文章编号:1007-5321(2017)02-0057-10 **DOI**:10.13190/j.jbupt.2017.02.009

基于重力模型生成假轨迹的隐私保护方法

张 翠

(1. 中国科学院信息工程研究所 信息安全国家重点实验室, 北京 100195;
2. 中国科学院大学 网络空间安全学院, 北京 100195)

摘要: 针对连续查询场景中用户实时位置的隐私保护问题,设计了一种基于客户端的假轨迹生成方法. 该方法使用网格划分地理空间,统计网格划分后每个网格内的历史查询数据. 通过分析网格内的历史查询数据构建实时预测用户移动轨迹的重力模型. 在重力模型基础上结合历史查询概率定义了轨迹熵度量轨迹隐私保护等级,并在最大运行速度限制下,提出了一种具有最大轨迹熵的基于 k-匿名的假轨迹隐私保护算法. 实验结果验证了所设计的假轨迹生成方法能够有效地保护真实轨迹的隐私.

关 键 词: 连续位置服务; 重力模型; 轨迹隐私; 推理攻击
中图分类号: TN929.5 **文献标志码**: A

Generating Dummies Based on Gravity Model for User's Trajectory Privacy

ZHANG Cui

(1. The State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100195, China;
2. School of Cyber Security, University of Chinese Academy of Sciences, Beijing 100195, China)

Abstract: The real-time location privacy preserving is a hotspot in continuous Location-Based Services (LBSs). A client-based dummy trajectory generation method is proposed. Based on the spatial grid partition, the history data in each cell of this grid is analyzed. Using the gravity model, a prediction model for users' movement pattern is built. Combined with the movement pattern model and the history query probability, the author defines a trajectory entropy to metric the trajectory privacy level. Based on k-anonymity principle, a limited velocity dummy trajectory generation algorithm with maximum trajectory entropy is proposed. Experiments from synthesis data and real-world data validate the effectiveness of our proposed method.

Key words: continuous location-based services; gravity model; trajectory privacy; inference attack

0 引言

伴随着移动互联网、智能终端设备的飞速发展以及定位系统的广泛使用,催生了基于位置服务(LBSs,location-based services)这一新型服务模式. 人们可以随时随地向位置服务提供商(LSP,location service provider)查询所需的服务,获取各种各样的 LBSs 服务. 同时,LBSs 衍生出新的服务模式,比如,基于位置的兴趣点推荐[1],交通预测[2]和众包服务[3]等,便于大数据技术的发展. 但是人们在享受 LBSs 的同时,自身的位置信息被不可信的 LSP 获悉. 而 LSP 可能将这些位置信息以盈利性目的提供给广告服务商,向用户收取额外的费用,也可能被敌手长期观察导致用户的兴趣爱好、健康状况、出行轨

收稿日期: 2016-10-25
基金项目: 国家高技术研究发展计划(863 计划)项目(2015AA016007); 国家自然科学基金青年基金项目(61502489)
作者简介: 张 翠(1985—), 女, 博士生, E-mail: zhangcui@iie.ac.cn; 李凤华(1966—), 男, 研究员, 博士生导师.

迹、社会关系等隐私信息泄露. 甚至在机器学习算法的帮助下,敌手容易推断出真实轨迹,判断出用户的生活规律,从而实施犯罪行为等[4]. 因而,位置信息的隐私问题亟待解决.

基于 k-匿名位置隐私保护方法[5-11]面临具有大数据分析能力获得人类真实移动模式的敌手的推理攻击[12-13]. 针对假位置方案未考虑相邻位置集合间的时空关联性导致位置隐私保护效果降低的问题,Liu 等[10]提出了一种连续请求中通过连续合理性检查和隐私增强技术构造假位置的隐私保护方法. 但是该文难以确保构造虚假位置的成功率且难以抵御具有人类移动模式的背景信息的推理攻击.

为了抵御具有这种背景信息的敌手攻击,解决连续移动的用户每个位置的隐私问题. 综合考虑用户静态所处位置的历史查询概率、连续查询位置之间的转移概率和连续查询的位置之间的时空可达性三个因素,提出了无需依赖任何匿名服务器的基于重力模型实时构造具有最大轨迹熵的假位置隐私保护方法,有效抵御具有大数据分析能力获得移动用户真实移动模式的敌手的推理攻击.

1 相关工作

本节主要综述现有位置隐私保护中基于虚假位置保护位置及轨迹隐私的研究工作,该类方法无需依赖可信第三方,直接在客户端上运行隐私保护功能模块,并且该方法不降低位置服务的查询精度,更加实用可靠. 但是,由于构造的虚假位置常常与真实位置之间存在差异,容易受到敌手的推理攻击.

Kido 等[14]首次提出通过客户端产生假位置保护用户位置隐私的方法,在普遍性、阻塞程度和均匀性三种匿名效果度量标准基础上,分别设计了 MN (moving in a neighborhood)假位置生成算法和 MLN (moving in a limited neighborhood)假位置生成算法,通过在位置请求中增加一些假位置达到迷惑位置服务提供商的目的. 但是该方法并未考虑敌手可能掌握的背景信息,容易被攻击者排除掉一些假位置,降低位置隐私保护的效果. 为了有效抵御基于背景信息的敌手的攻击,Niu 等[9]提出一种划分网格的假位置选择算法实现基于客户端的 k-匿名隐私保护(DLS,dummy location selection)算法. 该算法精心地选择与用户所处真实位置具有相同历史查询概率的网格位置. 然而该算法忽略了移动用户位置之间的时空相关性,只考虑用户单个静态位置的隐私,并不适合持续查询中轨迹的隐私保护. 而在实际应用中,用户常常会持续向 LSP 发出位置查询,用户更加注重连续查询中的轨迹隐私.

针对连续查询中的轨迹隐私,Lei 等[15]提出基于虚假轨迹的隐私保护方法,自定义了轨迹泄露概率,通过对用户产生的真实轨迹进行旋转获得与真实轨迹交叉的虚假轨迹,从而使得敌手无法通过轨迹的几何形状和方向区分出用户产生的真实轨迹. 但是,他们并未采用 k-匿名保护轨迹,而且也未考虑敌手拥有的背景信息,因而,容易遭到具有地图背景信息的敌手发起的推理攻击. 后来 Li 等[16]为了防止拥有用户所处位置的特征的敌手的推理攻击,采用基于 k-匿名的产生虚假轨迹的隐私保护方法,通过旋转和偏移产生 k 条虚假轨迹,有效降低轨迹泄露概率,并避免了敌手的推理攻击. 但是该方法并未考虑虚假轨迹上连续 2 个位置之间的转移概率情况. Niu 等[17]针对连续查询场景下用户路径隐私问题,综合考虑最小隐匿区域和背景信息,提出了一种高效的轨迹隐私保护机制 Dummy-T,确保轨迹中构造的每个假位置尽可能地接近真实位置. 虽然他们考虑了很多位置的真实因素,但是尚未考虑用户真实的移动模式. 因而对于掌握一些人类移动模式的敌手,其方法难以确保隐私保护效果.

由于大数据预测分析技术的发展,敌手容易获得人类的移动模式[18-19]. 而现有的位置隐私保护方法尚未考虑人类真实的移动模式,容易遭到推理攻击,Bindschaedler 等[20]提出了规范系统地构造语义真实的假轨迹隐私保护方法. 在该方法中首先提出了两种度量人类移动模式的标准,从人类移动特征的地理和语义维度上真实地形式化一条合成轨迹. 然后基于提出的度量标准,以用户真实轨迹为种子构建概率生成模型合成虚假轨迹,其中人类移动特征由马尔可夫模型构建,但是马尔可夫模型不适合长期处理数据的场景且计算复杂度较高. 而重力模型作为另一种有效模拟人类移动的模型,文献[21]采用重力模型对人类移动的一致性和规律性进行建模. 计算简单,预测效果较好. 因而,运用重力模型模拟人类的移动模式,在获得人类移动模式的基础上,根据自定义的度量标准设计假轨迹隐私保护方案.

虽然 Alicia 等[22]综合考虑了人类的移动模式,定义了信息熵度量隐私保护方法,但是它们并未考虑移动用户的最大速率的限制. 而 Alfalayleh 等[23]

提出了一种通用信息熵度量方法衡量隐私保护方法的性能. 但是这些度量隐私保护方法的信息熵尚未综合用户所处的地理特征和移动模式.

因而,综合考虑用户所处位置的地理特征、移动模式、最大运行速度限制等因素,分别采用历史查询概率和重力模型量化用户所处位置的地理特征和移动用户的移动特征,提出了轨迹熵度量方法,在此度量方法基础上,构造了具有最优轨迹熵的 $k-1$ 条虚假轨迹.

2　预备知识

下面定义所需的背景信息、k-匿名隐私保护等级、轨迹等概念,并详细说明攻击模型及研究动机.

2.1　相关定义

定义 1（背景信息）

背景信息是指敌手根据现有的技术条件获得的一些信息. 这些信息辅助敌手识别某个对象的真实身份,特指用户的历史查询记录和用户的移动模式.

定义 2（轨迹）

轨迹是指一段时间内由单点位置依时间先后组成的集合,其形式化定义为: $L=\{x_1,\cdots,x_m\}$,其中,$x_i=\langle a_i,b_i,t_i\rangle$,$i\in[1,m]$.

其中:x_i 表示第 i 个时刻用户所在位置,i 表示用户第 i 次发出的位置查询请求,m 表示用户产生的轨迹长度,a_i 表示定位系统定位出的经度,b_i 表示定位系统定位出的纬度.

定义 3（k-匿名隐私保护等级）

k-匿名位置(或者轨迹)隐私保护满足敌手从包含 k 个位置(或者轨迹)集合中区分出用户位置(或者轨迹)的概率不超过 $1/k$. 其数学定义为

$$\forall x_i\in A,p(x_i=U)\leqslant 1/k,i=1,2,\cdots,k$$

其中:A 为位置(或者轨迹)匿名集,x_i 为匿名集合某个位置(或者某条轨迹),$p(x_i=U)$ 为敌手判断出匿名集中的某个位置(某条轨迹)为用户真实位置(或者轨迹)U 的概率.

定义 4（位置熵）

熵概念源自信息论,用来表征一个系统内的不稳定性. 在位置服务的隐私保护中,对于敌手,设计的位置隐私保护模块可以看作一个系统,敌手从该系统(常包括 k 个对象的系统)中识别出某个对象是真实用户的不确定性,可以定义为熵. 熵的标准定义为

$$H=-\sum_{i=1}^{k}p_i\mathrm{lb}(p_i)\tag{1}$$

其中:k 表示隐私保护系统满足 k-匿名条件,p_i 表示敌手从 k 个对象中识别出真实用户的概率(假设用户的位置与用户的身份对应).

2.2　攻击模型

针对位置服务中位置信息的攻击模型分为主动攻击者和被动攻击者. 主动攻击者主要采取的攻击形式是主动截取信息、合谋攻击(合谋中的攻击者可能是事务的参与方,或者是通过买通事务的若干参与方共同欺骗其余的参与方以获得非法权限或利益)、推理攻击(通过分析数据非法获得对象或数据库的知识,使得敌手以高概率推理出对象的一些敏感信息). 而被动攻击主要是收集信息而非进行信息的访问,数据的合法用户对该攻击无所觉察. 被动攻击包括嗅探、信息收集等攻击方法.

假定敌手是主动攻击者,它们通过合谋手段获得 LBSs 服务器上保存的有关移动用户的所有信息,包括任意用户的当前和历史请求等;敌手具有大数据分析能力,通过大数据处理技术,判断出用户的最大运行速度、移动模式和移动概率转移模型等,有利于推理出用户真实身份、真实位置甚至真实的运行轨迹等信息. 另外,敌手拥有地图信息、黄页电话号簿等,可以判断出地理区域的特征,甚至区域内人口密度等信息.

2.3　研究动机

尽管目前已有一些基于假轨迹的隐私保护方法,但是忽略了轨迹的移动模式等背景信息,容易遭到基于移动模式的背景信息的推理攻击. 敌手不仅仅能够得到某个位置区域用户发出位置服务请求的情况,而且还可以分析出用户从一个位置到达下一个位置的可能性,即敌手具有利用大数据挖掘技术分析出用户可能的移动路径的能力.

在用户的历史查询和移动模型预测的基础上,针对这种敌手设计一种新颖的轨迹隐私保护方法. 用户可以在设备上产生 k 个不确定性最高的假位置. 假如用户持续运动产生了一条 $m=3$ 的轨迹,按照现有的轨迹隐私保护算法,用户只能保护处于当前位置之前经历过的所有位置的隐私,也就是说,产生 $k-1$ 条从几何形式上与真实轨迹相似的轨迹,达到混淆真实轨迹的目的. 但是如果用户需要对每个时刻都进行隐私保护,该算法的计算量非常大,并且没有从出现的概率性角度获得与用户真实的运动模

式相似的假轨迹. 如图1所示的场景,用户从下班地点出发的真实轨迹是先去附近的医院,然后到达购物中心.

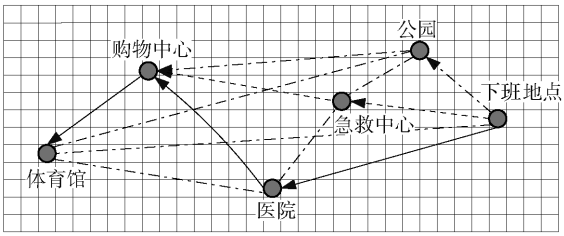

图1 抽象用户运动轨迹示意图

如图1所示,用户从下班地点出发,到达附近的医院,再到购物中心,构成真实轨迹,用实线表示. 其中医院是敏感位置,用户不希望被LSP获得此位置. 此时,用户通过伪造2条假轨迹保护真实轨迹中的敏感位置信息,其中一条假轨迹是由下班地点到公园,再到购物中心构成,用点式虚线表示;另一条假轨迹是由下班地点到急救中心,再到购物中心构成,用虚线表示. 由于用户到达这些位置点的可能性存在差异,用户从下班地点到达急救中心的可能性较低,而从下班地点到达公园的可能性较高. 从这种差异上来看,敌手可以排除假轨迹中可能性较低的轨迹,从而降低了轨迹隐私保护效果. 因而,在构造假轨迹过程中充分考虑用户在位置之间的转移可能性(转移概率),可以有效抵御基于移动模式的背景信息的推理攻击.

2.4 系统结构

如图2所示,系统架构包括位置服务提供商、移动互联网、具有隐私保护功能(Dummy生成系统)的移动设备、定位系统以及移动模型处理部分. 在该架构中各个部分LSP常见的有百度、腾讯和谷歌等. 移动互联网由蜂窝网络、WiFi和蓝牙等无线通信设备表示. 智能设备包括智能手机、平板电脑、穿戴设备等. 移动模型处理部分由移动用户在发起位置服务请求之前,向大数据处理服务器获得所需的移动通用模型.

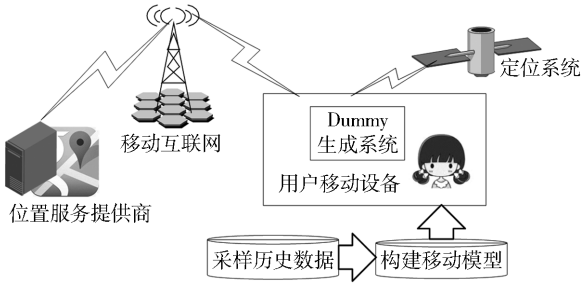

图2 系统架构示意图

位置服务请求中用户端处理流程大致可以描述为:① 用户预先获得所处区域的历史查询概率分布及移动模型;② 用户在本地产生假位置,将这些假位置与真实位置混淆一起;③ 用户每隔一段时间向LSP发起位置服务查询请求;④ 用户等待LSP的响应结果.

3 基于重力模型假轨迹生成算法

假轨迹方法作为一种无需增加第三方服务器达到隐私保护目的的方法,可以确保用户获得精准的查询结果. 下面详细描述该方法的设计过程:如何构造出k个假轨迹集合满足轨迹熵值最大,有效抵御拥有人类移动模式和区域历史查询记录的攻击. 假如区域按照网格划分为$N \times N$个网格,用户在这些网格内移动. 每个网格具备自身的特点,如有些地方是城市闹区,有些地方是偏僻的郊区,有些是水域,并且用户移动满足人类共有的移动模式. 在这些考虑因素的范畴内,设计假轨迹生成算法.

3.1 历史查询概率

历史查询概率定义为

$$q_i = \frac{Q_i}{\sum_{i=1}^{N^2} Q_i} \tag{2}$$

其中:Q_i指第i个网格内出现位置查询请求的次数,N^2是指整个研究区域包含的网格数目. 第i个网格的历史查询概率q_i就是第i个网格内出现位置查询请求的次数占整个研究区域内出现位置查询请求次数的比值,历史查询概率q_i表示第i个网格区域内人们的到访情况,可以通过多种方法获得历史查询概率. 如通过移动接入点统计在网格内出现的位置查询请求的次数,或者根据移动设备运营商基站转发的位置服务统计位置查询请求的次数. 对于这里选取的k个位置需要对这些位置的历史查询概率进行归一化处理,以便于计算.

3.2 移动模型建模

下面对研究区域中的用户位置构建移动模型,求解用户位置的转移概率. 首先采用重力模型对位置之间移动用户的流通量进行建模,然后根据历史查询数据拟合重力模型求解调整系数,最后根据建立的重力模型计算位置的转移概率.

重力模型的思想源自牛顿提出的万有引力定律. 2个位置之间人员流动情况可以看作2个位置

之间相互吸引程度的强弱，与 2 个位置之间转移的可能性正好吻合. 该重力模型广泛应用在交通、人口迁移、商业贸易等领域[24-26]. 而在基于位置服务的隐私保护中，将该模型用来辅助设计隐私保护算法研究尚不多. 现在重点介绍如何用该模型对移动用户移动模式进行预测. 首先，定义重力模型框架：

$$T_{i,j} = c\frac{(O_i)^a (D_j)^b}{\exp(rs_{i,j})} \tag{3}$$

其中：$T_{i,j}$表示从第 i 到第 j 个位置之间的出行量，O_i 为离开第 i 个位置的交通发生量，D_j 为到达第 j 个位置的交通吸引量，$s_{i,j}$为第 i 到第 j 个位置之间的距离，a、b、c、r 为该模型的调整系数，i 和 j 分别在网格空间内取值，$i, j \in \{1,2,\cdots,N\}$.

根据历史查询数据拟合重力模型求解调整系数. 采用多元线性回归方法对该模型进行拟合，求解该模型中的调整系数. 对式(3)左右两端同时取以 e 为底的对数，可以得到重力模型的对数表达式，即

$$\ln T_{i,j} = \ln c + a\ln O_i + b\ln D_j - rs_{i,j}, i,j \in \{1,2,\cdots,N\} \tag{4}$$

其中：$\ln O_i$、$\ln D_j$ 和 $s_{i,j}$是自变量，$\ln T_{i,j}$是因变量，$\ln c$、a、b 和 r 是系数. 将自变量 $\ln O_i$、$\ln D_j$ 和 $s_{i,j}$($i,j \in \{1,2,\cdots,N\}$)表示成矩阵 $\boldsymbol{X}$，即

$$\boldsymbol{X} = \begin{pmatrix} 1 & X_{11} & X_{12} & X_{13} \\ 1 & X_{21} & X_{22} & X_{23} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{I1} & X_{I2} & X_{I3} \end{pmatrix} \tag{5}$$

其中：$X_{(i-1)N+j,2} = \ln O_i$，$X_{(i-1)N+j,2} = \ln D_j$，$X_{(i-1)N+j,4} = s_{i,j}$，$I = N^2$. 将因变量 $\ln T_{i,j}$($i,j \in \{1,2,\cdots,N\}$)表示成矩阵 $\boldsymbol{Y}$，即

$$\boldsymbol{Y} = (Y_1, Y_2, \cdots, Y_{N^2})^{\mathrm{T}}$$

其中 $\boldsymbol{Y}_{(i-1)N+j} = \ln T_{i,j}$. 将系数 $\ln c$、a、b 和 r 表示成矩阵 $\boldsymbol{\beta} = (\ln c, a, b, r)^{\mathrm{T}}$，用 ε 表示 $\boldsymbol{Y}$ 和 $\boldsymbol{X}$ 两个向量之间的偏差. 将式(4)表示成线性方程组的形式，得到

$$\boldsymbol{Y} = \boldsymbol{\chi\beta} + \varepsilon \tag{6}$$

然后采用最小二乘法估计系数向量 $\hat{\boldsymbol{\beta}}$. 按照偏差平方和最小原则估算 $\boldsymbol{Y}_{(i-1)N+j}$的误差 E 有

$$E = \sum_{r=1}^{N^2} (Y_r - \hat{Y}_r)^2 \tag{7}$$

其中：Y_r 表示真实的数据；$\hat{Y}_r$ 表示样本数据，该样本数据通过分析历史轨迹数据中人员流量获得. X 由历史轨迹数据中人员离开量和到达量及统计的位置之间的距离获得，因而系数估计值 $\hat{\boldsymbol{\beta}}$ 为

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y} \tag{8}$$

通过式(8)计算出的系数矩阵 $\hat{\boldsymbol{\beta}}$ 可以得到系数 $\ln c$、a、b 和 r，进一步可以得到重力模型中的调整系数 a、b、c、r，将调整系数代入式(3)可以得到重力模型的表达式.

将所需要的 O_i、D_j 和 $s_{i,j}$代入重力模型的式(3)可以计算出第 i 个位置和第 j 个位置之间的出行量 $T_{i,j}$. 用第 i 个位置和第 j 个位置之间的转移概率 $P_{i,j}$ 表示第 i 个位置和第 j 个位置的转移情况，将求得的第 i 和第 j 个位置之间的交通流通量 $T_{i,j}$代入式(9)计算第 i 个位置和第 j 个位置之间的转移概率为

$$P_{i,j} = \frac{T_{i,j}}{\sum_k T_{i,k}} \tag{9}$$

3.3　轨迹熵

轨迹熵作为一种度量隐私保护等级的方法，Alicia 等[18]详细地说明了位置熵扩展到连续位置熵的过程，但是他们只考虑连续位置之间的到访概率，并未考虑两个位置之间的转移概率. 因而针对连续查询中基于 k-匿名的位置隐私保护，定义轨迹熵. 而轨迹熵，表征一段轨迹泄露给敌手的可能性. 因为轨迹上每两个位置之间存在一个转移可能情况，将该性质反映到熵中，定义为一个新的概念：连续查询实时轨迹熵

$$H_j = -\sum_{j=1}^{k} p_j \mathrm{lb}(p_j) \tag{10}$$

$$p_j = q(x_j^{i-1}) R(x_j^i | x_j^{i-1}) q(x_j^i) \tag{11}$$

其中：x_j^{i-1} 表示第 j 条轨迹上第 $i-1$ 个时刻用户所处的网格，$R(x_j^i | x_j^{i-1})$ 表示第 j 条轨迹中用户从第 $i-1$ 个时刻所处的网格 x_j^{i-1} 跳转到第 i 时刻所处网格 x_j^i 的概率，$q(x_j^i)$ 表示第 j 条轨迹中第 i 个时刻用户在网格 x_j^i 发起查询请求的概率，$q(x_j^{i-1})$ 表示第 j 条轨迹中第 $i-1$ 个时刻用户在网格 x_j^{i-1} 发起查询请求的概率，p_j 表示第 j 条轨迹上第 $i-1$ 个时刻用户在网格 x_j^{i-1} 发起查询请求并且在第 i 个时刻的用户在网格 x_j^i 发起查询请求的概率. $P_j^i = q(x_j^{-1}) R(x_j^i | x_j^{i-1})$ 称为轨迹概率. 当所有的 p_j 取值相等时，轨迹熵 H 取得最大值，即 $\max(H_j) = \mathrm{lb}(k)$.

用户的历史查询概率和移动倾向性从信息熵中体现出来，一般是整个 k 个数据集的概率值越接近，熵值往往更大. 这从隐私保护的角度看，k 个数据集的可区分性越低，对于隐私保护的效果越好. 因而，按照该隐私量化标准进行下面的 k 个数据集的设计.

3.4 生成假轨迹

从轨迹的角度,将轨迹上的每个位置进行有效的隐私保护,达到连续观察位置查询请求的用户轨迹的不可区分.

针对实时保护轨迹隐私场景,实时构造 k 个假位置保护每条轨迹上每个位置的隐秘性. 综合考虑用户的最大移动范围,在此限定条件下,采用贪心算法达到产生的 k 条轨迹的轨迹熵的最优.

假设用户在一个 $N \times N$ 网格区域内移动,移动速度为 v,最大的运动速度为 $v_{\max}$. 用户每隔 Δt 发出一次位置查询请求,该用户对自身的隐私要求限制为 k. 为了避免遍历所有空间网格内,降低算法的复杂度,在此要求穷举法获得的数据集中每个元素满足:

$$\zeta = \text{argmax } H_j,\ \text{s. t. },\ s \leqslant v_{\max}\Delta t \tag{12}$$

其中:s 指选择的第 i 个时刻的假位置到下一个 $i+1$ 时刻位置之间的距离,$\Delta t = t_{i+1} - t_i$. 详细过程如算法 1 所示.

算法 1 限制速度地实时选择假轨迹算法

输入:区域网格空间 $N \times N$,真实轨迹 $L_u\{x_1, x_2, \cdots, x_m\}$,最大运行速度 $v_{\max}$

输出:k 条轨迹的匿名集合、轨迹熵

步骤:

1) 根据网格空间,获取到对应的历史查询概率 $q(x_j^i)$ 和转移概率 $R(x_j^i \mid x_j^{i-1})$

2) 取出真实轨迹 L_u 中第一个位置 x_1,调用 DLS 算法[9]

3) 以 k 个位置为中心,$v_{\max}\Delta t$ 为半径划定圆形区域,计算圆形区域内对应的网格

4) 计算用户处于网格处的实时轨迹熵值,将这些熵值进行排序,选出实时轨迹熵最大的 k 条轨迹

5) 重复第 1) 步到第 4) 步直到选出具有最大实时轨迹熵的 $k-1$ 条轨迹第 m 时刻的网格

6) 输出由上述步骤选出的第 1 个时刻到第 m 个时刻 $k-1$ 条假轨迹

3.5 安全性分析

假设敌手的目的是为了得到用户的准确位置,并通过准确位置获得与用户位置相关的私密信息. 敌手具有的背景信息包括:敌手知晓用户提交到 LSP 的位置查询请求的时间和位置信息;长期观察收集到用户在区域内发送位置查询请求的概率;敌手拥有用户地图信息,知晓两个位置之间的距离和可达时间;具有大数据分析出用户移动模式的能力. 但是,敌手并不能匹配出用户与所知晓的信息.

定理 1 本文方法抵御合谋攻击.

如果多个用户组成一个集合 S,而且彼此之间互相关联,则存在合谋攻击.

$$\exists u_1, \cdots, u_i \in S, \text{cov}(u_i, u_j) \neq 0,$$
$$i \neq j \text{ 且 } i, j = 1, 2, \cdots$$

而本方案只在客户端构造并运行,不考虑与其他客户端进行合作的情况. 如果考虑 LSP 与其他敌手合谋,则敌手也将进行推理攻击.

定理 2 本文方法抵御推理攻击.

针对 k-匿名的推理攻击,假设用户真实的轨迹为 $\langle v_1, v_2, \cdots, v_i \rangle$,LSP 接收到的集合 $\bigcup_{j=1}^{k} \langle v_j^1, v_j^2, \cdots, v_j^i \rangle$,每条轨迹具有 m 个位置点,从其中 i 位置的历史查询概率 q_i,可以判断 i 位置是不是真实存在的位置,另外,通过判断 $i-1$ 位置转移到 i 位置的概率 $R(v_j^i \mid v_j^{i-1})$,如果 $R(v_j^i \mid v_j^{i-1}) \simeq 0$,则可以推断该段轨迹是伪造的轨迹.

然而,本方案正是从这两个量的角度出发,选择出于真实轨迹尽可能接近的轨迹. 根据上文定义的轨迹概率 $P_j^i = q_i R(x_j^i \mid x_j^{i-1})$,计算由 k 个位置构成的集合的熵值 H,采用贪心算法获得 H 最大的情况的最优解,所以很显然敌手难以分辨出其中哪段轨迹是虚构的轨迹.

4 实验及仿真

为了验证基于重力模型的假轨迹生成算法的有效性,采用微软亚洲研究院 Geolife 收集的真实数据集进行实验[27],分析出重力模型. DUMMY 生成系统调用百度地图的 API 接口,进行真实环境的测试.

4.1 基本设置

开发环境采用 Matlab 编程,硬件环境为 Intel Core i5-4258U@2.4 GHz 处理器,4 G 内存;软件环境采用 Matlab R2010b 软件.

实验数据部分采用微软亚洲研究院 Geolife 项目收集的 2007-04-01—2012-08-01 期间 182 个用户的 GPS 轨迹数据. 整个数据集由 17 621 条轨迹,总行驶距离为 1 292 951 km,总耗时 50 176 h. 这些数据是由不同的 GPS 日志记录器和 GPS 手机采集. 从中随机选取 1 000 条轨迹,大致反应出北京市各个区域的人口流动情况.

实验分为两部分,第 1 部分是根据已有数据集拟合出重力模型,分别计算出两地之间的交通流量、

进入某地的交通量、离开某地的交通量和两地之间距离；第 2 部分是利用合成的数据和真实数据，实现轨迹生成算法.

4.1.1　拟合重力模型

从 Geolife 数据集中选用 E116° ~ E116°14′，北纬 N39°85′ ~ N39°94′，100 km × 100 km 区域内的轨迹数据，并划分该区域为 100 × 100 个网格，模拟出任意 2 个网格之间的吸引度情况，得到下面 4 个矩阵：每 2 个网格之间的轨迹数量 T；指定轨迹方向后从第 i 个网格出发的轨迹数量 O（交通发生量）；到达第 j 个网格的轨迹数量 D（交通吸引量）；第 i 个网格到达第 j 个网格的距离 s.

假设用户连续发起 3 次查询服务，产生一条由网格位置(1,2)→(2,2)→(3,2)组成的轨迹. 为了便于验证算法的性能优劣，选用一些实例数据进行实验. 这些数据是从大数据中抽样出来的部分数据，用来分析重力模型. 选用的实例数据如表 1 和表 2 所示. 参数 D_j 表示到达第 j 个网格的轨迹数量，参数 O_i 表示从第 i 个位置离开的轨迹数量. 如表 1 中坐标为(1,2)的网格为第 1 个位置，坐标为(2,2)的网格为第 2 个位置，这 2 个网格之间的距离 s 值为 17，到达第 2 个位置的车辆数量 D_j 值为 7，从第 1 个位置离开的车辆数量 O_i 值为 7. 根据表 1、表 2 中的已知数据，经过多元线性回归方法，标定出重力模型中的系数，求得重力模型中系数 $\ln c$、a、b 和 r. $\ln c = 2.181\,3$，$a = 1.303$，$b = 1.008\,9$，$r = 2.1$. 最后得到重力模型为

$$T_{i,j} = 0.78\,\frac{(O_i)^{1.303}(D_j)^{1.008\,9}}{\exp(2.1 s_{i,j})}$$

表 1　两个位置之间的距离 $s_{i,j}$

j	i		
	1	2	3
1	8	17	22
2	17	15	23
3	22	23	7

表 2　到达 D_j 及离开 O_i 的车辆数量

D_j	O_i		
	1	2	3
1	17	7	4
2	7	38	6
3	4	5	17

在此模型基础上，假定给定 O_i 和 D_j，根据上述模型（该模型通过验证理论结果与真实结果吻合），可以获得如表 3 所示的 $T_{i,j}$.

表 3　两地之间的出行量

j	i		
	1	2	3
1	6.922 779 38	0.580 826 24	0.192 117 648
2	0.447 419 11	0.237 730 62	0.068 995 34
3	0.125 571 49	0.046 727 2	1.390 826 01
合计	7.495 769 98	0.865 284 04	1.651 939

为了刻画出两个网格之间的来往流量的转移关系，将 $T_{i,j}$ 转换为归一化的转移概率 $R_{i,j}$. 该转移概率计算结果如表 4 所示.

表 4　归一化转移概率

j	i		
	1	2	3
1	0.691 379 64	0.058 007 26	0.019 186 84
2	0.044 683 85	0.023 742 21	0.006 890 581
3	0.012 540 85	0.004 666 657	0.138 902 13

4.1.2　历史查询概率

假设根据上述提出的方法可以获得网格位置对应的历史查询概率 $q(x_j^i)$. 本实验中根据合成的数据计算出 $q(x_j^i)$，假设用户发出请求的概率满足高斯分布，统计网格内分布到的个数并进行归一化处理，获得各个网格单元内的概率值. 实验选用 1 万个高斯分布的点进行实验，所得结果如表 5 所示.

表 5　历史查询概率

j	i		
	1	2	3
1	0.002 9	0.005 7	0.009 6
2	0.009 7	0.014 9	0.021 8
3	0.009 6	0.016 4	0.022 5

在已知历史查询概率和转移概率基础上，根据轨迹熵的定义计算轨迹熵，比较选择的 k 条轨迹是否是最优的. 假设真实位置处于网格(1,2)→(2,2)→(3,2)，用这些位置随机组合成一条长度为 3 的轨迹，计算所有的熵值，对这些熵值排序，取出排在前 2 的两条轨迹.

在获得这些信息后，应用该算法产生假轨迹，在

该区域内选定某个位置作为用户真实的发送请求的位置,采用 DLS 算法[9]产生 $k-1$ 个假位置,然后按照提出的算法,产生 $k-1$ 条假轨迹.

由于该算法增加了历史查询概率及移动转移概率等限定条件,并且基于假位置的方法不会降低用户的查询精度,因而性能方面可以考虑算法的隐私保护等级和运行时间.

4.2 实验结果

实验所需要的网格索引,历史查询概率值和转移概率值与网格的映射关系已知. 分别在合成的轨迹数据集和真实轨迹数据集情形下,对本文方案具有的隐私保护效果进行实验,为了简化实验,本次实验选用一条轨迹长度为 3 的轨迹作为种子轨迹. 在合成数据集的情形下,选用了改进版的 DLS 算法、随机选择算法、最优选择算法作为参考;而在真实轨迹集情形下,由于数据集有限,为了实验效果,选用了随机选择算法和最优算法,用来与本文算法进行轨迹熵的比较.

4.2.1 合成数据集下轨迹熵与隐私等级关系

利用合成的轨迹,分别对比随机模式、连续运行 DLS 模式、最优轨迹模式、重力模型预测模式(本文方案)下,按照轨迹熵计算公式计算出的熵值如图 3 所示,其中横坐标是匿名等级 k,纵坐标是轨迹熵,带星号曲线为最优的轨迹熵,标记菱形的曲线为随机走动时轨迹熵,标记圆形的曲线是 DLS 在连续情况下熵值变换情况,标记方形的曲线为本文提出的方案.

从图 3 中可以看出,轨迹熵随着 k 值增加而不断增大,采用选择与真实位置具有相同的背景信息的 k 个位置,可以获得比较大的熵值,也就是说具有更好的隐私保护能力,具有更强的不可区分性. 而在实验所选的区域内选择连续的假位置,构成 k-匿名的轨迹集,根据这些假轨迹计算轨迹熵,由于选择的随机性,造成选择出的假轨迹对应的背景信息值差异较大. 因而轨迹熵较小,明显低于所提的方案. 另外,图中 DLS 方案是改进的 DLS 算法(在选择单点位置时采用 DLS 算法,而选用连续两个位置时为两个位置之间的转移概率随机分配值). 从图中发现本文的方案较改进的 DLS 算法具有更高的轨迹熵,表明轨迹隐私保护效果更好. 与最优方案对照,由于本文的方案限制在有效的运行区域内,造成选择的轨迹并不是整个区域内轨迹熵最大的区域,而只是局部区域轨迹熵最优,所以轨迹熵低于最优方案.

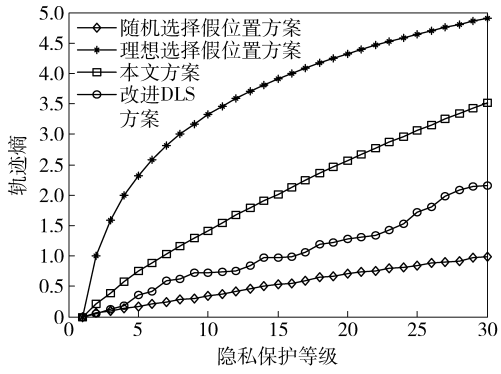

图 3 合成数据环境下轨迹熵比较

4.2.2 真实轨迹数据隐私保护等级比较

主要实现了通过真实轨迹集拟合重力模型,产生转移概率值,将本文方案与随机选择算法、最优算法进行比较. 从图 4 中可以看出,本文算法比随机选择的位置(baseline)具有更高的隐私性,但是由于真实数据的模拟比较有限和时间复杂度的问题,选择隐私保护等级 k 在 3 ~ 7 之间,轨迹熵值的变化趋势如图 4 所示. 分析数据变化可知,随机模式下选择假轨迹,选择的位置具有的转移模型和历史查询概率不确定,而提出的算法以轨迹熵最大为前提选择假位置,因而,在轨迹熵方面明显比随机选择的方法更优. 由于真实轨迹集在有限的 100×100 网格内,能够统计出的所需数据有限,在未来工作中,应该扩大网格空间,细粒度化单元网格,这样可以获得更精准的背景信息.

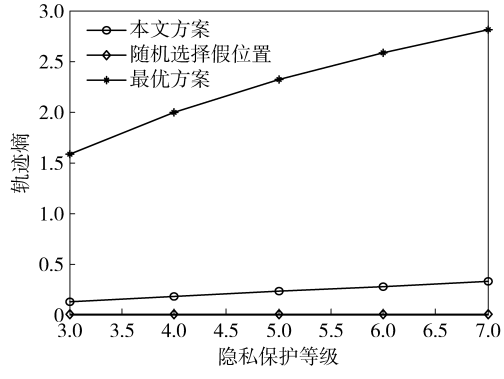

图 4 真实环境下轨迹熵比较

4.2.3 算法复杂度及其运行时间

除了验证算法的隐私保护效果之外,还需要分析算法复杂度. 该算法的时间复杂度主要在于遍历网格空间对应的背景信息值,经过对这些背景信息值进行排序选择轨迹熵值排在前 k 位的位置,用于

构成最终的 k-匿名的轨迹集. 由于笔者提出的算法增加了最大运行范围的限制,大大降低了所需的遍历空间和算法的运行时间. 在本实验给定的环境下,笔者提出的算法复杂度为 $O((\delta n)^k)$,其中 δ 取决于运行速度及查询周期,$\delta<1$. n 表示网格数目,k 表示匿名集合大小. 用作比较的算法的复杂度分别为:随机选择算法($O(n^k)$)、最优选择算法($O(n^{2k})$)和 DLS 算法($O(2n^k)$).

这些算法的运行时间差异如图 5 所示. 由于随机选择算法不需要进行遍历、排序等处理,直接在给定的位置点中选择 k 个位置,然后重复运行 3 次即可完成 k-匿名轨迹集的构造. 而最优方案需要遍历给定的所有位置点,查询位置点对应的背景信息值,所以时间开销方面大大增加,而本文方案在给定的位置点中根据可达到区域限制条件进行位置选择,减少了查询数量. 因而较最优方案,在时间运行方面可以节约一些时间. 而对应改进的 DLS 方案(DLS 方案运行 3 次),因为算法需要二次排序,所以算法运行时间大大增加.

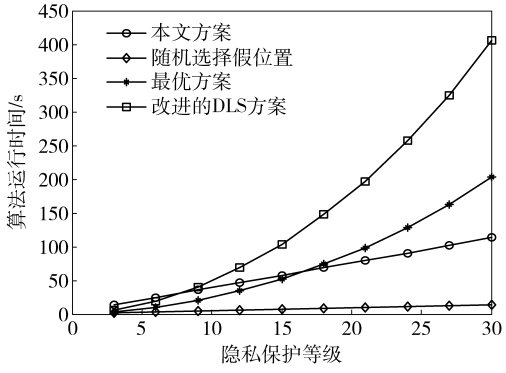

图 5　算法运行时间比较

5　结束语

针对基于位置服务中移动用户持续发出位置请求产生的轨迹隐私保护问题,提出了一种考虑真实轨迹运动模式的生成虚假轨迹的隐私保护方法. 本方法直接可以运行在移动设备上,实时地保护移动用户每个位置处的隐私,通过综合考虑用户所处区域的历史查询概率和用户在空间上移动转移情况,构建 $k-1$ 条每个位置处都满足熵值最大的虚假轨迹,使得拥有历史查询概率和移动模式预测能力的敌手无法区分真假轨迹. 从安全性和实验仿真进一步验证了该方法的有效性. 另外,如何将提出的模型应用在发布轨迹的隐私保护中和考虑逆高斯模型的应用将是下一步工作.

参考文献:

[1] Yu Yonghong, Chen Xingguo. A survey of point-of-interest recommendation in location-based social networks[C]// Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence. Austin: AAAI, 2015: 53-60.

[2] Wang Hongtao, Li Qiang, Yi Feng, et al. Influential spatial facility prediction over large scale cyber-physical vehicles in smart city[J]. EURASIP Journal on Wireless Communications & Networking, 2016, 2016(1): 1-12.

[3] He Zongjian, Cao Jiannong, Liu Xuefeng. High quality participant recruitment in vehicle-based crowdsourcing using predictable mobility[C] // INFOCOM 2015. Hong Kong: IEEE, 2015: 2542-2550.

[4] Bellovin Steven M, Hutchins Renee M, Jebara Tony, et al. When enough is enough: location tracking, mosaic theory, and machine learning[J]. NYUJL & Liberty, 2013(8): 556.

[5] Gedik Bugra, Liu Ling. Protecting location privacy with personalized k-anonymity: architecture and algorithms [J]. IEEE Transactions on Mobile Computing, 2008, 7(1): 1 - 18.

[6] Chow Chi-Yin, Mokbel Mohamed F, Aref Walid G. Casper*: query processing for location services without compromising privacy[J]. ACM Transactions on Database Systems, 2009, 34(4): 1 - 48.

[7] Beresford Alastair R, Stajano Frank. Location privacy in pervasive computing[J]. IEEE Pervasive Computing, 2003, 2(1): 46-55.

[8] Palanisamy Balaji, Liu Ling. Attack-resilient mix-zones over road networks: architecture and algorithms[J]. IEEE Transactions on Mobile Computing, 2015, 14(3): 495-508.

[9] Niu Ben, Li Qinghua, Zhu Xiaoyan, et al. Achieving k-anonymity in privacy-aware location-based services[C] // INFOCOM 2014. Toronto: IEEE, 2014: 754-762.

[10] 刘海, 李兴华, 王二蒙, 等. 连续服务请求下基于假位置的用户隐私增强方法[J]. 通信学报, 2016, 37(7): 140-150.
Liu Hai, Li Xinghua, Wang Ermeng, et al. Privacy enhancing method for dummy-based privacy protection with continuous location-based service queries[J]. Journal on Communications, 2016, 37(7): 140-150.

[11] Wang Yong, Peng Jing, He Long-ping, et al. LBSs privacy preserving for continuous query based on semi-honest third parties[C] // International Performance Com-

puting and Communications Conference. Austin: IEEE, 2012: 384-391.

[12] Götz Michaela, Nath Suman, Gehrke Johannes. MaskIt: privately releasing user context streams for personalized mobile applications[C] // ACM SIGMOD International Conference on Management of Data. Scottsdale: ACM, 2012: 289-300.

[13] Feng Zhenni, Zhu Yanmin. A survey on trajectory data mining: techniques and applications[J]. IEEE Access, 2016, 4(10): 2056-2067.

[14] Kido Hidetoshi, Yanagisawa Yutaka, Satot Tetsuji. An anonymous communication technique using dummies for location-based services[C] // International Conference on Pervasive Services. Santorini: IEEE, 2005: 88-97.

[15] Lei Po-Ruey, Peng Wen-Chih, Su Ing-Jiunn, et al. Dummy-based schemes for protecting movement trajectories[J]. Journal of Information Science & Engineering, 2012, 28(2): 335-350.

[16] 李凤华, 张翠, 牛犇, 等. 高效的轨迹隐私保护方案[J]. 通信学报, 2015, 36(12): 114-123.
Li Fenghua, Zhang Cui, Niu Ben, et al. Efficient scheme for user's trajectory privacy[J]. Journal on Communications, 2015, 36(12): 114-123.

[17] Niu Ben, Gao Sheng, Li Fenghua, et al. Protection of location privacy in continuous LBSs against adversaries with background information[C] // International Conference on Computing, Networking and Communications. Kauai: IEEE, 2016: 1-6.

[18] Xu Fengli, Tu Zhen, Li Yong, et al. Trajectory recovery from ash: user privacy is not preserved in aggregated mobility data[Z]. 2017, arXiv: 1702.06270[cs.CY].

[19] Noulas Anastasios, Scellato Salvatore, Lathia Neal, et al. Mining user mobility features for next place prediction in location-based services[C] // IEEE International Conference on Data Mining. Brussels: IEEE, 2012: 1038-1043.

[20] Bindschaedler Vincent, Shokri Reza. Synthesizing plausible privacy-preserving location traces[C] // IEEE Symposium on Security and Privacy. San Jose: IEEE, 2016: 546-563.

[21] Wang Yingzi, Yuan Nicholas Jing, Lian Defu, et al. Regularity and conformity: location prediction using heterogeneous mobility data[C] // ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Paris: ACM, 2015: 1275-1284.

[22] Rodriguez-Carrion Alicia, Rebollo-Monedero David, Forné Jordi, et al. Entropy-based privacy against profiling of user mobility[J]. Entropy, 2015, 17(6): 3913-3946.

[23] Alfalayleh Mousa, Brankovic Ljiljana. Quantifying privacy: a novel entropy-based measure of disclosure risk[C] // Combinatorial Algorithms. [S. l.]: Cham, Springer, 2014: 24-36.

[24] Haynes Kingsley E, Fotheringham A Stewart. Gravity and spatial interaction models[C] // Scientific Geography Series. Beverly Hills: Sage Publications, 1984: 9-13.

[25] Chew Yong Huat, Nanba Shinobu, Peng Keong, et al. On the verification of the gravity model used for mobility modeling[C] // IEEE International Conference on Communications. Glasgow: IEEE, 2007: 5642-5647.

[26] Zhang JiaDong, Chow Chi-Yin. Spatiotemporal sequential influence modeling for location recommendations: a gravity-based approach[J]. Transactions on Intelligent Systems & Technology, 2015, 7(1): 1-25

[27] Zheng Yu, Li Quannan, Chen Yukun, et al. Understanding mobility based on GPS data[C] // International Conference on Ubiquitous Computing. Seoul: IEEE, 2008: 312-321.

[28] 谢香君. 重力模型标定方法及比较分析[J]. 交通标准化, 2008(8): 17-20.
Xie Xiangjun. Calibration method and comparison of gravity model[J]. Communication Standardization, 2008(8): 17-20.