计算机研究与发展
Journal of Computer Research and Development
ISSN 1000-1239/CN 11-1777/TP
50(3): 578-593, 2013

基于聚类杂交的隐私保护轨迹数据发布算法

吴英杰[1,2,3] 唐庆明[1] 倪巍伟[2] 孙志挥[2] 廖尚斌[1]
[1](福州大学数学与计算机科学学院 福州 350108)
[2](东南大学计算机科学与工程学院 南京 211189)
[3](福州大学网络系统信息安全福建省高校重点实验室 福州 350108)
(yjwu@fzu.edu.cn)

A Clustering Hybrid Based Algorithm for Privacy Preserving Trajectory Data Publishing

Wu Yingjie[1,2,3], Tang Qingming[1], Ni Weiwei[2], Sun Zhihui[2], and Liao Shangbin[1]
[1](*College of Mathematics and Computer Science, Fuzhou University, Fuzhou* 350108)
[2](*College of Computer Science and Engineering, Southeast University, Nanjing* 211189)
[3](*The Higher Educational Key Laboratory for Network System Information Security of Fujian Province, Fuzhou University, Fuzhou* 350108)

Abstract Recently, privacy preserving trajectory data publishing has become a hot topic in data privacy preserving research fields. Most previous works on privacy preserving trajectory data publishing adopt clustering techniques. However, clustering based algorithms for trajectory data publishing only consider preserving the privacy of each single trajectory, ignoring the protection of the characteristics of trajectory clustering groups. Therefore, the publishing trajectory data by clustering are vulnerable to suffer re-clustering attacks, which is verified by theoretical analysis and simulated experiments. In order to avoid re-clustering attacks, a (k,δ,Δ)-anonymity model and a clustering hybrid based algorithm CH-TDP for privacy preserving trajectory data publishing are presented. The key idea of CH-TDP is to firstly hybridize between clustering groups, which are generated by the (k,δ)-anonymity model and the related algorithms, and then adopt perturbation within each clustering group. The aim of CH-TDP is to avoid suffering re-clustering attacks effectively while assuring the data quality of the released trajectory data not less than a threshold Δ. CH-TDP and the traditional algorithms are compared and experimental results show that CH-TDP is effective and feasible.

Key words privacy preserving; trajectory data publishing; re-clustering attack; clustering; hybrid

摘 要 传统关于轨迹数据发布的隐私保护研究大多采用聚类技术,其相关算法只关注每条轨迹的隐私保护,忽视对轨迹聚类组特征的保护.通过理论分析和实验验证发现,对采用聚类发布技术产生的轨迹数据进行二次聚类,可得到原始轨迹数据在发布之前的聚类组特征,从而可能导致隐私泄露.为了有效预防二次聚类攻击,提出一种(k,δ,Δ)-匿名模型和基于该模型的聚类杂交隐私保护轨迹数据发布算法CH-TDP,算法CH-TDP对采用(k,δ)-匿名模型及相关算法处理得到的聚类分组先进行组间杂交,而后再进行组内扰乱,其目标在防止出现二次聚类攻击的前提下,保证发布轨迹数据的质量不低于阈值Δ.实验对算法CH-TDP的可行性及有效性与同类算法进行比较分析,结果表明算法CH-TDP是有效可行的.

收稿日期:2011-05-05;**修回日期**:2011-08-29
基金项目:国家自然科学基金项目(61003057);福建省自然科学基金项目(2010J01330);福州大学科技发展基金项目(2012-XQ-27)

关键词 隐私保护;轨迹数据发布;二次聚类攻击;聚类;杂交

中图法分类号 TP311.13

在现实生活中,出于决策制定和科学研究的需要,许多研究机构或学术组织都会对外发布数据.如何保证所发布的数据既是可用的,又不会泄露数据中所包含个体的隐私信息是一个迫切需要解决的研究课题.目前关于隐私保护数据发布的研究工作大多是面向关系型数据的.近年来的研究显示,非关系型数据的发布也存在隐私威胁和敏感信息泄露的问题.轨迹数据发布中的隐私保护问题由于移动位置服务(LBS)的盛行而开始得到学术界的关注.然而,由于移动轨迹数据具有时间相关、位置相关和大规模高维的特点,传统基于关系型数据发布的隐私保护模型将无法直接用于解决移动轨迹数据发布中可能存在的隐私泄露问题.由此,设计有效的轨迹数据发布隐私保护模型及相关算法将有助于在保证个体隐私安全的前提下,提高移动轨迹数据的价值,从而为相关应用提供决策支持.

1 相关工作

Samarati 和 Sweeny 首先提出面向关系型数据发布的隐私保护 k-匿名模型[1-2].k 匿名模型要求表中的每一条记录至少和其他 $k-1$ 条记录关于准标识符(QI)的值相同.经过多年研究,k-匿名模型已日趋成熟.与关系型数据不同的是,在移动轨迹数据中,很难界定准标识符和敏感属性.目前关于移动轨迹数据的准标识符定义仍然是一个开放问题.因此,k-匿名以及近年来提出的其他一些针对关系型数据库的隐私保护模型无法直接用于预防高维的移动轨迹数据发布中可能存在的隐私泄露[3-4].

近年来,人们对隐私保护移动轨迹数据发布所进行的研究大多是在传统的 k-匿名模型基础上,先对移动轨迹进行聚类,而后对聚类组进行扰乱、概化或特征发布[5-10].文献[5-6]率先提出将轨迹匿名问题转化成一个约束聚类问题.利用移动轨迹数据抽样和定位系统的不精确性,提出(k,δ)-匿名模型.该模型要求对于每一条移动轨迹,在其周围距离为 δ 的不确定区域内,至少存在 $k-1$ 条其他移动轨迹.文献[6]采用欧式距离作为轨迹聚类的度量函数,先对移动轨迹进行聚类,而后对每个轨迹聚类组进行点扰乱发布.文献[10]针对文献[6]中欧式距离要求轨迹之间要有相同时间区间的不足,将文献[11]提出的 EDR 作为轨迹聚类的度量函数,使得不同时间范围的轨迹可以聚在一起.文献[7]采取先聚类后概化的策略,先对移动轨迹进行聚类,而后对每个轨迹聚类组进行概化发布.此外,文献[7]还针对轨迹聚类组概化发布可能存在的概化区域相交攻击,提出对轨迹聚类组进行组内边扰乱的发布策略.文献[8]将时间作为准标识符,并假设不同的移动轨迹有不同的准标识符,提出基于位置的空间概化发布方式,确保发布数据中每条轨迹均至少和其他 $k-1$ 条轨迹关于准标识符的取值相同.文献[9]则在路网约束下提出一个基于聚类的轨迹匿名发布算法.

上述发布模式可有效预防攻击者使用特定的位置信息对发布轨迹数据进行隐私攻击,同时保证发布轨迹数据具有较高的质量.然而,以上算法只关注每条轨迹所对应个体的隐私保护,却忽视了轨迹聚类组特征的保护.本文通过理论分析和实验验证,发现通过对上述基于聚类的算法处理后发布的匿名轨迹数据进行二次聚类,可得到原始轨迹数据在发布之前的聚类组特征,从而可能存在二次聚类攻击.本文提出了抵御二次聚类攻击的(k,δ,Δ)-匿名模型和基于聚类杂交的隐私保护轨迹数据发布算法,并对模型及算法的可行性及有效性与同类算法进行实验比较分析.

2 基础知识与相关隐私保护技术

2.1 相关定义

定义 1. 轨迹.在实际应用中,轨迹可表示为三维空间(两维坐标加时间维)中的一条折线,记为 $tr=\{p_1,p_2,\cdots,p_m\}$,其中 $p_k=(x_k,y_k,t_k)$ 表示轨迹 tr 在 t_k 时刻的位置为(x_k,y_k),$t_1<t_2<\cdots<t_m$,m 为轨迹 tr 的点数.轨迹数据库 D 是轨迹的集合,记 $D=\{tr_1,tr_2,\cdots,tr_n\}$,$n$ 为轨迹数据库中轨迹的条数.

在现实生活中,移动轨迹数据抽样和定位系统(如 GPS)由于无法精确定位,导致采集到的轨迹往往是一条不确定的轨迹.例如,行驶在同一条道路上的汽车经过两个不同坐标(100,100)和(101,100.5),但是这两个不同坐标表达的却很可能是相同的位置信息.因此文献[6]提出了关于轨迹的不确定性模型.

定义 2[6]. 轨迹的不确定性模型. 给定在时间 t_1 和 t_n 之间的轨迹 tr 和不确定性阈值 δ,对于 tr 上的每一个点 (x,y,t),其不确定区域是以 (x,y,t) 为圆心,以 δ 为半径的水平圆盘,其中 (x,y) 为 tr 在时间 $t\in[t_1,t_n]$ 的期望位置. 记轨迹 tr 关于 δ 的不确定性模型为 $\langle tr,\delta\rangle$,$\langle tr,\delta\rangle$ 形成的轨迹圆柱记为 $Vol(tr,\delta)$,表示在时间 $t\in[t_1,t_n]$ 内所有圆盘的集合.

轨迹的不确定性模型如图 1 所示:

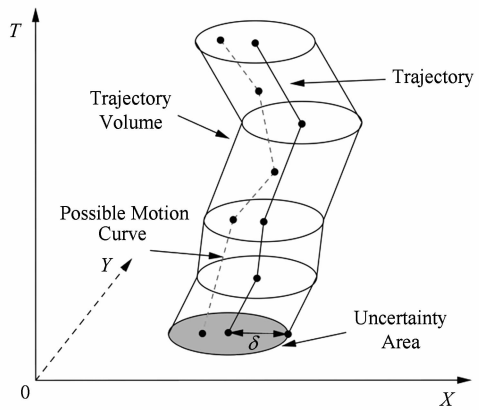


Fig. 1 Uncertain trajectory model.

图 1 轨迹的不确定性模型

定义 3. 轨迹之间的距离. 本文采用欧几里德距离度量轨迹之间的距离. 定义两条轨迹 tr_1,tr_2 在时刻 t 的距离为

$$Dist(tr_1[t],tr_2[t])=\sqrt{(tr_1[t].x-tr_2[t].x)^2+(tr_1[t].y-tr_2[t].y)^2},$$

则在时间范围 $[t_1,t_n]$ 内,两条轨迹 tr_1,tr_2 的距离为

$$Dist(tr_1,tr_2)=\sum_{t=1}^{n}Dist(tr_1[t],tr_2[t]).$$

定义 4. 相近轨迹. 在时间范围 $[t_1,t_n]$ 内,称两条轨迹 tr_1 和 tr_2 为相近轨迹当且仅当对于 tr_1 内的每一个点 (x_1,y_1,t) 和 tr_2 内的每一个点 (x_2,y_2,t),它们之间的距离 $Dist((x_1,y_1),(x_2,y_2))\leqslant\delta$,记为 $Coloc_\delta(tr_1,tr_2)$.

定义 5. (k,δ)-匿名模型. 给定不确定性阈值 δ 和匿名阈值 k,轨迹集合 D 满足 (k,δ)-匿名当且仅当 $|D|\geqslant k$,且 D 中任意两条轨迹 tr_1 和 tr_2 均满足 $Coloc_\delta(tr_1,tr_2)$.

一般地,不确定性阈值 δ 越大,说明产生的聚类组越大,发布数据的安全性越高但数据的可用性却越低. 相反地,不确定性阈值 δ 越小,说明产生的聚类组越精细,发布数据的可用性越高而数据的安全性却越低.

2.2 基于聚类的轨迹数据隐私保护技术

传统的基于聚类的轨迹数据隐私保护技术一般可以分为 2 个步骤:1)将待发布的轨迹数据根据相似性进行分组;2)对每个聚类分组进行匿名化处理. 而在第 2 个步骤中,处理的方法一般包括扰乱、概化和特征发布. 如图 2 所示:

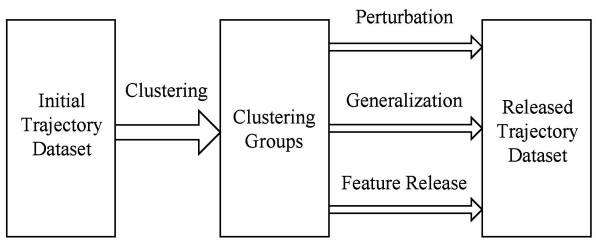


Fig. 2 Privacy preserving trajectory data publishing.

图 2 轨迹数据隐私保护的一般策略

所谓扰乱,是指对于每个聚类组,重新构造组内的轨迹,一般包括基于点的扰乱和基于边的扰乱. 如图 3 所示. 基于点的扰乱是指重构某个时刻的位置点. 例如,在图 3(a)中,某个聚类组在某个时刻有一个实心离群点在不确定区域外,经过扰乱后,该点被移入不确定区域内. 基于边的扰乱是指重构某两个连续采样时刻的位置点之间的边. 图 3(b) 左侧是某个聚类组在某两个连续采样时候的轨迹片段示意图,而右侧则是重构两个时刻的位置点之间的边后的结果.

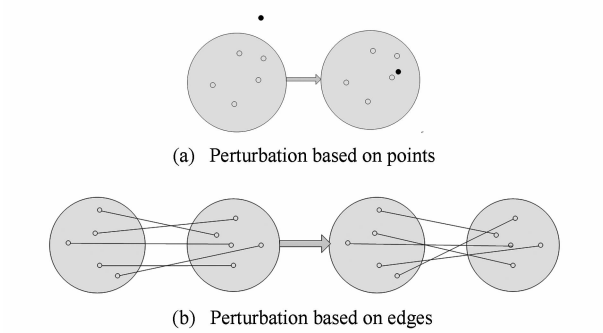


Fig. 3 Trajectory perturbation.

图 3 轨迹扰乱

概化一般是指将每个聚类分组中每个采样时刻的所有位置点用一个矩形区域[12]表示,如图 4 所示. 特征发布则是将每个聚类组抽象出一条有代表性的轨迹,然后每个聚类组只发布这条轨迹. 一般地,这条代表轨迹可能是由某些概化区域的中心所形成,例如图 4 中用虚线表示的轨迹 $rep1$ 与 $rep2$,也可能是在不确定模型下某个轨迹圆柱的中心,如图 1 所示.

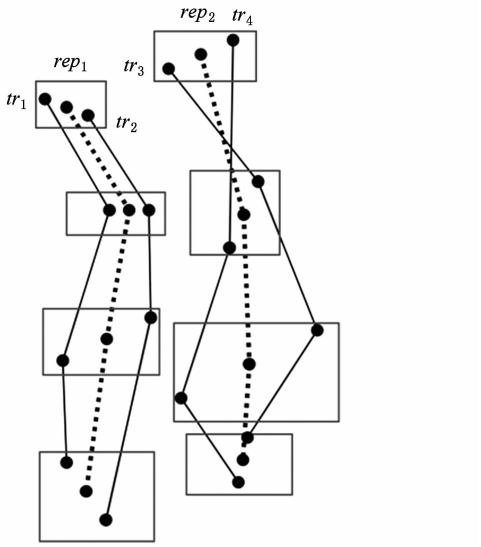


Fig. 4 Trajectory generalization and feature release.

图 4 轨迹概化与特征发布

3 二次聚类攻击

3.1 问题提出

本节将从整个聚类组的角度来分析可能存在的二次聚类攻击.在关系型数据库中,除了能识别记录所对应个体的显示标识符外,其余属性一般可分为准标识符和敏感属性,且经常假设准标识符是固定的.而轨迹数据集则很难区分准标识符和敏感信息.一般认为用户轨迹的每一个点都可能包含敏感信息.

假设经过传统概化操作处理后得到的某条轨迹概化示意图如图 5 所示.

很明显,该条轨迹经概化处理后,每个位置点的信息都是模糊的.然而,该条轨迹对于数据统计和决策制定可能仍然是有用的,而恶意用户也可以通过该条轨迹得知有用的信息.例如,恶意用户 Alice 知道 Bob 处于某个移动轨迹数据集中,并且经过了图 5 中的区域 A,D 和 E,则 Alice 很可能通过这些信息将上述这条轨迹($A\rightarrow B\rightarrow C\rightarrow D\rightarrow E$)同其他轨迹区别开来,从而得知 Bob 肯定也去了区域 B 和 C.从某种意义上来说,如果发布后的数据对于数据使用者是有用的,那么对于数据攻击者也是同样有用的.

同样地,对于非概化发布的数据,如果 Alice 能够得知原始数据在匿名操作之前的聚类组,那么 Alice 同样可以得知其目标攻击对象的某些信息.

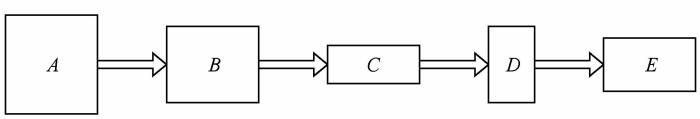


Fig. 5 Trajectory generalization.

图 5 轨迹概化

例如,假设 Alice 的攻击对象的实际轨迹为图 6 中的曲线,而 Alice 已经知道其目标轨迹中的第 1、第 4 和最后一个点.同时假设 Alice 可以通过某种渠道得到实际数据发布之前的聚类分组($A\rightarrow B\rightarrow C\rightarrow D\rightarrow E$).如果此时仅仅只有一个聚类组和 Alice 的已知信息完全匹配,那么 Alice 就可以大致确定出其目标轨迹也经过了区域 B 和区域 C.更严重地,若 Alice 对 B 和 C 中的点的位置分布进行更深入的分析,可能进一步得出更有用的信息.而在实际的应用中,为了保证数据的有效性,A,B,C,D 和 E 这 5 个区域必然不会太大,因而,Alice 得到的信息也将是有效的信息.

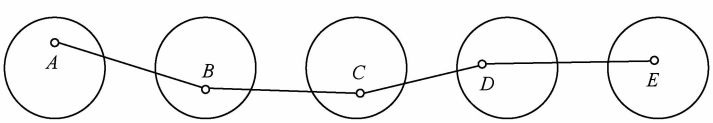


Fig. 6 Trajectory clustering.

图 6 轨迹聚类

通过以上的分析,可知恶意用户无需知道原始轨迹数据集,只需知道轨迹在匿名发布之前的聚类组信息,就可能进一步发现用户的个人隐私.现实的问题是恶意用户是否能够从已发布的数据中发现原始的聚类分组呢?本文研究的目的就是说明这种威胁性的存在,并给出相应的预防措施.以下先给出二次聚类攻击的定义.

定义 6. 二次聚类攻击.对于经过基于聚类和隐匿技术处理后的轨迹数据,利用原来的聚类算法和/或相同的聚类参数进行二次聚类,可能发现原始数据在发布之前的聚类组特征,从而得到用户有用信息的攻击模式称为二次聚类攻击.

以下将分别从基于点扰乱、基于边扰乱和基于特征发布3个方面来分析从已发布的轨迹数据中发现原始聚类分组的可能性.

3.2 基于点扰乱的二次聚类攻击

在某个给定的时刻 t,给定一个以点(x,y)为圆心,以 δ 为半径的不确定区域 $Circle(x,y,\delta)$,现采用如图3(a)所示的基于点扰乱的技术将不确定区域 $Circle(x,y,\delta)$ 外的点(x_r,y_r)随机移入 $Circle(x,y,\delta)$后得到点(x',y'). 为了说明二次聚类攻击存在的可能性,以下先假设在海量数据背景下不确定区域 $Circle(x,y,\delta)$中的点的分布是相对均匀的,由此有如下性质1成立.

性质 1. 对于不确定区域 $Circle(x,y,\delta)$中的任意一个点(x_1,y_1),(x_1,y_1)到(x',y')的距离小于(x_1,y_1)到(x_r,y_r)的距离的概率大于(x_1,y_1)到(x',y')的距离大于(x_1,y_1)到(x_r,y_r)的距离的概率.

证明. 如图7所示,设(x',y')和(x_r,y_r)连线的中垂线是 $Line\_v$,它和 $Circle(x,y,\delta)$相交于点 i_left 和 i_right. 不妨假设(x',y')在 $Line\_v$ 的左侧,而(x_r,y_r)在 $Line\_v$ 的右侧,则处于 $Line\_v$ 左侧的点到(x',y')的距离必然小于到(x_r,y_r)的距离. 由于 $Dist((x,y),(x',y'))<Dist((x,y),(x_r,y_r))$,必然有圆心$(x,y)$处于 $Line\_v$ 的左侧. 也即 $Line\_v$ 和 $Circle(x,y,r)$相交形成的两个弓形中,左侧((x',y')所在一侧)的弓形的面积比右侧的弓形的面积要大. 从而有性质1成立. 证毕.

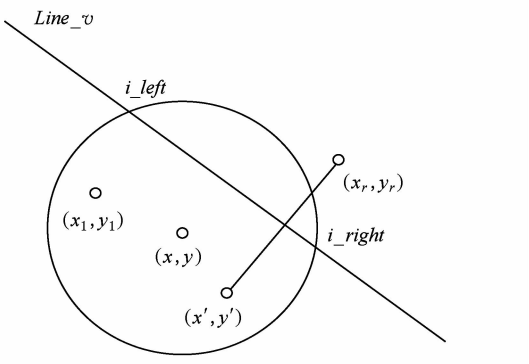


Fig. 7 Model 1 of points perturbation.

图7 点扰乱模型图1

性质1说明了在给定的不确定模型下以及本文假设前提下,移动离群点到不确定性区域之内,在更大概率上将减少原来所对应的离群点到不确定区域内的任意一个点的距离. 基于性质1,可得出如下结论:

结论 1. 假设不确定性区域 $Circle(x,y,\delta)$中的所有点的集合为$\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$,对于某个不确定区域外的点$(x_r,y_r)$,将它移入不确定区域后对应的点为$(x',y')$,则 $\sum_{i=1}^{n}Dist((x',y'),(x_i,y_i))$很有可能小于$\sum_{i=1}^{n}Dist((x_r,y_r),(x_i,y_i))$.

证明. 如图8所示,A 是 $Line\_v$ 的右侧区域. 根据性质1,$Line\_v$ 左侧区域比右侧区域大,因此,可在 $Line\_v$ 左侧做 A 的对称区域 B. C 为 $Line\_v$ 左侧扣除 B 后的区域. 当(x_r,y_r)被移入 $Circle(x,y,\delta)$中后,A 中的所有点到(x_r,y_r)的距离反而小于到(x',y')的距离. 而 B 中的所有点到(x_r,y_r)的距离大于到(x',y')的距离.

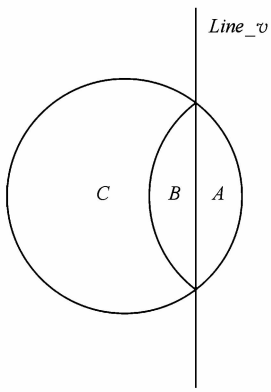


Fig. 8 Model 2 of points perturbation.

图8 点扰乱模型图2

如图9所示,在海量数据均匀分布的假设下,若弓形区域 A 中存在点 P_A,则可寻找 B 中和 P_A 基本对称的点 P_B,满足 $Dist(P_A,(x',y'))-Dist(P_A,(x_r,y_r))\approx Dist(P_B,(x_r,y_r))-Dist(P_B,(x',y'))$,即 A 导致的增加距离和 B 导致的减少距离可能相等或相近. 而 C 中的点到(x_r,y_r)的距离都大于到(x',y')的距离,从而极有可能导致

$$\sum_{i=1}^{n}Dist((x',y'),(x_i,y_i))<\sum_{i=1}^{n}Dist((x_r,y_r),(x_i,y_i)).$$

证毕.

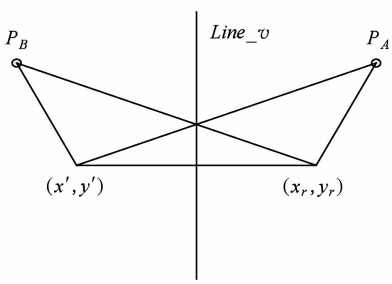


Fig. 9 Model 3 of points perturbation.

图9 点扰乱模型图3

特别需要指出的是,上述分析的目的是为了说明二次聚类攻击发生的可能性.在现实应用中,数据分布不可能如此理想化.然而,当不确定区域内的点数量足够多时,A 和 B 导致的距离变化有可能相对接近.即便不是接近的,C 和 B 导致的距离减少也很可能大于 A 导致的距离增加.

结论 1 是在给定时刻 t 的情况下考察每个离散采样时刻位置点之间的聚合情况.而在相同的模型和假设下,则可得出如下结论:

结论 2. 对于给定的轨迹数据集合 TR,其中任意一个聚类组为 $X:(tr_1,tr_2,\cdots,tr_n)$,假设 X 经过点扰乱后产生的对应的组为 $X':(tr'_1,tr'_2,\cdots,tr'_n)$.则对于任意的 $1\leqslant i_1<i_2\leqslant n$,有下述不等式成立:$Dist_{1\leqslant i_1<i_2\leqslant n}(X'.tr_{i_1},X'.tr_{i_2})\leqslant Dist_{1\leqslant i_1<i_2\leqslant n}(X.tr_{i_1},X.tr_{i_2})$.

证明. 利用结论 1 可知,对于任意给定的时刻 t,必然有下述不等式成立:$Dist_{1\leqslant i_1<i_2\leqslant n}(X'.tr'_{i_1}[t],X'.tr'_{i_2}[t])\leqslant Dist_{1\leqslant i_1<i_2\leqslant n}(X.tr_{i_1}[t],X.tr_{i_2}[t])$,进而,对所有时刻,有不等式 $\sum_t Dist_{1\leqslant i_1<i_2\leqslant n}(X'.tr'_{i_1}[t],X'.tr'_{i_2}[t])\leqslant\sum_t Dist_{1\leqslant i_1<i_2\leqslant n}(X.tr_{i_1}[t],X.tr_{i_2}[t])$ 成立,亦即 $Dist_{1\leqslant i_1<i_2\leqslant n}(X'.tr'_{i_1},X'.tr'_{i_2})\leqslant Dist_{1\leqslant i_1<i_2\leqslant n}(X.tr_{i_1},X.tr_{i_2})$ 成立.其中,等号仅当轨迹数据中不存在离群点时成立. 证毕.

结论 2 表明,在经过基于点的扰乱技术处理后,原始轨迹聚类组 X 中将有部分轨迹之间的距离更加接近,从而使扰乱后发布的轨迹在二次聚类时更容易聚合到一起.

以上是在一种合理假设的前提下分析采用点扰乱技术发布的轨迹数据中二次聚类攻击存在的可能性.从聚类的角度可以理解为:基于点扰乱的轨迹数据发布技术并没有破坏原始聚类组,由此发布的轨迹数据有高可用性,但同时很可能存在高隐私威胁.

3.3 基于边扰乱的二次聚类攻击

以下将从数学角度较为严格地分析基于边扰乱技术的轨迹聚类发布中二次聚类攻击存在的可能性.

根据定义 2,在不确定性模型下,可以将一个轨迹聚类组想象成为一系列移动的圆盘.在每个离散的时刻 t_i,每个轨迹聚类组对应于平面上的一个单独的圆盘.对于两个给定的轨迹聚类组 X 和 Y,假设它们在 t_i 时刻对应的圆盘分别是 $Circle(X,t_i)$ 和 $Circle(Y,t_i)$.不妨假设 $Circle(X,t_i)$ 的圆心 P_X 在 (x_1,y_1) 并且半径为 δ_1,而 $Circle(Y,t_i)$ 的原心 P_Y 在 (x_2,y_2) 并且半径为 δ_2,如图 10 所示,则有如下性质 2 成立.

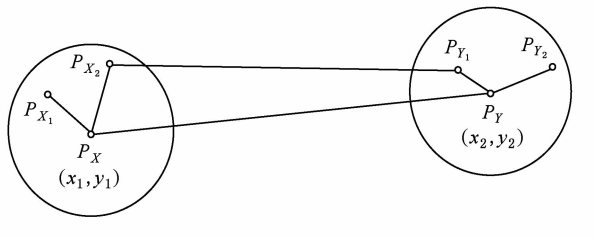


Fig. 10 Model 1 of edges Perturbation.

图 10 边扰乱模型图 1

性质 2. 对于 $Circle(X,t_i)$ 中的任意两个点 P_{X_1},P_{X_2} 和 $Circle(Y,t_i)$ 中的任意两个点 P_{Y_1},P_{Y_2},若 $Dist(P_X,P_Y)>\mathrm{Max}(3\delta_1+\delta_2,3\delta_2+\delta_1)$,则必然有如下式子成立:

$$\begin{cases}Dist(P_{X_1},P_{X_2})<Dist(P_{X_1},P_{Y_1}),\\ Dist(P_{X_1},P_{X_2})<Dist(P_{X_1},P_{Y_2}),\\ Dist(P_{X_1},P_{X_2})<Dist(P_{X_2},P_{Y_1}),\\ Dist(P_{X_1},P_{X_2})<Dist(P_{X_2},P_{Y_2}).\end{cases}\tag{1}$$

和

$$\begin{cases}Dist(P_{Y_1},P_{Y_2})<Dist(P_{X_1},P_{Y_1}),\\ Dist(P_{Y_1},P_{Y_2})<Dist(P_{X_1},P_{Y_2}),\\ Dist(P_{Y_1},P_{Y_2})<Dist(P_{X_2},P_{Y_1}),\\ Dist(P_{Y_1},P_{Y_2})<Dist(P_{X_2},P_{Y_2}).\end{cases}\tag{2}$$

式(1)中的第 3 个式子可证明如下:

$Dist(P_X,P_{X_2})+Dist(P_{X_2},P_{Y_1})+Dist(P_{Y1},P_Y)>Dist(P_X,P_Y)\rightarrow Dist(P_{X_2},P_{Y_1})>Dist(P_X,P_Y)-Dist(P_X,P_{X_2})-Dist(P_{Y1},P_Y)\rightarrow Dist(P_{X_2},P_{Y_1})>Dist(P_X,P_Y)-\delta_1-\delta_2\rightarrow Dist(P_{X_2},P_{Y_1})>\mathrm{Max}(2\delta_1,2\delta_2)>2\delta_1\geqslant Dist(P_{X_1},P_{X_2})$.

同样地,其余 7 个式子可以类似证明.

根据性质 2,可推出如下结论:

结论 3. 对于给定的聚类组 $X:(tr_1,tr_2,\cdots,tr_n)$ 以及聚类组 $Y:(tr_1,tr_2,\cdots,tr_m)$,经过扰乱后得到的对应聚类组为 $X':(tr'_1,tr'_2,\cdots,tr'_n)$ 和聚类组 $Y':(tr'_1,tr'_2,\cdots,tr'_m)$.则任选 X' 和 Y' 中的两条轨迹 $X'.tr'_{i_1}$,$X'.tr'_{i_2}$,$Y'.tr'_{i_3}$ 和 $Y'.tr'_{i_4}$.必然有:

$$\begin{cases}Dist(X'.tr'_{i_1},X'.tr'_{i_2})<Dist(X'.tr'_{i_1},Y'.tr'_{i_3}),\\ Dist(X'.tr'_{i_1},X'.tr'_{i_2})<Dist(X'.tr'_{i_1},Y'.tr'_{i_4}),\\ Dist(X'.tr'_{i_1},X'.tr'_{i_2})<Dist(X'.tr'_{i_2},Y'.tr'_{i_3}),\\ Dist(X'.tr'_{i_1},X'.tr'_{i_2})<Dist(X'.tr'_{i_2},Y'.tr'_{i_4}).\end{cases}\tag{3}$$

$$\begin{cases} Dist(Y'.tr'_{i_3}, Y'.tr'_{i_4}) < Dist(X'.tr'_{i_1}, Y'.tr'_{i_3}), \\ Dist(Y'.tr'_{i_3}, Y'.tr'_{i_4}) < Dist(X'.tr'_{i_1}, Y'.tr'_{i_4}), \\ Dist(Y'.tr'_{i_3}, Y'.tr'_{i_4}) < Dist(X'.tr'_{i_2}, Y'.tr'_{i_3}), \\ Dist(Y'.tr'_{i_3}, Y'.tr'_{i_4}) < Dist(X'.tr'_{i_2}, Y'.tr'_{i_4}). \end{cases} \tag{4}$$

下面以不等式(3)中的第1个不等式为例,说明上述8个式子是必然成立的.

根据性质2,$X'.tr'_{i_1}$ 和 $X'.tr'_{i_2}$ 在任意时刻的距离都比 $X'.tr'_{i_1}$ 与 $Y'.tr'_{i_3}$ 之间的距离要小,从而 $X'.tr'_{i_1}$ 和 $X'.tr'_{i_2}$ 必然是彼此更加接近,也更容易被聚到一起的轨迹.同样地,其他7个式子也是成立的.从而,对于给定的移动数据集 TR,当采用基于边扰乱的轨迹聚类发布技术时,若任意两个给定的聚类组 X 和 Y 满足 $Dist(P_X, P_Y) > \mathrm{Max}(3\delta_1+\delta_2, 3\delta_2+\delta_1)$ 时,则二次聚类攻击是必然可能存在的.

以上所有分析都是在 $Dist(P_X, P_Y) > \mathrm{Max}(3\delta_1+\delta_2, 3\delta_2+\delta_1)$ 这个条件下进行的.事实上,即便去掉 $Dist(P_X, P_Y) > \mathrm{Max}(3\delta_1+\delta_2, 3\delta_2+\delta_1)$ 这个假设前提,虽然无法用严格的数学推导来说明二次聚类攻击的存在性,但是依然可以合理怀疑基于边扰乱的轨迹聚类算法所发布的数据很可能遭受二次聚类攻击.如图11所示,假设 tr_{X_1} 是聚类组 X 中实际存在的一条轨迹,而 tr_X 和 tr_Y 分别是聚类组 X 和聚类组 Y 的特征(或聚类组的中心).由于 tr_{X_1} 是 X 中的轨迹,因此,一般可以认为 tr_{X_1} 到 tr_X 的距离小于 tr_{X_1} 到 tr_Y 的距离.虽然不能保证在每一个时刻 tr_{X_1} 到 tr_X 的距离一定比到 tr_Y 的距离小,但是,对于任意选定的 tr_{X_1},以及任意选定的时刻 t_1,可认为 tr_{X_1} 到 tr_X 的距离在期望上应该比 tr_{X_1} 到 tr_Y 的距离更小(否则,极有可能产生更好的聚类分组).同样地,经过边扰乱操作后所得到的新轨迹,其在每个时刻的位置点都来自原来隶属于 X 的轨迹.因此,这样构造出来的轨迹到 tr_X 的距离在期望上应该比到 tr_Y 的距离近.

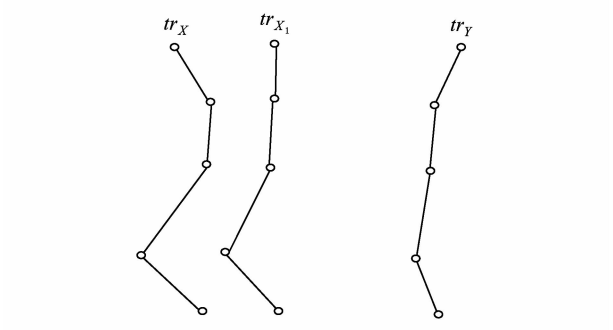


Fig. 11 Model 2 of edges perturbation.

图11 边扰乱模型图2

3.4 基于特征发布的二次聚类攻击

先前人们关于特征发布的研究工作更多地关注于如何保证 tr_X 能够代表聚类组 X 中足够多条的轨迹,而忽略了 tr_X 本身所可能带来的隐私威胁.事实上,基于特征发布的保护方式与概化方式类似,不需要经过二次聚类,就会有隐私泄露的威胁.对于给定的轨迹集合 TR,假设其中某个聚类组为 X:(tr_1, tr_2, …, tr_n),对应的特征轨迹为 tr_X.若恶意用户Alice已经知道 tr_1 中某些时刻的位置信息,则其可利用已知的 tr_1 的信息片段去匹配发布后的各个组特征.tr_X 必然是其中会被匹配到的一个.如果 tr_X 是唯一一个被匹配到的轨迹,则聚类组 X 的聚类特征 tr_X 将被检索到,从而可能导致聚类组的某些隐私信息泄露.

例如,在图10中,假设特征发布为 $P_X \rightarrow P_Y$,所代表的轨迹是 $\{P_{X_1} \rightarrow P_{Y_1}, P_{X_2} \rightarrow P_{Y_2}\}$.如果Alice知道第1条轨迹经过点 P_{X_1},通过特征 $P_X \rightarrow P_Y$,Alice极有可能知道第1条轨迹也通过以 P_{Y_2} 为圆心,r_2 为半径的某个区域.特别地,如果这个区域对于数据使用者是有意义的,对于Alice将是同样有用的.

4 基于聚类杂交的轨迹数据发布算法CH-TDP

4.1 聚类组特征保护的改进(k,δ,Δ)-匿名模型

通过之前在合理假设下的理论分析可以看出,传统的先聚类后匿名的算法只关注每条轨迹所对应个体的隐私保护,却忽视了轨迹聚类组特征的保护.本文以(k,δ)-匿名模型为基础,提出针对聚类组特征保护的改进(k,δ,Δ)-匿名模型.其中,Δ 是一个质量阈值.

(k,δ,Δ)-匿名模型的基本思想是对采用(k,δ)-匿名模型及相关算法处理得到的聚类分组先进行组间杂交,而后再进行组内扰乱.通过控制组间扰乱的程度,来达到既保护聚类组的共同特征,又避免具体的轨迹信息泄漏的目的.(k,δ,Δ)-匿名模型的目标是在预防发布轨迹数据遭受二次聚类攻击的前提下,保证发布轨迹数据的质量不低于质量阈值 Δ.以下将给出一个基于(k,δ,Δ)-匿名模型的轨迹聚类杂交发布算法.

4.2 算法描述

算法1. CH-TDP.

输入:原始轨迹数据,k,δ,Δ,聚类的度量函数;

输出:满足(k,δ,Δ)-匿名模型的轨迹发布数据.

1) 使用某个聚类方式和给定的度量函数,对原始轨迹进行聚类操作;

2) 设定初始杂交比例 $\alpha_l=0$,$\alpha_r=100$,$\alpha_{\text{mid}}=50$;

3) 若第 1)步得到 n 个聚类组$\{TR_1,TR_2,\cdots,TR_n\}$,则设定 N 个空的容器$\{S_1,S_2,\cdots,S_n\}$,对于任意的 i,将 TR_i 抽取$\alpha_{\text{mid}}\%$条轨迹,分别加入到 S_1,$S_2,\cdots,S_{i-1},S_{i+1},\cdots S_n$ 中;

4) 将$\{TR_1,TR_2,\cdots,TR_n\}$分别加入到$\{S_1,S_2,\cdots,S_n\}$中;

5) 检查$\{S_1,S_2,\cdots,S_n\}$是否满足给定的 Δ;

6) 若不满足 Δ,则 $\alpha_l=\alpha_{\text{mid}}$,$\alpha_{\text{mid}}=\dfrac{(\alpha_l+\alpha_r)}{2}$;否则 $\alpha_r=\alpha_{\text{mid}}$,$\alpha_{\text{mid}}=\dfrac{(\alpha_l+\alpha_r)}{2}$;

7) 若 α_{mid}已测试过,或者计算已达到一定的精度,则程序结束;

8) 返回第 3)步.

CH-TDP 算法的核心思想是对于已经得到的聚类组先进行组间扰乱,而后再进行组内扰乱.组内扰乱的具体技术可以引用之前的工作的思想,而本文主要探讨如何有效进行组间扰乱.为了实现组间扰乱,CH-TDP 算法让每个聚类组参杂来自其他聚类组的轨迹.正如算法第 3 步所示,最终扰乱后的每个聚类组,都包含来自每个原始聚类组的轨迹.而这个扰乱的程度是通过 $\alpha_{\text{mid}}\%$来控制.很明显,当 $\alpha_{\text{mid}}\to 0$ 时,聚类组之间几乎没有扰乱,质量最高.而当 $\alpha_{\text{mid}}\to 100$ 时,扰乱后的聚类组最安全.因此,在给定的质量阈值下,可采取二分搜索的框架,来寻找合适的 $\alpha_{\text{mid}}\%$.

采用这种基于二分杂交的策略,可找到合适的杂交比率 α_{mid},使得每个聚类分组在进行匿名化操作之前既保留了本身的特征,又具有整个数据集的特征,以达到预防二次聚类针对聚类组特征的攻击.第 5 节将通过大量仿真实验和数据分析来说明(k,δ,Δ)-匿名模型及 CH-TDP 算法的有效性.

5 实验结果与分析

本节将从隐私保护有效性和匿名数据的质量两个方面进行实验研究.本文 CH-TDP 算法的比对对象是在之前工作中具有代表性的 NWA 算法.首先,对 CH-TDP 算法和 NWA 算法以及使用了不同聚类策略的 NWA 算法得到的匿名轨迹分别进行二次聚类,并分别比较所得聚类组与原聚类组的相似度,以此来验证传统基于聚类的轨迹发布算法所生成的轨迹数据存在被二次聚类攻击的可能性,同时也说明本文的 CH-TDP 算法能够有效抵御二次聚类攻击;然后,从轨迹相似度、区域查询结果和频繁项查询结果 3 个方面体现本文算法所发布的轨迹数据具有较高的质量.需要说明的是,下文所注的 CH-TDP 算法实际上是其单次运行实例,并不是多次二分搜索的结果,通过 CH-TDP 的单次结果来和之前的算法进行比较.

5.1 实验数据和环境

实验数据由 Brinkhoff 基于网络的移动对象数据生成器[13]生成,生成的数据表示由德国奥尔登堡(Oldenburg)市的道路网络状况一天的移动数据,该数据生成器可以在 http://www.fh-oow.de/institute/iapg/personen/brinkhoff/generator/网址中下载.实验中使用了两组数据集,OLDEN 由 100 000 条轨迹组成,且是文献[6]的标准数据集.而 OLDEN2 是随机生成的,由 21 000 条轨迹组成.表 1 为两个数据集的统计信息.其中 $MBB\ radius(D)$ 表示数据集 D 中最小覆盖矩形对角线的一半,$|D|$ 表示轨迹条数,$|D_{\text{pre}}|$ 表示预处理后等价类的个数,$MaxTime$ 表示数据集中的最大时间间隔,$Size$ 表示数据的大小.实验环境为:Intel® Pentium® Dual E2200@2.20 GHz;2 GB 内存;Windows XP 操作系统;算法由 Visual C++ 6.0 编写.

Table 1 Dataset Statistics

表 1 数据集的统计信息

Dataset	$MBB\ radius(D)$	$\lvert D\rvert$	$\lvert D_{\text{pre}}\rvert$	$MaxTime$ /min	$Size$/Mb
OLDEN	35 779.3	100 000	435	141	350
OLDEN2	35 779.3	21 000	15	201	15.8

5.2 隐私保护有效性

实验采用文献[5]提及的两种聚类算法,第 1 种聚类方法和 NWA 算法相同,即贪心聚类方法,每次贪心的选取距离最近的 $k-1$ 条轨迹.第 2 种采用层次聚类方法中的自底向上聚类方法,每次选取距离最近的两个类组合成新的类,新的类中轨迹条数要求小于 $2k$.

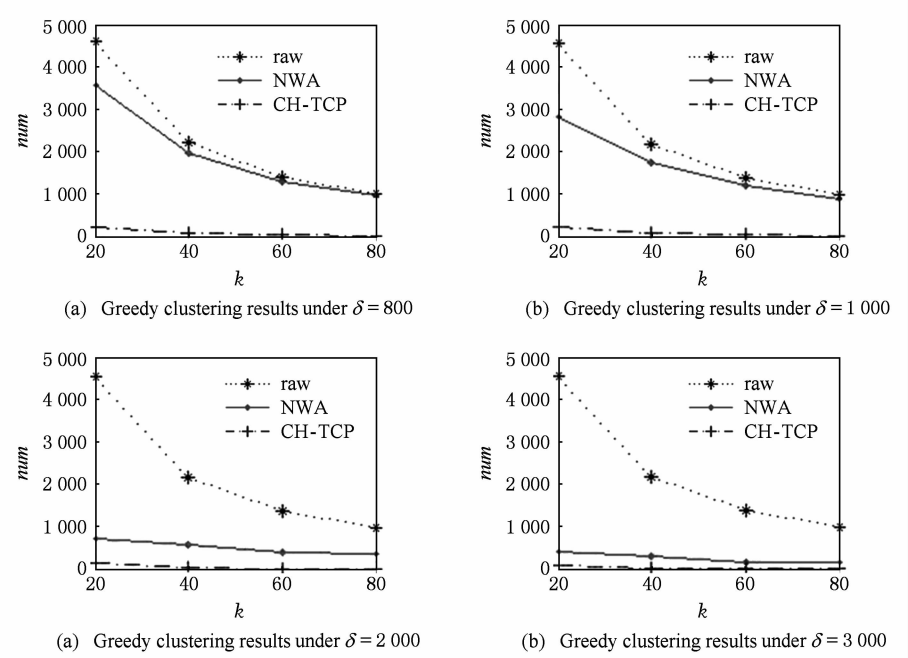
实验过程中,先使用 NWA 算法将数据集处理成匿名数据集 D',计算数据集 D' 中每一个等价类的聚类中心 rep',并对匿名数据集 D' 使用相同的方法进行二次聚类得到匿名数据集 D'',计算数据集 D'' 中每一个等价类的聚类中心 rep''.如果聚类中心

上的两个对应点之间的欧几里德距离小于等于给定的值 β,那么认为这两个点是相近的,如果两条轨迹上对应点相近的比例超过 π,那么认为这两条特征轨迹是相似的.实验采用二分图匹配[14]算法计算出攻击者可以从 D''数据集中重新聚类出多少条特征轨迹与数据集 D'相似.

同样地,使用 CH-TDP 算法将数据集处理成匿名数据集 D^+,计算数据集 D^+ 中每一个等价类的聚类中心 rep^+,并对匿名数据集 D^+ 进行二次聚类得到匿名数据集 D^{++},计算数据集 D^{++} 中每一个等价类的聚类中心 rep^{++}. 而后亦使用二分图匹配算法处理. 此后,重复之前的实验过程,唯一不同的是,比较对象成为了层次聚类方法和 CH-TDP.

首先,设定参数为 $\beta=100$,$\pi=90\%$,杂交系数 $\alpha=20\%$,在 k 和 δ 变化的情况下分别在 2 个数据集上进行实验.

图 12 和图 13 中的 NWA 标明的曲线分别为对 OLDEN 数据集在贪心聚类以及自底向上聚类后的

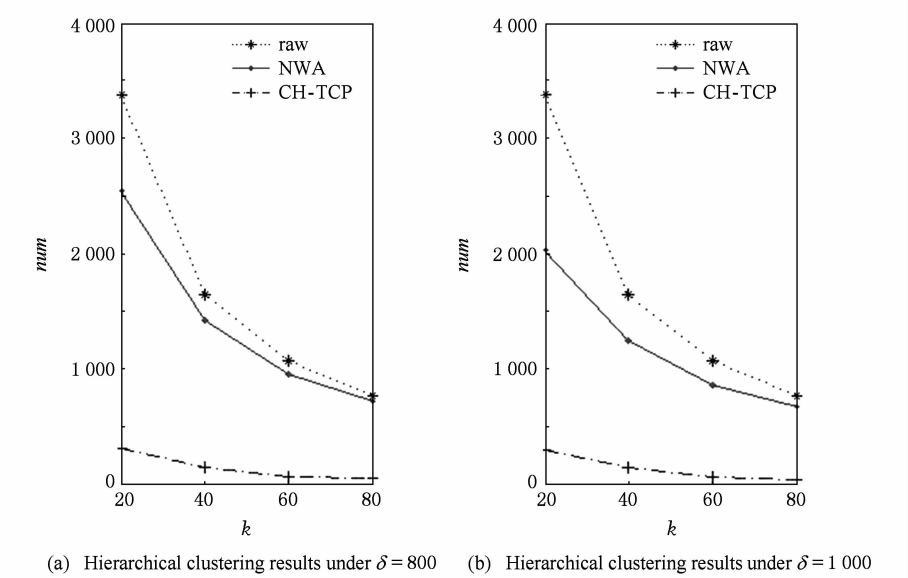


(a) Greedy clustering results under $\delta=800$ (b) Greedy clustering results under $\delta=1\,000$

(a) Greedy clustering results under $\delta=2\,000$ (b) Greedy clustering results under $\delta=3\,000$

Fig. 12 Greedy clustering results with respect to OLDEN.

图 12 关于 OLDEN 数据集的贪心聚类结果



(a) Hierarchical clustering results under $\delta=800$ (b) Hierarchical clustering results under $\delta=1\,000$

Fig. 13 Hierarchical clustering results with respect to OLDEN.

图 13 关于 OLDEN 数据集的自底向上聚类结果

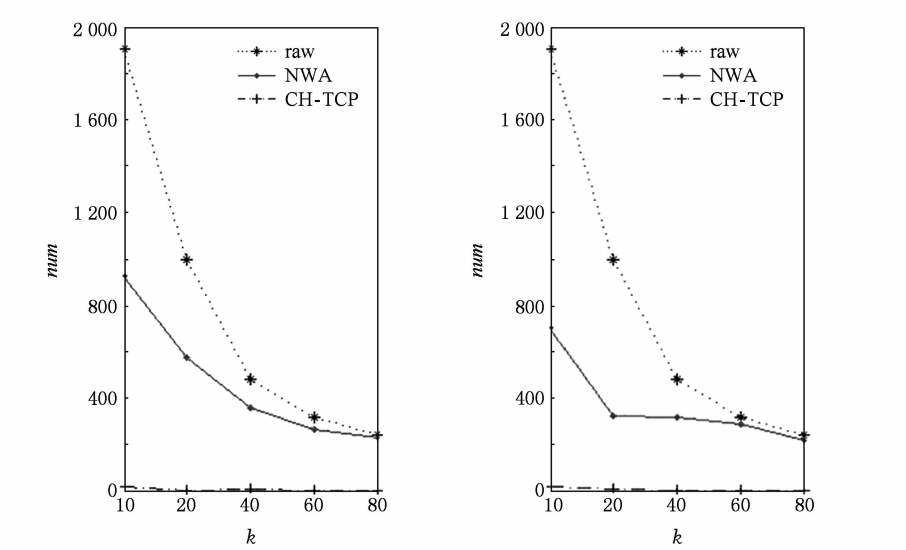
扰乱数据进行二次聚类所得的特征轨迹条数. 而raw标明的曲线为原始数据中进行聚类,但不扰乱的特征轨迹条数, CH-TDP标明的曲线为用CH-TDP算法处理OLDEN后亦进行二次聚类所得数据的特征轨迹条数. 可以看出,CH-TDP标明的曲线中特征轨迹条数明显较少. 随着k值的增加,聚类组逐渐减少,3条轨迹也都呈现出下降趋势. 而CH-TDP算法则明显呈现出对于二次聚类攻击的抵御.

图14和图15的实验和图12与图13类似,只是使用了OLDEN2数据集. 其实验结果也同之前类似,很好说明了CH-TDP算法对于二次聚类攻击的抵御.

接下来,在这一部分设置的参数为$\delta=1\,000$, $\pi=90\%$,也在k和α变化的情况下进行实验比较.

图16和图17分别是在OLDEN和OLDEN2上使用不同杂交比例α进行的实验. 可以看到,在给定的α变化范围内,两组图中CH-TDP标明的曲线几乎都接近于0水平线. 因而,实验结果表明了CH-TDP算法对于聚类组特征的有效保护.

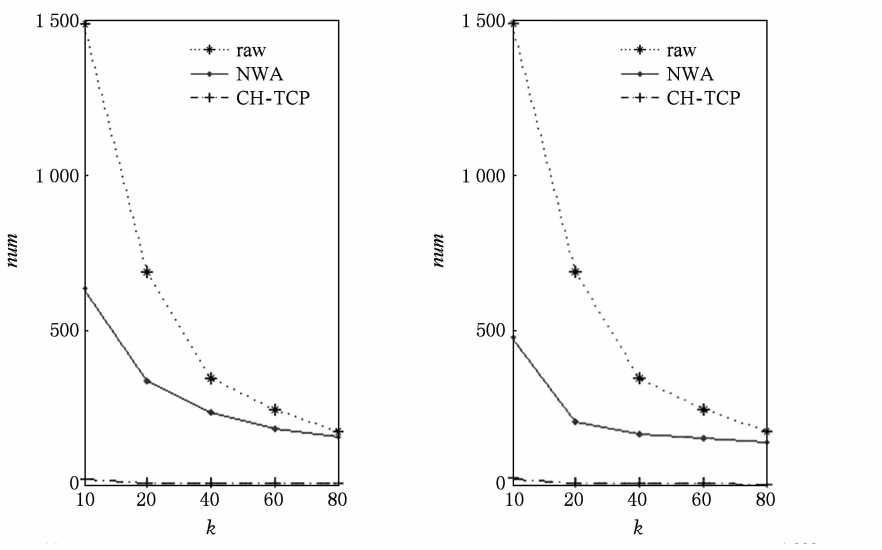
以上实验可以有效说明本文CH-TDP算法有效地抵御了二次聚类攻击.



(a) Greedy clustering results under $\delta=800$ (b) Greedy clustering results under $\delta=1\,000$

Fig. 14 Greedy clustering results with respect to OLDEN2.

图14 关于OLDEN2数据集的贪心聚类结果



(a) Hierarchical clustering results under $\delta=800$ (b) Hierarchical clustering results under $\delta=1\,000$

Fig. 15 Hierarchical clustering results with respect to OLDEN2.

图15 关于OLDEN2数据集的自底向上聚类结果

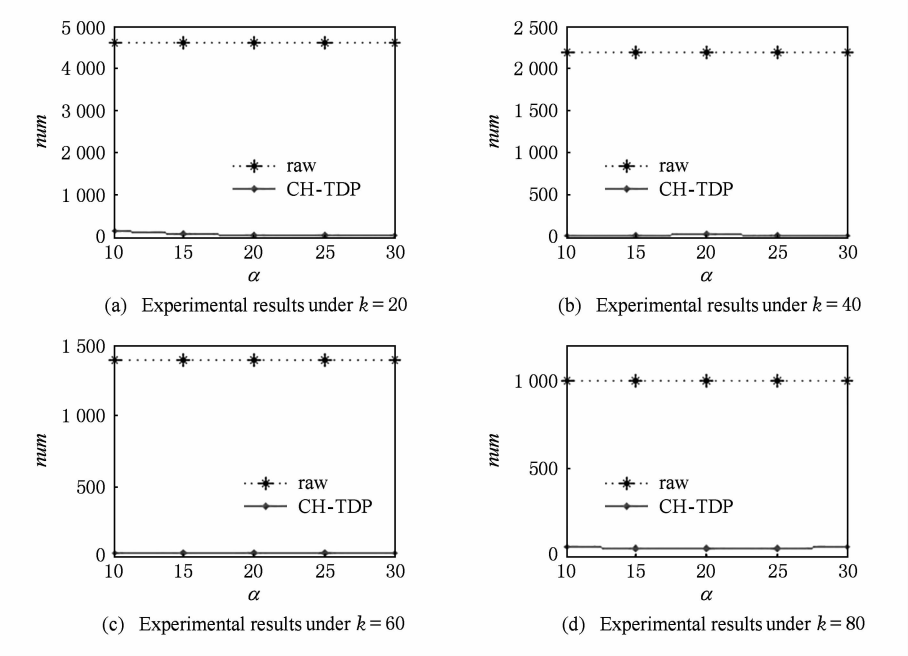


Fig. 16 Experimental results with respect to α under OLDEN.

图 16 OLDEN 数据集在不同杂交比例的实验结果

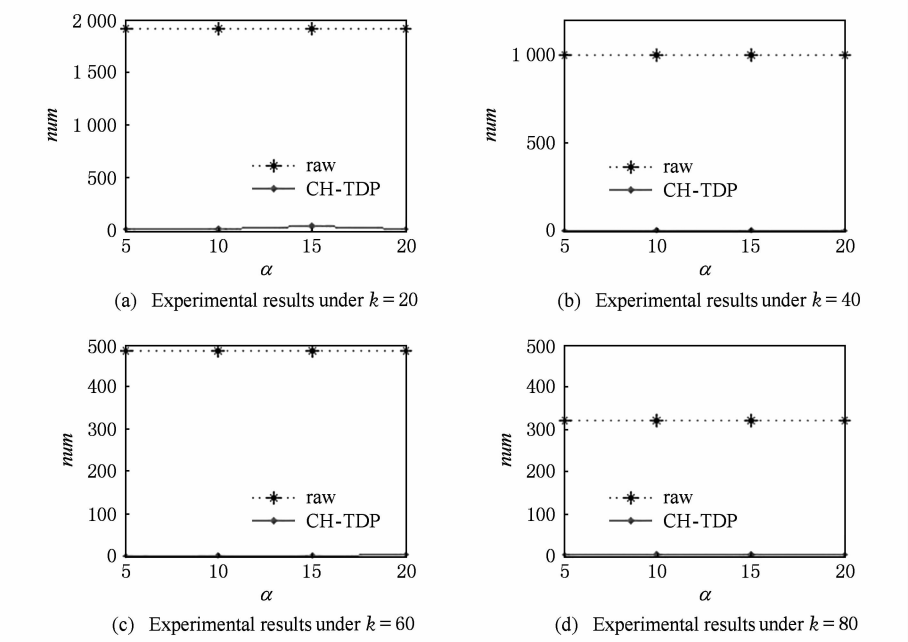


Fig. 17 Experimental results with respect to α under OLDEN2.

图 17 OLDEN2 数据集在不同杂交比例的实验结果

5.3 匿名数据质量

5.3.1 轨迹相似度比较

先分别采用 NWA 算法和 CH-TDP 算法对给定的数据集 D 进行处理,得到匿名轨迹数据集 D' 和 D^+. 而后,分别计算 NWA 算法和 CH-TDP 算法所得到的匿名数据集的匹配率. 设定如果平面区域上两个点之间的距离不大于给定阈值 β,就认为这两个位置点在扰乱之后是相近的,可能表示同样的位置信息. 而对于来自 D'(或 D^+)中的任意两条轨迹,如果该两条轨迹之间相近点的比例超过给定的另一阈值 π,就认为这两条轨迹是匹配的. 本文在实验中将使用经典最大二分图匹配算法来分别找出 D 和 D',以及 D 和 D^+ 之间的最大匹配. 如果两个最大匹配在数量上十分接近,就可以从轨迹相似度的角度来说明 CH-TDP 算法对于原始数据 D 的破坏被控制在有效范围内. 在这一部分实验中,参数设置

如下所示:$\beta=\delta$,$\pi=90\%$,杂交系数 $\alpha=10\%$.

图 18 和图 19 分别为在 OLDEN 和 OLDEN2 上进行的轨迹相似度实验.根据实验,代表 NWA 经典算法和 CH-TDP 算法的两条曲线在多个参数设置的情况下都相当接近.从而,本实验从一定程度上说明了 CH-TDP 算法所产生数据的有效性.

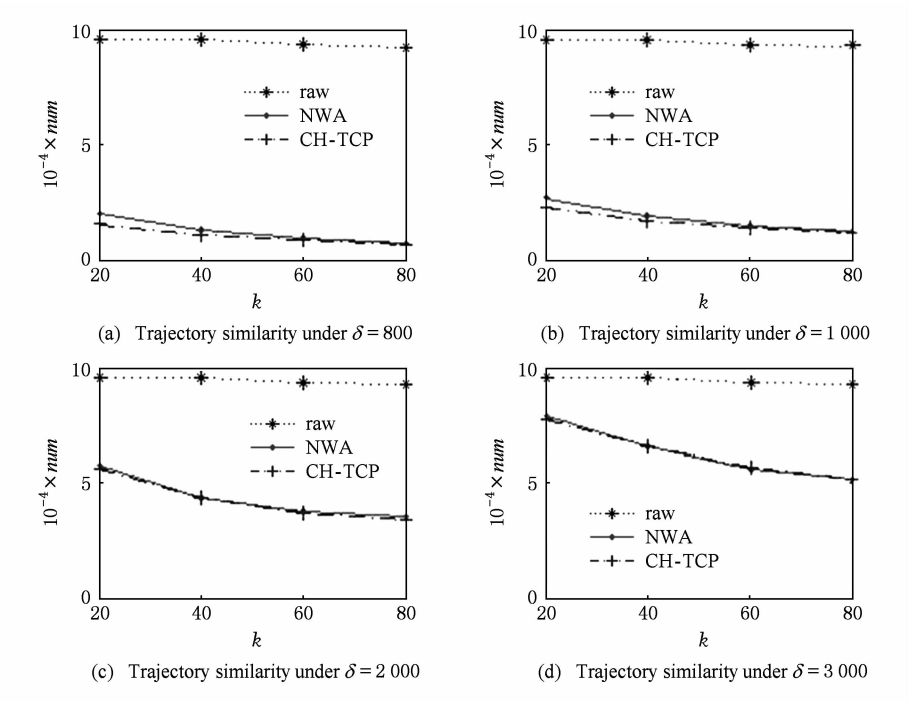


Fig. 18 Trajectory similarity with respect to OLDEN.

图 18 关于 OLDEN 数据集的轨迹相似度

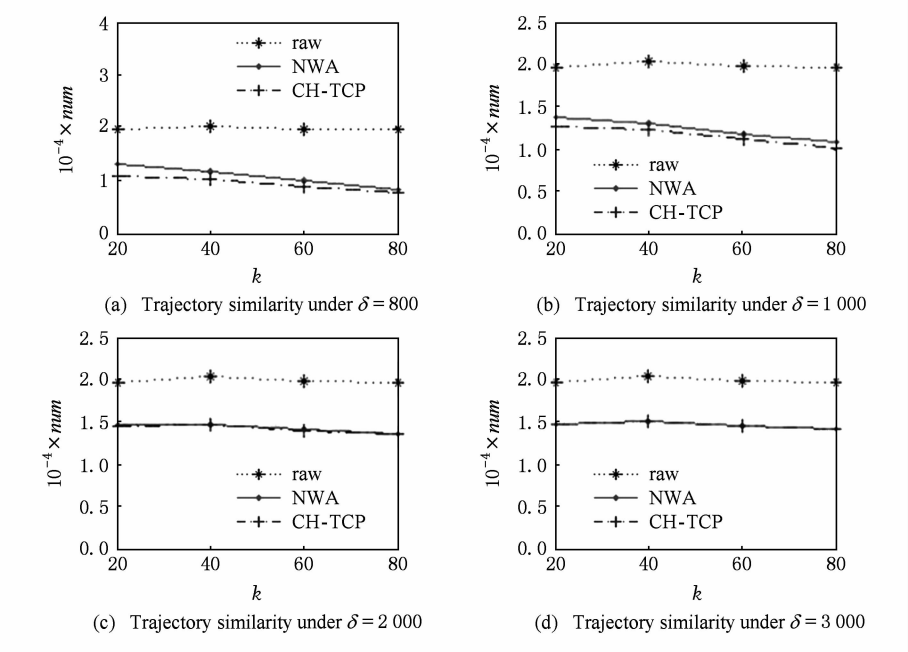


Fig. 19 Trajectory similarity with respect to OLDEN2.

图 19 关于 OLDEN2 数据集的轨迹相似度

5.3.2 区域查询结果比较

发布数据的目的是为了查询和分析,所以度量轨迹数据有用性的最好的方法之一是比较不同算法所发布数据集的区域查询结果.这里采用与文献[12]相同的基于时空范围的不确定性查询.在实验过程中,参数设置为 $\beta=\delta$,杂交系数 $\alpha=20\%$.

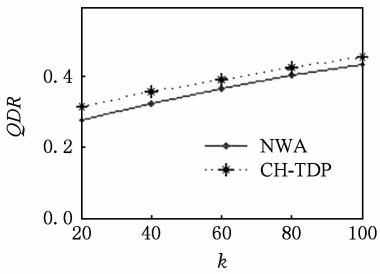
在文献[12]中,定义了在给定的时间范围[t_b,t_e]内,移动对象 tr 和区域 R 之间的位置关系.特别

的,我们关心是否有关系 $inside(R,tr)$(即轨迹 tr 是否在区域 R 之中).因为移动对象的位置是连续变化的,人们经常要询问在$[t_b,t_e]$时间范围内有时(sometime)还是一直(always)都存在 $inside(R,tr)$ 关系.如果存在一些可能的移动轨迹 tr 在时刻 t 时在区域 R 内,那么该移动对象就有可能将 tr 作为它实际的移动轨迹,并且可能在时间 t 时在区域 R 内.然而,移动对象可能选择其他的运行轨迹作为它实际的移动轨迹.同样的,如果每一条可能的移动轨迹 tr 在时刻 t 时都在区域 R 内,那么可以保证该对象在时间 t 时一定在区域 R 内.这里考虑两个区域查询的度量函数:

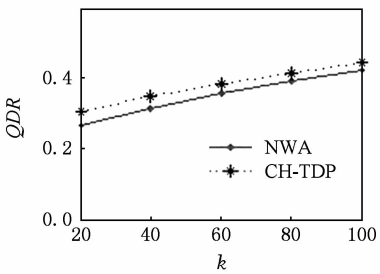
$$Q_1: Possibly\_Sometime\_Inside(\tau,R,t_b,t_e) \equiv (\exists tr_\tau)(\exists t \in [t_b,t_e])inside(R,tr_\tau(t),t),$$

$$Q_2: Definitely\_Always\_Inside(\tau,R,t_b,t_e) \equiv (\forall tr_\tau)(\forall t \in [t_b,t_e])inside(R,tr_\tau(t),t).$$

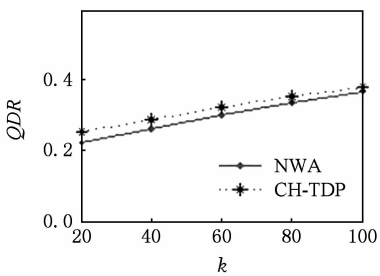
定义查询误差率(query distortion rate)$QDR=|Q(D)-Q(D')|/\max(Q(D),Q(D'))$.随机生成半径为 500,1 000,2 000,3 000,4 000,5 000 的圆各 1 000 组查询数据,查询数据的时间范围为随机选取的时



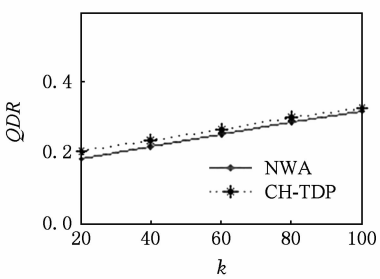
(a) Query distortion rate Q_1 under $\delta=800$



(b) Query distortion rate Q_1 under $\delta=1\,000$



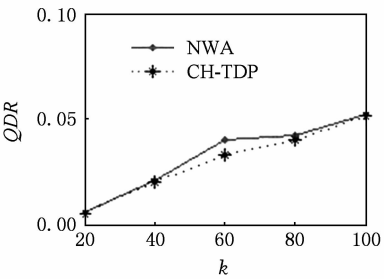
(c) Query distortion rate Q_1 under $\delta=2\,000$



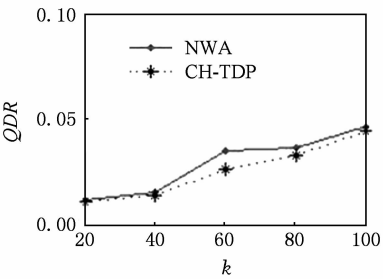
(d) Query distortion rate Q_1 under $\delta=3\,000$

Fig. 20 Query distortion rate Q_1 with respect to OLDEN.

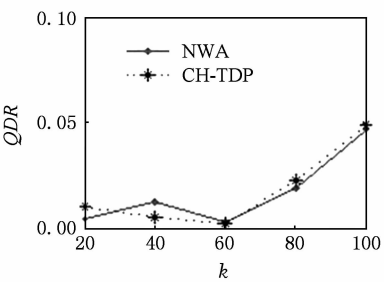
图 20 关于 OLDEN 数据集的 Q_1 查询误差率



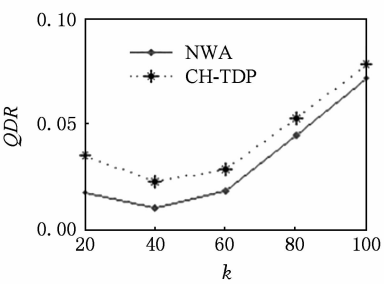
(a) Query distortion rate Q_2 under $\delta=800$



(b) Query distortion rate Q_2 under $\delta=1\,000$



(c) Query distortion rate Q_2 under $\delta=2\,000$



(d) Query distortion rate Q_2 under $\delta=3\,000$

Fig. 21 Query distortion rate Q_2 with respect to OLDEN.

图 21 关于 OLDEN 数据集的 Q_2 查询误差率

间间隔$[t_b,t_e]$,并在数据集最大时间间隔范围之内.

图 20 和图 21 分别为针对 OLDEN 上的 Q_1 和 Q_2 查询误差率实验.对于利用 NWA 算法和 CH-TDP 算法处理 OLDEN 后得到的数据,分别进行之前所述的 Q_1 和 Q_2 查询实验.可以看出,代表 NWA 和 CH-TDP 形成的两条曲线相当接近,从而从查询误差率角度说明了 CH-TDP 算法的有效性.

图 22 和图 23 分别为针对 OLDEN2 上的 Q_1 和 Q_2 查询误差率实验.其实验结果和 OLDEN 上的 Q_1 和 Q_2 查询误差率实验结果类似,也说明了CH-TDP

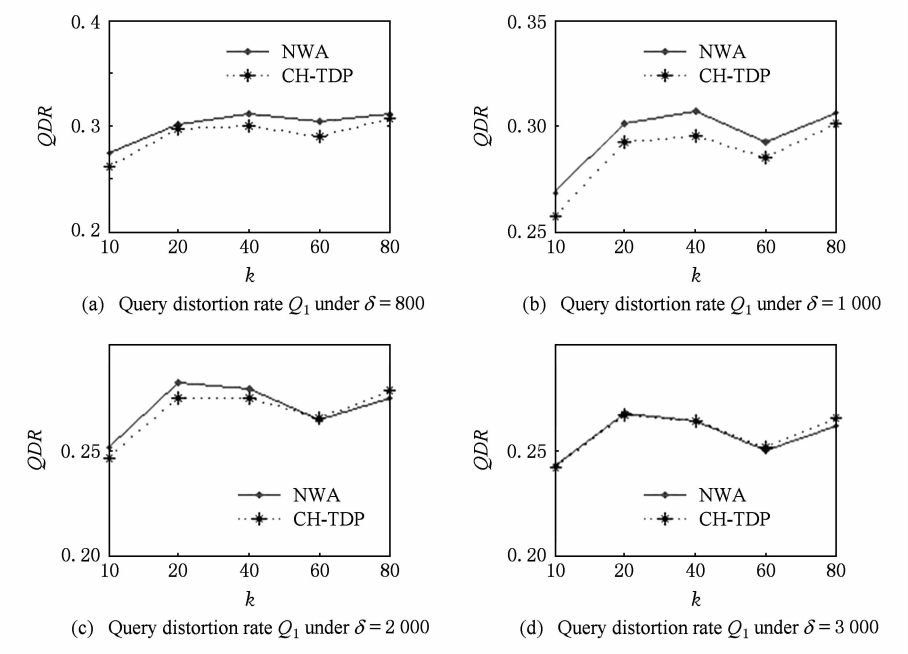


Fig. 22 Query distortion rate Q_1 with respect to OLDEN2.

图 22 关于 OLDEN2 数据集的 Q_1 查询误差率

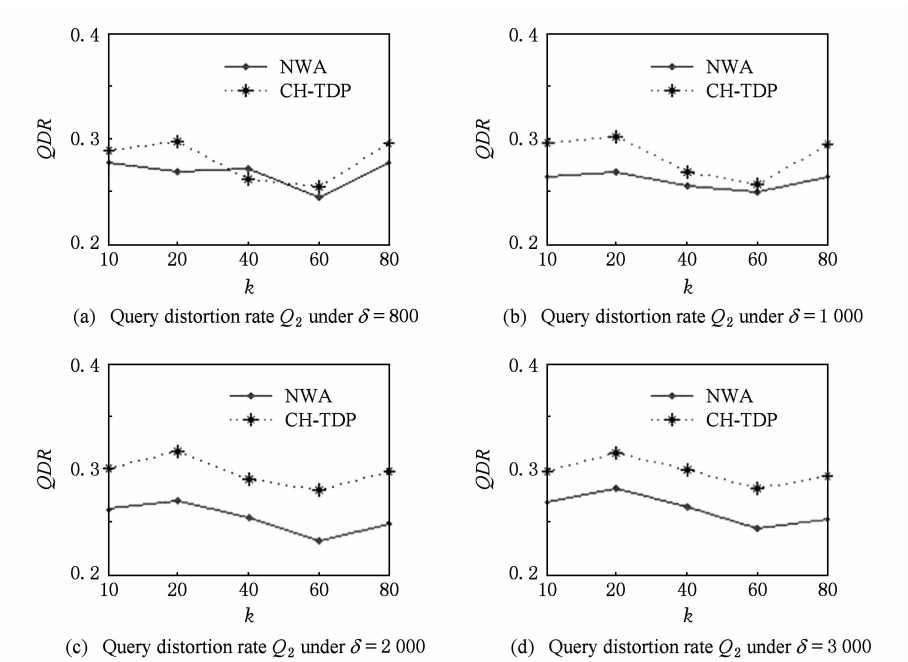


Fig. 23 Query distortion rate Q_2 with respect to OLDEN2.

图 23 关于 OLDEN2 数据集的 Q_2 查询误差率

算法的有效性.

5.3.3 频繁项查询结果比较

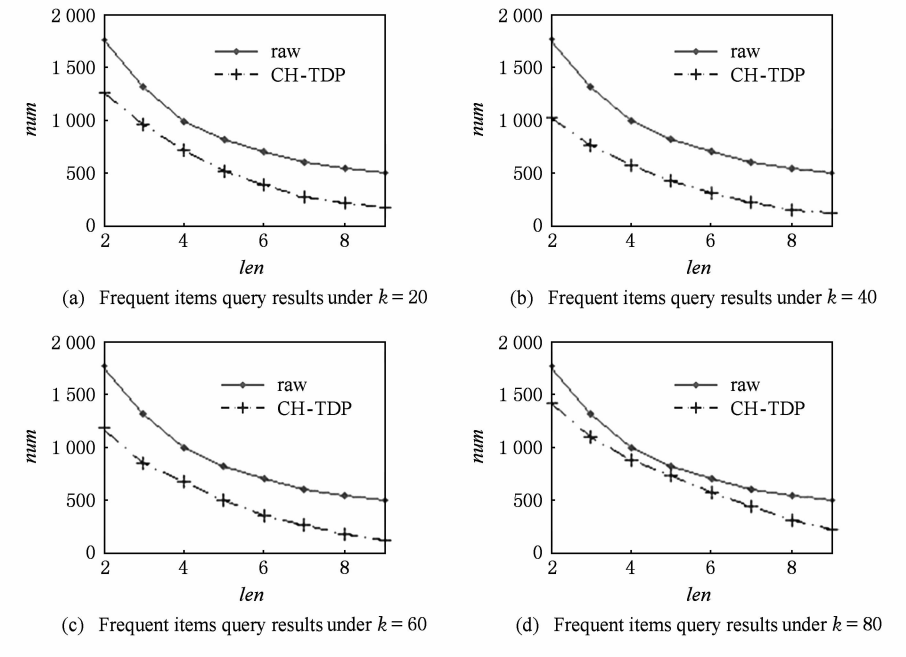
在实际应用中,轨迹数据可能被用于决策制定或数据挖掘.这要求发布后的轨迹数据仍然能够进行频繁项挖掘.一般而言,短的项往往频繁,并且经常是决策挖掘的对象,例如像 $A\rightarrow B$,$A\rightarrow B\rightarrow C$ 这样

的连续或者不连续的轨迹片段. 而长的项,例如 $A\rightarrow B\rightarrow C\rightarrow D\rightarrow E\rightarrow F\rightarrow G$ 往往能使某条或者某些少量轨迹同其他轨迹区别开来,往往属于必须被保护的隐私范畴. 本节从频繁项(或长短项查询)的角度,来分析 CH-TDP 算法所产生的匿名数据的有效性. 希望通过实验来说明如下事实:

当使用短查询时,CH-TDP 算法所产生的数据依然是很好的数据查询的对象. 而当使用长查询时,CH-TDP 算法所产生的数据很难提供有效查询. 即便查询到一些轨迹,也是不可信和不可靠的轨迹.

实验中分别随机生成长度为 2～10 的轨迹各 100 条,并使用这些轨迹在数据集中进行查询. 在这一部分设置的参数为 $\beta=500$, $\pi=90\%$, 杂交系数 $\alpha=20\%$, $\delta=1\,000$.

图 24 和图 25 分别为在 OLDEN 和 OLDEN2 上的频繁项查询实验结果. 从图中可以看出,随着查

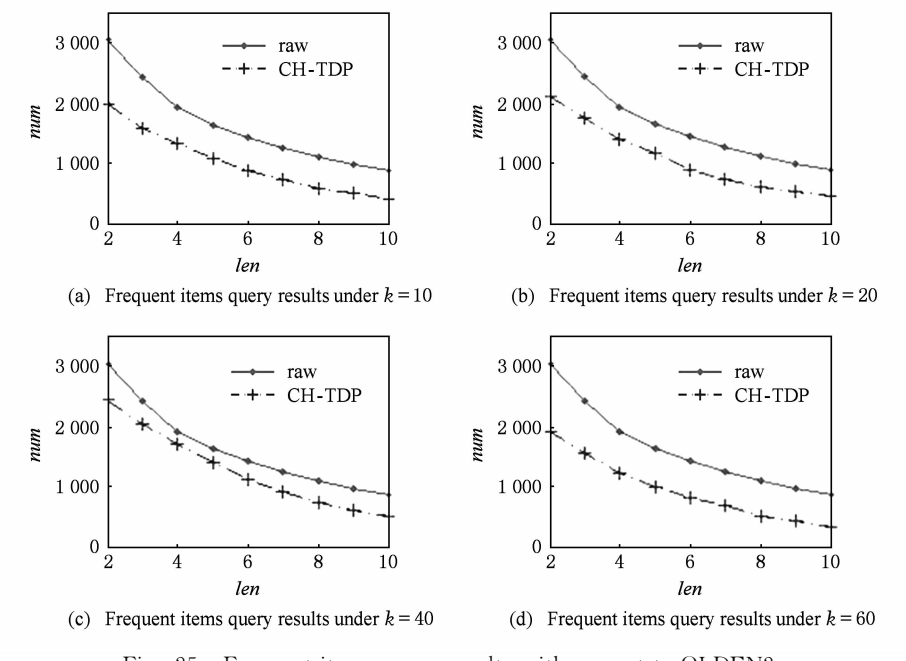


(a) Frequent items query results under $k=20$ (b) Frequent items query results under $k=40$

(c) Frequent items query results under $k=60$ (d) Frequent items query results under $k=80$

Fig. 24 Frequent items query results with respect to OLDEN.

图 24 关于 OLDEN 数据集的频繁项查询结果



(a) Frequent items query results under $k=10$ (b) Frequent items query results under $k=20$

(c) Frequent items query results under $k=40$ (d) Frequent items query results under $k=60$

Fig. 25 Frequent items query results with respect to OLDEN2.

图 25 关于 OLDEN2 数据集的频繁项查询结果

询长度的增加,raw 标明的曲线和 CH-TDP 标明的曲线分别呈现下降趋势.而 CH-TDP 算法在短查询(例如长度为 2,3 或 4)时,依然能够查出大量轨迹.而在长查询(例如长度为 8,9 或 10)时,查出的轨迹数量十分稀少,基本和之前的实验期望符合.

6 结论与进一步工作

本文针对传统基于聚类的轨迹数据发布算法只关注单条轨迹的隐私而忽视对轨迹聚类组特征保护的不足,发现轨迹数据聚类发布后可能存在二次聚类攻击,并提出抵御二次聚类攻击的(k,δ,Δ)-匿名模型和基于该模型的聚类杂交隐私保护轨迹数据发布算法.理论分析和仿真实验结果表明,本文的模型及算法是有效可行的.在将来的研究工作中,我们将进一步针对其他轨迹距离度量和聚类算法进行实验研究,以期设计出更合理有效的隐私保护轨迹数据发布模型及算法.

参 考 文 献

[1] Samarati P. Protecting respondent's identities in microdata release [J]. IEEE Trans on Knowledge and Data Engineering, 2001, 13(6): 1010-1027

[2] Sweeney L. K-anonymity: A model for protecting privacy [J]. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 2002, 10(5): 557-570

[3] Michael B, Tom Z J. A face is exposed for aol searcher no. 4417749 [N]. New York Times, 2006-08-09(8)

[4] Zhou Shuigeng, Li Feng, Tao Yufei, et al. Privacy preservation in database applications: A survey [J]. Chinese Journal of Computers, 2009, 32(5): 847-858 (in Chinese)
(周水庚, 李丰, 陶宇飞, 等. 面向数据库应用的隐私保护研究综述[J]. 计算机学报, 2009, 32(5): 847-858)

[5] Abul O, Bonchi F, Nanni M. Never walk alone: Trajectory anonymity via clustering, ISTI-007/2007 [R]. Pisa: Institute of Information Science and Technologies (ISTI), Italian National Research Council (CNR), 2007

[6] Abul O, Bonchi F, Nanni M. Never walk alone: Uncertainty for anonymity in moving objects databases [C] //Proc of IEEE ICDE'08. Piscataway, NJ: IEEE, 2008: 376-385

[7] Saygin Y, Nergiz E, Atzori M. Towards trajectory anonymization: A generalization-based approach [C] //Proc of the SIGSPATIAL ACM GIS 2008 Int Workshop on Security and Privacy in GIS and LBS. New York: ACM, 2008: 52-61

[8] Yarovoy R, Bonchi F, Lakshmanan V S, et al. Anonymizing moving objects: How to hide a MOB in a crowd? [C] //Proc of the 12th Int Conf on Extending Database Technology: Advances in Database Technology. New York: ACM, 2009: 23-26

[9] Lin Dan, Gurung S, Jiang Wei, et al. Privacy-preserving location publishing under road-network constraints [G] // LNCS 5982: Proc of the 15th Int Conf on Database Systems for Advanced Applications. Berlin: Springer, 2010: 17-31

[10] Abul O, Bonchi F, Nanni M. Anonymization of moving objects databases by clustering and perturbation [J]. Information Systems, 2010, 35(8): 884-910

[11] Chen L, Özsu M T, Oria V. Robust and fast similarity search for moving object trajectories [C] //Proc of ACM SIGMOD'05. New York: ACM, 2005: 491-502

[12] Trajcevski G, Wolfson O, Hinrichs K, et al. Managing uncertainty in moving objects databases [J]. ACM Trans on Database Systems, 2004, 29(3): 463-507

[13] Brinkhoff T. Generating traffic data [J]. IEEE Data Engineering Bulletin, 2003, 26(2): 19-25

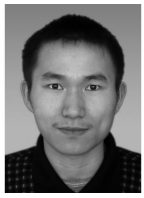
[14] Lovasz L, Plummer M D. Matching Theory [M]. New York: North-Holland, 1986: 1-40

Wu Yingjie, born in 1979. PhD, associate professor. His main research interests include data mining and privacy preserving.

Tang Qingming, born in 1985. Master candidate. His main research interests include data mining and privacy preserving (tqm2004@gmail.com).

Ni Weiwei, born in 1979. PhD and associate professor. His current research interests include data mining and data privacy preserving(wni@seu.edu.cn).

Sun Zhihui, born in 1941. Professor and PhD supervisor of Southeast University. Senior member of China Computer Federation. His main research interests include data mining and complicated information system integration(sunzh@seu.edu.cn).



Liao Shangbin, born in 1986. Master candidate. His main research interests include data mining and privacy preserving (liaoshangbin@gmail.com).