·安全技术· 文章编号：1000-3428(2018)08-0142-09 文献标志码：A 中图分类号：TP399

基于用户真实轨迹的虚假轨迹生成方法

林邓伟[1]，王云峰[2]

（1. 焦作大学 信息工程学院，河南 焦作 454000；2. 北京邮电大学 网络空间安全学院，北京 100876）

摘 要： 现有的轨迹隐私保护方法在对用户进行 K-匿名保护时，较难防御拥有背景信息的攻击。为此，提出一种利用用户的真实轨迹构建虚假轨迹的方法。采用真实轨迹构建（$K-1$）条虚假轨迹实现 K 匿名，解决敌对者通过随机性识别出虚假轨迹的问题，将敌对者的背景信息融入用户运动轨迹的马尔科夫模型，防止敌对者通过背景信息识别出虚假轨迹。实验结果表明，与轨迹替换、轨迹旋转、随机行走等方法相比，该方法具有更高的虚假轨迹生成效率和较好的轨迹隐私保护效果。

关键词： 轨迹隐私保护；虚假轨迹；真实轨迹；背景信息；K-匿名

中文引用格式： 林邓伟，王云峰. 基于用户真实轨迹的虚假轨迹生成方法[J]. 计算机工程，2018，44(8)：142-150.

英文引用格式： LIN Dengwei，WANG Yunfeng. False trajectory generating method based on user's true trajectory[J]. Computer Engineering，2018，44(8)：142-150.

False Trajectory Generating Method Based on User's True Trajectory

LIN Dengwei[1]，WANG Yunfeng[2]

（1. College of Information Engineering，Jiaozuo University，Jiaozuo，Henan 454000，China；
2. College of Cyber Space Security，Beijing University of Posts and Telecommunications，Beijing 100876，China）

【Abstract】 Existing methods for trajectory privacy-preserve can't protect K-anonymous protection from attacks with background knowledge. To this end，a method generating dummy trajectories with user's true trajectories is proposed. By generating（$K-1$）dummy trajectories to achieve K-anonymous with true trajectories，adversaries can't identify them through randomness. By modeling the Markov model of user's trajectories with considering background information，dummy trajectories can't be broken because of background information compared with methods of trajectory replacement，trajectory rotation and random walk，the experimental results show its efficiency of dummy trajectories generation and effectiveness of privacy-preserving.

【Key words】 trajectory privacy-protection；false trajectory；true trajectory；background information；K-anonymous

DOI： 10.19678/j.issn.1000-3428.0049930

0 概述

近年来，基于位置的服务（Location-based Service，LBS）开始在交通运输、社交网络等多个领域得到应用。以 LBS 应用较为广泛的车载导航和手机地图行业为例，我国 2015 年第 4 季度的前装车载导航市场出货量达 125.1 万台，2016 年的中国手机地图覆盖用户规模则达到了 4.57 亿，预计到 2018 年将达到6.42 亿。用户可以利用各种 LBS 应用获取查询餐饮、地图导航等多样化的服务。

然而，用户在获取多样化的服务的同时，却面临着位置隐私泄露的风险。当前，这一问题正随着位置服务应用的日益广泛而成为一个普遍存在的安全隐患，如 2012 年的苹果公司擅自收集用户位置信息事件、2016 年上海广升的预装软件事件等。调查显示，约有 9% 的免费 Android 应用会向第三方广告商透漏用户的手机号，另有 15% 应用则会将设备 ID 信息泄露出去[1]。文献[2]的分析结果显示，一些位置隐私的泄露会危及用户的财产甚至于生命安全。因此，对位置隐私的保护刻不容缓。

目前，位置隐私保护方法致力于避免攻击者或不可信第三方获取用户身份信息和位置信息的关联性等研究。K-匿名技术[3-4]是应用较为广泛的技术之一。针对单个查询的 K-匿名技术通过将用户身份和位置信息的对应关系由 1∶1 转化为 K∶1，达到隐匿用户位置信息的目的。该技术通过假设的可信第三方匿名服务器完成位置信息匿名。该方法虽然能保护单个查询隐私，但无法保护用户的轨迹隐私，也无法

基金项目： 教育部博士点基金（20124116120004）。

作者简介： 林邓伟（1972—），男，副教授、硕士，主研方向为网络安全、虚拟现实；王云峰，博士研究生。

收稿日期： 2018-01-02 **修回日期：** 2018-02-02 **E-mail：** jzlindengwei@163.com

防御具有背景信息的位置攻击,如位置链接攻击[5]、身份匹配攻击[6]等。为此,研究人员提出了轨迹隐私保护方法[7-8]。这类方法主要通过生成($K-1$)条虚假轨迹保护用户真实轨迹不被识别。相比于前一种技术,这类方法考虑了位置间的关联性,能够防御针对单个位置的攻击。但是,虚假轨迹上的位置总是随机生成,具有随机性,往往存在一些与现实不符合的位置[9]。这导致攻击者很容易结合背景信息排除掉虚假轨迹,影响轨迹隐私保护效果。

本文针对轨迹隐私保护方法中攻击者容易通过背景信息、虚假轨迹的随机性获取轨迹隐私的问题,考虑背景信息、轨迹相似性,提出一种基于用户真实轨迹的虚假轨迹生成方法。以用户的真实轨迹为基础,通过分析用户的行为模式,建立由真实轨迹构建的虚假轨迹。虚假轨迹和用户的真实轨迹有相同的运动模式,能够保证与用户真实轨迹的相似性,同时也能够应对背景信息攻击,以实现对用户轨迹的 K-匿名保护。

1　相关研究

针对轨迹隐私保护问题,当前研究主要集中在基于可信第三方服务器构建 K-匿名空间、合成虚假轨迹、轨迹隐私度量 3 个方面。

构建 K-匿名空间方法的主要思路是可信第三方服务器根据需要保护的用户轨迹构造一个至少包括 K 个用户的区域作为匿名区域,并把匿名区域发送给 LBS 服务器实现轨迹隐私保护。

针对 K-匿名空间构建过程中虚假位置的选择问题,文献[10]提出了一种 DLS 算法。算法通过考虑可能被利用的信息,选择熵作为度量来选择虚假位置,达到构建匿名区域的目的,且能够扩大隐藏区域,从而高效地实现 K-匿名。DLS 算法在构建匿名区域时需要按照时间构建虚假位置,容易遭受计时攻击。文献[11]提出 R-匿名时间模糊算法。算法引入时间混淆技术来打破用户查询顺序,使得攻击者无法预测用户轨迹。但是未考虑攻击者掌握的背景信息,攻击者能够利用背景信息区分匿名区域的不合理区域,预测用户的真实轨迹。文献[12]提出了基于位置语义的 SALS 匿名区域构建方法。SALS 方法通过对不同位置语义赋予不同的权值,使用户根据位置语义的权值计算 MAXDEN 来确定自己的匿名区域。这使得用户能够以对等(P2P)方式相互配合生成隐藏区域,而不是实际位置。

文献[13]提出了一种基于移动终端的虚假轨迹生成方法。该方法依据用户真实轨迹中位置的先后顺序,依次选择与对应位置呈一定角度的方向上的一个点生成虚假位置,直至生成虚假轨迹。方法生成的虚假轨迹避免了与真实轨迹的重合,起到了保护轨迹隐私的效果。但是无法防御拥有背景信息的位置隐私攻击。文献[14]提出了一种基于车辆移动轨迹的虚假轨迹生成方法。该方法在生成虚假轨迹时,综合考虑了车辆移动的轨迹运动模式,因而能够生成与真实轨迹相似的虚假轨迹,能够在一定程度上防御拥有背景信息的位置隐私攻击。但是所生成的轨迹会存在一些现实中无法到达的位置,影响轨迹隐私保护效果。同时,这些方法虽然能够保护轨迹隐私,但是由于虚假轨迹上均为虚假位置,容易被攻击者使用特定的攻击手段识别。为此,文献[15-16]提出基于真实用户轨迹的虚假轨迹生成方法。文献[15]提出使用用户的历史轨迹构建虚假轨迹,文献[16]提出使用真实存在的用户轨迹构建虚假轨迹的方法。这些方法在构建虚假轨迹时,只是基于真实用户轨迹来构建,但是在特定区域中寻找相似度较高的用户轨迹存在较高难度,难以达到 K-匿名。

在轨迹隐私保护效果的衡量方面,使用较为普遍的度量标准是 K-匿名度量。现有的轨迹隐私的 K-匿名度量,主要是选取($K-1$)条虚拟轨迹和用户真实轨迹一起由可信的第三方服务器发送到位置服务器,从而提高用户真实轨迹预测的不确定性。但是,现有的基于 K-匿名度量的度量方法往往忽略了背景信息的影响或者实现难度较高。因而轨迹隐私的 K-匿名度量需要考虑背景信息和实现的可行性。

针对上述问题,本文提出一种基于用户真实轨迹的虚假轨迹生成方法。方法通过分析用户行为模式,通过聚类选取与用户真实轨迹具有相同行为模式的其他用户轨迹作为虚拟轨迹,并通过 EMD 距离计算所生成的虚拟轨迹中与用户真实轨迹具有最大相似性的($K-1$)条轨迹,实现用户轨迹的 K-匿名保护。相比于上述文献,本文所提方法有以下优点:

1)生成的虚假轨迹基于真实用户位置,能够防止出现不符合现实的虚假轨迹。

2)生成的虚假轨迹与真实轨迹具有相同的行为模式,能够较为可靠地实现轨迹的 K-匿名。

3)本文方法构建的马尔科夫模型融入了背景信息,能够防备具有背景信息的攻击。

2　基于虚假轨迹的隐私保护问题分析

如前所述,现有基于虚假轨迹的隐私保护方法存在的一些问题影响了轨迹隐私保护的效果。本节则对这些问题进一步剖析,力图找出影响轨迹隐私保护的内部机制,并给出解决这些问题的基本思路。

2.1　背景信息

现有面向轨迹隐私的保护方法主要以位置服务中的用户身份信息和位置信息的关联性为研究对象。然而,在现实中,用户使用其他服务时仍然会泄露身份信息。攻击者也可以通过一些公共信息,如地图上的实际路径、道路限制等推断用户身份信息和位置信息间的关联性。类似这些和用户轨迹信息无直接关联,却有助于攻击者获取用户隐私的信息就是背景信息。在大数据时代,攻击者能够通过各种途径轻易得

到各种背景信息,并结合轨迹信息对用户的身份等敏感信息进行推测。图 1 概率分布攻击显示了攻击者利用背景信息推测用户真实位置的情况。

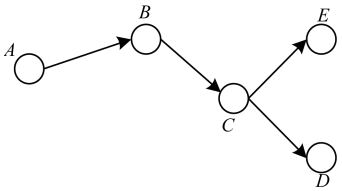


图 1 概率分布攻击

概率分布攻击指攻击者根据用户在不同位置的分布情况推测用户的位置隐私。在图 1 中,A、B、C、D 为用户的真实轨迹。用户对位置 D 进行了 $K=2$ 的匿名保护,E 为生成的虚假位置。假设攻击者能够根据用户的历史轨迹推测其有 80% 的概率会从 C 位置到达 D 位置,则攻击者获知用户在 C 位置后,就能根据历史轨迹这一背景信息推测出用户的真实位置,使 D 位置匿名失败。

攻击者利用历史轨迹的目的是根据用户在各个位置的服务请求信息得到用户的行为模式,并最终得到各个位置间的关联性。因此,本文将图 1 所示情况下的攻击者的背景信息表达为用户在某个位置发送服务请求的概率。同时,参考文献[17],将每个位置的服务请求概率表示为网格地图中每个网格的服务请求概率。

2.2 虚假位置的约束条件

在现有的虚假轨迹生成方法中,存在的问题是采用随机方法生成的虚假轨迹中经常会存在一些不满足约束条件的虚假位置,最终导致用户的真实轨迹以较高概率被识别。

图 2 是采用随机方法生成虚假轨迹的示例。其中 T 为由 A、B、C、D 4 个位置组成的真实轨迹,T'为由 A'、B'、C'、D' 4 个位置组成的虚假轨迹。假设用户以允许的最大速度在此区域内唯一的一条高速公路上行驶,则满足的约束条件为:1)高速公路的唯一性;2)用户的最大行驶速度。对比轨迹 T 和 T'上同时刻的位置,可以发现在 $A'\rightarrow B'$和 $C'\rightarrow D'$这两段距离,生成的虚拟轨迹不满足第 2 个约束条件,使得 T'以极大概率被排除,影响隐私保护效果。

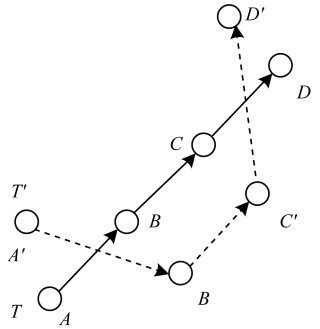


图 2 虚假位置约束条件示例

本文选取与用户轨迹相似的其他用户的真实轨迹作为虚假轨迹来解决该问题。获取区域内满足条件的虚假轨迹后,通过 EMD 距离公式选取出与真实轨迹具有最高相似度的 K 条虚假轨迹,实现轨迹 K-匿名。

2.3 用户行为模式

目前,利用用户的真实轨迹构建虚假轨迹是一种有效的轨迹隐私保护方法。这种选取与用户轨迹相似的其他用户的真实轨迹作为虚假轨迹。由于组成这些虚假轨迹的位置全部为真实位置,因此能够避免不合理位置的存在。然而,考虑到一条轨迹上的位置足够多时,现实中任意 2 个人的轨迹难以完全相似,容易导致匿名失败。

针对这一问题,本文采用基于用户行为模式的方法构建虚假轨迹。假设被保护用户 a 往返于家庭 Ha 和工作 Wa 2 个位置,同时在特定区域内用户 b 也往返于家庭 Hb 和工作 Wb 2 个位置,则 a 和 b 有相同的行为模式 $<H,W>$。此时虽然 $Ha\neq Hb$,$Wa\neq Wb$,但是仍然可以用 Hb 和 Wb 代替 a 对应的位置。考虑到具有相同行为模式的两条轨迹不一定具有数目相等的位置点,则基于被保护轨迹的行为模式构建虚假轨迹。

3 虚假轨迹生成方法

本节根据第 2 节的问题,拟从攻击者角度出发,给出虚假轨迹的生成方法。

3.1 用户运动轨迹模型

对用户的运动轨迹建模是合成虚假轨迹的基础。本文拟从攻击者角度出发,分析用户轨迹被重构的可能性,从而从整体方面为用户轨迹隐私保护提供保证。

定义 1 用户 U 在统计时长 t 内的轨迹记录为 $T(U)=\{(x_0,y_0),(x_1,y_1),\cdots,(x_m,y_m)\}$。其中,点 (x_i,y_i)为 U 在 t_i 时刻的位置,表示为 $U^{t_i}_{(x_i,y_i)}$,m 表示轨迹 $T(U)$上位置的个数。

定义 2 轨迹 $T(U)$满足马尔科夫假设。

证明:根据定义 1,用户 U 的位置变量 $U^{t_i}_{(x_i,y_i)}$是一个随时间变化的随机变量。假设(x_i,y_i)前后时刻的位置分别为(x_{i-1},y_{i-1})和(x_{i+1},y_{i+1}),且用户 U 在 t_i 时刻的位置为(x_i,y_i),则无论用户在 t_{i-1}时刻位于(x_{i-1},y_{i-1})的可能性如何,都不会影响用户 U 在 t_{i+1}时刻位于(x_{i+1},y_{i+1})的可能性,即用户在 t_{i+1}时刻的位置只依赖于其在 t_i 时刻的位置。同样地,轨迹 $T(U)$中任一时刻的位置只依赖于其前一刻位置。从而,用户运动轨迹模型满足马尔科夫假设。

通常,位置语义对应的是由多个位置点组成的

区域。如图 3 所示,轨迹 T 上存在 6 个位置点、2 个位置语义。其中由 A、B、C 3 个位置组成的区域表示超市 M,由 D、E、F 3 个位置组成的区域表示医院 N。因此,为了简化用户轨迹模型,在定义 1 的基础上给出用户运动轨迹的定义。

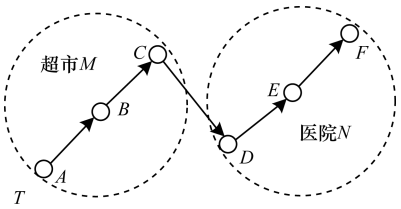


图 3　位置语义和区域

定义 3　用户 U 在统计时长 τ 内的轨迹记录为 $T(U)=\{r_0,r_1,\cdots,r_n\}$。其中,组成轨迹的 r_j 为 U 在 τ_j 时间段内所在的位置区域。

根据定义 1 ~ 定义 3 所定义的 $T(U)$ 满足马尔科夫假设,即是用户在下一个区域出现的概率只与其前一个区域及由前一区域向该区域运动的转移概率有关。据此,给出定义 4 的用户运动轨迹的马尔科夫模型。

定义 4　用户 U 在统计时长 τ 内的轨迹记录 $T(U)$ 的位置状态空间集合 $\{r_0,r_1,\cdots,r_n\}$ 则用户运动轨迹的马尔科夫模型定位为 $\vartheta=<\rho(u),\pi(u)>$ 二元组。其中,$\rho(u)$ 为用户位置运动的转移概率集合,$\pi(u)$ 为用户位置的联合概率集合,也即:

$$\rho_{r_i}^{r_j}(U)=P(U^{t+1}=r_j\mid U^t=r_i;\Gamma^t=\tau_i,\Gamma^{t+1}=\tau_j) \tag{1}$$

$$\pi_{r_i}^{\tau_i}(U)=P(U^t=r_i,\Gamma^t=\tau_i) \tag{2}$$

其中,$\pi_{r_i}^{\tau_i}(U)$ 为用户 U 在 τ_j 时间段内位于区域 r_i 的概率,$\rho_{r_i}^{r_j}(U)$ 为用户 U 由区域 r_i 移动到 r_j 的条件转移概率,U^t、Γ^t 分别表示 2 个相邻区域中前一个区域的区域变量和时间变量,则 U^{t+1}、Γ^{t+1} 分别为后一个区域的区域变量和时间。

3.2　基于行为模式的虚假轨迹构建

基于真实用户轨迹构建虚假轨迹,难以找到与匿名轨迹完全匹配的其他用户轨迹,导致匿名失败。然而,当用户轨迹数据发布时,通常发布的是关键的位置,而非全部轨迹数据。因而,可以通过用户行为模式构建虚假轨迹。图 4 显示了这种情况下的虚假轨迹构建方法。图 4(a)采用完全匹配方法为轨迹 1 构建虚假轨迹 2。由于轨迹 2 中 $C2$ 位置与 $C1$ 方向不匹配导致匿名失败。图 4(b)则选择与轨迹 1 具有相同行为模式(Wal 超市-麦当劳)的轨迹 2 作为虚假轨迹,成功进行匿名。因此,选择具有同一行为模式的轨迹作为虚假轨迹能够解决完全匹配匿名方法的不足。

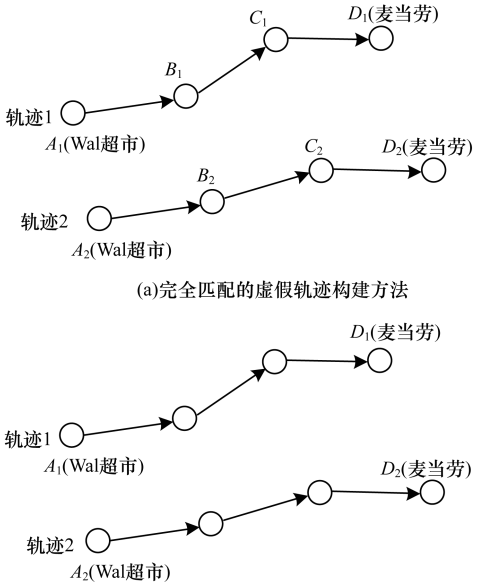


图 4　基于行为模式的虚假轨迹

3.3　用户背景信息

在概率分布攻击中,攻击者推测用户位置隐私的依据是位置分布概率。在图 1 中,不考虑背景信息情况下,用户由 C 到 D 位置的概率,即 $P(D\mid C)=0.5$。假设存在背景信息“80% 的用户选择去往 D 位置”,则此时 $P(D\mid C)=0.8$,攻击者将有 80% 概率推测出用户会从 C 位置到达 D 位置,极大地增加位置泄露的概率。因此,用户的历史轨迹这一背景信息中蕴含着用户的位置分布情况。从而,可以用历史轨迹作为背景信息。

定义 5　假设存在 h 条历史轨迹,且分布的位置数目分别为 $\lambda_1,\lambda_2,\cdots,\lambda_h$。这些位置分布于一个划分为 $\eta\times\zeta$ 个网格的区域中,每个网格单元分布的位置数目为 $a_{c,d}$。其中,c、d 分别表示沿 η、ζ 方向的列索引、行索引。则每个网格单元的位置分布概率为:

$$P_{c,d}=\frac{a_{c,d}}{\sum_{l=1}^{h}\lambda_l} \tag{3}$$

假设轨迹 $T(U)$ 中的 r_j 区域包含 e 个位置,且这些位置分布于网格单元 (c,d) 中的数目为 $e_{c,d}$。则 r_j 区域的位置分布概率,即用户由区域 r_i 移动到 r_j 的条件转移概率为:

$$\rho_{r_i}^{r_j}(U)=\max_{(c,d)\in(\eta,\zeta)}\left\{\frac{e_{c,d}}{a_{c,d}}\times P_{c,d}\mid e_{1,1}+e_{1,2}+\cdots+e_{\eta,\zeta}=e\right\} \tag{4}$$

3.4　轨迹相似性

对于不同的虚假轨迹,其区分难度是不同的。为了衡量这种难度,本文采用相似性度量函数 EMD(Earth Mover’s Distance)度量所合成的虚假轨迹和真实轨迹的相似程度。

对于任意 2 个分布 y、z,$EMD(y,z)$ 表示分布 y

转化为分布 z 的最小代价。y 和 z 相似程度越高，$EMD(y,z)$ 越小，因而可以度量 2 个分布间的相似性。

定义 6 设 Y 和 Z 分别为定义在状态空间 $\Omega_Y=\{y_\omega|\omega=0,1,\cdots,n_\omega\}$ 和 $\Omega_Z=\{z_\theta|\theta=0,1,\cdots,n_\theta\}$ 的离散型随机变量。P_Y、P_Z 分别是 Y 和 Z 位于 Ω_Y、Ω_Z 上的概率分布，即 $P_Y(Y=y_\omega)=\rho_{y_\omega}$，即 $P_Z(Z=z_\theta)=\rho_{z_\theta}$，$f$ 为变量 Y 和 Z 的联合概率分布，ρ 为边缘概率分布。则分布 P_Y 和 P_Z 的 EMD 距离定义为：

$$EMD(P_Y,P_Z)=\min\{\sum_{\omega=1}^{n_\omega}\sum_{\theta=1}^{n_\theta}f_{\omega\theta}\cdot d(y_\omega,z_\theta)|(Y,Z)\sim f,y_\omega^*\sim P_Y,z_\theta\sim P_Z\} \tag{5}$$

其中，$f_{\omega\theta}$ 为 $Y=y_\omega$ 和 $Z=z_\theta$ 的联合概率分布，$d(y_\omega,z_\theta)$ 为 $Y=y_\omega$ 和 $Z=z_\theta$ 间的距离，且满足以下约束条件：

$$\sum_{\omega=1}^{n_\omega}\sum_{\theta=1}^{n_\theta}f_{\omega\theta}=\min\{\sum_{\omega=1}^{n_\omega}\rho_{y\omega},\sum_{\theta=1}^{n_\theta}\rho_{z\theta}\} \tag{6}$$

$$\sum_{\omega=1}^{n_\omega}f_{y_\omega}\leqslant\rho_{z_\theta},0\leqslant\omega\leqslant n_\omega \tag{7}$$

$$\sum_{\theta=1}f_{\omega\theta}\leqslant\rho_{y_\omega},0\leqslant\theta\leqslant n_\theta \tag{8}$$

$$f_{\omega\theta}\geqslant0,0\leqslant\omega\leqslant n_\omega,0\leqslant\theta\leqslant n_\theta \tag{9}$$

假设轨迹 $T(U)$ 和 $T(V)$ 在 $\Gamma^t=\tau_i$ 和 $\Gamma^{t+1}=\tau_j$ 2 个时间段内位置区域分别为 $U^t=r_i$、$U^{t+1}=r_j$ 和 $V^t=r_\vartheta$、$V^{t+1}=r_\chi$。若用 r_χ 作为 r_j 的虚假位置区域，则 $EMD(\pi_{r_\chi}^{\tau_j}(U),\pi_{r_j}^{\tau_j}(V))$ 将随着 r_χ 和 r_j 相似度的增大而降低，因而能够用 $EMD(\pi(u),\pi(v))$ 距离度量 $T(U)$ 和 $T(V)$ 轨迹的相似性。考虑到 EMD 距离和变量相似性的关系，定义轨迹 $T(U)$ 和 $T(V)$ 相似度。

定义 7 假设用户 U、V 在统计时长 τ 内的轨迹记录分别为 $T(U)=\{r_0,r_1,\cdots,r_{n_U}\}$、$T(V)=\{r_0,r_1,\cdots,r_{n_V}\}$，且对应的位置概率分布为 $\pi(u)$、$\pi(v)$ 则两者的 EMD 距离为：

$$EMD(\pi(u),\pi(v))=\min\{\sum_{T(U)}\sum_{T(V)}[\pi_{r_\chi}^{\tau_j}(U)\cdot\pi_{r_j}^{\tau_j}(V)]\cdot d(r_\chi,r_j)\} \tag{10}$$

其中，$d(r_\chi,r_j)$ 为位置区间 r_χ 和 r_j 在定义 5 所定义的 $\eta\times\zeta$ 网格上的欧氏距离。因此，$T(U)$ 和 $T(V)$ 轨迹的相似度 $sim(T(V),T(U))$ 为：

$$sim(T(V),T(U))=1-\frac{EMD(\pi(u),\pi(v))}{Z(\pi(u),\pi(v))} \tag{11}$$

其中，$Z(\pi(u),\pi(v))=\max\{\sum_{T(U)}\sum_{T(V)}[\pi_\chi^{\tau_j}(U)\cdot\pi_{r_j}^{\tau_j}(V)]\cdot d(r\chi,r_j)\}$，保证了 2 条轨迹的相似度 $sim\in[0,1]$。

3.5 轨迹泄露概率

轨迹泄露概率是通过计算虚假轨迹和真实轨迹相似程度来衡量。

假设用户 U 的轨迹 $T(U)$ 及对应虚假轨迹集合为 $R=\{T(U(fake_\sigma))|\sigma=0,1,\cdots,n_\sigma\}$，$T(U(fake_\sigma))$ 表示 $T(U)$ 的第 σ 条虚假轨迹。为了实现轨迹 K-匿名，必须从 R 中选择 $(K-1)$ 条与 $T(U)$ 具有足够相似度的虚假轨迹，最优的情况是所选轨迹与 $T(U)$ 的相似度全部为 0，否则 U 的轨迹隐私将泄露。满足要求的轨迹相似度 $sim<\Delta$，Δ 为轨迹相似度阈值，是用户自定义的常数。

另外，有效虚假轨迹越多，真实轨迹的 K-匿名效果越好，但是匿名成功率越困难，且真实轨迹上的位置泄露情况也将越发严重。如果所有虚假轨迹所有位置中包含全部真实轨迹上所有的位置，那么真实轨迹将面临着全面泄露的风险。

基于上述两方面因素，轨迹隐私泄露概率表示为：

$$L(U)=\frac{\sum_{\sigma=1}^{K-1}sim(T(U),T(U(fake_\sigma)))}{K} \tag{12}$$

通常，轨迹隐私泄露概率 $L(U)$ 越大，用户隐私保护效果越差。

3.6 轨迹的 K-匿名

3.6.1 系统架构

考虑到轨迹的 K-匿名对设备的计算能力、存储能力、实时性等有较高的要求，本文提出的轨迹匿名方法采用集中式 3 层系统架构，如图 5 所示。架构在可信的第三方匿名服务器上完成轨迹的匿名，具有较高的计算能力、存储能力、实时性等；同时由于用户的轨迹隐私数据全部保存在匿名服务器而非 LBS 服务器，具有较高的安全性。

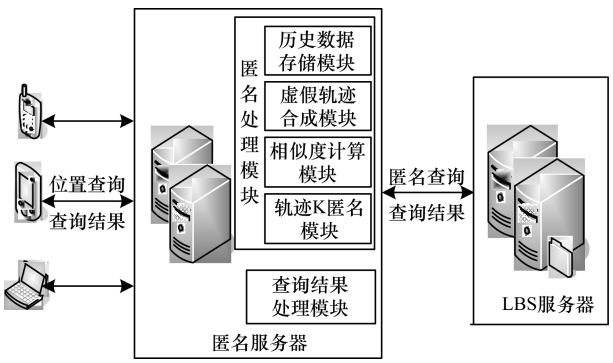


图 5 本文系统架构

实施轨迹匿名时的主要工作流程如下：

1）匿名服务器接收到用户的位置查询请求后，通过历史数据存储模块，实时地建立用户轨迹的马尔科夫模型，并存储用户数据和位置请求。

2）依据建立的马尔科夫模型，通过虚假轨迹合成模块中的算法合成符合要求的虚假轨迹。

3）依据相似度计算模块中的算法分别计算虚假轨迹和真实轨迹的相似度，便于对轨迹进行匿名。

4）依据相似度选取符合要求的虚假轨迹完成

K-匿名,并将虚假轨迹和真实轨迹发送给 LBS 服务器。

5)LBS 服务器接收到匿名服务器发送的匿名位置请求后,将查询结果发送给匿名服务器。

6)匿名服务器接收到 LBS 服务器的查询结果后,依据所存储的用户信息和查询请求,将查询结果返回给对应的用户。

3.6.2　轨迹 K-匿名算法

本文给出了轨迹 K-匿名算法。算法能够根据接收到的用户位置查询请求合成虚假轨迹,并返回符合要求的($K-1$)条虚假轨迹,完成轨迹的 K-匿名。结合图 5 所示的系统架构,该算法由 3 个算法组成。

算法 1 用来生成与匿名轨迹同类别的其他轨迹作为虚假轨迹。假设 t_i、$U^{t_i}_{(x_i,y_i)}$、E、$V^{t_i}_{(x_i,y_i)}(m)$,分别表示当前时刻及当前时刻用户 U 的位置、发送位置查询的其他用户集合及 E 中第 m 个用户 V_m 的位置,$C(U)$ 表示包含轨迹 $T(U)$ 及与其同类别轨迹的轨迹集合,D_i、$D_i^{V_m}$、Γ_i、$T(U)$、Γ'为临时列表,分别表示 t_i 时刻定义 1 所示的用户 U 的轨迹、用户 V_m 的轨迹、集合 E 中所有用户轨迹集合、定义 3所示用户 U 的轨迹、集合 E 中所有用户的轨迹集合,$LIST(D_i,l)$、$SET(\Gamma_i,l)$分别为 D_i 和 Γ_i 集合进行的操作。其详细描述如下所示。

算法 1　虚假轨迹生成算法

输入　t_i、$U^{t_i}_{(x_i,y_i)}$、E、$V^{t_i}_{(x_i,y_i)}(m)$

输出　$C(U)$

1. $D_i \leftarrow \varnothing, D_i^{V_m} \leftarrow \varnothing, \Gamma_i \leftarrow \varnothing, T(U) \leftarrow \varnothing, \Gamma' \leftarrow \varnothing$
2. for(初始时刻 t0 到 t_i 时刻的每一时刻 t)
3. if $t_0 = t_i$
4. $D_0 \leftarrow U^{t_0}_{(x_0,y_0)}$
5. else
6. $D_i \leftarrow D_{i-1} \cup \{U^{t_i}_{(x_i,y_i)}\}$
7. end if
8. for (时刻 t_i 时集合 E 中每一用户 V_m)
9. if $t_0 = t_i$
10. $D_0^V \leftarrow \{V^{t_0}_{(x_0,y_0)}\}$
11. $\Gamma_0 \leftarrow$初始时刻所有用户轨迹
12. else
13. $D_i^V \leftarrow D_{i-1}^V \cup \{V^{t_i}_{(x_i,y_i)}(m)\}$
14. $\Gamma_i \leftarrow \Gamma_{i-1} \cup \{D_i^{V_m}\}$
15. end for
16. end for
17. $T(U) \leftarrow LIST(D_i,1), \Gamma' \leftarrow SET(\Gamma_i,1)$
18. $T' \leftarrow \Gamma' \cup T(U)$
19. C←K-means(T')//调用聚类算法分类 T'
20. return C(U)//输出含有用户 U 轨迹的类

算法 1 通过接收从初始时刻 t_0 到当前时刻 t_i 时间段内待匿名用户 U 和其他位置查询用户的位置,生成对应的轨迹 D_i 和轨迹集合 V(1 行 ~18 行),并依据 K-means 算法找出与用户 U 同类别轨迹作为 $T(U)$的虚假轨迹(19 行)。

算法 1 生成的虚假轨迹具有不同的识别难度,需要计算轨迹的相似度。因此,算法 2 基于算法 1 输出结果 $C(U)$计算虚假轨迹和真实轨迹的相似度。假设 $C(U)$、$T(U)$、$T(V)$、h,分别表示算法 1 的输出轨迹集合、$C(U)$中用户 U 轨迹、$C(U)$中的虚假轨迹和历史轨迹,Sim 表示 $T(U)$及与 U 同类别的虚假轨迹的相似度集合,map、G、H、C'分别表示轨迹所在地经纬度范围、网格元素列表、历史轨迹列表、位置联合概率列表。其详细描述如下所示。

算法 2　轨迹相似度度量算法

输入　$C(U)$、$T(V)$、h

输出　Sim

1. map←Ø,G←Ø,H←Ø,C'←Ø,Sim←Ø
2. map←{(LAmin,LAmax),(LOmin,LOmax)}//输入//轨迹所在地的经纬度范围
3. G←GridZoom(m,n,1,map)//将经纬度范围依照比//例 l缩放,并创建 m × n 网格
4. for(G中每一个表格g)//计算每个表格中分布的历史//位置数目
5. for(H 中每一个历史位置 h)
6. if h 位于表格 g
7. Count(g) = Count(g) + 1
8. P(g) = Count(g)/len(H)
9. end for
10. end for
11. for(C(U)中每一个虚假轨迹T(V))//计算每个//T(V)的转移概率和位置联合概率
12. ρ(V)←Ø,π(V)←Ø//ρ(V)为T(V)转移概率集合,//π(V)为 T(V)位置联合概率集合
13. for(T(V)中每一个位置区域 r)
14. for(r 中每一个位置(x,y))
15. R←Ø//R 是区域 r_i 移动到 r_j 的条件转移概率集合
16. for(G 中每一个表格 g)
17. if(x,y)位于表格 g
18. Count(rg) = Count(rg) + 1
19. P(rg) = Count(rg)/Count(g)
20. R←R∪{P(rg)}
21. end for
22. end for
23. end for
24. $\rho^{r_j}_{r_i}(U)$←从区域 r_j 形成的集合 R 选取值最大的元素,计算方法见式(4)
25. $\rho(V) \leftarrow \rho(V) \cup \{\rho^{r_j}_{r_i}(V)\}$
26. if $r_i = r_0$
27. $\pi(V) \leftarrow \{\rho^{r_j}_{r_i}(V)\}$
28. else
29. $\pi(V) \leftarrow \pi(V) \cup \{\rho^{r_j}_{r_i}(V) * \pi(V)\}$
30. end for
31. C'←C'∪{π(V)}
32. Sr←Ø,Sπ←Ø,//Sr为T(U)和T(V)的位置区域//矩阵,Sπ 两者的位置联合概率矩阵
33. for(C(U)中每个虚假轨迹T(V))//计算相似度,计//算方法见式(11)

34. Sr←distmatix(T(U),T(V))//建立 Sr 矩阵
35. Sπ←flowmatix(T(U),T(V))//建立 Sπ 矩阵
36. max(T(U),T(V))←EMD(T(U),T(V))//计//算 T(U)和 T(V)最大 EMD 距离
37. min(T(U),T(V))←EMD(T(U),T(V))//计算//T(U)和 T(V)最小 EMD 距离
38. sim(T(U),T(V))←sim(min(T(U),T(V)),max(T(U),T(V)))
39. Sim←Sim∪{sim(T(U),T(V))}
40. end for
41. return Sim

算法 2 在算法 1 基础上,首先创建网格,并计算网格位置分布概率,然后依据网格位置分布概率计算输入的各个轨迹的条件转移概率和位置联合概率(1 行 ~32 行),并计算虚假轨迹和真实轨迹的相似度(33 行 ~40 行)。

轨迹的 K-匿名需要选出($K-1$)条符合用户需求的虚假轨迹。算法 3 基于算法 2 计算用户的轨迹隐私泄露概率。假设 Sim、Δ、K 分别表示算法 2 输出的轨迹相似度集合、用户自定义相似度常数、匿名等级,$L(U)$表示轨迹隐私泄露概率,R 表示符合用户自定义相似度阈值 Δ 的虚假轨迹集合,$Sim(K)$表示与 R 中轨迹对应的相似度。则其详细描述如下所示。

算法 3 轨迹 K-匿名算法
输入 Sim、Δ
输出 $L(U)$
1. R←Ø,Sim(K)←Ø
2. for(Sim 中每一个相似度 sim(T(U),T(V)))
3. if sim(T(U),T(V)) < Δ
4. R←R∪{T(V)}
5. R←Sim(K)∪{sim(T(U),T(V))}
6. end for
7. if len(R) < k - 1
8. return("本次匿名失败")
9. else
10. for(Sim(K)所有相似度 sim(T(U),T(V)))
11. SUM←SUM + sim(T(U),T(V))
12. end for
13. L(U)←SUM/K
14. return L(U)

为了选出($K-1$)条符合用户需求的虚假轨迹。算法 3 从算法 2 的输出结果中,选出与用户自定义相似度阈值 Δ 相符的轨迹相似度(1 行 ~9 行),然后选出至少($K-1$)条轨迹进行 K-匿名(10 行 ~12 行),并返回成功匿名后的轨迹隐私泄露概率(13 行 ~14 行)。

4 实验结果与分析

4.1 实验参数设置

为验证本文方法的有效性和高效性,实验中用户移动轨迹数据由 Thomas Brinkhoff 生成器模拟产生。模拟数据来自于奥尔登堡地区用户的真实移动轨迹,所记录的用户位置具有持续性和全面性,已被成功用于多项研究工作的验证。因而可用于本文方法的验证工作。本文选取 24 km × 27 km 区域内 2 000 个时间片内的 10 004 条轨迹共计 299 601 个采样点构成实验数据集,并依照 1∶1 000 的比例模拟到 2 400 × 2 700 个单元网格中。

实验环境为 Intel i5 7500 3.4 GHz,4 GB 内存,Windows8 64 bit 操作系统,算法在 Pycharm 环境下基于 Python3.5 语言实现的。

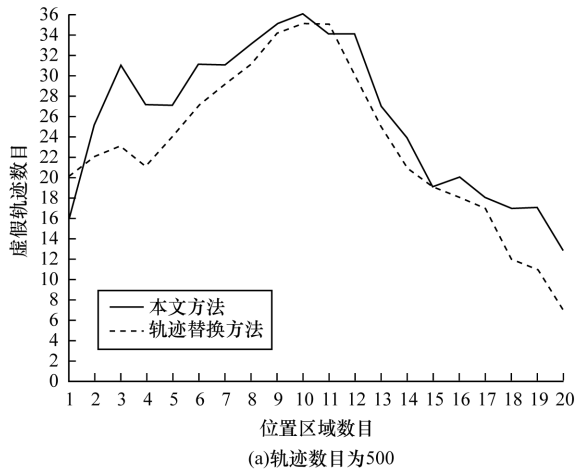
4.2 结果分析

实验主要采用 3.6.2 节的算法,从 3 个方面对本文方法作对比分析:不同方法对生成虚假轨迹数目的影响、背景信息对轨迹隐私泄露情况的影响、不同方法对用户服务质量的影响。

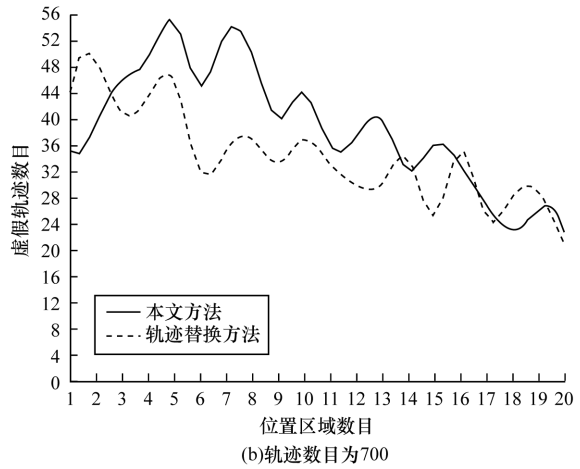
实验涉及到的参数包括:1)K 为匿名等级,通常为 $3 \leqslant K \leqslant 30$;2)$m$ 为位置区域包含的位置数目,$1 \leqslant m$;3)Δ 为虚假轨迹和真实轨迹相似度的阈值,$0 \leqslant \Delta \leqslant 1$。参与比较的方法包括轨迹替换方法[16]、轨迹旋转方法[17]、随机行走方法[18]。为了保证结果准确性,所有实验结果均为运行 500 次的平均值。

4.2.1 虚假轨迹生成数目的对比

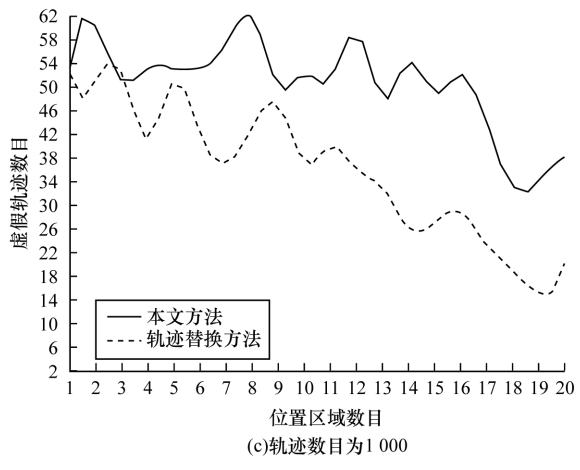
为了验证本文方法生成虚假轨迹的效果,本文设定在匿名等级 $K=3$ 情况下,分别从数据集中选取 500、700、1 000 条轨迹考察生成虚假轨迹数目随着位置区域数目增加呈现出的变化。由于文献[17-18]方法随机生成虚假轨迹,且虚假轨迹数目仅和匿名等级有关,因此本文选取文献[16]的方法作为对比。实验结果如图 6 所示。从图 6 可以看出,在不同的轨迹数目情况下,随着轨迹上位置区域数目的增加,2 种方法所生成的虚假轨迹数目存在3 个共同的变化趋势:1)位置区域数目为 1 时,轨迹替换方法生成的虚假轨迹数目多于本文方法;2)2 种方法生成的虚假轨迹数目均呈现出先增加再逐渐减少的趋势;3)轨迹上位置区域数目相同时,本文方法生成的虚假轨迹数目多于轨迹替换方法。之所以如此,原因在于位置区域数目为 1 时(初始时刻),多数轨迹上均存在唯一的位置区域数目。由于轨迹替换方法选取位置区域数目相同的轨迹作为虚假轨迹,而本文方法则考虑用户的行为模式,因此所生成的虚假轨迹数目小于轨迹替换方法。然而,随着位置区域数目的增加,不同轨迹开始具有不同的位置区域数目,由于本文提出的方法基于用户行为模式生成虚假轨迹,因此保证了对具有多个位置的轨迹匿名时,生成虚假轨迹的效果优于轨迹替换方法。同时,本文方法构建的虚假轨迹均来源于真实的位置,因而能够防止轨迹因位置的随机性而被识别。



(a)轨迹数目为500



(b)轨迹数目为700



(c)轨迹数目为1 000

图 6　不同方法下的虚假轨迹数目曲线

4.2.2　背景信息对轨迹隐私泄露概率的影响

轨迹隐私保护方法的效果体现于轨迹隐私泄露概率。实验分别使用 5×10^4、10×10^4、15×10^4 类采样点作为历史轨迹数据验证背景信息对轨迹隐私泄露概率的影响,实验结果如图 7 所示。从图 7 可以看出,在不同采样点情况下,轨迹隐私泄露概率是随着匿名等级的增大而降低的。且总体上,在同一匿名等级下,轨迹隐私泄露概率随着采样点增加而减小,即隐私保护效果越好。

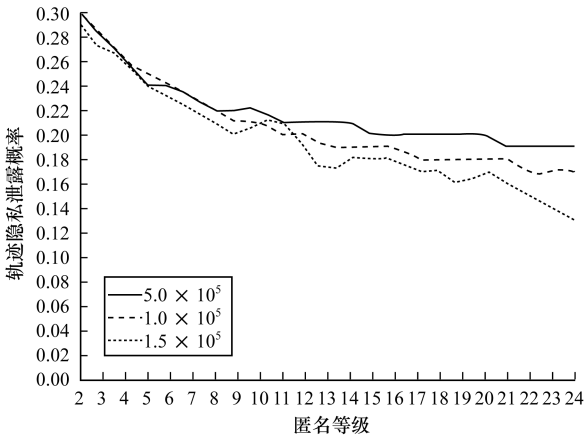


图 7　不同背景信息下的轨迹隐私泄露概率曲线

4.2.3　用户服务质量的对比

为评估不同方法对用户服务质量的影响,本文使用用户隐私保护效果这一度量对其评价。通常,轨迹隐私泄露概率越大,用户隐私保护效果越差。因此,本节基于上述实验,对比不同方法的轨迹隐私泄露概率。

实验选取数据集中的 10 000 条轨迹作为历史轨迹,相似度阈值 Δ 为 0.3。在参与对比的方法中,随机行走和轨迹旋转方法随机生成虚假轨迹,轨迹旋转方法和本文方法具有相同的背景信息,轨迹替换方法和本文方法使用 4.2.1 节中轨迹数目为 1 000 时,生成的虚假轨迹作为运行算法时的虚假轨迹。实验结果如图 8 所示。

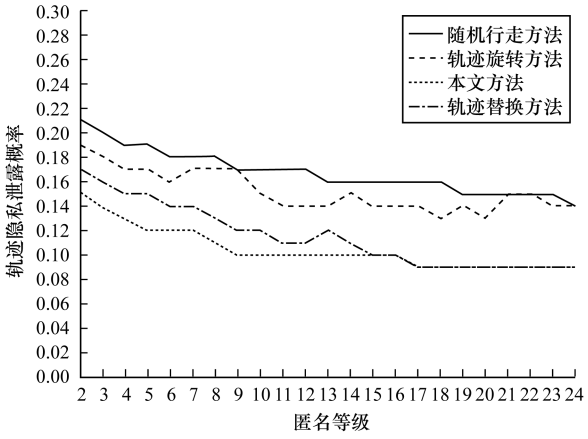


图 8　不同方法下的轨迹隐私泄露概率曲线

在图 8 中,对于同一匿名等级,随机行走方法的轨迹隐私泄露概率最高,轨迹旋转方法次之,本文方法最低。原因在于,随机行走方法采用随机法生成的轨迹中,存在许多不符合实际运动规律的位置点,容易通过背景信息识别,导致隐私保护效果较差;轨迹旋转方法虽然考虑了背景信息,但本质上仍然采用随机法生成轨迹,因而难以兼顾每个位置的背景信息,导致该方法的隐私保护效果强于轨迹旋转方法,但是弱于本文方法。轨迹替换方法隐私保护效

果仅次于本文方法,但是随着匿名等级的增大,容易匿名失败。原因在于,该方法本身采用了真实轨迹构建虚假轨迹,避免了随机性方法的不足,但是采用的虚假轨迹生成方法效率较低,难以生成符合要求的虚假轨迹数目,如图 8 匿名等级大于 15 时,其数值高于虚假轨迹数目,导致匿名失败。因而,该方法的隐私保护效果难以达到较高水平。本文方法由于采用基于用户行为模式的虚假轨迹生成方法,在一定程度上减少了轨迹替换方法容易匿名失败的不足,同时基于真实轨迹和背景信息生成虚假轨迹,使得其难以通过背景信息被识别,因而降低了隐私泄露概率,具有比上述方法更好的隐私保护效果。

5 结束语

本文针对随机性和背景信息导致的虚假轨迹容易被识别的问题,提出了一种基于用户真实轨迹的虚假轨迹生成方法。选择具有相同行为模式的用户轨迹构建虚假轨迹,并设计了用户运动轨迹的马尔科夫模型。由于马尔科夫模型融入了攻击者掌握的背景信息,因此能够用于计算轨迹的相似性,并从构建的虚假轨迹中选择($K-1$)条进行 K-匿名,通过虚假轨迹和真实轨迹的相似程度衡量轨迹隐私泄露水平。实验结果表明,该方法能获得较好的隐私保护效果。

参考文献

[1] 贾金营. 移动计算环境下基于位置服务的位置隐私保护技术研究[D]. 成都:电子科技大学,2015.

[2] 张 浩. 基于位置服务的信息隐私保护技术研究[D]. 合肥:中国科学技术大学,2014.

[3] LEE H, CHANG J W. Density-based k-anonymization scheme for preserving users' privacy in location-based services[C]//Proceedings of International Conference on Grid and Pervasive Computing. Berlin, Germany: Springer-Verlag,2013:536-545.

[4] PALANISAMY B,LIU Ling. Attack-resilient mix-zones over road networks: architecture and algorithms[J]. IEEE Transactions on Mobile Computing,2015,14(3): 495-508.

[5] DING Yuan, PEDDINTI S T, ROSS K W. Stalking Beijing from Timbuktu:a generic measurement approach for exploiting location-based social discovery[C]// Proceedings of the 4th ACM Workshop on Security and Privacy in Smartphones & Mobile Devices. New York, USA:ACM Press,2014:75-80.

[6] HOCHREINER C, HUBER M, MERZDOVNIK G, et al. Towards practical methods to protect the privacy of location information with mobile devices[C]// Proceedings of International Conference on Security of Information and Networks. Washington D. C., USA: IEEE Press,2014:17-24.

[7] MANO K, MINAMI K, MARUYAMA H. Pseudonym exchange for privacy-preserving publishing of trajectory data set[C]//Proceedings of Consumer Electronics. Washington D. C., USA:IEEE Press,2015:691-695.

[8] SHEN Hang, BAI Guangwei, YANG Mei, et al. Protecting trajectory privacy:a user-centric analysis[J]. Journal of Network & Computer Applications, 2017, 8(2):128-139.

[9] POULIS G,SKIADOPOULOS S,LOUKUDES G,et al. Apriori-based algorithms for km-anonymizing trajectory data[J]. IEEE Transactions on Data Privacy, 2014, 7(2):165-194.

[10] NIU Ben, LI Qinghua, ZHU Xiaoyan, et al. Achieving k-anonymity in privacy-aware location-based services[C]// Proceedings of IEEE INFOCOM'14. Washington D. C., USA:IEEE Press,2014:754-762.

[11] HWANG R H, HSUEH Y L, CHUNG H W. A novel time-obfuscated algorithm for trajectory privacy[C]// Proceedings of International Symposium on Pervasive Systems, Algorithms and Networks. Washington D. C., USA:IEEE Computer Society,2012:208-215.

[12] CHE Yanzhe, CHIEW K, HONG Xiaoyan, et al. SALS: semantics-aware location sharing based on cloaking zone in mobile social networks[C]//Proceedings of ACM Sigspatial International Workshop on Mobile Geographic Information Systems. New York, USA: ACM Press, 2012:49-56.

[13] WU Xichen, SUN Guangzhong. A novel dummy-based mechanism to protect privacy on trajectories[C]// Proceedings of IEEE International Conference on Data Mining Workshop. Washington D. C., USA: IEEE Press,2015:1120-1125.

[14] HARA T, ARASE Y, YAMAMOTO A, et al. Location anonymization using real car trace data for location based services[C]//Proceedings of the 8th International Conference on Ubiquitous Information Management and Communication. New York, USA:ACM Press,2014:34.

[15] PHAN T N,KUNG J,DANG T K. K UR-algorithm:from position to trajectory privacy protection in location-based applications[C]//Proceedings of International Conference on Database and Expert Systems Applications. Berlin, Germany:Springer-Verlag,2015:82-89.

[16] BINDSCHAEDLER V, SHOKRI R. Synthesizing plausible privacy-preserving location traces[C]//Proceedings of IEEE Symposium on Security and Privacy. Washington D. C., USA:IEEE Press,2016:546-563.

[17] 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.

[18] KATO R, IWATA M, HARA T, et al. A dummy-based anonymization method based on user trajectory with pauses[C]//Proceedings of International Conference on Advances in Geographic Information Systems. Washington D. C., USA:IEEE Press,2012:249-258.

编辑 刘 冰