·安全技术· 文章编号:1000-3428(2016)12-0133-06 文献标志码:A 中图分类号:TP393

移动轨迹数据去匿名化攻击方法

钟建友,常 姗,刘晓强,宋 晖

(东华大学 计算机科学与技术学院,上海 201620)

摘 要: 为保护移动对象轨迹隐私,轨迹数据集发布前常使用假名对轨迹进行匿名化处理。然而,假名用户的匿名轨迹仍面临隐私泄露风险。为此,提出一种新的去匿名化攻击方法。攻击者若获得其攻击对象当前或未来任意时段的若干轨迹片段,则可以此比对匿名历史轨迹数据集,从中识别出攻击对象的历史轨迹。对 2 组真实移动轨迹数据进行特征分析,给出基于轨迹特征相似度的去匿名方法。采用改进的词频-逆文档频率方法提取历史轨迹的特征向量,通过主成分分析降维后,对历史轨迹和攻击者所获得的轨迹片段进行特征匹配,识别出与攻击者所持有轨迹特征相似度最高的历史轨迹。实验结果表明,所提方法可获得较高的去匿名准确率。

关键词: 移动轨迹;假名;轨迹隐私;去匿名化;特征提取

中文引用格式: 钟建友,常 姗,刘晓强,等. 移动轨迹数据去匿名化攻击方法[J]. 计算机工程,2016,42(12):133-138.

英文引用格式: Zhong Jianyou, Chang Shan, Liu Xiaoqiang, et al. De-anonymization Attack Method for Mobile Trace Data[J]. Computer Engineering, 2016, 42(12):133-138.

De-anonymization Attack Method for Mobile Trace Data

ZHONG Jianyou, CHANG Shan, LIU Xiaoqiang, SONG Hui

(School of Computer Science and Technology, Donghua University, Shanghai 201620, China)

【Abstract】 To protect the trace privacy of mobile objects, pseudonym is used to the anonymous processing of trace before the release of the trace dataset. However, the anonymous trace of pseudonym users still faces the risk of privacy leakage. This paper proposes a new de-anonymization attack method. If an attacker obtains several track segments of his attack target at present or in any future period, comparing the traces with the anonymous historical trace dataset, the historical traces of the attack target are identified. The characteristics of the real moving track data of the two groups are analyzed, and a de-anonymization method based on characteristic similarity is presented. The feature vectors of history trace are extracted based on improved Term Frequency-Inverse Document Frequency(TF-IDF) method. The dimension is reduced by Principal Component Analysis(PCA), and the feature matching is performed on the track segments obtained by the historical track and the attacker, to recognize the historical trace with the highest degree of similarity with the trace characteristics of attackers. Experimental results show that the proposed method can obtain higher accuracy.

【Key words】 mobile trace; pseudonym; trace privacy; de-anonymization; feature extraction

DOI: 10.3969/j.issn.1000-3428.2016.12.024

0 概述

移动终端和定位技术的发展,使得随时随地获取移动对象的精确位置成为可能。将单个移动对象的一系列时间上相关的位置信息联系起来就形成了移动轨迹。移动轨迹中通常包含丰富的时空信息,通过合理的挖掘和分析可获得有价值的信息。例如,通过对参与车辆 GPS 轨迹数据的分析,交管部门可获得有关交通信息,如通过某个路段车辆的行驶速度判断交通拥挤情况[1]、路面条件检测[2]。再比如,通过分析城市居民(参与者)的日常移动行为轨迹,可分析城市各板块的功能,从而对未来城市规划

基金项目: 国家自然科学基金(61300199,61402101);中央高校基本科研业务费专项资金(2232014D3-21,2232014D3-42);上海自然科学基金(14ZR1400900)。

作者简介: 钟建友(1989—),男,硕士研究生,主研方向为移动网络隐私保护;常 姗(通讯作者),副教授、博士;刘晓强、宋 晖,教授、博士。

收稿日期: 2015-12-07 **修回日期:** 2016-01-13 **E-mail:** changshan@dhu.edu.cn

提供指导依据。然而,由于移动轨迹中可能包含参与者的许多隐私信息,恶意攻击者可根据非法获取的移动轨迹推测出各类其感兴趣的事件和位置。例如,推测攻击目标的生活周期或敏感位置,从而可能严重威胁到参与者的人身和财产安全。然而,基于对轨迹数据分析的需要,阻止这些信息的访问是不现实的,同时也无法完全保证数据访问者的合法性。例如,交管部门可能将车辆 GPS 轨迹发布给第三方机构进行数据分析,从而导致轨迹数据进一步泄漏给恶意攻击者。

本文提出一种移动轨迹数据去匿名攻击方法,以验证此类攻击的有效性,从而揭示匿名轨迹数据的隐私风险。对匿名轨迹的特征进行分析,给出一种改进的词频-逆文档频率方法构造轨迹特征向量,用于攻击者所持有轨迹与匿名轨迹集合中轨迹的比对,并使用真实轨迹数据集进行实验。

1 研究背景

为保护轨迹数据隐私,在轨迹数据发布前,需使用适当的隐私保护技术对轨迹数据进行预处理。目前常用的方法分为 2 大类:1)修改原始轨迹,降低轨迹在空时中的精度(例如,降低记录轨迹的分辨率或在轨迹中插入噪声),以达到保护隐私的目的,缺点是数据失真严重、可用性低。2)对轨迹匿名化处理,即使用假名(pseudonym,具有唯一性的随机标示符)替代参与者的真实身份,且参与者的真实身份无法通过任何方式与假名相关联。这种匿名化处理方法具有容易实现、计算开销低、不改变原始轨迹数据、可获得最大数据可用性的优点,因而被广泛采用。

然而,尽管假名技术消除了所发布轨迹中参与者的身份,却不能够有效地保护参与者的位置隐私。这是因为:1)每个参与者的运动轨迹具有其固有特征(模式),且短期内不会发生巨大变化。2)匿名轨迹发布后,参与者的运动仍然会持续发生,其在公共场所的运动或者踪迹可以通过各种方式被他人观察到。例如,攻击者可以对其攻击目标实施一段时间的跟踪,或从社交网络、博客等边信息中推断出攻击目标的位置。之后,攻击者将其获得的攻击目标的轨迹或位置与其可访问的匿名轨迹集合中的轨迹进行特征比对,就可从匿名轨迹中唯一或高概率地识别出其攻击目标的轨迹。例如,攻击者获得了攻击目标本周的若干段轨迹及位置,就可能据此比对上个月发布的匿名轨迹集合,并从中识别出其攻击目标的历史轨迹。

近年来,研究者们在位置或轨迹隐私保护、访问控制、风险发现和评估、隐私度量等方面的研究取得了一些进展[3-9]。k-匿名(k-anonymity)[4,10-11]是一种常见的轨迹隐私保护技术,即对任意一条轨迹,需要至少 $k-1$ 条其他轨迹被转换成完全相同的匿名轨迹来构成一个匿名轨迹集合。攻击者在没有背景知识的情况下只有 $1/k$ 的概率猜到参与者的真实轨迹。隐藏技术(cloaking)[12-13]通过降低记录数据的时空精度或添加噪声数据等措施削弱轨迹中连续点的依赖性,在一定程度上保证了轨迹的真实性。然而,这些技术往往导致匿名过程中不必要的信息损失,降低轨迹数据的可用性。假名技术[14]使用唯一的随机标示符替代参与者的真实身份,并确保随机标示符与参与者的真实身份间不存在关联关系。

轨迹隐私风险发现方面,文献[15]提出,攻击者可能从边信息中获得攻击目标的若干位置信息(这些位置发生在待识别匿名轨迹所在时间段内),并据此从匿名轨迹集中推断出攻击目标的完整历史轨迹。文献[16]提出了一种车载自组网节点中轨迹隐私攻防博弈模型,给出攻击和防御策略,分析了攻防双方之间的博弈过程。文献[17]从大规模移动通信数据中分析了匿名位置的泛化程度与用户隐私信息泄漏的关系以及边信息,特别是社交网络对缩小匿名集合、增加隐私风险的影响。文献[18]针对流行的用户轨迹隐私保护方法 Silent Cascade,提出一种新的轨迹隐私度量方法,将用户运动轨迹用带权无向图描述,从信息熵的角度计算用户的轨迹隐私水平。

2 基本定义与问题描述

给出本文所面对的移动轨迹数据的基本结构。其来源主要为车联网、移动社交网络等新兴互联网应用,有如下描述:

定义 1(移动轨迹数据集) 存储大量移动对象在不同时间采样点的位置信息,构成移动轨迹数据集 $D_{\mathrm{mt}}=\{V,\Gamma,L\}$,其中,$V=\{v_1,v_2,\cdots,v_n\}$ 表示数据集中的移动对象集合;位置点在时间上的有序集合称为轨迹,$\Gamma=\{TR_1,TR_2,\cdots,TR_n\}$ 表示移动对象所产生轨迹的集合,TR_i 表示 Γ 中移动对象 v_i 所生成的轨迹;$L=\{p_1,p_2,\cdots\}$ 为位置采样点的集合。

定义 2(位置数据记录) 单个位置数据记录 p 主要包含移动目标 v、地理坐标(x,y)(经度和维度)和记录时间 t,可用四元组表示为 $p=\langle v,x,y,t\rangle$。一般地,将移动对象 v_i 的第 j 条位置记录为 $p_j^{(v_i)}$,在不影响理解的情况下可直接写作 p_j。

定义 3(轨迹序列) 移动对象 v_i 的原始轨迹序列 TR_i 由其移动中的所有位置数据记录构成时空序列 $p_1\rightarrow p_2\rightarrow\cdots\rightarrow p_{len_i}(1\leqslant i\leqslant n)$,其中,$len_i$ 表示 TR_i 的长度;$p_j\rightarrow p_{j+1}\rightarrow\cdots\rightarrow p_{j+k}(1\leqslant j\leqslant\cdots\leqslant j+k\leqslant len_i)$ 称为 TR_i 的子轨迹。

定义 4(路段)　路段 r 是地图上的一条单向边,包含一对端点($r.start, r.end$),可简记为($r.s, r.e$)。

定义 5(路线)　路线 R 是由首尾相接的路段构成的序列,即 $R: r_1 \rightarrow r_2 \rightarrow \cdots \rightarrow r_m$,其中,$r_k.end = r_{k+1}.start, 1 \leqslant k < m$。$R$ 的起始和终点可以分别表示为 $R.s = r_1.s$ 和 $R.e = r_m.e$。

2.1　轨迹匿名

为保障移动对象的位置隐私,移动轨迹数据集对外公布之前需进行匿名化处理。对任意轨迹 TR_i,使用假名 $o_i \in O$ 替换每条轨迹产生者的真实身份 v_i。经匿名后的轨迹 TR_i 中所包含的所有位置数据 $p_j^{(v_i)} = \langle v_i, x_j, y_j, t_j \rangle$ 被更改为 $p_j^{(o_i)} = \langle o_i, x_j, y_j, t_j \rangle$。假名的使用需满足以下 2 个原则:

1)移动对象的真实身份无法通过任何方式与假名相关联;

2)每个移动对象对应唯一假名。

2.2　去匿名化攻击

在去匿名化攻击中,攻击者试图从匿名轨迹数据集中识别出其攻击目标对象 v_i 的移动轨迹 TR_i,即找出 $o_i \rightarrow v_i$ 的映射。具体地,去匿名化攻击分为 3 步:

1)攻击者可访问一组移动轨迹数据集 D_{mt},其中包括一个或多个攻击目标的移动轨迹。

2)攻击者通过观察或者其他方式获得攻击目标未来任意时间段的若干移动轨迹片段(非移动轨迹数据集中的子轨迹)。

3)分析匿名轨迹数据集中不同移动对象移动轨迹的时空特征,并与第 2)步所得移动轨迹的特征相比较,从而从匿名轨迹集合中识别出攻击目标的轨迹。

3　轨迹数据预处理

3.1　轨迹数据集与路网

本文采用上海和深圳的出租车 GPS 报告数据,作为原始移动轨迹数据集,这些数据被称为驾驶状态报告(Driving Status Report, DSR)。DSR 包括的具体信息有:车辆的编号 ID,当前位置的经纬度,时间戳,行驶速度,车辆的行驶角度和运动状态(比如出租车是否载人或者公交车是否到达公交站)。报告大约 1 min 被发送一次。两数据集的具体特征如表 1 所示。

表 1　数据集统计信息

地点	车辆数	数据开始时间	时间长度/d	间隔时间/s	轨迹数
上海	906	2007-02-01	28	15,60	25 368
深圳	1 945	2009-10-01	31	60	60 295

所采用的上海和深圳路网分别包含 66 459 条和 89 578 条路段。每条路段在各自的路网中有唯一编号。

3.2　轨迹分割及地图匹配

在上述数据集中,每个 GPS 日志文件中记录了一辆出租车一整天的运动轨迹。因此,车辆 v_i 的完整轨迹 TR_i 被以天为单位进行划分,得到若干子轨迹 $TR_{i,j}(1 \leqslant j \leqslant num_{\mathrm{day}})$,$TR_{i,j}$ 表示车辆 v_i 在第 j 天的轨迹。

由于 GPS 具有定位误差,因此需要借助电子地图中的路网信息进行位置矫正,从而将轨迹恢复到道路之上,并确定车辆相对于地图的位置。目前已存在一些经典地图匹配算法,本文应用 ST-Matching 地图匹配算法[19]将轨迹的 GPS 点映射到相应的路段上。因此,一条轨迹 TR_i(或 $TR_{i,j}$)被转换成路线 R_i(或 $R_{i,j}$)。

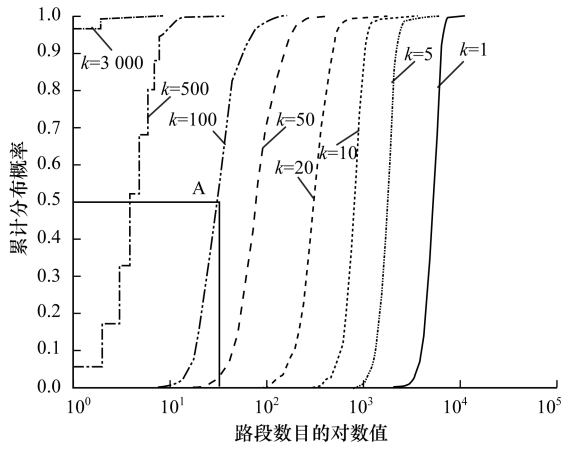
4　轨迹数据特征分析

对轨迹特征进行以下分析:1)车辆轨迹的路段偏好;2)不同车辆轨迹所包含路段的差异性。

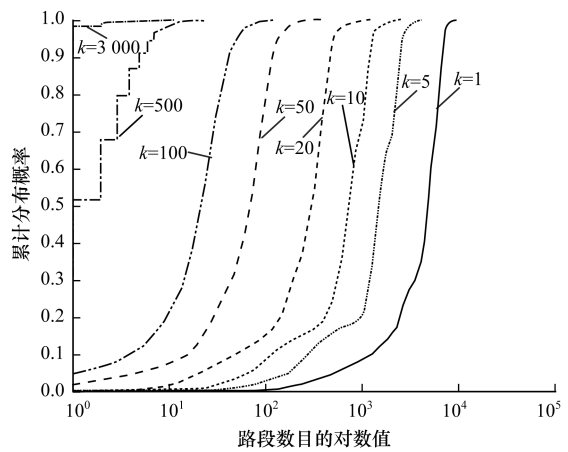
4.1　轨迹中的路段偏好

车辆在城市中道路上行驶时,并非机会均等地驶过地图中的每一个路段。相反,由于驾驶员所在区域的不同、道路在路网中的重要程度不同、驾驶习惯差异等因素,造成车辆 v_i 的轨迹 TR_i 所对应的路线 R_i 中,不同路段被经过的次数也不同,存在高频路段,可能成为潜在的轨迹特征。以下实验验证了路段偏好的存在:

首先,统计数据集中每条路线经过各路段的次数(例如,路段 r_j 在路线 R_i 中出现的次数记为 $t_{i,j}$)。设定阈值 k,将一条线路(对应一辆车)中出现次数大于 k 的路段数视为随机变量,画出其累积分布函数(Cumulative Density Function, CDF)图,如图 1(a)和图 1(b)所示。在图 1 中,横坐标表示满足阈值条件 $t_{i,j} \geqslant k$ 的路段数,纵坐标表示满足阈值条件的路段数小于等于某个数值(相应横坐标)的轨迹百分比。例如,在图 1(a)中,A 点表示当 $k = 100$ 时,在所有车辆轨迹中,50% 的轨迹具有不超过 30 条满足 $t_{i,j} \geqslant 100$ 的路段。曲线顶端则表示 100% 的轨迹具有不超过 200 条满足 $t_{i,j} \geqslant 100$ 的路段。从图 1(a)、图 1(b)中可看出,随着 k 值的增大,符合阈值条件的路段数急剧减少。例如,在图 1(a)中,$k = 1$ 时,每条满足阈值的路段数有千余条,而当 $k = 100$ 时,路段数就减少到 200 条以内。

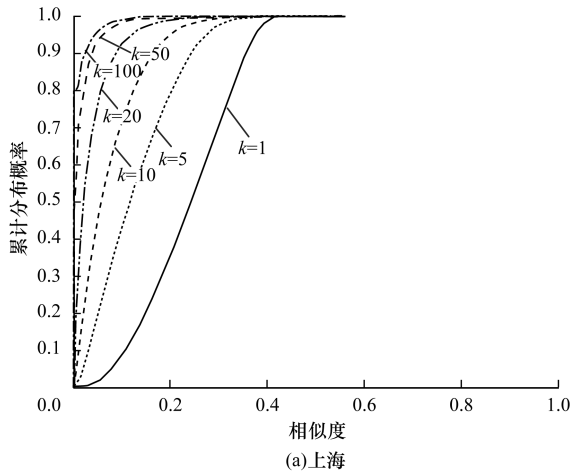


(a)上海

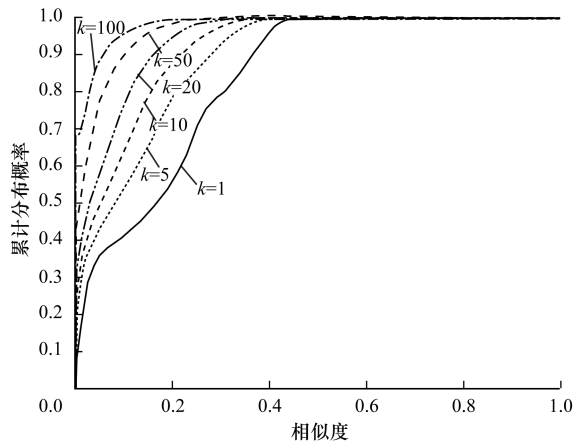


(b)深圳

图 1 路段偏好 CDF 图



(a)上海



(b)深圳

图 2 轨迹间相似度 CDF 图

4.2 轨迹间差异性

本节通过一组实验证明不同车辆行驶中的路段偏好具有显著的差异性。为此,有如下定义:$S_i=\{r_j \mid$路段 r_j 在 R_i 中出现的次数 $t_{i,j}\geqslant 0\}$,即 S_i 为 R_i 中所包含路段的集合。

$S_i^{(n)}=\{r_j \mid$路段 r_j 在 R_i 中出现的次数 $t_{i,j}\geqslant n\}$,即 $S_i^{(n)}$ 为 R_i 中出现大于 n 次的路段集合。

$$S_i^{(n)}\cap S_j^{(n)}=\{r_1 \mid r_1\in S_i^{(n)} \text{ and } r_1\in S_j^{(n)}\}$$

$$S_i^{(n)}\cup S_j^{(n)}=\{r_1 \mid r_1\in S_i^{(n)} \text{ or } r_1\in S_j^{(n)}\}$$

因此,轨迹 TR_i 和 TR_j 间的相似度为:

$$sim_{i,j}^{(n)}=\frac{|S_i^{(n)}\cap S_j^{(n)}|}{|S_i^{(n)}\cup S_j^{(n)}|}$$

其中,$|\cdot|$表示求集合的元素个数。

根据上述定义,设定阈值 k,可求出不同轨迹间的相似性。图 2(a)和图 2(b)分别为上海和广州轨迹数据相似度的 CDF 图。由图可见,上海和深圳的轨迹间相似度都在 0.45 以下,说明不同车辆轨迹具有各自路段的偏好特征。此外,随着 k 值的增大,相似度还进一步减小,说明去除了低频路段的噪音干扰后,路段偏好特征将更加显著。可以看到,当 k =100时,90% 以上轨迹对间的相似度小于 0.1,即不同轨迹间具有较大的路段偏好差异。

5 去匿名化攻击

由上节分析可知,可抽取轨迹中所包含各路段的频数,构造轨迹特征向量,用于比较轨迹间的相似性。本节给出一种攻击者实施去匿名化攻击的策略。假定攻击者可访问匿名轨迹集 $\Gamma=\{TR_1, TR_2,\cdots,TR_n\}$。同时,还获得了某个攻击目标 v_σ 的一段轨迹$\widetilde{TR}_\sigma$。攻击者的目的是依据$\widetilde{TR}_\sigma$从 Γ 中识别出TR_σ。攻击者首先进行地图匹配,将 Γ 转换为 $\Omega=\{R_1,R_2,\cdots,R_n\}$,并将$\widetilde{TR}_\sigma$转换为路线$\tilde{R}_\sigma$。然后使用一种改进的词频-逆文档(Term Frequency-Inverse Document Frequency,TF-IDF)来衡量路网中的每个路段在每条路线 R_i 中的重要程度,从而提取 R_i 中的关键路段。根据路段的 TF-IDF 值构建特征向量,然后使用余弦相似度进行特征匹配。

5.1 特征向量提取

攻击者对 Γ 中的轨迹构建特征向量。选取轨迹 TR_i,通过以下方法构建特征向量:

$$\boldsymbol{f}_i=(w_{i,1},w_{i,2},\cdots,w_{i,\phi})^{\mathrm{T}}$$

其中,$w_{i,j}$是路段 $r_j(1\leqslant j\leqslant\phi)$的改进 TF-IDF 值;$\phi$ 是整个路网中路段的数目。$w_{i,j}$可用下式进行计算:

$$w_{i,j}=\frac{t_{i,j}}{\mathcal{L}_i}\times\log_a\left(c_{i,j}\times\frac{n}{c_i}\right)$$

其中，$t_{i,j}$是路段 r_j 在轨迹路线 R_i 中出现的次数；$\mathcal{L}_i$ 是 R_i 的长度（即 R_i 中所包含路段的个数）；$\frac{t_{i,j}}{\mathcal{L}_i}$表示 r_j 在轨迹 TR_i 中出现的频率；c_i 表示 Γ 中包含路段 r_j 的轨迹数；$c_{i,j}$表示在 TR_i 中包含路段 r_j 的单日轨迹数。

5.2　特征匹配

构造 $\widetilde{TR}_\sigma$ 的特征向量 $\tilde{\boldsymbol{f}}_\sigma=(\tau_{\sigma,1},\tau_{\sigma,2},\cdots,\tau_{\sigma,\phi})^{\mathrm{T}}$，$\tau_{\sigma,j}$可用下式计算：

$$\tau_{\sigma,j}=\frac{\tilde{t}_{\sigma,j}}{\sum_{j=1}^{\phi}\tilde{t}_{\sigma,j}}$$

其中，$\tilde{t}_{\sigma,j}$是 r_j 在路线$\tilde{R}_\sigma$中出现的次数。

使用余弦相似度计算公式，分别计算 $\boldsymbol{f}_i$（$1\leqslant i\leqslant n$）与$\tilde{\boldsymbol{f}}_\sigma$ 间的匹配程度（M_{score}），如下：

$$M_{\mathrm{score}}(i,\sigma)=\frac{\boldsymbol{f}_i\cdot\tilde{\boldsymbol{f}}_\sigma}{\|\boldsymbol{f}_i\|\ \|\tilde{\boldsymbol{f}}_\sigma\|}=\frac{\sum_{j=1}^{\phi}w_{i,j}\times\tau_{\sigma,j}}{\sqrt{\sum_{j=1}^{\phi}(w_{i,j})^2}\times\sqrt{\sum_{j=1}^{\phi}(\tau_{\sigma,j})^2}}$$

5.3　特征向量降维

上述方法所提取的特征向量 $\boldsymbol{f}_i$ 的维度较高，这是由于 $\boldsymbol{f}_i$ 的维度取决于路网中的路段数量。同时，$\boldsymbol{f}_i$ 也是稀疏向量，可采用主成分分析法（Principle Components Analysis，PCA）对特征向量进行维度约简，步骤如下：

1）用 Γ 中轨迹的特征向量构造 $n\times\phi$ 矩阵 $\boldsymbol{F}=[\boldsymbol{f}_1,\boldsymbol{f}_2,\cdots,\boldsymbol{f}_n]$。

2）将 $\boldsymbol{F}$ 进行零均值化，即每一行减去该行的均值。

3）计算 $\boldsymbol{F}$ 的协方差矩阵：$\mathrm{cov}(\boldsymbol{F})=\frac{1}{n}\boldsymbol{F}\boldsymbol{F}^{\mathrm{T}}$。

4）将 $\mathrm{cov}(\boldsymbol{F})$进行特征值分解：$\mathrm{cov}(\boldsymbol{F})=\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^{\mathrm{T}}$。其中，$\boldsymbol{\Lambda}=\mathrm{diag}(\lambda_1,\lambda_2,\cdots,\lambda_\varphi)$为 ϕ 阶对角矩阵，其对角元素为各特征向量对应的特征值，且按照从大到小依次排列；矩阵 U 中每一行是 $cov(\boldsymbol{F})$的一个特征向量。

5）选取前 k 个特征值，使得下式成立：

$$\frac{\sum_{i=1}^{k}\lambda_i}{\sum_{i=1}^{\phi}\lambda_i}\geqslant\alpha,0\leqslant\alpha\leqslant1$$

即前 k 个特征值中包含的信息量至少为 α。

6）选取 $\boldsymbol{U}$ 中最大的 k 个特征向量（矩阵前 k 行）构成投影矩阵 $\boldsymbol{U}_k$，对原始矩阵 $\boldsymbol{F}$ 进行投影，得到降维后的新矩阵为：$\boldsymbol{F}_k=\boldsymbol{U}_k\boldsymbol{F}$。

7）对轨迹$\widetilde{TR}_\sigma$的特征向量$\tilde{\boldsymbol{f}}_\sigma$ 进行零均值化，即每行相应减去 $\boldsymbol{F}$ 中相应行均值，降维后的特征向量为：$\tilde{\boldsymbol{f}}_k=U_k\,\tilde{\boldsymbol{f}}_\sigma$。

6　实验结果及分析

6.1　实验方法

使用第 4 节中所描述的数据集进行实验，分别从上海和深圳轨迹数据集中选取前 20 d 和 22 d 构造匿名轨迹集合 Γ。攻击者可获得剩余轨迹（上海：8 d 轨迹，深圳：9 d 轨迹）中的任意一段（已知此段轨迹所对应车辆的 ID）。攻击目标是通过比对已知轨迹与匿名轨迹集，尽可能多地将 Γ 去匿名化。

使用轨迹匹配的准确率衡量去匿名化攻击的效果，计算公式如下：

$$a_{\mathrm{Accuracy}}=\frac{n_{\mathrm{crt}}}{n}$$

其中，n_{crt}表示匹配正确的轨迹条数。

6.2　轨迹长度对准确率的影响

本节研究攻击者所持有轨迹$\widetilde{TR}_\sigma$的长度对轨迹匹配准确率的影响。实验中，对上海和深圳 2 个数据集，从构造匿名轨迹集所剩余的轨迹中任意选取 d（$1\leqslant d\leqslant8(9)$）天构成$\widetilde{TR}_\sigma$。

实验结果如图 3 所示，当 $d=1$ 时，上海和深圳的轨迹匹配准确率可分别达到 71% 和 52%，即较短的$\widetilde{TR}_\sigma$就可获得较高准确率。随着攻击者获得轨迹长度的增加，准确率会进一步提高，但随着 d 的增大，上升趋于平缓。上海和深圳的准确率最高能达到 95% 和 70%。

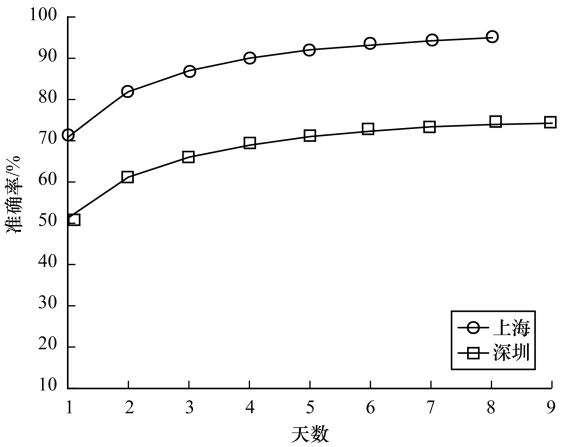


图 3　不同轨迹长度对准确率的影响

6.3　匿名集合尺寸与准确率的关系

上节实验的特征匹配中，在匿名轨迹集合中选取 M_{score}最高的轨迹作为待识别轨迹的历史轨迹，即从匿名轨迹集合中唯一确定待识别轨迹。然而，在实际应用中，若能够有效地缩小待识别轨迹的匿名集合也十分有意义。即确定待识别轨迹的 k 匿名集合，使得可以确信待识别车辆的轨迹在原集合的 k 条轨迹范围内（上节实验的 k 匿名集合尺寸为 1）。

因此,本节研究匿名集合尺寸与准确率的关系。实验中,对任意一条待识别轨迹,将匿名轨迹集中所有轨迹与其计算的 M_{score} 值进行降序排列,选取前 k 个构造 k 匿名集合,并计算匹配准确率。图 4 绘出了 k 的取值(即匿名集合尺寸)与准确率的关系。可以看出,$k=5$ 时,上海和深圳的准确率可分别达到 93% 和 75%;随着 k 值的增大,准确率逐渐上升。从 1 ~ 100 的增速达到最快,后面开始放缓。当候选集合 $k=100$ 时,上海和深圳的匹配准确率可分别达到 98% 和 94%。

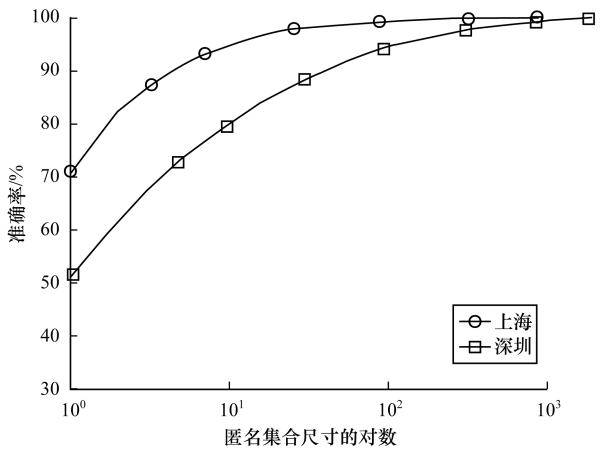


图 4 匿名集合尺寸与准确率的关系

7 结束语

假名技术被广泛用于轨迹数据集合的匿名化处理。本文针对基于假名的匿名轨迹,提出一种新的去匿名攻击方法。攻击者将历史匿名轨迹与其通过跟踪或侧信道获得的某移动节点的若干轨迹片段进行特征比对,可准确地识别出该移动节点的历史轨迹。使用上海和深圳出租车的真实轨迹数据,验证了所提出去匿名方法的准确性。实验结果表明,上海和深圳的去匿名准确率最高可达 95% 和 70%。未来工作将针对移动轨迹匿名发布方法展开研究。

参考文献

[1] Mohan P, Padmanabhan V N, Ramjee R. Nericell: Rich Monitoring of Road and Traffic Conditions Using Mobile Smartphones[C]//Proceedings of the 6th ACM Conference on Embedded Network Sensor Systems. New York, USA: ACM Press, 2008: 323-336.

[2] Eriksson J, Girod L, Hull B, et al. The Pothole Part-ol: Using a Mobile Sensor Network for Road Surface Monitoring[C]//Proceedings of the 6th International Conference on Mobile Systems, Applications, and Services. New York, USA: ACM Press, 2008: 29-39.

[3] Beresford A R, Stajano F. Location Privacy in Pervasive Computing[J]. IEEE Pervasive Computing, 2003, 2(1): 46-55.

[4] Gruteser M, Grunwald D. Anonymous Usage of Location-based Services Through Spatial and Temporal Cloak-ing[C]//Proceedings of the 1st International Con-ference on Mobile Systems, Applications and Services. New York, USA: ACM Press, 2003: 31-42.

[5] Gruteser M, Liu Xuan. Protecting Privacy in Continuous Location-tracking Applications[J]. IEEE Security & Privacy, 2004, 2(2): 28-34.

[6] Krumm J. Inference Attacks on Location Tracks[C]// Proceedings of the 5th International Conference on Pervasive Computing. Berlin, Germany: Springer, 2007: 127-143.

[7] Krumm J. A Survey of Computational Location Privacy[J]. Personal and Ubiquitous Computing, 2009, 1(6): 391-399.

[8] Kulik L. Privacy for Real-time Location-based Services[J]. SIGSPATIAL Special, 2009, 1(2): 9-14.

[9] 霍 峥,孟小峰. 轨迹隐私保护技术研究[J]. 计算机学报, 2011, 34(10): 1820-1830.

[10] Samarati P, Sweeney L. Protecting Privacy When Disclosing Information: k-anonymity and Its Enforcement Through Generalization and Suppression: SRI-CSL-98-04[R]. 1998.

[11] Sweeney L. k-anonymity: A Model for Protecting Privacy[J]. International Journal of Uncertainty, Fuzziness and Knowledge-based Systems, 2002, 10(5): 557-570.

[12] Gedik B, Liu Ling. Location Privacy in Mobile Systems: A Personalized Anonymization Model[C]//Proceedings of the 25th IEEE International Conference on Distributed Computing Systems. Washington D. C., USA: IEEE Press, 2005: 620-629.

[13] Hoh B, Gruteser M, Xiong Hui, et al. Preserving Privacy in GPS Traces via Uncertainty-aware Path Cloaking[C]//Proceedings of the 14th ACM Conference on Computer and Communications Security. New York, USA: ACM Press, 2007: 161-171.

[14] Pfitzmann A, Köhntopp M. Anonymity, Unobservability, and Pseudonymity——A Proposal for Terminology[C]// Proceedings of International Workshop on Design Issues in Anonymity and Unobservability. Berlin, Germany: Springer, 2001: 1-9.

[15] Ma C Y T, Yau D K Y, Yip N K, et al. Privacy Vulnerability of Published Anonymous Mobility Traces[J]. IEEE/ACM Transactions on Networking, 2013, 21(3): 720-733.

[16] 杨卫东,何云华,孙利民. 车载自组网节点轨迹隐私攻防博弈模型[J]. 通信学报, 2013, 34(z1): 240-245.

[17] Zang Hui, Bolot J. Anonymization of Location Data does not Work: A Large-scale Measurement Study[C]// Proceedings of the 17th Annual International Conference on Mobile Computing and Networking. New York, USA: ACM Press, 2011: 145-156.

[18] 王彩梅,郭亚军,郭艳华. 位置服务中用户轨迹的隐私度量[J]. 软件学报, 2012, 23(2): 352-360.

[19] Lou Yin, Zhang Chengyang, Zheng Yu, et al. Map-matching for Low-sampling-rate GPS Trajectories[C]// Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. New York, USA: ACM Press, 2009: 352-361.

编辑 顾逸斐