软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software, 2012,23(2):352−360 [doi: 10.3724/SP.J.1001.2012.03946]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

位置服务中用户轨迹的隐私度量*

王彩梅, 郭亚军+, 郭艳华

(华中师范大学 计算机科学系,湖北 武汉 430079)

Privacy Metric for User's Trajectory in Location-Based Services

WANG Cai-Mei, GUO Ya-Jun+, GUO Yan-Hua

(Department of Computer Science, Huazhong Normal University, Wuhan 430079, China)

+ Corresponding author: E-mail: ccnugyj@126.com

Wang CM, Guo YJ, Guo YH. Privacy metric for user's trajectory in location-based services. *Journal of Software*, 2012,23(2):352–360. http://www.jos.org.cn/1000-9825/3946.htm

Abstract: This paper proposes a trajectory privacy measure for Silent Cascade, which is a prevalent trajectory privacy preserving method in LBS (location-based services). In this measure, the user's trajectory is modeled as a weighted undirected graph, and the user's trajectory privacy level is calculated through the use of information entropy. It is pointed out in literatures that any privacy preserving methods will be subject to privacy threats once the attacker has new background knowledge. Therefore, adversarial background knowledge is hierarchically integrated into this measure. The privacy metric result composes of the assumptive background knowledge and the corresponding trajectory privacy level. ($K_{UL}(K_{i+},K_{i-}),K_L(K_{i+},K_{i-})$) association rules is also proposed to describe the assumptive background knowledge. Simulation results show that this metric is an effective and valuable tool for mobile users and the designers of trajectory privacy preserving methods to measure the user's trajectory privacy level correctly, even the attacker has variable background knowledge.

Key words: location-based service; trajectory privacy; privacy metric; background knowledge; association rule

摘 要: 针对一种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.该度量方法将用户运动轨迹用带权无向图描述,并从信息熵的角度计算用户的轨迹隐私水平.已有文献指出,当攻击者拥有新的背景知识时,任何一种隐私保护方法都会受到隐私威胁.因此,将攻击者的背景知识分级融入到度量方法中,隐私度量的结果由对背景知识的假设和相应的轨迹隐私水平值组成,并提出($K_{UL}(K_{i+},K_{i-}),K_L(K_{i+},K_{i-})$)联系规则的方法来描述对背景知识的假设.模拟实验结果表明,此度量方法为移动用户和轨迹隐私保护方法的设计者提供了一个有价值的工具,能够准确地评估在攻击者具有可变背景知识情况下,用户的轨迹隐私水平.

关键词: 位置服务;轨迹隐私;隐私度量;背景知识;联系规则

中图法分类号: TP301 **文献标识码**: A

计算机先进技术越来越多地、无形地融入到人们的日常生活中,为我们提供各种信息服务.传感定位技术和移动通信技术的相结合,使得基于位置的服务(location-based service,简称 LBS)引起人们极大的关注.近年来,

* 基金项目: 国家自然科学基金(61170017)
收稿时间: 2010-04-30; 定稿时间: 2010-09-29

位置服务不仅成为国际研究的热点,而且已成为国内外相关企业研发投入的重点之一.在基于位置的服务中,用户通过向服务器提供其所在的位置信息而享受到与位置有关的服务,诸如查找到离自己最近的宾馆、医院、饭店等.然而,恶意攻击者可以将位置信息和发出查询请求的内容联系到人们的私人生活、健康状况、政治立场和宗教倾向等,或者通过联系额外知识确定一个人的真实身份.一个人的身份一旦确定,其所有其他敏感信息都将泄露.这些攻击行为的存在,阻碍了位置服务的市场发展和商业前景.所以,在给用户提供服务的同时,对用户位置信息加以保密显得尤为重要.轨迹隐私保护是位置隐私保护中相当重要的一个方向,恶意攻击者有可能将用户时间顺序上的多个位置信息连接起来,从而得到用户在某一段时间内的运动轨迹.一旦用户的运动轨迹暴露,那么用户更多的敏感信息将会受到威胁.

研究者们提出了各种保护用户轨迹隐私的方法.然而当这些方法运用到实际中时,并不能达到理论上的隐私保护效果,用户需要及时得到关于他们当前的隐私保护程度的一个反馈.同时,为了评估保护隐私的技术水平是否有所提高,需要建立一种隐私度量机制来评估服务系统的隐私保护效果.这对这个问题的研究具有十分重要的意义.

本文第 1 节介绍相关工作.第 2 节针对 Silent Cascade 轨迹隐私保护方法提出一种新的轨迹隐私度量方法,包括用户运动轨迹的建模、攻击者背景知识的描述和轨迹隐私度量机制的建立.第 3 节通过模拟实验验证该轨迹隐私度量方法的有效性.第 4 节总结并展望下一步工作.

1 相关工作

在隐私度量方面,最早在匿名系统中使用匿名集的大小或有效匿名集的大小来度量用户匿名性[1].林欣等人[2]认为,当匿名集中各个用户的概率不相等时,匿名集的大小则不再能够正确地反映每个用户的真正的匿名性.他们提出连续查询攻击算法,根据集合中 k 个用户分别可能是真正查询发送者的概率计算出熵值,由熵值计算出查询匿名度.Xu 等人[3]认为,在基于连续位置服务的 k-匿名中,一个模糊区域中的用户约束了它在下一个模糊区域中的位置,这给攻击者提供了相关信息.因此,简单地保证每个模糊区域中包含至少 k 个用户,并不能提供给用户 k-匿名的保护.他们提出了一种基于模糊区域大小的熵度量机制,不仅考虑了模糊区域内实体的数量,而且考虑其匿名概率分布来量化系统匿名性.文献[4,5]分别使用使用最大跟踪时间和路径混淆来反映交通监控系统中 GPS 跟踪的车辆的隐私水平.

Ma 等人[6,7]提出 V2X 通信系统中的一种基于 trip 的度量机制来量化每个用户的位置隐私水平.采用信息理论方法,将隐私水平量化为位置信息与特定的个人相联系的不确定性,并考虑了攻击者在一个更长的时间内(如几天或几星期内)所获得的与隐私相关的累积信息对隐私水平的影响.文献[8,9]提出一种针对匿名通信系统的匿名度量方法,将匿名性量化为攻击者得到的信息量与攻击者想要完全知道系统的通信模式所需要的信息量之比来反映匿名通信系统的匿名度.Shokri 等人[10]提出一种基于扭曲的隐私度量方法,通过比较攻击者观察得到的跟踪用户的运动轨迹与用户真实运动轨迹之间的差异来反映用户的隐私水平.文献[2,3]的度量方法是针对一种具体的隐私保护模型和攻击者具有特定背景知识的情况下提出的评估方法,而实际生活中,攻击者所拥有的背景知识很可能是不断变化的.文献[6−10]只是提出了一种隐私度量的框架,普遍适用于较多的隐私保护模型,但是过于抽象.目前的位置隐私度量方法主要侧重于用户某个时刻的位置隐私度量,很少有专门针对用户轨迹的隐私度量方法,并且现有的位置隐私度量方法常常忽略了攻击者的背景知识,而攻击者的背景知识与隐私度量是紧密相关的[11].

Silent Cascade 是 Huang 等人[12]在 Silent Period[13,14]的基础上提出的一种轨迹隐私保护方法.在 Silent Period 方法中,将用户运动的空间区域分为混合区域和应用区域.用户在混合区域中既不发送任何服务请求信息,也不接受任何服务信息,直到换用假名从混合区域中出去.用户进入混合区域前后使用不同的假名,这样增加了将用户的前后两个连续的位置连接起来的难度,降低了用户的两个或多个位置信息之间的可连接性,避免了恶意攻击者得到用户的运动轨迹.但由于在混合区域不进行任何通信,所以损失了通信时间,降低了服务质量. Silent Cascade 方法对 Silent Period 方法加以改进,从时间和空间两个方面对用户的位置信息进行匿名处理,

通过平衡用户在混合区域和应用区域中的停留时间,在不降低服务质量的前提下,使得恶意攻击者难以将用户的两个或多个位置信息连接起来,从而对用户的轨迹隐私提供了更强的保护.

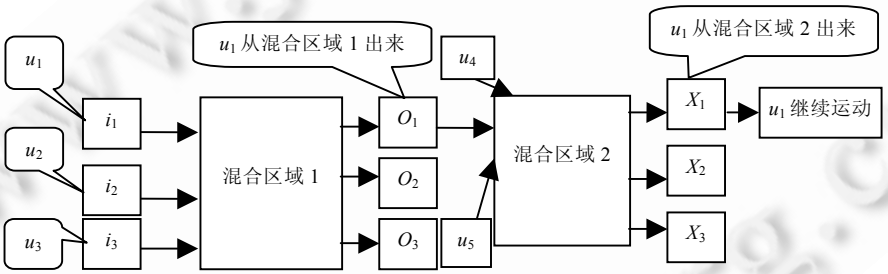
本文针对 Silent Cascade 轨迹隐私保护方法提出一种有效的轨迹隐私度量方法,将轨迹隐私量化为经过混合区域前后用户假名之间的可联系性,实时地评估这种轨迹隐私保护方法下移动用户的轨迹隐私水平.并将攻击者背景知识融入到度量方法中,从而度量在攻击者拥有可变背景知识下移动用户的轨迹隐私水平和轨迹隐私暴露率,以正确地反映这个轨迹隐私保护系统真正的价值.

2 轨迹隐私度量方法

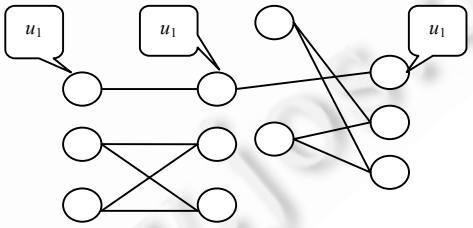
2.1 用户运动轨迹建模

对用户的运动轨迹进行建模是隐私度量的基础.我们拟从攻击者的角度来分析用户的运动轨迹被跟踪和被重构的可能性,从而能够为准确地量化评估轨迹隐私提供保证.下面以 Silent Cascade 方法为例来说明用户运动轨迹建模的方法.

将进入某个混合区域中的所有用户的集合表示为 $I=\{i_j\}$,从这个混合区域中出去的所有用户的集合表示为 $O=\{o_j\}$.当恶意攻击者对某个用户 u_1 进行跟踪时,可能观察到 u_1 经过多个混合区域,并多次更换假名.用户在某时刻 t 经过的混合区域表示为 $M_t(u_1)=(I_t,O_t)$.集合 I 和集合 O 中的用户假名存在一一对应的关系.因此,我们模拟一个用户运动空间,用带权的无向图来表示,$G=(V_i,V_o,E_{io})$.其中,该图中的顶点分别代表从混合区域中进入和出去的用户假名的集合,边代表一个用户在进出一个混合区域所使用的两个用户假名之间的可联系性,边的权重表示可联系性的概率.图 1 以用户 u_1 的运动为例,描述了在 Silent Cascade 轨迹隐私保护方案下,用户实际的运动过程和相对应的带权无向图表示的运动轨迹.这里,我们假设了知道用户 u_1 的运动过程作为例子,而实际情况下,用户 u_1 的运动过程是不可知的.



(a) 用户在 Silent Cascade 隐私保护方案下的实际运动过程



(b) 与(a)相对应的带权无向图

Fig.1 Modeling of user's trajectory

图 1 用户运动轨迹建模

2.2 攻击者背景知识

对每种隐私保护模型,攻击者所拥有的背景知识[15]的内容及多少直接地影响攻击这种模型的难易程度.在 Silent Cascade 轨迹隐私保护方案下,攻击者的能力直接影响到跟踪用户的每条可能轨迹的概率大小.攻击者得

到的有用背景知识越多,就越有可能将跟踪用户进出混合区域前后的用户假名联系起来,以对用户进行进一步的跟踪,从而得到用户的轨迹信息.所以,只有将攻击者的背景知识融入到度量方法中,当攻击者背景知识变化时,才能正确地度量用户的轨迹隐私水平.

将攻击者背景知识融入到隐私度量方法中是一个非常具有挑战性的问题,因为它需要预测攻击者可能知道多少背景知识或者知道哪些背景知识.然而这是不可实行的,因此只能对攻击者所可能拥有的背景知识作出各种假设.在本文中,对每次度量过程做出 3 种不同的假设,这 3 种假设的背景知识所包含的有用信息是逐级增多的.因此,隐私度量的结果应该是由 3 组数据组成,每组数据都是一个二元组:由对背景知识的假设和相应的轨迹隐私水平值组成.只有这样,用户才可以直观地感受到在不同的背景知识假设下他们的隐私受到的威胁.然后才能根据度量结果来判断哪一个假设更符合实际情况,再判断在这种假设下的轨迹隐私水平是否是可以接受的.如果是可以接受的,那么由轨迹隐私保护系统对用户提出的服务请求信息进行保护后,再由位置服务器提供服务;否则,用户可以延迟享受服务的时间.因此,对背景知识进行假设并不是要求准确地预测攻击者可能知道哪些信息,而是当用户将要提出服务时,让他们更全面地认识到享受这次服务可能会受到的隐私威胁.

(1) 攻击者背景知识的表达

Silent Cascade 轨迹隐私保护方法主要是通过切断用户连续位置之间的连接性来达到轨迹隐私保护的目的,攻击者的目的就是从经过轨迹隐私保护方法保护后的用户运动模式中得到位置之间的连续性,最终得到用户与运动轨迹的可联系性.因此,度量轨迹隐私、将轨迹隐私量化为进出混合区域时用户假名之间的可联系性,实际上需要从推算位置之间的连续性和位置与用户之间的可联系性出发.将这两种可联系性描述为条件概率,分别为 $P(L_B|L_A)$,$P(L_C|u_k)$.其中,$L_A,L_B,L_C\in L,u_k\in U,L$ 为在跟踪时间内与跟踪用户构成混合区域的所有用户(包括跟踪用户)向服务器提出服务请求时提供的所有位置信息的集合,U 为用户标识符的集合.现将所有的 $P(L_B|L_A)$ 和 $P(L_C|u_k)$都看作变量,而将背景知识表达为这些变量的约束,即条件概率的值.

将攻击者可以得到的信息分为两类:公共信息和跟踪得到的信息.公共信息是攻击者不经过对用户运动进行跟踪就可以得到的信息,例如,地图上的实际路径、攻击者之前积累的经验知识(如某用户经常走某条道路)、用户所走道路的车速限制等.跟踪得到的信息即为攻击者对用户进行跟踪所截取的服务请求信息,包括用户假名、查询内容、时间信息、位置信息.从这两类信息里可以得到 $P(L_B|L_A)$,$P(L_C|q_k)$的值,$q_k\in Q,Q$ 为用户准标识符的集合.准标识符(quasi-identifiers)通常是指用户的个人信息,如性别、年龄、家庭住址等,这些个人信息通常可以通过其他公共资源获得.攻击者的背景知识则是来自公共信息.攻击者将背景知识与跟踪得到的信息进行关联攻击,得到 $P(L_B|L_A)$,$P(L_C|u_k)$的值,就将背景知识表达成了变量 $P(L_B|L_A)$,$P(L_C|u_k)$的约束.它们通常以各种不同的形式出现,有等式 $P(L_2|L_1)=0.8$,$P(L_2,L_3|L_1)=0$,$P(\neg L_4|L_1)=0.5$,不等式 $0.3<P(L_1|u_1)<0.6$ 等.

(2) 攻击者背景知识的量化

由于隐私度量的结果由 3 组对背景知识的假设和相应的轨迹隐私水平值组成,因此必须找到一种方法来描述对背景知识的假设,让用户对所假设的背景知识有一个直观的认识.最为直接的方法是列举所有可能的背景知识的组合,通过计算得到每一个组合在度量轨迹隐私中所起作用的大小来衡量背景知识的强弱.因为背景知识的组合太多了,所以这种方法实际上是不可行的.

$A\Rightarrow B$ 表示 A 发生而引起 B 发生,此类型的关联规则为正关联规则;其他如$\neg A\Rightarrow B$ 或$\neg A\Rightarrow\neg B$ 之类的关联规则为负关联规则.所有的背景知识(如上面的等式或不等式)都可以表达为正关联规则和负关联规则,其中,条件概率的大小表达为规则的强度.例如,等式 $P(L_2|L_1)=0.8$ 可以表达为正关联规则:$L_1\Rightarrow L_2$,规则强度为 0.8;等式 $P(\neg L_4|L_1)=0.5$ 可以表达为负关联规则:$L_1\Rightarrow\neg L_4$,规则强度为 0.5.文献[16]正是基于这种思想,提出使用(K_+,K_-)最强联系规则的方法来量化背景知识,其中,K_+表示 K_+个最强的正关联规则,K_-表示 K_-个最强的负关联规则.对两类规则,分别根据其强度进行排序,然后挑选出 K_+个最强的正关联规则和 K_-个最强的负关联规则,使用这两个规则的集合大小来描述背景知识的强弱.但这种方法仅使用了关联规则的数量来衡量背景知识的强弱,这样是不太准确的,因为它忽略了关联规则的强度.

因此,本文对上述方法进行改进,使用属于不同强度区间的关联规则的数量来描述所假设的攻击者拥有的

背景知识的强弱,提出($K_{UL}(K_{i+},K_{i-})$,$K_L(K_{i+},K_{i-})$)联系规则的方法来量化背景知识.首先,通过关联攻击找出所有 u_k 和 L_C 以及 L_A 和 L_B 之间的正关联规则和负关联规则,规则的强度则是规则关联性的大小.然后将所有关联规则根据其规则强度分为 n 类,所以,i=1,2,…,n;概率大小总区间为[0,1].显然,n 越大,各关联规则之间的区分就越细致,对背景知识的量化也就越准确.同时,K_{UL} 表示所有 u_k 和 L_C 之间的关联规则的集合的大小,K_L 表示所有 L_A 和 L_B 之间的关联规则的集合的大小,K_{i+}表示集合中第 i 类正关联规则的数量,K_{i-}表示集合中第 i 类负关联规则的数量,$K_{UL}(K_{i+},K_{i-})$中,$\sum K_{i+}+\sum K_{i-}=K_{UL}$,$K_L(K_{i+},K_{i-})$中,$\sum K_{i+}+\sum K_{i-}=K_L$.这样,我们不仅考虑了背景知识中所包含的关联规则的数量,而且考虑其概率分布,对背景知识的量化更加准确.

2.3 轨迹隐私度量机制

下面建立一种轨迹隐私度量机制,根据所假设的攻击者拥有的背景知识并结合对用户运动轨迹的建模进行推理,计算出在 Silent Cascade 轨迹隐私保护方法下用户的轨迹隐私水平.

以前面对用户运动轨迹建模的分析为例,每个用户进入一个混合区域时,有一个唯一假名来标识这个用户,在出混合区域时,有且仅有 1 个假名,也就是说,进入前的所有用户假名和出去时的所有用户假名具有一一对应的映射关系.假设有 N 个用户进入混合区域,那么一共有 $N!$种映射关系,攻击者很难正确地判断它们之间的配对关系,在没有任何背景知识的情况下,攻击者认为出现这 $N!$种映射关系的概率是相等的.

对于每一个混合区域所对应的单个带权无向图 $G=(V_i,V_o,E_{io})$,我们用一个 $m\times m$ 的矩阵加以描述.这里,m 为进入混合区域的用户数量,即 $m=|V_i|=|V_o|$.其中,$E_{io}=\{e_{io}|i\in V_i,o\in V_o\}$,$e_{io}\in E_{io}$ 的值的大小由 i 和 o 之间的关联概率 $p(i,o)\in[0,1]$来决定,它表示无向图的边的权值.当 $p(i,o)$的值为 0 时,表示攻击者确定 i 和 o 之间没有联系性,它们不是对应同一个用户.当攻击者确定这两个用户名是指同一用户时,其概率值则为 1.另外,从每个顶点 $i\in V_i$ 或 $o\in V_o$ 发出的边的概率之和为 1,$\sum_{k=1}^{n}p(i_j,o_k)=1$,$\sum_{k=1}^{n}p(i_k,o_j)=1$.也就是说,矩阵的每一列或每一行之和均为 1.

对于每个用户,攻击者对用户进行跟踪,可能观察到用户经过时间顺序上的多个混合区域.在没有背景知识的情况下,攻击者认为从混合区域中出去的任何一个假名所代表的用户是攻击者所跟踪用户的概率相等.在经过一些混合区域之后,假设会形成 X 种可能的用户运动轨迹,这些属于单个用户的可能的多条运动轨迹则是树形结构,每条轨迹的概率也相等,均为 1/X,其中,X 为树的叶子节点的个数.攻击者在获得和利用背景知识进行推理后,每条可能的运动轨迹是用户真实运动轨迹的概率不再相等.而且攻击者可以根据用户经过每个混合区域的概率值最终确切地计算出每条运动轨迹的概率值,再通过信息理论方法计算每个用户的轨迹隐私水平.

信息熵常常用于表示某种特定信息的出现概率,一个系统越是有序,信息熵就越低;反之,一个系统越是混乱,信息熵就越高.攻击者对用户轨迹隐私攻击的主要目的是要挑选出概率最大的轨迹,如果攻击者认为特定用户每条可能的运动轨迹的概率相等,则表示系统是混乱的;如果每条可能的运动轨迹的概率差别很大,则表示系统是有序的.因此,选用信息熵来度量轨迹隐私是可行的.

随机变量 X 的熵表示通过学习 X 获得的信息量,用来量化一个变量 X 的不确定性的程度,计算如下:

$$H(X)=-\sum_{x\in X}p(x)\log p(x) \tag{1}$$

当所有可能的情况集中在同一个值上时,X 的熵取最小值 0;当所有值等概率时,X 的熵取最大值(N 为样本空间的大小).

攻击者根据所拥有的背景知识推理并计算出某个用户每条可能的运动轨迹的概率值,每条轨迹的概率由无向图中组成这条轨迹的各条边的权重相乘得到.将用户 u_i 每条可能的轨迹记为 $T_1,T_2,\ldots,T_X$,其概率值分别为 $P_1,P_2,\ldots,P_X$,则

$$P_\chi=\prod_{e_{io}\in T_\chi}e_{io}=\prod_{e_{io}\in T_\chi}p(i,o),\ \chi=1,2,...,X \tag{2}$$

这里,$e_{io}=\begin{cases}1/Size(M_t(u_i))!, & \text{当攻击者没有背景知识时}\\ P(L_i/L_o)\text{ or }P(L_i/u_i)\cdot P(L_o/u_i), & \text{当攻击者具背景知识时}\end{cases}$,$Size(M_t(u_i))$为进入混合区域 $M_t(u_i)$的用户数量,i 从混合区域 $M_t(u_i)$进入,o 从混合区域 $M_t(u_i)$出去,L_i 为假名为 i 的用户发出的位置信息,L_o 为假名为 o 的用

户发出的位置信息.

根据公式(1),我们可以计算出特定用户 u_i 的熵为

$$H(u_i)=-\sum_{x=1}^{X}p_x\log p_x \tag{3}$$

在攻击者没有任何知识背景的情况下,每条运动轨迹的可能性是等概率的.这时,这个用户的最大熵值为

$$H_{\mathrm{Max}}(u_i)=-\log\frac{1}{X} \tag{4}$$

我们用 $D(u_i)$来衡量用户轨迹隐私水平.它表示特定用户的轨迹隐私保护程度.

$$D(u_i)=\frac{H(u_i)}{H_{\mathrm{Max}}(u_i)}\% \tag{5}$$

因此,$1-D(u_i)$表示用户 u_i 的轨迹隐私泄漏率.$D(u_i)$的值越大,表示用户 u_i 的轨迹隐私受保护程度越高、隐私泄漏率越低、攻击者的能力越小;反之亦然.

3 实验结果与分析

在模拟实验中,对于 Silent Cascade 轨迹隐私保护方法满足通信质量需求一定的情况下,通过假设攻击者拥有不同的背景知识,然后在 Silent Cascade 轨迹隐私保护方法下的同一个实例中分别度量用户的轨迹隐私水平,评估当攻击者所拥有的背景知识不同时同一用户的轨迹隐私保护水平.实验以 UCI 机器学习数据集里的 Adult 数据集和城 Oldenburg 的交通网络图作为初始数据,将 Adult 数据集里的个人作为用户模拟到 Oldenburg 市的交通网络图中,并且用户的轨迹隐私受到Silent Cascade方法的保护.将攻击者背景知识作为输入,测试用户的轨迹隐私水平作为输出.测试时,所输入的攻击者背景知识取自攻击者背景知识库,并通过 ME method[16]表达成联系规则,攻击者背景知识库由 Adult 数据集、Oldenburg 交通网络图以及用户发出的查询请求信息共同构成.模拟实验在普通微机的 Windows 平台下使用 Visual C++ 6.0 开发实现,测试机器基本参数为 Pentium4 CPU 3.00GHz,内存 2GB.实验着重从以下几个方面分析了攻击者所拥有的背景知识如何影响用户的轨迹隐私:

(1) 攻击者有无背景知识对轨迹隐私度量结果的影响

在基于位置的服务中,已有的隐私度量方法基本上是假设攻击者没有任何背景知识或者只是拥有某些特定的背景知识.而现实生活中,攻击者都是掌握一定的背景知识的,而且攻击者所掌握的背景知识很可能是不断变化的.根据实验结果描绘了两条曲线,Γ_0 是假设攻击者没有任何背景知识,Γ_1 是假设攻击者具有一定的背景知识.实验时,从攻击者背景知识库的 Adult 数据集、交通网络图和用户发出的查询请求记录中分别取记录量为 3 000,2 000,1 000,可表达为大概 3×10^4 个联系规则,如图 2 所示.可以看出,当攻击者没有任何背景知识时,隐私没有暴露;而当攻击者具有一定的背景知识时,随着跟踪时间的变长,隐私则逐渐趋于暴露.因此,在轨迹隐私度量方法中考虑攻击者所拥有的背景知识是必不可少的.同时,也表明了轨迹隐私度量的结果不仅应该包含轨迹隐私水平值,而且应该包含对背景知识的假设,即度量结果应该由这两个元素组成.这体现了在帮助用户理解他们的隐私受保护程度方面,让用户知道攻击者所拥有的背景知识起着至关重要的作用.

(2) K_{UL} 和 K_L 对轨迹隐私度量结果的影响

在实验中对比分析了两类背景知识对轨迹隐私的影响:用户与位置之间的关联规则、两个位置之间的关联规则,如图 3 所示.描绘了 3 条曲线:K_{UL} 曲线使用了 $3K/4$ 个用户与位置之间的关联规则和 $K/4$ 个位置与位置之间的关联规则;K_L 曲线使用了 $K/4$ 个用户与位置之间的关联规则和 $3K/4$ 个位置与位置之间的关联规则;(K_{UL},K_L)曲线使用了 $K/2$ 个用户与位置之间的关联规则和 $K/2$ 个位置与位置之间的关联规则.从图中我们可以清楚地看出,随着攻击者得到更多的背景知识,用户的轨迹隐私水平变得越来越低.尤其是当 K 很小时,隐私水平下降得很快;当攻击者得到越来越多的背景知识时,下降速率变慢.这是因为在背景知识量增大的同时也包含了更多的冗余信息,相对而言,对攻击者有用的信息变少,所以隐私水平的下降速率变慢.

对比分析这 3 条曲线可以看出,不同类型的背景知识对隐私暴露的影响大小不一样,隐私水平下降的速率不同.(K_{UL},K_L)曲线下降得最快.这表明,即使背景知识的量是相同的,但这两类背景知识的组合比单独的一类背

景知识包含更多的有用信息,攻击者得到这两类背景知识的组合对其更加有用.所以,我们在量化背景知识时,要同时考虑 K_{UL} 和 K_L.

同时,对比 K_{UL} 和 K_L 两条曲线,当背景知识的量相同而用户与位置之间关联的背景知识所占比例更大时,用户轨迹隐私水平的下降速率更快.可以看出,用户与位置之间关联的背景知识比两个位置之间关联的背景知识对攻击者得到用户的轨迹隐私起到更大的作用.所以,在我们设计轨迹隐私保护方法时,应该着重在切断用户准标识符与位置之间的可关联性方面做工作,也应该着重避免泄漏关于用户准标识符与位置之间关联的背景知识.

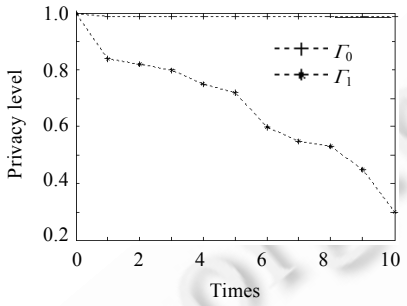


Fig.2 Effect of adversarial background knowledge on privacy metric result

图 2 攻击者有无背景知识对轨迹隐私度量结果的影响

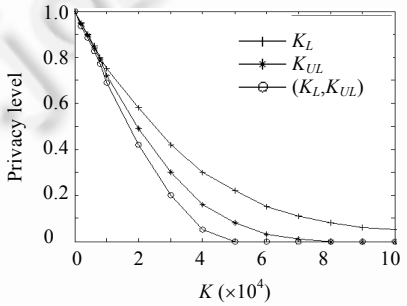


Fig.3 Effect of the value K_{UL} and K_L on privacy metric result

图 3 K_{UL} 和 K_L 对轨迹隐私度量结果的影响

(3) n 的取值对轨迹隐私度量结果的影响

相对于文献[16]中提出的量化背景知识的方法,本文提出了($K_{UL}(K_{i+},K_{i-})$,$K_L(K_{i+},K_{i-})$)联系规则的方法,不仅考虑了背景知识中的正关联规则和负关联规则的数量,而且考虑了这些关联规则的概率分布.在本次模拟实验中,我们选取了包含 $K/2$ 个 K_{UL} 关联规则和 $K/2$ 个 K_L 关联规则的背景信息(这里,$K=2\times10^4$),这两类关联规则分别所包含的正关联规则和负关联规则的数量相等.

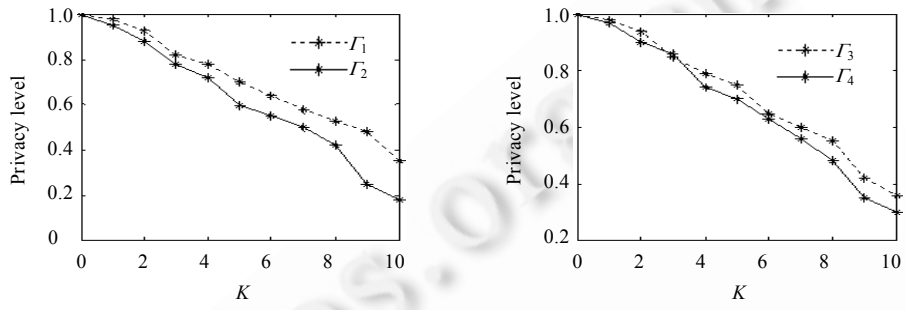
图 4(a)中描绘了两条曲线,一条曲线的背景知识为 Γ_1,另一条曲线的背景知识为 Γ_2.Γ_1,Γ_2 所包含的关联规则在数量上相同,但概率分布不同.当 n=0,即不考虑关联规则的概率分布时,对它们的描述相同.当 n=5 时,i=1,2,3,4,5,将概率分为 5 类:[0,0.2],(0.2,0.4),(0.4,0.6),(0.6,0.8),(0.8,1),对它们的描述分别为 $\Gamma_1=(K_{i+}=K_{i-}=0.1\times10^4, i=1,2,3,4,5)$;$\Gamma_2=(K_{1+}=K_{1-}=0.05\times10^4, K_{2+}=K_{2-}=0.05\times10^4, K_{3+}=K_{3-}=0.05\times10^4, K_{4+}=K_{4-}=0.15\times10^4, K_{5+}=K_{5-}=0.2\times10^4)$. 为了简洁,这里只是描述了($K_{UL}(K_{i+},K_{i-})$,$K_L(K_{i+},K_{i-})$)中的($K_{i+},K_{i-}$),下面对 Γ_3 和 Γ_4 的描述也是如此.

当 n=0,即不考虑关联规则的概率分布时,在轨迹隐私度量结果中,对背景知识假设的描述相同,都为 Γ ($\Gamma=\Gamma_1=\Gamma_2$),但隐私水平值是不一样的;当 n=5 时,因为考虑了关联规则的概率分布,轨迹隐私度量结果中对背景知识假设的描述不同,分别为 Γ_1 和 Γ_2,自然分别对应不同的隐私水平值.从实验结果可以看出,当只使用关联规则的数量来量化背景知识时,并不能准确地描述背景知识,就会出现这样的情况:所描述的背景知识一样,但是

隐私水平值不同.因此,将关联规则的概率分布融入到量化背景知识的方法中是有必要的.

图 4(b)中同样描绘了两条曲线,一条曲线的背景知识为Γ_3,另一条曲线的背景知识为Γ_4.我们分别取这样的数据:当 n=5 时,对背景知识的描述是相同的,都是$\Gamma'(\Gamma'=\Gamma_3=\Gamma_4)$,$\Gamma'=(K_{1+}=K_{1-}=0.1\times10^4,K_{2+}=K_{2-}=0.1\times10^4,K_{3+}=K_{3-}=0.1\times10^4,K_{4+}=K_{4-}=0.1\times10^4,K_{5+}=K_{5-}=0.1\times10^4)$;当 n=10 时,把概率区间分为 10 类,分别为[0,0.1],(0.1,0.2),(0.2,0.3),(0.3,0.4),(0.4,0.5),(0.5,0.6),(0.6,0.7),(0.7,0.8),(0.8,0.9),(0.9,1),对背景知识的描述不同,分别为Γ_3 和Γ_4,$\Gamma_3=(K_{i+}=K_{i-}=0.05\times10^4,i=1,2,\ldots,10)$;$\Gamma_4=(K_{i+}=K_{i-}=0.02\times10^4,i=1,3,5,7,9,K_{i+}=K_{i-}=0.08\times10^4,i=2,4,6,8,10)$.

当 n=5 时,对背景知识的描述是相同的,都是$\Gamma'(\Gamma'=\Gamma_3=\Gamma_4)$,而隐私度量结果不一样;当 n=10 时,对背景知识的描述不同分别为Γ_3和Γ_4,分别对应不同的隐私水平值.从实验结果可以看出,n 越大,关联规则的概率分布区分得越细致,所描述的背景知识就越准确.



(a) n=0, n=5 (b) n=5, n=10

Fig.4 Effect of the value n on privacy metric result

图 4 n 的取值对隐私度量结果的影响

4 结论与展望

本文针对现有的 Silent Cascade 轨迹隐私保护方法,提出基于熵理论的轨迹隐私度量方法,并将攻击者背景知识融入到度量方法中.实验结果表明了该度量方法的有效性,当攻击者所拥有的背景知识变化时,也能准确地度量用户的轨迹隐私.同时,对实验结果的分析给轨迹隐私保护方法的设计者提供了一些参考,有助于他们设计出更好的轨迹隐私保护方法.

随着对位置服务中轨迹隐私度量的研究,如何将攻击者所拥有的背景知识更贴切地融入到度量方法中,将是我们下一步要进行的重点工作.

References:

[1] Kelly DJ, Raines RA, Grimaila MR, Baldwin RO, Mullins BE. A survey of state-of-the-art in anonymity metrics. In: Antonatos S, ed. Proc. of the 1st ACM Workshop on Network Data Anonymization. Alexandria: ACM, 2008. 31−40. [doi: 10.1145/1456441.1456453]

[2] Lin X, Li SP, Yang CH. Attacking algorithms against continuous queries in LBS and anonymity measurement. Journal of Software, 2009,20(4):1058−1068 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/3428.htm [doi: 10.3724/SP.J.1001.2009.03428]

[3] Xu T, Cai Y. Location anonymity in continuous location-based services. In: Samet H, ed. Proc. of the 15th Annual ACM Int'l Symp. on Advances in Geographic Information Systems. Seattle: ACM, 2007. 1−8. [doi: 10.1145/1341012.1341062]

[4] Gruteser M, Hoh B. On the anonymity of periodic location samples. In: Hutter D, ed. Proc. of the 2nd Int'l Conf. on Security in Pervasive Computing. LNCS 3450, Heidelberg: Springer-Verlag, 2005. 179−192. [doi: 10.1007/978-3-540-32004-3_19]

[5] Hoh B, Gruteser M, Xiong H, Alrabady A. Preserving privacy in GPS traces via uncertainty-aware path cloaking. In: Ning P, ed. Proc. of the 14th ACM Conf. on Computer and Communications Security. Alexandria: ACM, 2007. 161−171. [doi: 10.1145/1315245.1315266]

[6] Ma ZD, Frank K, Michael M. A location privacy metric for V2X communication systems. In: Manousakis K, ed. Proc. of the 2009 IEEE Sarnoff Symp. Princeton: IEEE, 2009. 1−6. [doi: 10.1109/SARNOF.2009.4850318]

[7] Ma ZD, Frank K, Michael M. Measuring location privacy in V2X communication systems with accumulated information. In: Ni LM, ed. Proc. of the 6th IEEE Int'l Conf. on Mobile Ad-Hoc and Sensor Systems. Macao: IEEE, 2009. 322−331. [doi: 10.1109/MOBHOC.2009.5336983]

[8] Edman M, Sivrikaya F, Yener B. A combinatorial approach to measuring anonymity. In: Proc. of the IEEE Intelligence and Security Information. New Brunswick: IEEE, 2007. 356−363. [doi: 10.1109/ISI.2007.379497]

[9] Gierlichs B, Troncoso C, Diaz C, Preneel B, Verbauwhede I. Revisiting a combinatorial approach toward measuring anonymity. In: Atluri V, ed. Proc. of the 7th ACM Workshop on Privacy in the Electronic Society. Alexandria: ACM, 2008. 111−116. [doi: 10.1145/1456403.1456422]

[10] Shokri R, Freudiger J, Jadliwala M, Hubaux JP. A distortion-based metric for location privacy. In: Al-Shaer E, ed. Proc. of the 8th ACM Workshop on Privacy in the Electronic Society. Chicago: ACM, 2009. 21−30. [doi: 10.1145/1655188.1655192]

[11] Riboni D, Pareschi L, Bettini C. Shadow attacks on users' anonymity in pervasive computing environments. Pervasive and Mobile Computing, 2008,4(6):819−835. [doi: 10.1016/j.pmcj.2008.04.008]

[12] Huang L, Yamane L, Mastsuura K, Sezaki K. Silent Cascade: Enhancing location privacy without communication QoS degradation. In: Clark JA, ed. Proc. of the 3rd Int'l Conf. on Security in Pervasive Computing. LNCS 3934, Heidelberg: Springer-Verlag, 2006. 165−180. [doi: 10.1007/11734666_13]

[13] Huang LP, Matsuura K, Yamane H, Sezaki K. Enhancing wireless location privacy using silent period. In: Pauly L, ed. Proc. of the IEEE Wireless Communications and Networking Conf. New Orleans: IEEE, 2005. 1187−1192. [doi: 10.1109/WCNC.2005.1424677]

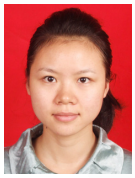
[14] Huang LP, Yamane H, Matsuura K, Sezaki K. Towards modeling wireless location privacy. In: Danezis G, ed. Proc. of the 5th Int'l Workshop on Privacy Enhancing Technology (PET). LNCS 3856, Heidelberg: Springer-Verlag, 2005. 59−77. [doi: 10.1007/11767831_5]

[15] Bettini C, Mascetti S, Wang XS, Jajodia S. Anonymity in location-based services: Towards a general framework. In: Proc. of the 8th Int'l Conf. on Mobile Data Management. Mannheim: IEEE, 2007. 69−76. [doi: 10.1109/MDM.2007.19]

[16] Du WL, Teng ZX, Zhu ZT. Privacy-MaxEnt: Integrating background knowledge in privacy quantification. In: Lakshmanan LVS, ed. Proc. of the 2008 ACM SIGMOD Int'l Conf. on Management of Data. Vancouver: ACM, 2008. 459−472. [doi: 10.1145/1376616.1376665]

附中文参考文献:

[2] 林欣,李善平,杨朝晖.LBS 中连续查询攻击算法及匿名性度量.软件学报,2009,20(4):1058−1068. http://www.jos.org.cn/1000-9825/3428.htm [doi: 10.3724/SP.J.1001.2009.03428]



王彩梅(1986－),女,湖北仙桃人,硕士生,主要研究领域为信息安全.

郭艳华(1985－),女,硕士生,主要研究领域为信息安全.

郭亚军(1965－),男,博士,教授,CCF 高级会员,主要研究领域为信息安全,普适计算,物联网.