteser M, Liu X. Protecting privacy in continuous location-tracking applications. IEEE Security and Privacy, 2004, 2(2): 28-34

[14] Zhao Jing, Zhang Yuan, Li Xing-Hua, Ma Jian-Feng. A trajectory privacy protection approach via trajectory frequency suppression. Chinese Journal of Computers, 2014, 37(10): 2096-2106(in Chinese)
(赵婧，张渊，李兴华，马建峰. 基于轨迹频率抑制的轨迹隐私保护方法. 计算机学报，2014，37(10)：2096-2106)

[15] Lee K C K, Lee W C, Leong H V, Zheng B. Navigational path privacy protection//Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM 2009). Hong Kong, China, 2009: 691-700

[16] Shokri R, Theodorakopoulos G, Troncoso C, et al. Protecting location privacy: Optimal strategy against localization attacks//Proceedings of the ACM Conference on Computer and Communications Security (CCS2012). Raleigh, USA, 2012: 617-627

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

[18] Bordenabe N E, Chatzikokolakis K, Palamidessi C. Optimal geo-indistinguishable mechanisms for location privacy//Proceedings of the ACM Conference on Computer and Communications Security (CCS 2014). Scottsdale, USA, 2014: 251-262

[19] Theodorakopoulos G, Shokri R, Troncoso C, et al. Prolonging the hide-and-seek game: Optimal trajectory privacy for location-based services//Proceedings of the 13th Workshop on Privacy in the Electronic Society (WPES2014). Scottsdale, USA, 2014: 73-82

[20] Huo Z, Meng X, Hu H, Huang Y. You can walk alone: Trajectory privacy-preserving through significant stays protection//Proceedings of the 17th International Conference on Database Systems for Advanced Applications (DASFAA 2012). Busan, South Korea, 2012: 351-366

[21] Dwork C, Kenthapadi K, McSherry F, et al. Our data, ourselves: Privacy via distributed noise generation//Proceedings of the 25th Annual International Conference on the Theory and Applications of Cryptographic Techniques (EUROCRYPT 2006). Petersburg, Russia, 2006: 486-503

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

[23] Dwork C, Roth A. The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science, 2014, 9(3-4): 211-407

[24] Chen R, Fung B C M, Desai B C, Sossou N M. Differentially private transit data publication: A case study on the Montreal transportation system//Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD2012). Beijing, China, 2012: 213-221

[25] Chen R, Acs G, Castelluccia C. Differentially private sequential data publication via variable-length n-grams//Proceedings of the ACM Conference on Computer and Communications Security (CCS 2012). Raleigh, USA, 2012: 638-649

[26] Shao D, Jiang K, Kister T, et al. Publishing trajectory with differential privacy: a priori vs. a posteriori sampling mechanisms//Proceedings of the 24th International Conference on Database and Expert Systems Applications (DEXA 2013). Prague, Czech Republic, 2013: 357-365

[27] Jiang K, Shao D, Bressan S, et al. Publishing trajectories with differential privacy guarantees//Proceedings of the Conference on Scientific and Statistical Database Management (SSDBM2013). Baltimore, USA, 2013: 12: 1-12

[28] He X, Cormode G, Machanavajjhala A, et al. DPT: Differentially private trajectory synthesis using hierarchical reference systems//Proceedings of the 41st International Conference on Very Large Data Bases (VLDB 2015). Hawaii, USA, 2015: 1154-1165

[29] Hua J, Gao Y, Zhong S. Differentially private publication of general time-serial trajectory data//Proceedings of the IEEE Conference on Computer Communications (INFOCOM2015). Hong Kong, China, 2015: 549-557

[30] Xiao Y, Xiong L. Protecting locations with differential privacy under temporal correlations//Proceedings of the 22nd ACM Conference on Computer and Communications Security (CCS2015). Denver, USA, 2015: 1298-1309

[31] You T H, Peng W C, Lee W C. Protecting moving trajectories with dummies//Proceedings of the 8th International Conference on Mobile Data Management (MDM2007). Mannheim, Germany, 2007: 278-282

[32] Kellaris G, Papadopoulos S, Xiao X, Papadias D. Differentially private event sequences over infinite streams//Proceedings of the 40th International Conference on Very Large Data Bases (VLDB 2014). Hangzhou, China, 2014: 1155-1166

[33] Gotz M, Nath S, Gehrke J. Maskit: Privately releasing user context streams for personalized mobile applications//Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD 2012). Scottsdale, USA, 2012: 289-300

[34] Shokri R, Theodorakopoulos G, Boudec J Y L, Hubaux J P. Quantifying location privacy//Proceedings of the 32nd IEEE Symposium on Security and Privacy (S&P 2011). Berkeley, USA, 2011: 247-262

[35] Li Hang. Statistical Learning Methods. Beijing: Tsinghua University Press, 2012(in Chinese)
(李航. 统计学习方法. 北京：清华大学出版社，2012)

[36] Platt J C. Sequential minimal optimization: A fast algorithm for training support vector machines. Seattle, USA: Microsoft Research, Technical Report: MSR-TR-98-14, 1998

《计算机学报》

WU Yun-Cheng, born in 1989, Ph. D. candidate. His research interest includes private data analysis and data management in IoTs.

CHEN Hong, born in 1965, Ph. D., professor, Ph. D. supervisor. Her research interests include database, data warehouse, Internet of Things and privacy preservation.

ZHAO Su-Yun, born in 1979, Ph. D., associate professor. Her research interests include fuzzy sets, rough sets and uncertainty data processing.

LIANG Wen-Juan, born in 1980, Ph. D. candidate. Her research interests include data privacy preservation and database.

WU Yao, born in 1990, Ph. D. candidate. His research interests include crowd sensing and big data management.

LI Cui-Ping, born in 1971, Ph. D., professor, Ph. D. supervisor. Her research interests include social network analysis, data mining and recommender systems.

ZHANG Xiao-Ying, born in 1987, Ph. D., lecturer. Her research interests include wireless sensor network and privacy preservation.

Background

Protecting location privacy and trajectory privacy in location-based services (LBS) has become a hot spot of research. There are two types of LBS, namely, snapshot and continuous LBS. For a snapshot LBS, a mobile user only needs to report its current location to a service provider once to get its desired information. For continuous LBS, a mobile user has to report its location in a periodic manner to obtain the desired information. Privacy preserving techniques for LBS can be classified into four categories. (1) Location Obfuscation. The basic idea is to send a cloak region instead of the real location to the service provider. (2) Mix Zones. This technique is to change the pseudonyms when several users enter a mix-zone, ensuring unlinkability between the incoming users and outgoing users. (3) Suppression. The basic idea is not to report the current location if user is in a sensitive area. (4) Perturbation. This technique sends a false location that is close to the real location to the service provider. At present, the existing studies mainly focus on privacy preserving snapshot queries. Differential privacy, a popular paradigm for providing privacy with strong theoretical guarantees, has recently gained significant attention in snapshot LBS. Several differentially private approaches that generates a false location according to the real location have been proposed, however, these snapshot approaches cannot be directly applied to the trajectory privacy protection scenario (continuous LBS), since they seldom consider the geo-spatial and temporal correlation of the locations between several timestamps.

Aiming at solving the trajectory privacy problem, this paper firstly propose CPL algorithm to calculate the privacy level of each region in the map according to geo-spatial correlation, and define a privacy model that integrates privacy level and differential privacy. In addition, we analyze the effect of released location on the true locations based on Markov chain, and present a differentially private location release mechanism DPLRM, to protect the privacy of true locations and trajectory. Experimental results on real datasets demonstrate the performance of proposed mechanism. Although this paper provides a possible way to protect trajectory privacy using the notion of differential privacy, there are two improvements that could be explored to enhance the performance of this approach. Firstly, the concrete mobility pattern of user, e. g., the way of transportation on a specific timestamp can be integrated into the calculation of privacy level, which might improve the accuracy of privacy level. Secondly, how to reduce the computational cost for DPLRM is also an important direction.

This work is supported by the National Natural Science Foundation of China (Grant Nos. 61532021, 61702522, 61772537, 61772536), the National High Technology Research and Development Program of China (863 Program) (No. 2014AA015204).

《计算机学报》