基于 PSO 优化的移动位置隐私保护算法

李 婕[1)] 白志宏[2)] 于瑞云[2)] 崔亚盟[3)] 王兴伟[2)]

[1)](东北大学计算机科学与工程学院 沈阳 110169)

[2)](东北大学软件学院 沈阳 110169)

[3)](去哪儿网无线事业群 北京 100080)

摘 要 在这个移动互联网技术和大数据技术快速发展的时代,基于位置的服务使移动用户的位置信息数据化,给人们的生活带来极大便利的同时也威胁到了移动用户的位置隐私.传统的位置隐私保护方法只对当前位置和当前时刻的隐私进行考虑,这类方法既没有严格的隐私度量标准,又无法应对在攻击者得到用户的历史时序位置信息的情况下进行概率推测攻击.针对传统方法的这些问题,该文基于概率推测模型设计了一种位置隐私保护算法 MaskK,首先通过隐马尔可夫模型(HMM)对用户的移动状态和位置发布情况进行建模,计算出用户移动位置的抑制发布概率向量,然后利用该概率向量中的概率对用户的部分位置进行抑制发布,使攻击者通过搜集到的用户历史位置数据得到的信息量尽可能的小,并引入 k-匿名思想和粒子群优化算法(PSO)进行优化,进一步提高算法的运行效率和服务质量.该文通过真实数据对提出的算法进行了科学的实验,验证了 MaskK 在隐私保护效果、服务质量和运行效率上的优越性.

关键词 位置隐私保护;k-匿名;粒子群优化算法;隐马尔可夫模型;移动通信网络

中图法分类号 TP311 **DOI 号** 10.11897/SP.J.1016.2018.01037

Mobile Location Privacy Protection Algorithm Based on PSO Optimization

LI Jie[1)] BAI Zhi-Hong[2)] YU Rui-Yun[2)] CUI Ya-Meng[3)] WANG Xing-Wei[2)]

[1)](*Department of Computer Science and Engineering, Northeastern University, Shenyang* 110169)

[2)](*Department of Software, Northeastern University, Shenyang* 110169)

[3)](*Wireless Business Group, Qunar.Com, Beijing* 100080)

Abstract In the era of rapid development of mobile Internet technology and big data technology, mobile users can access a variety of location-based services through mobile smart devices, which becomes an indispensable part of people's daily life. However, location-based services make the location information into data records, which brings a great convenience to people's daily life, at the same time, also brings threats to location privacy of people. Further, this may be detrimental to the user's personal and property safety. Traditional privacy protection methods only consider protecting location privacy of the current time and the current location, and this kind of methods don't define strict mathematical model of privacy measurement. Moreover, if there is a probabilistic inference attack based on the user's timing position data, these methods can't work very well. To this issue, an improved location privacy protection algorithm MaskK is proposed, which is based on probabilistic inference. Firstly, we use Hidden Markov Model (HMM) and the history

收稿日期:2016-11-15;在线出版日期:2017-11-29.本课题得到国家自然科学基金(61572123,61502092,61672148)、国家杰出青年基金(71325002)、中央高校基本科研业务费专项资金资助项目(N151604001)、中国博士后科学基金(2016M591449)、教育部-中国移动科研基金(MCM20160201)、辽宁省百千万人才工程项目(201514)资助.**李 婕**,女,1982 年生,博士,副教授,主要研究方向为机会网络通信、移动大数据分析、社会计算等. E-mail: lijie@mail.neu.edu.cn. **白志宏**,男,1992 年生,硕士,主要研究方向为大数据分析、隐私保护方法. **于瑞云**(通信作者),男,1974 年生,博士,教授,博士生导师,主要研究领域为参与式感知系统、普适与移动计算、大数据分析等. E-mail: yury@mail.neu.edu.cn. **崔亚盟**,男,1991 年生,工程师,主要研究方向为大数据分析. **王兴伟**,男,1968 年生,博士,教授,博士生导师,主要研究领域为未来互联网、云计算、网络安全和信息安全.

location data of the mobile user to establish a model, which shows the user's mobile state and location releasing state. And we define a privacy metric δ-privacy, we can use it to measure the level of privacy by calculating the difference between the posterior probability and the prior probabilities of the user transfer according to the user's location data released. Secondly, based on δ-privacy, in order to protect the user's location privacy, we need to suppress the release of some of the user's locations, so that the attackers get the posterior probability as little as possible by the user's locations released. That is, according to the user's new location data released, the attackers get the amount of information that is as little as possible. So, we need to compute the suppressed release probability vector of the user's location. Meanwhile, we introduce k-anonymity and Particle Swarm Optimization (PSO). In the process of user locations release, the k-anonymity is used to guarantee the user's sensitive locations which are difficult to be distinguished. So as to achieve a fuzzy for user's sensitive location, which increases the difficulty of the attacker to guess the sensitive locations of the user. In the process of calculating the suppressed release probability vector, the PSO algorithm is used to improve the efficiency of the algorithm, which finds the probability vector that satisfies the δ-privacy requirement by multiple iterations. In addition, this paper defines a fitness function of the particle, we can find the relatively optimal suppression release probability vector to protect the user's sensitive locations privacy by finding the extremum that satisfies the fitness function of the particle. Moreover, to prevent an attacker from intercepting the LBS request of user between the user and the trusted center server, we design a kind of user-centric system model. Thus the trusted server only needs to compute the release probability vector. Through the above process, we can make sure that the users' sensitive locations could be protected well, and the efficiency of algorithm is high. At last, we use the real data to conduct a scientific experiment for MaskK, which shows that the MaskK algorithm performs well on the effect of privacy, quality of service, and the efficiency of execution.

Keywords location privacy protection; k-anonymity; particle swarm optimization; hidden Markov model; mobile communication networks

1 引 言

随着移动智能设备和智能应用的广泛传播,人类生活与移动智能应用紧密结合.移动智能终端的数据通信和传感设备等位置感知技术的发展将人和物体的地理位置数据化,例如:手机、PDA 和车载系统等智能移动设备中配置了 GPS、WIFI、陀螺仪、加速度计等与设备位置状态相关的模块,可以随时随地获取移动用户的位置信息[1].然而,位置数据不但直接包含了用户的隐私信息,同时隐含了其它敏感信息.例如,通过数据挖掘技术得到用户的移动轨迹,进而可以推测出其经常访问的地理位置,从而泄露用户的工作地点或者家庭住址等敏感信息[2],因此,位置数据的不恰当使用会严重威胁用户的隐私[3].

位置隐私保护技术经历了长时间的发展,从早期利用访问控制技术保护用户位置信息,到针对单个精确位置进行匿名化[4],再进一步通过匿名算法对用户的敏感位置实现隐匿.多年来,位置隐私保护技术取得了大量成果.但是,随着大数据时代的到来,攻击者即使无法直观地获取用户的敏感位置数据,也可以通过获得移动用户的大量位置相关数据,从其他角度分析用户的历史位置数据,对用户敏感位置进行推测.所以,传统的隐私保护方法已经不能有效地保护用户隐私.

许多 LBS 中,移动用户分享相关位置的风险大于好处[5].这是由于很多移动应用程序通过移动网络收集到的用户个人信息(其中必然包含一些隐私信息)远超过应用程序本身的需求.另外,应用程序并未提供有效的隐私保护策略避免用户的隐私泄露,同时未清楚地说明用户的敏感信息被用于何处,与何种方式与第三方共享.虽然应用提供商通过"知情与同意"的方式向用户发出提示信息,用户可

以选择不安装这些应用或者在使用应用时关闭一些功能(例如 GPS、WiFi 等),但是这样会导致用户体验降低.为了实现用户隐私和实用效率之间的均衡,可以让用户在更细小的粒度上控制自己位置数据的发布,例如用户在交通道路上可能会直接发布位置信息得到相应的服务,而如果用户在银行或者医院等敏感的位置,则考虑是否发布位置信息.因此,通过对位置信息更细粒度的描述,让用户在使用位置服务时,可以选择用户隐私和服务质量的折中点,来选择性地发布位置,并享受应用提供的服务.

为了支持用户对位置信息发布更细粒度地控制,首先应该计算出用户在何时何地何种情况下需要抑制位置信息的发布.针对此问题,早期的研究让用户在指定的敏感位置抑制位置信息的发布,而在其他位置正常发布位置信息.这种算法通过直观的方式保护了用户的位置隐私,但在大数据时代下,攻击者可以通过历史数据推测用户的敏感信息,导致该类算法保护强度明显变弱.

目前,位置隐私保护机制主要有如下 3 类[6]:(1)基于启发式隐私保护度量的位置隐私保护机制[7-8],通常是基于位置 k-匿名的模糊化、空间匿名和时间匿名等隐私机制;(2)基于隐私信息检索的位置隐私保护机制,这类隐私机制是在确保位置服务可用的情况下,使攻击者不能区分用户对不同位置发起的查询访问;(3)基于概率预测的位置隐私保护机制,这类技术主要通过限制攻击者根据获得的用户位置数据对用户出现在某个敏感位置进行的推测的有效性[9].

基于启发式隐私度量的算法种类颇多,如 Beresford 和 Stajano 提出了一种基于假名的位置隐私保护机制[10].定义了一个叫 mix zone 的混合空间,用户发送 LBS 请求后,在 mix zone 中按照一定的规则更换用户之间的假名,使攻击者无法准确地跟踪用户的位置;Mokbel 和 Chow 等人提出了 Casper 模型[11],该算法采用四叉树结构,从用户所在的当前网格开始根据四叉树的索引向上递归,找到满足匿名条件的匿名区域停止,以一个空间替换用户的当前位置发送请求到 LBS 提供商.

基于隐私信息检索的位置隐私保护机制受到设备计算能力的限制,Mouratidis 等人最先提出将隐私信息检索协议(Private Information Retrieval,PIR)用于研究交通路网的最短路径相关查询隐私的问题[12],并且设计了一种简明的索引机制,从而使检索时间更为合理,通过最短路径的计算优化,保护用户的敏感位置信息.

基于概率预测的位置隐私保护机制逐渐受到广泛关注,成为当今位置隐私问题研究的热点之一.Tsai 等人提出了一种简单隐私保护方法 Mask-Sensitive[13],当用户移动到系统或者用户标记的敏感位置时,系统会抑制用户当前的位置信息的发布,在普通位置 $r(r\in R, R=L-S, L$ 表示所有历史位置集合,S 为用户敏感位置集合),用户可以发布当前位置.这个方法虽然能够保证用户不会受到一般的攻击者的攻击,但是如果攻击者收集到大量的用户数据,攻击者能够推断出,当用户的位置发布被抑制时,其所在的位置就是敏感位置,再据此进一步推断当前用户所处的位置是哪一个敏感位置.

Götz 等人提出了改进的 MaskIt 算法[14],该算法通过求取用户的每个位置发布位置信息的概率,对他的敏感位置进行保护.每当用户到达一个位置,该位置有一个对应的发布位置的概率,用户按照这个概率发布或者抑制发布位置信息.在 MaskIt 算法中,假设用户抑制发布当前位置,攻击者即使得不到用户的具体位置信息,但是能够得知用户抑制了发布的事件.在这个假设前提下,攻击者可以采用一些恶意攻击方式(如植入病毒、跟踪攻击等)窃取用户的敏感位置信息.

针对基于上下文的隐私推测攻击,Wang 和 Zhang[15]提出一种基于上下文感知的隐私保护方法,用于抵御那些具有调整攻击策略能力的恶意敌手.通过马尔可夫决策过程(Markov Decision Process,MDP)对用户与攻击者之间的竞争进行建模,量化用户与攻击者之间的竞争关系.另外,提出一种有效的极小极大学习算法去获得最优策略.

Zhang 等人[16]提出位置隐私保护算法 FakeMask 去降低位置信息中时间相关性.在 FakeMask 中发布的位置可能不是真实用户位置(例如可能是用户历史经过的位置)去混淆攻击者,增加了发布的真实位置,从而提高了服务质量,但该方法采用暴力搜索的方式寻找最优解,那需要耗费巨大计算量,因此限制了它的效率.

Yang 等人[17]在差分隐私的基础上,设计了一种 TDPS 模型.差分隐私是一种隐私模型可以用于保护发布的位置数据,能够在攻击者了解用户背景知识的情况下提供强大且独立的隐私保证,并维持数据的可用性,TDPS 模型在它的基础上,利用压缩感知技术,通过为实物数据库构造一个完整的检索树,并在压缩感知的基础上,为检索树添加噪音获得

噪音检索树,最后实现频繁相集挖掘任务.通过以上措施去保护用户数据的敏感位置信息并保证较好的数据可用性.当前,差分隐私方面的研究主要面向怎样加入假信息从而影响分析的结果,并且主要针对离线数据,且差分隐私对于攻击者背景知识的假设非常保守.

另一种典型的概率计算方式是基于图模型,它是 HMM 的一般化形式.在图模型中,用户在某时刻的精确位置可以和他的历史位置数据有关.Parate 等人[18]利用图模型对移动用户的位置轨迹进行建模,并获得了相对准确的位置数据在时间与空间上的关系.并且,其计算先验概率与后验概率的方法和 HMM 的方法类似.

在上述的 3 种方法中,基于启发式度量的位置隐私保护机制在一定程度上容易出现位置的不正确性,而且很多此类隐私技术没有考虑到位置数据的时序性;第 2 种基于信息检索的位置隐私保护,虽然一定程度上可以保证移动用户敏感位置的安全和确保位置数据的准确性,但是算法复杂度高、开销大,目前的硬件环境尚不能很好地支持此类技术的应用;然而,基于概率预测的位置隐私保护机制可以更好地平衡位置隐私算法中隐私度和执行效率之间的关系.

所以,由基于概率推测的位置隐私保护技术的快速发展可知,用户敏感位置信息的发布问题已经成为当前研究的热点.

本文通过对现有的位置隐私保护的体系结构和匿名算法的分析和总结,基于 MaskIt,提出了一种防止概率推测攻击的位置隐私保护机制 MaskK.并结合辽宁移动通信有限公司提供的用户数据进行实验,验证 MaskK 的优越性.主要的创新点如下:

(1)本文利用改进的中心服务器体系结构,通过 HMM 对用户的移动模型进行建模,引入了 k-匿名的思想,保证了在发布位置的过程中,用户的敏感位置难以区分.

(2)在计算用户移动位置的抑制发布概率向量过程中,为了提高运行的效率,使用 PSO 算法对发布概率向量进行求解计算.

(3)自定义粒子适应度函数,通过找到满足粒子适应度的极值,得到保护用户敏感位置隐私的最优抑制发布概率向量.

本文第 2 节主要介绍位置隐私保护体系结构以及攻击模型;第 3 节详细介绍 MaskK 算法的设计,主要有用户的移动模型建立,基于 k-匿名的改进和基于粒子群算法(Particle Swarm Optimization,PSO)的发布概率向量的计算;第 4 节主要介绍实验仿真并分析结果;第 5 节主要对现有工作的总结和对未来的展望.

2 相关工作

2.1 理论基础

为了提高位置大数据中用户隐私保护的性能,本文用到了 k 匿名技术和 PSO 优化算法,下面对其进行具体介绍:

k-匿名:为了保证用户的位置隐私,很多隐私保护技术向 LBS 提供商提供非精确的位置信息,或者直接提供假的信息.k-匿名方法最早是由 Sweeney 在 2002 年提出的,该方法首先应用在数据发布中,用来保护关系型数据库中的数据,k-匿名技术是泛化法的一种,它对每条记录的非敏感属性做泛化处理,使得发布后的数据中任意一条数据都有至少 $k-1$ 条其它数据与之不能区分[19].

PSO 优化:粒子群优化算法[20-22]是一种新兴的进化算法(Evolutionary Algorithm,EA),和模拟退火算法[23]相类似,也是从随机解或指定初始解开始,进行多次循环迭代,逐步找到相对最优解.与其他算法不同的地方在于,它可以通过适应度函数来评价每个解的优良水平.PSO 算法首先要初始化一定数量的随机粒子,再循环迭代寻找满足相应条件的相对最优解.PSO 搜索速度快、效率高,易于工程实现,适合于实值型处理.基于本文的数据量较大的背景能更好地提高算法效率.

2.2 位置隐私保护系统的拓扑结构

目前主流的位置隐私保护体系结构主要包括以下 3 种[23-25]:节点独立的结构、自组织结构和可信的中心服务器结构.

可信的中心服务器体系结构,如图 1,由用户、可信的中心服务器和 LBS 提供商构成,是当前较为

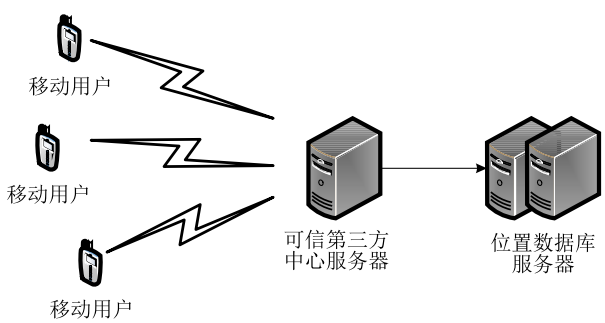

图 1 基于中心服务器的系统拓扑结构

常用的一种体系结构. 在用户与 LBS 提供商之间加入可信的第三方中间件，可以防止 LBS 提供商的不可信从而使得用户位置数据丢失的情况. 可信的中心服务器结构对用户的要求非常少，符合用户便捷化的享受服务的心理.

然而，在大数据背景下，信息变得更加透明，传统位置隐私保护机制无法有效适应当前环境. 而且这种架构下，用户与中心服务器这两个部分非常容易受到攻击者的攻击. 尤其是在下面这种情况：攻击者通过黑客入侵等方式在用户的客户端中植入木马，该木马记录用户所有发出的位置请求，然后将相应数据传给攻击者，此时攻击者就可以在用户和可信第三方中心服务器之间截获用户的 LBS 请求.

本文使用的 MaskK 方法的前提假设是攻击者知道用户的所有信息，第三方中心服务器如图 2 所示，负责计算用户在某地点时是否发布当前的位置信息. 每个位置是否发布当前位置的信息有一个发布概率，攻击者得到该概率也不能判断用户是否发布了当前的位置，只能通过概率推测的手段进行攻击.

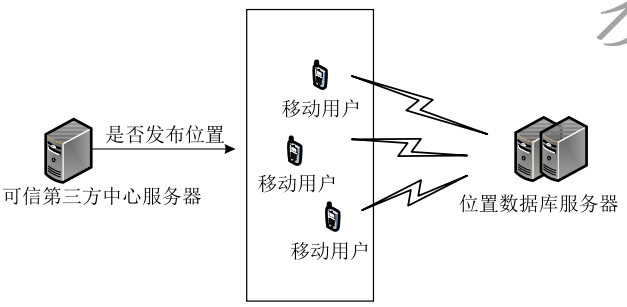

图 2　MaskK 运行系统的拓扑结构

2.3　基于概率推测的攻击模型

在大数据环境下，攻击者通常使用各种数据（包括但不仅限于位置数据）来推测用户在某一时刻 t 的隐私. 与在传统的位置隐私保护机制中关联攻击具有相似的特点，但是其并不能有效地概括大数据时代全方位推测对隐私泄露的威胁. 基于位置大数据的攻击方式通常都会采用相同的量化和度量方法，从而对其攻击的效果进行评估.

本文的攻击模型中，攻击者可以获得移动用户的数据 D，且攻击者希望通过对 D 进行建模，进而可以推测出在 t 时刻移动用户 U 处于某敏感位置 s_i 的概率 P[4].

这里给出一个定义对攻击效果进行量化，即无论何时用户发布当前位置，都不泄漏用户出现在某敏感位置 i 的 δ-隐私.

定义 1.　关于用户敏感位置 δ-隐私[6]. 在任意时刻 t，移动用户处于位置 s_i 的概率为 $P\{U_i^t\}$，且攻击者获取的移动用户发布的历史位置数据用 L_t 表示，对于 δ 有

$$P\{U_i^t|L_t\}-P\{U_i^t\}\leq\delta \tag{1}$$

其中，δ 是用户的隐私需求的度量，δ 越小则用户对自己隐私保护的需求程度越大. U_i^t 表示移动用户 U 在时刻 t 出现在一个敏感地理位置 s_i. $P\{U_i^t\}$ 表示用户 U 在时刻 t 处于位置 s_i 的先验概率，而攻击者通过其所获取的移动用户历史位置数据集合 L_t 推断出用户 U 在 t 时刻在敏感位置 s_i 的后验概率为 $P\{U_i^t|L_t\}$.

从信息熵[6]的角度来解释定义 1：攻击者利用最新发布的用户移动位置数据，为推测某时刻用户出现在某敏感位置提供的信息量要足够的小. 式(2)表示根据移动用户发布的最新的位置数据，能够得到的信息量计算方式：

$$H=\sum_i P\{U_i^t\}\log P\{U_i^t\}-\sum_i P\{U_i^t|L_t\}\log P\{U_i^t|L_t\} \tag{2}$$

根据定义，假设用户定义或系统认定的敏感位置的个数为 n，则攻击者对用户位置数据隐私的最大攻击效果 $n\delta$，而本文中基于概率预测的位置隐私保护方法只需要满足 δ-隐私，即在任意时刻，为了达到用户的隐私保护要求，文中的位置隐私保护方法要满足定义中的 δ/n 隐私.

3　算法设计

3.1　用户移动模型建立

本文通过 Markov 模型[26-27]刻画用户移动模型，当然用户的移动模型有其它描述更深刻的模型，但是 Markov 模型相对简单且可以很好地描述一个人的移动模式，更适合当前的大数据背景[28].

为了方便进行实验，我们将用户移动台连接到的基站位置，近似地作为用户的所处的位置，从中选取部分位置差异性相对较大的用户作为实验用户，假设一个用户到达的所有历史位置的集合为 L，位置点的总数为 n，然后通过以下步骤建立每个用户的马尔可夫链移动模型.

初始化一个 n 维的概率向量 $\boldsymbol{\pi}$，向量 $\boldsymbol{\pi}$ 由 t_0 时的位置 L_i 的初始概率 π_i 组成，如式(3)：

$$\pi_i = P(X(t_0) = L_i),\ i = 1, \cdots, n \tag{3}$$

用一个 $n \times n$ 的概率转移矩阵 $\boldsymbol{A}$ 来表示所有位置点向下一个时刻的位置点转移的概率. $\boldsymbol{A}$ 中元素 a_{ij} 表示某一时刻 t_n 时用户由所在位置 L_i 在 t_{n+1} 时刻移动到 L_j 的概率,如式(4):

$$a_{ij} = P(X(t_{n+1}) = L_j \mid X(t_n) = L_i),\ i, j = 1, \cdots, n \tag{4}$$

如图 3 所示是一个用户移动的马尔可夫链模型. 每个节点都有自身的转移方向和对应的概率,表示当前位置用户可能以对应的概率转移或继续停留. 根据图 3 中所示,可以获得该用户的马尔可夫模型的转移概率矩阵如下:

$$\boldsymbol{A} = \{a_{ij}\} = \begin{Bmatrix} a_{11} & a_{12} & 0 & a_{14} & 0 & 0 \\ 0 & a_{22} & a_{23} & a_{24} & a_{25} & 0 \\ 0 & 0 & a_{33} & 0 & 0 & a_{36} \\ 0 & 0 & 0 & a_{44} & a_{45} & 0 \\ 0 & 0 & 0 & 0 & a_{55} & 0 \\ a_{61} & 0 & 0 & a_{64} & a_{65} & a_{66} \end{Bmatrix} \tag{5}$$

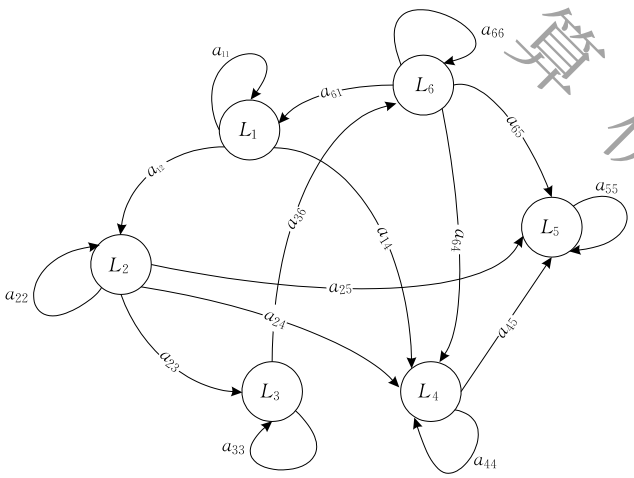

图 3 用户在 6 个状态下移动的马尔可夫链模型

这样我们可以通过一个初始向量和一个概率转移矩阵来组成马尔可夫模型,结合用户的所有位置点的集合来构建出用户的移动模型.

3.2 用户位置发布的 HMM

用户在使用基于位置的服务时,其移动和发布位置信息的行为可以用一个隐马尔可夫模型[29-30]近似地表示.

如图 4 所示,为某用户位置发布的隐马尔可夫模型示意图. $Y = \{y_1, y_2, y_3, y_4, y_5\}$ 表示某用户的历史位置的集合, $X = \{x_1, x_2, x_3, x_4, x_5\}$ 为观察状态的集合. 实线表示各隐藏状态间的状态转移情况,虚线表示隐藏状态发射为可见状态的情况.

其中用户的真实位置(即隐藏状态)的转移为

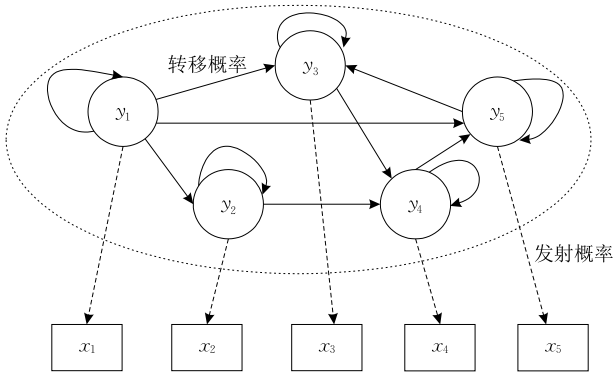

图 4 某用户位置发布的 HMM

Markov 模型. 根据马尔可夫性质假设,有

$$P(y_t \mid y_1, y_2, \cdots, y_{t-1}) = P(y_t \mid y_{t-1}) \tag{6}$$

式(6)表示用户每到达一个位置 y_{t-1},他要到达的下一个位置 y_t 只和当前的位置有关系,与之前的位置序列无关. 此外对于 HMM 还有另外一个假设条件如式(7)所示:

$$P(x_1, x_2, \cdots, x_t \mid y_1, y_2, \cdots, y_t) = \prod_{t=1}^{T} P(x_t \mid y_t) \tag{7}$$

式(7)说明了移动用户于 t 时刻发布的位置,即观察状态 x_t,仅仅与其在 t 时刻所处的位置 y_t 有关,而与之前用户所处的位置和用户所发布的位置无关.

以如下方式表示一个移动用户发布位置的隐马尔可夫模型:

$$\lambda = (L, O, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi}),$$

其中,

$L = \{l_1, l_2, \cdots, l_n\}$:表示状态的有限集合,这里是指用户所有的历史发布位置的集合,其中的元素 l_i 表示用户经过该地点并在这个位置点上发送过位置服务请求.

$O = \{o_1, o_2, \cdots, o_n\}$:表示发布的用户位置集合.

$\boldsymbol{A} = \{a_{ij}\}$:记录用户位置之间转移概率矩阵,其中 $a_{ij} = P[X_{t+1} = l_j \mid X_t = l_i]$ 表示由位置 l_i 转移到 l_j 的转移概率.

$\boldsymbol{B} = \{b_{ij}\}$:是观察值概率发布矩阵,也称作发射矩阵或者混淆矩阵,其中 $b_{ij} = P[O_t = o_j \mid X_t = l_i]$ 表示由位置 l_i 得到发射状态 o_j 的发射概率.

$\boldsymbol{\pi} = \{\pi_i\}$:表示了初始状态下概率分布的向量.

此外,我们定义需要保护的用户敏感位置集合为 $S = \{s_1, s_2, \cdots, s_k\}$,其中 $k < n$,即 $S \subsetneq L$. 攻击者可以通过截取位置服务请求中用户发布的位置信息来推断用户下一时刻处于某敏感位置的概率.

根据以上建立的 HMM,我们能够根据用户的

k 步转移概率矩阵得到某用户在 t 时刻处于敏感位置 s_i 的概率：

$$P(X_t=s_i)=(A^{(1)}\cdot A^{(2)}\cdots A^{(t)})_{s_i} \tag{8}$$

一系列状态的联合概率为

$$P(X_1,\cdots,X_T)=\prod_{t=1}^{T}a^{(t)}_{X_{t-1},X_t} \tag{9}$$

并且，由式(8)我们可以得到用户处于位置 l_i 到 l_j 的转移概率：

$$P(X_{t_2}=l_j\,|\,X_{t_1}=l_i)=(\boldsymbol{e}_iA^{(t_1+1)}\cdot A^{(t_1+2)}\cdots A^{(t_2)})_j \tag{10}$$

其中 $\boldsymbol{e}_i$ 是一个单位向量，其中的第 i 个元素为 1，其余元素都为 0.

后验概率是通过观察序列和已知用户移动的 Markov 模型等条件推断用户在 t 时刻位于 l_i 位置的概率. 对于一个已经发布的观察状态的序列 $O_T=(o_1,o_2,\cdots,o_T)$，我们能够计算出某用户 t 时刻出现在位置 l_i 的条件概率如下：

$$P(X_t=l_i\,|\,\boldsymbol{o})=\frac{P(X_t=l_i,o_1,o_2,\cdots,o_T)P(o_1,o_2,\cdots,o_T\,|\,X_t=l_i)}{P(\boldsymbol{o})} \tag{11}$$

根据前向算法我们可以计算上面公式中的一部分

$$\begin{aligned}\alpha_i(t)&=P(X_t=l_i,o_1,o_2,\cdots,o_T)\\&=\sum_{j=1}^{n}\alpha_j(t-1)a^{(t)}_{ji}.b^{t-1}_{i,o_{t-1}}\end{aligned} \tag{12}$$

而对于另一部分 $P(o_1,o_2,\cdots,o_T\,|\,X_t=l_i)$，就类似地使用后向算法. 我们称这一部分为后向变量，和前向变量对应地表示为 $\beta_i(t)$.

$$\begin{aligned}\beta_i(t)&=P(l_t,l_{t+1},\cdots,l_T\,|\,X_t=l_i)\\&=\sum_j b^t_{i,l_t}a^{t+1}_{i,j}\beta_j(t+1)\end{aligned} \tag{13}$$

我们假设 $T+1$ 时刻的后向变量均为 1，然后我们可以通过前向后向算法[31-32] (Forward-Backward Algorithm)得到在观察序列 $O_T=o_1,o_2,\cdots,o_T$ 的条件下，t 时刻位于位置 l_i 的概率：

$$P_{posterior}=P(X_t=l_i\,|\,O_T)=\frac{\alpha_i(t)\beta_i(t)}{\sum_j\alpha_j(t)\beta_j(t)} \tag{14}$$

通过式(14)可知，攻击者可以利用用户当前发布的位置数据推测其 t 时刻出现在位置 l_i 的后验概率，随着发布的位置数据的增加，攻击者将推测出更加准确的位置信息，从而泄露用户敏感位置的隐私信息. 因此，为了有效地保护移动用户的敏感位置隐私，需要对移动用户的敏感位置进行抑制发布，同时对相应的非敏感位置进行一定数量的抑制，使攻击者根据当前发布位置向量得到的后验概率满足 δ-隐私，即 $P_{posterior}-p(X_t=s_i)\leq\delta$. 如算法 1 所示为 δ-隐私检验算法.

算法 1.　强攻击背景下 δ-隐私检验算法.

1. 输入：抑制发布概率向量 $\boldsymbol{p}$，观察序列 O_T，Markov 模型 M，敏感位置集合 S，参数 k、δ
2. 初始化抑制发布向量 $\boldsymbol{p}$ 获得一个观察序列 O_T flag=false
3. FOR EACH $s\in S$ DO
4. FOR $t\in[T]$ DO
5. IF o_t is NaN THEN
6. 计算先验概率 $prior=P(X_t=s)$
7. 计算后验概率 $posterior=P(X_t=s|O_T)$
8. IF $|posterior-prior|\leq\delta$ THEN
9. RETURN flag=true
10. ELSE
11. RETURN flag=false
12. END IF
13. END IF
14. END FOR
15. END FOR
16. 输出：flag

3.3　MaskK 算法描述

基于 k-匿名思想和 PSO 优化的位置隐私保护机制 MaskK 包括两个过程，首先利用 PSO 计算出用户移动位置的抑制发布概率向量，然后根据每个位置的发射概率(发射概率=1−抑制发布概率)决定对当前位置是否进行发布. 同时引入 k-匿名思想，假设用户敏感位置为 m 个，为了进一步保护其隐私信息，可以由用户指定或系统设定抑制发布的位置个数 $k(k>m)$. 在确保 δ-隐私的情况下，确保一定数量的非敏感位置抑制发布，达到混淆攻击者的目的，使用户的敏感位置更加难以区分.

根据攻击假设，攻击者能够得到某一时刻 t 用户的位置服务请求受到抑制的信息，例如攻击者在用户手机植入一段程序去获取用户每次进行位置服务请求时的信息，在位置发布未受到抑制的情况下，攻击者得到的信息可以表示为以下三元组 $(id,location,query)$，其中 id 是用户的标识，$location$ 是用户发布的位置数据，$query$ 表示用户的位置服务请求，当位置发布受到抑制的情况下，攻击者会得知用户要发起一次基于位置的服务请求，但是由于发布受到抑制，因此攻击者得到的三元组可能为 (id,NaN,NaN)，虽然攻击者没有得到用户的位置和具体的请求，但是他依然获得到了用户的一个观察状态. 所以用户所有的观察状态集合 O 为

$$O=(L-S)\bigcup\{NaN\} \tag{15}$$

这个观察序列的状态首先包含了除去敏感位置外的所有位置,然后还包含了一个 NaN 状态.当用户发布的位置信息为 NaN 时表示用户在时刻 t 的位置发布受到了抑制.

MaskK 算法首先初始化用户的状态,通过对用户历史位置的统计计算,得出用户的 Markov 模型,再初始化用户的 Markov 模型中参数,即用户的第一个位置为出发点,用户在第一个时刻 t_0 所处的位置就是第一个位置.然后对于以后的每一个经过的位置,判断该位置是否能够发布位置信息,如果该位置能够发布位置就直接发布,如果系统判定该位置不能发布位置信息,则在信息发布时就会受到抑制,发布出的信息就为 NaN,最后将发布的信息加入观察序列(算法 2 第 7~11 行).

3.3.1 基于 PSO 优化的抑制发布概率向量计算

为了使 MaskK 能够更好地保护用户隐私,使用定义 1 中的 δ-隐私来限制攻击者获得最小的信息量增益.我们对于一些具有 HMM 知识背景的攻击者,称为强攻击,这一类的攻击者通常掌握了 HMM 的所有背景知识,能够解决在 HMM 的有关问题中的评估问题、解码问题和学习问题.攻击者通过已知用户 Markov 移动模型,即获得了用户的初始位置和用户在不同位置之间的转移概率,然后调整 HMM 模型 λ 的参数,在给定的一组观察序列时,推断用户在该观察序列的某个或某 n 个敏感位置的概率.这显然是一个 HMM 的学习问题.此时,我们只需要找到一组合适的抑制发布的概率向量,使得这个抑制发布的概率向量能够保证定义中的 δ-隐私,也就是通过所得的抑制发布的概率向量发布位置,使攻击者根据获取的位置计算的后验概率与先验概率之差小于 δ,从而使攻击者不能有效推测出用户的敏感位置信息.

我们将一个用户的所有历史发布位置分为两类:一类是用户的敏感位置,为了保护用户的敏感位置信息,尽可能使这些位置的抑制发布概率为 1,即观察状态为 NAN.另一类为普通位置,通过一定概率对这些位置进行抑制发布,从而攻击者难以区分抑制发布的位置当中哪些是敏感位置,使攻击者推出的后验概率和其获得的用户到达某敏感位置点的先验概率之差小于 δ.

例如图 5,S 为用户的敏感位置,S'为用户的非敏感位置,用户在初始位置出发,经过 S 或 S'到达终点位置.根据对用户的历史信息建模,已知用户到敏感位置 S 的先验概率为 0.2,到达非敏感位置 S' 的概率是 0.8.此时攻击者得到 t 时刻用户的发布状态为 NAN(当前位置未发布),则根据可见状态,用户在 t 时刻到达 S 的后验概率

$$p(S|NAN)=\frac{p(NAN|S)p(S)}{p(NAN)}=\frac{1\times 0.2}{1\times 0.2+p(NAN|S')\times 0.8} \tag{16}$$

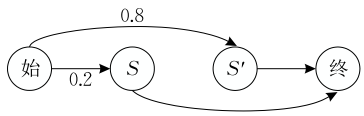

图 5 简单的用户移动模型

为了满足 1/5-隐私,所以有

$$p(S|NAN)-p(S)\leq\delta \tag{17}$$

可以得到 S'的抑制发布概率 $p(NAN|S')\geq 3/8$,也就是说,只有当 S'的抑制发布概率大于 3/8 时,且 S 的抑制发布概率为 1,满足 δ-隐私条件,从而可以有效地保护用户位置隐私.

由此可知,要保护用户的敏感位置隐私,提高隐私保护度,对非敏感位置也要以一定的概率进行抑制发布.所以,我们要得到一个抑制发布概率向量:

$$\boldsymbol{p}=\{p_1,p_2,\cdots,p_n\} \tag{18}$$

对用户的所有位置进行发布和抑制发布,其中,p_n 表示在第 n 个位置的信息发布受到抑制的概率.在满足 δ-隐私的同时,要最大化用户发布位置的效用值,使得用户的服务质量达到最高.因此我们定义如下适应函数:

$$fitness=-|privacy-\delta|+utility \tag{19}$$

$$privacy\leq\delta \tag{20}$$

式(19)的第一部分如算法 1 所示是隐私检验($privacy-\delta$),第二部分是服务质量 $utility$,该公式涵盖了隐私要求与服务质量需求,它的最大值即为在满足隐私要求的条件下服务质量的最大化.其中 $privacy$ 表示如下:

$$privacy=\max\{privacy_1,privacy_2,\cdots,privacy_n\} \tag{21}$$

$privacy_i$ 表示通过抑制发布概率向量发布的观察序列中攻击者能够推断出某个时刻位于某敏感位置的后验概率和先验概率之差:

$$privacy_n=|P_{posterior}-P_{prior}| \tag{22}$$

其中,$P_{posterior}$ 表示后验概率,P_{prior} 表示先验概率.而 $privacy$ 表示所有这些后验概率和先验概率之差的最大值,$privacy$ 与 δ 越接近,服务质量的损失越小.

适应函数的第二部分是衡量服务质量的函数,这里我们定义抑制发布概率向量的效用函数:

$$utility(\boldsymbol{p})=\sum_{O_T}P(O_T)\,|\{i\,|\,o_i\neq NaN\}|$$
$$=\sum_{t\in[T],i\in[n]}P(X_t=l_i)(1-p_i^t)\quad(23)$$

其中，p_i^t表示 t 时刻所处的位置点 l_i的抑制发布概率，即发布位置的观察状态为 NaN 的概率. 通过定义 $utility(\boldsymbol{p})$，即发布位置的数学期望，表示在满足 δ-隐私的情况下，发布的用户位置越多，利用位置大数据应用(如道路推荐应用)提供的服务将更精确，服务质量越好.

但是，要找到使 $fitness$ 达到最大值的抑制发布向量，即要确定用户所有移动位置的抑制发布概率，其组成的向量使 $fitness$ 达到最大，这个过程需要遍历每位取值为(0,1)的所有 N(N 表示用户移动位置个数)维向量，是 NP 难问题. 因此，我们使用优化的 PSO 算法去计算抑制发布概率向量 $\boldsymbol{p}=\{p_1,p_2,\cdots,p_n\}$. PSO 算法中的每个粒子即一个抑制发布的概率向量.

通过以上论述，粒子适应度值越大说明粒子的适应度越好，另外我们计算后验概率和先验概率之差(差值越小隐私保护效果越好)，并将其与 δ 的差取负值和效用值 $utility$ 相加作为粒子适应度函数，表示当后验概率和先验概率之差越接近 δ，且 $utility$ 越大时，适应度值就越大，说明粒子(即抑制发布概率向量)就越优秀. 确定了适应函数后，初始化 $1.5\times n$ 个 n 维向量作为 PSO 的搜索粒子，n 表示用户的总位置个数.

由以上知识可知，抑制发布概率向量 $\boldsymbol{p}$ 决定了用户位置隐私水平，当 $\boldsymbol{p}$ 中的概率变大时，发布的用户位置可能减少，由式(14)得到的后验概率将减少，从而根据式(2)中可知攻击者可以得到的用户的信息量变少，则隐私水平增大. 所以，隐私度是 $\boldsymbol{p}$ 的单调增函数，则 $-|privacy-\delta|$ 是 $\boldsymbol{p}$ 的减函数. 反之，根据式(23)$utility(\boldsymbol{p})$是 $\boldsymbol{p}$ 的单调减函数，$utility$ 随 $\boldsymbol{p}$ 的增大而减小.

根据隐私度的单调性，我们可以对 PSO 算法进行进一步的优化. 现实生活中，为了确保用户基本的位置服务质量需求，在用户经过的所有位置中，发布的位置数量肯定要占用户总位置数 n 的绝大部分. 另外，敏感位置的抑制发布概率较大，可以提前将部分敏感位置的抑制发布概率置为 1，降低粒子维度. 所以在抑制发布概率向量中多数的抑制概率范围是(0,0.5)的. 所以我们可以初始化所有粒子的概率为(0,0.5). 每个粒子以向量中的一个非敏感位置元素为方向开始搜索. 通过逐渐增大每个粒子，并根据其中最优粒子对其它粒子进行位置更新，向最优解靠近. 直到得到一个相对最优粒子 $\boldsymbol{p}_{\max}$，那达到满足 δ-隐私的临界值或预设定的迭代次数，停止搜索过程，将这个粒子作为最终的抑制发布概率向量. 此时，效用值较大，达到了保证位置隐私条件下的相对最优.

经过以上过程，我们可以得到一个相对最优的抑制发布概率向量 $\boldsymbol{p}_{\max}$，使 $fitness$ 达到相对最大值，从而使用户的敏感位置隐私和服务质量达到了最优的均衡. 可能有其它的向量可以使值达到最优，但那需要遍历(0,1)范围的所有 n 维向量，那是难以实现的.

3.3.2　基于 k-匿名的用户位置发布

本节描述 k-匿名在 MaskK 算法中的应用，MaskK 算法的核心功能部分见算法 2. MaskK 算法的输入为用户的 Markov 模型 M，参数 k、δ，输出为用户的观察序列.

算法 2.　MaskK 用户观察序列生成算法.

1. 输入：用户 Markov 模型 M，位置集合 L，敏感位置集合 S，参数 k、δ
2. $initialize(\delta,k,M,L,S)$
3. 初始化 M 的参数 $\boldsymbol{\pi}=(1,0,0,\cdots,0)$
4. 观察序列 $O=\{l_0\}$
5. FOR 当前时刻 $t=1$ TO T DO
6. $\quad l_t=getUserLocation(t)$
7. $\quad$IF $isSuppressed(t,l_t,O)$ is true THEN
8. $\quad\quad o_t=NaN$
9. $\quad\quad Num++$
10. $\quad$ELSE $o_t=l_t$
11. $\quad$END IF
12. $\quad O=O+\{o_t\}$
13. END FOR
14. IF $Num<k$
15. ELSE FOR $maxSuppressed(t,O)$ is true THEN
16. $\quad\quad o_t=NaN$
17. END IF
18. 输出：观察序列 O

其中函数 $getUserLocation(t)$ 表示获取用户在 t 时刻他所处的位置点. $isSuppressed(t,l_i,O)$ 表示用户当前时刻 t 处于位置 l_i，系统根据之前的观察序列，来判定当前位置当前时刻对位置信息的发布是否会受到抑制，并且这个函数需要保证最终输出的观察序列中位置发布受到抑制的点至少有 k 个. 在这 k 个点中至少有 k' 个普通位置的信息发布为 NaN，以此使得这 k' 个普通位置的观察状态和敏感位置难以区分，其中：

$$k=m+k' \tag{24}$$

m 表示抑制发布的敏感位置的个数. 当对 k 的数量进行限定后,根据上述得到的抑制发布概率向量,在算法运行中可以得到 Num 个抑制发布的位置,如果 $Num<k$,不满足限定条件,则要在用户的移动位置中找出 k-Num 非敏感位置进行抑制发布. 已知某个位置 i 的发射概率

$$b_{i,o}=1-p_i \tag{25}$$

其中 p_i 表示位置 i 的抑制发布概率. 由式(12)~(14)知,抑制发布概率越大,攻击者得到的后验概率越小. 从而攻击者利用观察序列得到的后验概率是抑制发布概率向量的反函数. 所以,我们找出抑制发布概率向量中概率最大的 k-Num 概率所对应的位置,对这 k-Num 个位置进行抑制发布操作,即抑制发布概率为 1,从而最小化推测的后验概率.

如果限定 $k=3$,敏感位置个数为 $m=1$,如图 6 所示,是 MaskSensitive 方法保护用户隐私的隐马尔可夫模型示例,只抑制用户的两个敏感位置.

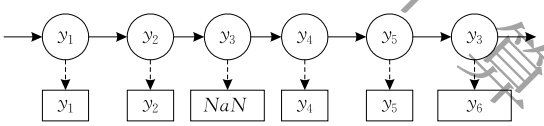

图 6 MaskSensitive 方法中用户的 HMM

而经过算法 2 的步骤,最后用户位置发布的 HMM 如图 7 所示,示例图中的用户观察序列有 $k=3$ 个位置被抑制发布,实现 k-匿名.

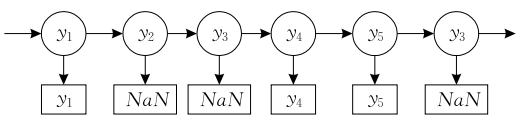

图 7 MaskK 方法中 k-匿名改进的用户 HMM($k=3$)

对于 MaskSensitive,通过统计观察序列中的 NaN 状态,攻击者很容易推测用户某时刻位于敏感位置的概率. 而在 MaskK 中,至少存在 k' 个无法和敏感位置区分的普通位置. 所以,k 值越大,抑制发布的位置当中非敏感位置越多,攻击者推测出用户的敏感位置越困难,但本文中 k 值过大会影响用户的服务质量,所有 k 值不宜过大. 本文中根据用户的隐私度需求适当设定 k 值,可以使 MaskK 方法发布的观察序列有效降低攻击者推断用户某时刻位于敏感位置的后验概率.

4 实验仿真

4.1 实验环境及数据集

实验的环境为64位Windows7系统,内存(RAM)为 8.00GB,处理器为 Intel(R) Core(TM) i7. 实验数据的来源主要是辽宁移动通信公司提供的 Abis 接口数据.

首先我们对用户的位置数据进行预处理,提取用户逗留的位置点,将用户的 LAC(移动通信系统中的位置区编码)和 CI(同一位置区内的小区编码)对应到移动公司提供的基站位置表中,如果位置点在连续时间内不发生改变,则认为用户在该位置点逗留.

如图 8 所示,通过对辽宁移动公司中沈阳地区五个主要市区的将近 200 万用户的位置切换记录统计,根据图中所示信息可以看出,用户在上午七点到晚上九点之间移动较为频繁,移动位置数量多,可以使实验效果更为明显,所以选取的实验数据为每天上午七点到下午九点之间的用户位置更新数据.

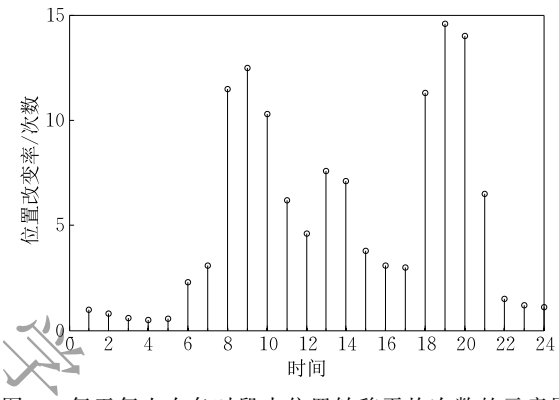

图 8 每天每人在各时段内位置转移平均次数的示意图

4.2 隐私保护效果及安全性分析

通过与已有的抵御上下文攻击类似的隐私保护算法对比,MaskSensitive 方法直接抑制敏感位置点的发布,而 MaskIt 方法则通过一组概率向量来对所有的点以一定的概率进行抑制,FakeMask 方法发布的位置可能为虚假的位置信息去降低攻击者推测的后验概率. 本文中的位置隐私保护方法 MaskK 处理对所有的点以一定的概率进行抑制外,还保证了被抑制的位置点满足 k-匿名,并且保证匿名点在 k 值的附近浮动,不会出现大部分位置点都受到抑制的情况.

如图 9 所示,是对 50 个用户取某一天的观察序列,使用 4 种方法进行隐私检验的结果柱状图. 每种方法有 3 个柱状图分别表示出现 3 种状态的位置所占比例(rate). 这 3 个柱状图中,第 1 个柱状是表示发布的位置点所占比例,其中,灰色部分为用户的敏感位置;第 2 个柱状图表示抑制发布的位置点所占比例,灰色部分是指抑制发布的位置点中敏感位置点所占比例,阴影部分是普通位置;第 3 个柱状图表

示违背了 δ-隐私的敏感位置占所有敏感位置的比例. 从图 9 中可以看出，MaskSensitive 方法抑制发布的位置点占用户位置点个数的 22%，尽管如此我们可以从该方法的第 3 个柱状图看出，敏感位置中不满足 δ-隐私的敏感位置点占 48%. MaskK 相对 MaskSensitive 方法对敏感位置点的保护，不会出现违背 δ-隐私的情况. 相对 MaskIt 方法，MaskK 算法抑制发布的位置点的个数会由于概率(即抑制发布概率向量)原因而出现较少，而发布的位置数量较多. 因此在保证了隐私检验的前提下，发布更多的位置数据，可以提高位置服务的精确性，从而提高用户的位置服务质量. FakeMask 中用假位置进行混淆，可以更多的发布用户的位置，但是没有对抑制发布的位置中用户的非敏感位置进行限制，容易造成抑制发布的位置中非敏感位置过少. 由以上可以看出，MaskSensitive 方法中大量敏感位置违背了 δ-隐私，即攻击者根据发布的位置推测的后验概率与先验概率之差远大于 δ，那泄露用户的敏感位置信息，而另外 3 种方法违背 δ-隐私的位置比例均为 0，从而有效的保护用户的敏感位置不被攻击者通过上下文推测攻击获取.

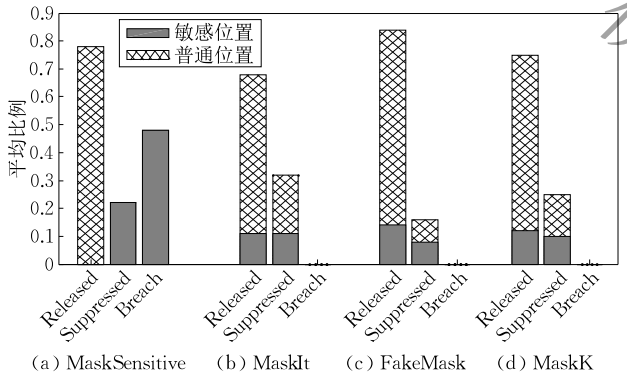

图 9　4 种方法隐私检验结果的柱状图

如图 10 所示是在基于 Forward-Backward 算法的攻击模型下，主要比较 MaskK 与文献[10]中方法的隐私保护效果.

这部分的实验使用了 85 个用户的位置数据，这些用户的数据移动位置覆盖范围比较完整，位置点的距离适中，容易区分，实验结果可以更为完整和明显. 通过这些数据首先得到抑制发布的概率向量，其中输入的参数 δ 的值取 0.1，然后通过抑制发布的概率向量和用户的 Markov 模型，计算同一个时刻的用户处于敏感位置的后验概率和先验概率之差，用来表示隐私保护效果，差值越小，说明推测出的后验概率涵盖的信息量就越少，隐私效果则更好.

从图 10 中可以看出，MaskSensitive 的方法在位置个数较少时能够基本满足用户的隐私需求，但是位置点增多时，无法满足 δ 隐私. 而本文中的方法和 MaskIt 方法在各种情况下均能够满足 δ 隐私，并且在位置点个数较少时，能够接近完美隐私即 δ 为 0 的情况. 本文中的方法在隐私检验方面相较 MaskIt 能够更好地为用户提供隐私保护.

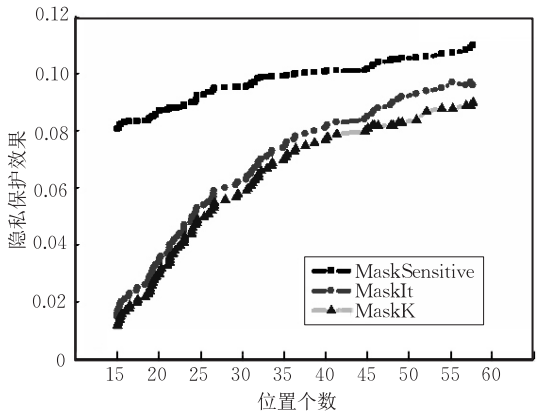

图 10　基于 Forward-Backward 算法的攻击模型下 3 种方法的隐私保护机制的安全性($\delta=0.1$)

综合上述的实验结果，本文算法通过定义合适的目标函数和 k 匿名思想的引入，可以在确保用户服务质量的前提下更好的保护用户的敏感位置信息.

4.3　服务质量分析

用户在使用位置隐私保护系统时，为了保护个人隐私，必然受到系统的约束，导致小部分位置发布收到抑制，无法提供完整的位置服务. 本文的 MaskK 方法在满足用户隐私需求的同时定义了隐私保护的服务质量(*utility*)，衡量用户发布的位置数量，*utility* 值越大，则意味着发布的用户位置数量相对越多，最大限度地保证用户的服务质量.

图 11 所示，在相同迭代次数(2000 次)的情况下，不同 δ 的取值时，用户使用位置隐私保护系统的效用值对比示意图. 图中可以看出，在不同的

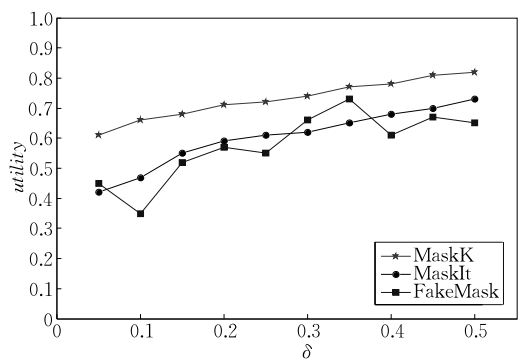

图 11　不同的 δ 服务质量对比图

δ的取值时，MaskK 的服务质量均优于 MaskIt 和 FakeMask 方法. 理论上 FakeMask 达到相对最优解时的 *utility* 要优于本文的方法，因为其混入假位置，发布的位置数量更多. 但是，在相同迭代次数限制的情况下，采用暴力搜索的方式很难找到相对最优解，且得到的结果不稳定. MaskK 的效用值在相同迭代次数下随着 δ 的变化逐渐增大，比较稳定，相较于另两种方法有明显的优势.

图 12 所示的是用户的敏感位置的数量不同时两种方法的服务质量对比图，3 种方法随着敏感位置的增多，效用值都有降低的趋势，因为敏感位置增加，校验抑制发布概率向量的成功率变低，找到满足 δ-隐私的向量变的困难. 并且，MaskK 方法在敏感位置个数较少时，由于要保证 k-匿名的隐私需求，因此要指定一定数量的位置点对位置信息的发布进行抑制，由此导致效用低于 MaskIt，但是随着敏感位置点的增加，效用渐渐优于 MaskIt. FakeMask 由于限定迭代次数，效用值下降比较明显，且不稳定.

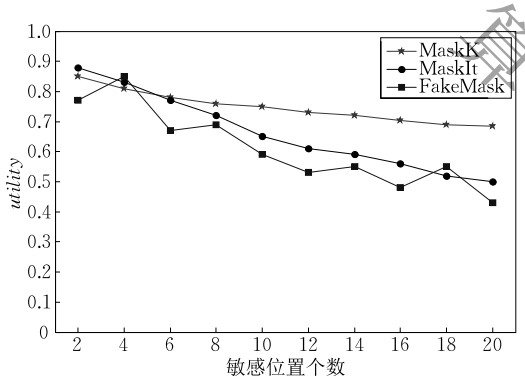

图 12 不同的敏感位置个数服务质量对比图($\delta=0.1$)

综上可以看出，本文算法在效用值上相较于其它两种方法存在优势，从而可以为用户提供更好的服务质量.

4.4 抑制发布概率向量的计算效率分析

这部分的实验分为两组. 第 1 组实验选取 8 个用户的数据进行对比实验，如表 1 中所示，用户的敏感位置个数相同，但是总的历史位置个数不同.

表 1 第 1 组实验数据表

用户编号	位置个数	敏感位置个数
01	80	15
02	100	15
03	120	15
04	140	15
05	160	15
06	180	15
07	200	15
08	220	15

该实验将 3 个方法使用同样的数据，为了使实验效果更明显，实验中数据的预处理和隐马尔可夫模型的建立的时间不计入考虑，只对得到一个满足 δ-隐私的抑制发布概率向量的计算速度比较，突出 3 种方法计算效率的好坏. 所以背景为用户的移动模型已经建立，在这个基础上计算获得用每一个抑制发布概率向量所需的时间，其中 k 的值设置为 25.

如图 13 所示，是 MaskIt、FakeMask 和本文提出的方法 MaskK 按照表 1 中数据进行实验后的计算抑制发布概率向量所需时间对比折线图. 由图中所示随着用户位置数的增加，3 种方法所花费的时间逐渐加长. 但 MaskK 方法由于采用 PSO 优化算法，迭代速度快，相比于其它两种方法在时间效率上拥有明显的优势.

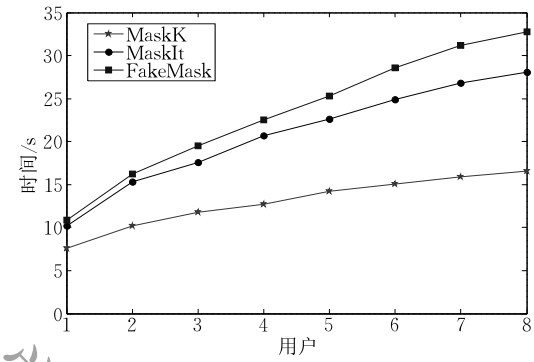

图 13 位置个数不同时 3 种方法计算抑制发布概率向量所需时间的对比图

第 2 组实验的数据如表 2 所示，选取 5 组位置个数为 160 的用户，在这 160 个位置当中选出不同数量的敏感位置去计算获得抑制发布概率向量所需的时间，然后取 5 组数据的平均值进行对比实验.

表 2 第 2 组实验数据表

编号	位置个数	敏感位置个数
01	160	10
02	160	15
03	160	20
04	160	25
05	160	30
06	160	35
07	160	40

如图 14 所示，本文方法 MaskK 在用户的敏感位置不同时，计算所需时长均低于 MaskIt 和 FakeMask 方法，由于 FakeMask 采用的是暴力搜索的方式寻找抑制发布概率向量，计算效率有明显的劣势. 并且，在同一个用户的相同位置数据中，历史位置点个数不变，随着敏感位置点的增多，MaskK 耗时有下降的趋势，这是由于根据 MaskK 中粒子的特性对

PSO 进行了一定的优化，当敏感位置越多时计算速度有一定的提高.

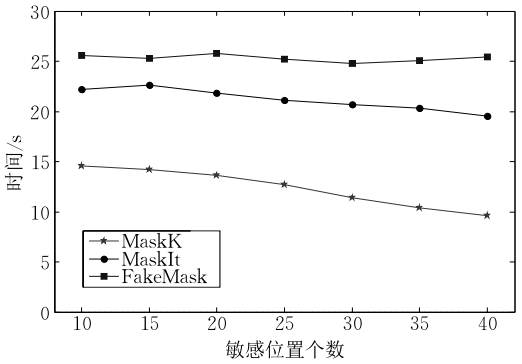

图 14　敏感位置个数不同时 3 种方法计算抑制发布概率向量所需时间

综上，在不同总位置数量和不同敏感位置数量两种情况下进行计算时间对比，本文算法在时间效率上有着明显的优势. 随着用户位置的增多，花费时间也随之变化，但这些过程都在第 2 节介绍的中心服务器上完成，用户只需访问服务器运算的结果，因此所需时间仍是毫秒级，对用户的服务质量影响较小. 本文算法在保证用户位置隐私的情况下具有较高的时间效率，同时可以为用户提供更好的位置隐私服务质量.

5　结论与展望

用户在使用基于位置的服务过程中，对位置隐私的要求将越来越高，所以位置隐私保护的研究是基于位置服务的应用基础，并且面临着很多挑战. 对于传统的位置隐私保护方法采用的 3 种常见的拓补结构：独立结构、自组织结构、中心服务器结构，本文首先基于中心服务器结构设计出了一种新的基于可信服务器的拓补结构，弥补了中心服务器成为攻击目标的不足. 在此基础上本文提出了 MaskK 算法，该算法使用 Markov 模型对用户的移动模式进行建模，引入了 k-匿名的思想，保证了在发布位置的过程中，有 $k-m$ 个普通位置点与隐私位置点无法区分. 另外，在 HMM 的建立中使用了 PSO 优化算法求取抑制发布概率向量，提高算法的性能及服务质量.

此外，在基于 PSO 算法的求解过程中，确定适应度函数时，引入基于前向后向算法(Forward-Backward Algorithm)的强攻击模式，使得最后求解得出的抑制发布的概率向量能够抵御强攻击的情况.

最后，本文在辽宁移动公司提供的用户真实数据集上进行了充分的实验，证明了 MaskK 算法在隐私保护效果和服务质量以及运行效率等方面的优越性.

在未来的研究工作中，尝试使用其他概率图模型进行用户移动模式的建模，使得该模型更能真实的描述用户移动模式. 另外，本文中的算法虽然能够较好地求取抑制发布的概率向量，但是在用户的位置出现极端的情况下(例如，用户为出租车司机，位置记录数异常多)可能会导致 PSO 优化求解中粒子的维度过大，在以后的工作中需要加入对极端因素的考虑.

参　考　文　献

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

[2] Wicker S B. The loss of location privacy in the cellular age. Communications of the ACM, 2012, 55(8): 60-68

[3] Jabeur N, Zeadally S, Sayed B. Mobile social networking applications. Communications of the ACM, 2013, 56(3): 71-79

[4] Beresford A R, Rice A, Skehin N, et al. Mockdroid: Trading privacy for application functionality on smartphones// Proceedings of the 12th Workshop on Mobile Computing Systems and Applications. Phoenix, USA, 2011: 49-54

[5] Cheng R, Zhang Y, Bertino E, et al. Preserving user location privacy in mobile data management infrastructures//Danezis G, Golle P eds. Privacy Enhancing Technologies. Berlin, Germany: Springer, 2006: 393-412

[6] Wang Lu, Meng Xiao-Feng. Location privacy preservation in big data era: A survey. Journal of Software, 2014, 25(4): 693-712(in Chinese)
(王璐，孟小峰. 位置大数据隐私保护研究综述. 软件学报，2014, 25(4): 693-712)

[7] Chakraborty S, Shen C, Raghavan K R, et al. ipShield: A framework for enforcing context-aware privacy//Proceedings of the Usenix Conference on Networked Systems Design and Implementation. Seattle, USA, 2014: 143-156

[8] Damiani M L. Location privacy models in mobile applications: Conceptual view and research directions. Geoinformatica, 2014, 18(4): 819-842

[9] Machanavajjhala A, Reiter J P. Big privacy: Protecting confidentiality in big data. XRDS: Crossroads. The ACM Magazine for Students, 2012, 19(1): 20-23

[10] Beresford A R, Stajano F. Location privacy in pervasive computing. IEEE Pervasive Computing, 2003, 2(1): 46-55
[11] Mokbel M F, Chow C Y, Aref W G. The new casper: Query proces sing for location services without compromising privacy//Proceedings of the International Conference on Very Large Data Bases, Seoul, Korea, 2006: 763-774
[12] Mouratidis K, Man L Y. Shortest path computation with no information leakage//Proceedings of the International Conference on Very Large Data Bases. Istanbul, Turkey, 2012: 692-703
[13] Tsai J, Kelley P G, Cranor L F, Sadeh N. Location sharing technologies: Privacy risks and controls. Journal of Law and Policy for the Information Societ, 2010, 6(2): 1-26
[14] Götz 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. Scottsdale, USA, 2012: 289-300
[15] Wang W, Zhang Q. Privacy preservation for context sensing on smartphone. IEEE/ACM Transactions on Networking, 2016, 24(6): 3235-3247
[16] Zhang L, Cai Z, Wang X. FakeMask: A novel privacy preserving approach for smartphones. IEEE Transactions on Network & Service Management, 2016, 13(2): 335-348
[17] Yang Yin, Zhang Zhenjie, Miklau G, et al. Differential privacy in data publication and analysis//Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data. Scottsdale, USA, 2012: 601-606
[18] Parate A, Chiu M C, Ganesan D, Marlin B M. Leveraging graphical models to improve accuracy and reduce privacy risks of mobile sensing//Proceedings of the International Conference on Mobile Systems, Applications, and Services. Taipei, China, 2013: 83-96
[19] Sweeney L. k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002, 10(5): 557-570
[20] Poli R, Kennedy J, Blackwell T. Particle swarm optimization. Swarm Intelligence, 2007, 1(1): 33-57
[21] Kennedy J. Particle swarm optimization//Sammut C, Webb G I eds. Encyclopedia of Machine Learning. New York, USA: Springer, 2010: 760-766
[22] Kennedy J, Eberhart R. Particle Swarm Optimization. New York, USA: Springer, 2011: 1942-1948
[23] Cheng R, Zhang Y, Bertino E, et al. Preserving user location privacy in mobile data management infrastrnctures//Proceedings of the Privacy Enhancing Technology Workshop (PET'06). Cambridge, UK, 2006: 393-412
[24] Gruteser M, Grnnwald D. Anonymous usage of location based services through spatial and temporal cloaking//Proceedings of the International Conference on Mobile Systems, Applications, and Services (MobiSys'03). San Francisco, USA, 2003: 163-168
[25] Chow C, Mokbel M F, Liu X. A peer-to-peer spatial cloaking algorithm for anonymous location-based setvices//Proceedings of the ACM Symposium on Advances in Geographic Information Systems (ACM GIS'06). Arlington, USA, 2006: 171-178
[26] Kim E, Helal S, Cook D. Human activity recognition and pattern discovery. IEEE Pervasive Computing, 2010, 9(1): 48-53
[27] Liang F, Liu C, Carroll R J. Bayesian inference and Markov chain Monte Carlo//Advanced Markov Chain Monte Carlo Methods: Learning from Past Samples. Imprint: Chichester, West Sussex, UK: John Wiley & Sons, Ltd, 2010: 1-26
[28] Zhao K, Tarkoma S, Liu S, et al. Urban human mobility data mining: An overview//Proceedings of the IEEE International Conference on Big Data. Washington, USA, 2016: 1911-1920
[29] Rabiner L R, Juang B H. An introduction to hidden Markov models. ASSP Magazine, 1986, 3(1): 4-16
[30] Teng Geer, He Chang-Zheng, Jiang Xiao-Yi. Research advancement of hidden Markov model and its application in management. Soft Science, 2012, 26(2): 122-126(in Chinese) (腾格尔, 贺昌政, 蒋晓毅. 隐马尔可夫模型研究进展及其管理领域应用. 软科学, 2012, 26(2): 122-126)
[31] Federgruen A, Tzur M. A simple forward algorithm to solve general dynamic lot sizing models with n periods in $O(n\log n)$ or $O(n)$ time. Management Science, 1991, 37(8): 909-925
[32] Baggenstoss P M. A modified Baum-Welch algorithm for hidden Markov models with multiple observation spaces. IEEE Transactions on Speech & Audio Processing, 2000, 9(4): 411-416

LI Jie, born in 1982, Ph.D., associate professor. Her research interests include opportunistic communication, mobile big data analysis, social computing, etc.

BAI Zhi-Hong, born in 1992, M.S. His research interests include big data analytics, privacy protection methods, etc.

YU Rui-Yun, born in 1974, Ph.D., professor, Ph.D. supervisor. His research interests include participatory sensing systems, pervasive and mobile computing, big data analysis, etc.

CUI Ya-Meng, born in 1991, engineer. His research interests focus on big data analytics.

WANG Xing-Wei, born in 1968, Ph.D., professor, Ph.D. supervisor. His research interests include future Internet, cloud computing, network security and information security.

Background

Location privacy in Location Based Services (LBS) is the ability to protect the association between user's identity, query sources, servers and database, thereby preventing an imminent attacker from easily linking users of LBS to certain locations. Development of mobile communication and sensing technologies forms location based big data, which brings new challenges for location privacy. Adversaries can speculate users' sensitive locations according to their history location information. So, Mobile users are facing the danger of privacy information leakage under the environment of big data.

In this paper, a location privacy protection algorithm named Location Privacy Protection Algorithm Based on K-anonymity and PSO in Mobile Networks (MaskK) is proposed. We utilize Markov model to describe the location transition of mobile users, and get users' transition probability matrixes. Next, we define a δ-privacy to limit the amount of information that Adversaries can obtain from users' history data. Thereby, we can utilize PSO algorithm to compute the vectors of suppression probabilities ($\boldsymbol{p}$) based on HMM and above definition, and release the transition states of users according to $\boldsymbol{p}$. In addition, K-anonymity is used in the released states that improve the level of privacy. The MaskK algorithm not only protects user's location privacy but also improve the quality of service and the efficiency of running.

This work is supported by the National Natural Science Foundation of China under Grant Nos. 61572123, 61502092 and 61672148, the National Science Foundation for Distinguished Young Scholars of China under Grant No. 71325002, the Fundamental Research Fund for the Central Universities under Grant No. N151604001, the General Financial Grant from the China Postdoctoral Science Foundation under Grant No. 2016M591449, and the Ministry of Education-China Mobile Research Fund under Grant No. MCM20160201.

One of main tasks of the abovementioned projects is to protect mobile users' location privacy under the environment of big data. Some published papers were focusing on users' location disturbance or conceal users' information, etc.

This paper addresses the problem of location privacy protection based on big data, and proposes the location privacy protection algorithm to limit adversaries speculate the users' sensitive locations through users' history data. The algorithm is used for restricting the amount of information that users' history data leak. It is beneficial to protect users' location privacy in the era of big data.