基于假位置和 Stackelberg 博弈的位置匿名算法

夏兴有[1)] 白志宏[1)] 李 婕[2)] 于瑞云[1)]

[1)](东北大学软件学院 沈阳 110169)

[2)](东北大学计算机科学与工程学院 沈阳 110169)

摘 要 移动互联网技术的快速发展、无线定位技术的进步和移动智能设备的普及使得移动用户可以通过移动智能设备获得各种基于位置的服务,这极大地方便了人们的生活.但是,随着移动用户对位置服务的依赖,位置隐私问题日益严重,用户随时面临着信息泄露的危险,严重时可能会危害到用户的人身及财产安全.本文分析了已有位置隐私保护系统结构和算法的优缺点,提出了一种基于半可信第三方服务的隐私保护系统结构,并设计了一种基于假位置和 Stackelberg 博弈的位置匿名算法.算法将整个位置空间网格化,通过真实的位置服务请求数据计算出网格内用户服务请求的概率分布,在发起过位置服务请求的网格中以一定的规则选择出位置匿名集,代替用户的真实位置发送给位置服务提供商;针对隐私保护要求更高的场景,算法基于 Stackelberg 博弈模型对位置匿名算法进行优化,通过将用户位置隐私的保护模型和攻击者的位置推测模型进行博弈,从而得到更为优化的匿名结果.文章基于滴滴打车数据集对算法进行了验证,实验结果表明,算法可以在满足服务质量要求的前提下保证用户的位置隐私,为用户提供更好的位置服务体验.

关键词 基于位置的服务;位置隐私;k 匿名;假位置;Stackelberg 博弈

中图法分类号 TP18 **DOI 号** 10.11897/SP.J.1016.2019.02216

A Location Cloaking Algorithm Based on Dummy and Stackelberg Game

XIA Xing-You[1)] BAI Zhi-Hong[1)] LI Jie[2)] YU Rui-Yun[1)]

[1)](*Software College, Northeastern University, Shenyang* 110169)

[2)](*School of Computer Science and Engineering, Northeastern University, Shenyang* 110169)

Abstract With the rapid development of mobile Internet technology, wireless location technology and mobile intelligent devices, mobile users can obtain many kinds of location based services (LBS), such as map navigation, shared bikes, and mobile social services, which bring great convenience to users' lives. However, due to the users' reliance on LBS, the problem of location privacy leakage is becoming more and more serious. Users are exposed to the danger of location leakage at any time, which may be detrimental to the users' lives and property safety in serious cases. In recent years, the location privacy protection algorithm (LPPA) has received widespread attention. The existing LPPA are mainly based on the user identity anonymity, the information encryption, and the location obfuscation and generalization, which serves the purpose of limiting the leakage of sensitive information while ensuring the quality of location services. Based on the analysis of the merits and drawbacks of the existing LPPA and related architectures, this paper first proposes a Semi-Trusted Third Party (STTP) based location privacy protection architecture. In the STTP architecture, the pseudonym server (PS) hides the identity information of the users,

收稿日期:2018-03-13;在线出版日期:2018-11-23.本课题得到国家自然科学基金项目(61672148,61502092)、辽宁省高校创新团队支持计划资助项目(LT2016007)、中央高校基本科研业务费专项资金项目(N182608004,N171702001,N171604016)、辽宁省"百千万人才工程"资助项目(201514)、教育部-中国移动科研基金项目(MCM20160201)资助.**夏兴有**,博士研究生,中国计算机学会(CCF)会员,主要研究方向为群智感知计算、位置隐私保护、移动大数据分析.E-mail:xiaxy_neu@163.com.**白志宏**,硕士,主要研究方向为位置隐私保护、群智感知计算.**李 婕**,博士,副教授,主要研究方向为机会网络通信、移动大数据分析、社会计算等.**于瑞云**(通信作者),博士,教授,博士生导师,主要研究领域为群智感知与计算、移动大数据分析、虚拟现实技术等.E-mail:yury@mail.neu.edu.cn.

the location anonymous server (LAS) performs location privacy protection algorithm, and the LBS provider (LBSP) provides service results according to the queries. The STTP architecture keeps the users' privacy information (the identity, the actual location, the search information and so on) in the PS, the LAS and the LBSP respectively. Therefore, it is impossible to confirm the users' complete privacy information, even if the attackers get some data from a part of servers. Then, with the STTP architecture, a query-History based Cloaking Location Selection algorithm (HCLS) is designed to achieve the optimal level of k anonymity. HCLS divides the whole location space into a lot of grids, and the historical query data is applied to calculate the probability distribution of the users' location requests. HCLS selects an obfuscation location and $k-1$ dummies to form an anonymity set based on the user's real position and the query distribution. The selected dummy are spread as far as possible, and the anonymity set will be sent to the location service provider instead of the user's real location. Finally, for higher requirements of location privacy protection, an improved HCLS model (HCLS_SG) is put forward against attackers implementing the optimal inference attack. HCLS_SG accounts for the fact that the strongest attackers not only observes the anonymity set from the LAS but also knows the algorithm implemented by the LAS. HCLS_SG formalizes the mutual optimization of user-attacker objectives (location privacy vs. services quality) by using the framework of Stackelberg games. The optimal result of anonymity sets in HCLS_SG can maximize the expected distortion (error) that the optimal adversary incurs in reconstructing the actual location of a user, while fulfilling the user's service-quality requirement. This paper adopts the real dataset to evaluate the proposed algorithm. The experimental results demonstrate that the algorithm performs well on the constraint of the location privacy and the service quality, and guarantees good experiences of LBS for the users.

Keywords　location-based services; location privacy; k anonymity; dummy; Stackelberg game

《计算机学报》

1　引　言

随着移动通信技术和定位技术的广泛应用，基于位置的服务(Location Based Service，LBS)已经快速地融入到人们的日常生活中[1-2]. 移动用户可以使用智能终端设备(如智能手机、智能平板电脑等)随时随地查询所需的与位置相关的服务信息，方便快捷地解决生活和工作中的各种问题[3].

然而，在享受位置服务所带来的便利的同时，移动用户也面临着位置隐私被泄漏出去的潜在危险[4-5]. LBS 提供商的数据可能被攻击者恶意窃取，攻击者通过对收集到的用户位置和查询内容进行分析[6]，挖掘出用户的个人信息(如生活作息、移动规律、身体状况、兴趣爱好等)，这将可能威胁到用户的人身及财产安全.

目前，LBS 隐私泄露问题大致分为 2 类：位置隐私(location privacy)泄露和查询隐私(query privacy)泄露. 位置隐私指与用户位置信息相关的敏感信息，包括用户访问过的历史位置、实时位置等，统计这些位置可以帮助攻击者准确定位用户，了解该用户的家庭地址和工作单位等基本信息. 查询隐私指与查询内容相关的兴趣点(Point Of Interest，POI)信息，包括附近餐馆、酒店、超市、购物中心、医院、银行等，借助查询内容攻击者可以分析用户的爱好和需求，甚至可以结合背景知识推断出用户健康状况、宗教信仰等其他敏感信息.

针对上述隐私泄露问题，学者们提出了大量的隐私保护方案，所保护的用户属性主要包括：用户身份、精确位置、时间信息以及查询内容. 隐私保护目标通常为 4 个属性的任意组合，针对不同的目标，目前的隐私保护方案大致可划分为以下 3 类：

(1) 基于假名的用户身份保护方法[7-10]. 对于任何 LBS 请求，此类方法利用可信的第三方服务器生成假名来替换用户的真实身份，用户通过被分配的假名从 LBS 提供商获取相关的服务信息，以此实现

个人身份的隐匿.

基于假名的隐私保护方法是最早被提出的用于对用户位置隐私进行保护的方法.然而,在这个大数据时代背景下,攻击者可以从多种途径获得用户的相关数据,借助数据分析、数据挖掘等技术手段,很容易重构出用户的真实身份及其所在位置.所以,只隐藏用户身份 ID 不足以保护用户的隐私信息.

(2)基于加密的查询内容保护方法[11-14].加密,是一种重要的用户隐私保护技术手段.该方法将发送的位置和查询信息以一定的规则进行加密,即使攻击者窃取到信道中的信息也无法对其进行解密,从而保护用户的信息安全.

隐私信息检索协议(Private Information Retrieval,PIR)是应用在对网络平台中的数据进行私密性处理的协议.Mouratidis 等人[11]将 PIR 用于研究交通路网最短路径查询的隐私保护问题中,对最短路径的计算过程实现了强隐私保护.Chen 等人[12]在 PIR 的基础上为二进制数据设计了一种新的加密协议 cPIR,cPIR 技术可以强化位置查询的隐私保护效果,同时保证计算开销最小化.基于加密的位置隐私保护技术效果较好,但通信及计算花销较大,因此会受到移动设备计算能力的限制.

(3)基于模糊和泛化的用户位置保护方法.基于模糊的位置保护方法[15-17]主要指在用户发起位置服务请求时,通过提交非精确的位置来避免泄露用户敏感的位置信息,所采用的技术主要包括位置偏移、位置扩张等.位置偏移指在用户真实位置周围,选出一个标识位置替代用户的真实位置发起位置服务请求;位置扩张将真实位置扩大成一个位置区域,通过加入噪音来降低位置精度.

为达到隐私保护的相关性要求,Ardagna 等人[15]提出了 3 种位置模糊基本操作:扩大半径、缩小半径以及圆心偏移.通过以上 3 种操作的组合,能够以多种方式实现用户要求,灵活完成位置模糊处理.为了防止攻击者通过统计大量数据获得位置偏移概率分布,Andrés 等人[17]利用偏移向量概率分布函数,使用户在隐匿区域内分布的概率密度为常数,从而达到均匀分布的效果.在基于模糊的位置保护方法中,位置模糊化程度越高,则算法隐私保护效果越好,但服务质量会受到相应影响.

基于泛化(匿名)的位置保护方法[18-21],将用户的精确位置扩充为一个包含 k 个真实用户的空间区域(空间匿名),或者产生 $k-1$ 个假位置与用户真实位置一起组成匿名集合(假位置匿名),利用空间区域或者匿名集合发起位置服务请求,使攻击者难以确定用户的真实身份和准确位置.

Gedik 等人提出了一种个性化的空间匿名方法 CliqueCloak[18],该方法利用无向图找到 $k-1$ 个位置,连同用户真实位置形成一个 Clique 图形匿名域,该匿名域可以有效地保护用户的位置隐私.Yigitog 等人[19]为了在用户移动情况下保护用户位置安全,根据用户的真实移动速度对已经提交的匿名域进行修正,增加生成匿名域的真实性,使攻击者难以推测用户的移动行为.在基于泛化(匿名)的位置保护方法中,假位置匿名简单容易,但隐私效果较差,相对而言,空间匿名具有较高的隐私保护水平,但是要以牺牲服务质量为代价.

结合了上述位置隐私保护算法的优缺点,本文提出了一种基于假位置和 Stackelberg 博弈的位置匿名算法.该算法复杂度较低,且匿名效果较好,可以在满足服务质量要求的前提下保证用户的位置隐私.本文的主要创新点如下:

(1)提出了一种基于半可信第三方服务的系统结构,该结构在原有匿名服务器系统结构的基础上加入了假名服务器,通过一定的机制将用户的隐私信息分别存放在三方的服务器上,攻击者即使窃取一方的信息仍然无法定位用户,获取用户的完整隐私信息.

(2)针对攻击者的单点攻击,提出了一种基于假位置和位置偏移的位置匿名算法,通过计算历史查询位置的概率分布生成位置匿名集,实现 k-匿名,以此防止位置过滤的问题,提高匿名成功率.

(3)针对攻击者的推测攻击,引入 Stackelberg Game 模型对匿名结果进行优化,通过设置服务质量阈值,限制服务质量损失的最大值,并通过定义相关的线性规划,在保证用户服务质量的情况下,最优化用户的位置隐私保护水平.

本文第 2 节主要介绍现有的假位置匿名算法、Stackelberg 博弈位置隐私保护算法以及相关系统结构的研究进展;第 3 节主要介绍本文所使用的基于半可信第三方服务的隐私保护系统结构;第 4 节详细介绍基于历史信息的假位置和偏移位置选择策略;第 5 节介绍如何通过 Stackelberg 博弈对假位置及偏移位置选择进行优化的过程;第 6 节在安全性和可用性方面,对系统算法进行性能分析;第 7 节主要介绍实验仿真及相关实验结果;第 8 节对现有工作进行总结.

2　相关工作

2.1　假位置匿名算法

假位置(dummy)匿名作为位置泛化的重要手段，具有部署简单且不影响服务质量的优点，得到了学者们广泛的关注.

Lu 等人[22]提出了一种 PAD 方法，该方法根据虚拟的网格和圆圈生成所需个数的假位置，利用这些假位置形成匿名集合，覆盖用户的真实位置. 该方案主要考虑两方面问题，一方面 dummy 分布不均匀时，零散的 dummy 很容易被攻击者排除掉，故方案采用均匀产生 dummy 的方法；另一方面考虑了 dummy 所形成的隐私保护区域应当超过某个阈值，以防止隐私保护区域太小直接泄露用户位置.

Zha 等人[23]在 PAD 的基础上，提出了基于虚拟圆的 CirDummy 算法和基于虚拟网格的 Grid-Dummy 算法，这两种算法将 k-匿名所需的 k 个位置全部用假位置代替(即提交的所有位置中不包含用户的真实位置)，进一步提高攻击者推测用户真实位置的难度.

Niu 等人[24]提出 DLS 方案，DLS 根据每个位置的查询概率来生成 dummy，并通过引入位置熵来度量 dummy 的隐私保护程度. 理想情况下，当所有 dummy 的查询概率与真实位置相同时，熵值最大，隐私保护水平最高. 随后，为了提高查询缓存命中率，Niu 等人[25]提出了一种基于缓存的 dummy 选择算法，该算法基于信息熵来度量缓存对隐私保护程度的影响，并通过最大化查询隐私和 dummy 缓存贡献度来增强位置隐私保护水平.

Pingley 等人[26]考虑了用户的运动模型和历史查询上下文信息，提出了一种虚拟查询生成方案 DUMMY-Q. 该方案使用 POOL-BUILDER 来生成虚拟查询，并基于四叉树的存储/检索算法降低 DUMMY-Q 的存储和计算开销.

2.2　Stackelberg 博弈位置隐私保护算法

在大数据环境下，攻击者可以利用搜集到的各种数据对用户的位置隐私信息进行推测[27]. 面对这种推测攻击，基于概率推理的位置隐私保护机制[28-29]逐渐受到了学者们的关注. 在请求服务过程当中，此类方法要对用户真实位置进行一定程度的偏移，以此限制攻击者进行位置概率推测的能力，而 Stackelberg 博弈也成为此类方法中平衡隐私保护级别与服务质量要求的重要手段.

Shokri 等人[30]提出基于 Stackelberg 博弈的保护策略，该策略假设攻击者已获取先验知识，让用户和攻击者轮流进行博弈：用户在确保服务质量损失小于给定阈值的情况下最大化隐私保护水平，而攻击者根据先验知识和偏移位置力求最小化隐私保护水平. 通过博弈，该策略最终可以在最优化隐私保护水平的同时确保服务质量损失小于给定阈值.

Bordenabe 等人[31]在 Shokri 等人方案的基础上，引入差分隐私(Differential privacy)，构建了一种最优化服务质量的隐私保护机制. 由于差分隐私可以不基于先验知识，因此该机制可以在满足位置不可分辨性的前提下，最小化服务质量损失.

Shokri 等人[32]进一步提出使用差分隐私和失真隐私(Distortion privacy)两个指标来优化基于 Stackelberg 博弈的隐私保护策略. 差分隐私限定了用户隐私泄露的程度，而失真隐私则度量了推测用户隐私的错误率，结合这两个标准，该隐私保护策略能够在保证服务质量的同时，抵御更多种类的推测攻击，满足隐私保护的要求.

2.3　位置隐私保护系统结构

随着大量的位置隐私保护技术被提出，各种隐私保护技术所依赖的系统结构呈现出明显的类别差异. 系统结构作为隐私保护技术执行的载体，得到广泛的研究和发展. 目前比较主流的系统结构有以下三种：用户独立的系统结构、依赖可信第三方的系统结构和用户自组织的系统结构.

用户独立的系统结构[32]，是一种仅包括客户端(即移动设备)和服务器端的 C/S 结构. 该体系结构将每个用户当作独立的个体，让用户设备执行位置隐私保护机制，直接向服务器端发送服务请求，并接收查询结果. 这种系统结构的优势在于部署十分简单，方便用户根据其隐私保护需求调整隐私保护粒度，但是隐私保护算法的实现受移动设备性能的限制，且无法获取全局用户位置信息.

依赖可信第三方的系统结构[33]由第三方可信匿名服务器获取所有用户的位置信息，并负责执行位置隐私保护机制，是目前较为常用的隐私保护系统结构. 这种系统结构的优势在于匿名服务器能够获取海量用户的位置信息，并且辅助用户过滤服务数据，其缺点在于匿名服务器可能成为攻击者的目标，在实际使用中它的可信性难以保障.

用户自组织的系统结构[34]利用用户组网来执行位置隐私保护机制. 在发起位置服务请求前，用户需要在一定范围内寻找合适的其他用户进行合作，以此来完成对个人精确位置的匿名. 这种系统结构的优势在于部署相对灵活，不需要额外的服务器等

《计算机学报》

其它中间件,其缺点在于恶意节点容易混入匿名组中直接窃取用户的隐私信息,同时也会存在着由于用户数量不足导致匿名失败的问题.

基于假位置的匿名算法一般需要依赖可信第三方匿名服务器,因此有由于第三方服务器被攻陷而导致的用户隐私泄露的风险;基于 Stackelberg 博弈的位置隐私保护算法一般使用用户独立(user centric)的系统结构,不需要依赖匿名服务器,但是缺乏对全局信息(例如所有用户的历史服务请求分布)的掌控.针对上述问题,本文提出了一种基于半可信第三方服务的隐私保护体系结构 STTP,并在 STTP 基础上设计了一种基于假位置和 Stackelberg 博弈的位置匿名算法.算法虽然依赖于第三方服务器,但是却假定第三方服务并不是完全可信的(第三方服务器可能被攻破),因此在最大程度上保护了用户的位置隐私.

3 基于半可信第三方服务的隐私保护系统结构

3.1 系统结构组成

针对目前位置隐私保护系统结构的不足,本文设计了一种基于半可信第三方服务(Semi-Trusted Third Party,STTP)的隐私保护系统结构,如图 1 所示,该结构主要由以下四个部分组成:

(1)用户(User).User 使用移动终端在需要的时候发起位置服务请求;

(2)假名服务器(Pseudonym Server,PS).PS 负责提供对应用户的假名 U_{pseu} 和查询信息的加密解密密钥 key_{public},$key_{privacy}$;

(3)位置匿名服务器(Location Anonymization Server,LAS).LAS 负责将用户的精确位置 l 转化成匿名集,并在 LBS 提供商返回查询结果后,抽取合适的服务信息返回给用户;

(4)LBS 提供商(Location Based Service Provider,LBSP).LBSP 根据位置查询请求,返回对应的服务结果.

STTP 方案设定 PS、LAS 和 LBSP 都是"负责且诚实"的,一方面它们不会破坏协议流程,能按照协议完成自己的工作;另一方面,它们只收集并分析自己职责范围内的用户数据.同时,更进一步设定 PS、LAS 和 LBSP 是不能监守自盗、相互勾结的(即它们不会同时被一个攻击者控制).这个设定是合理的,因为如果三方共谋,对于用户将没有任何秘密可言.

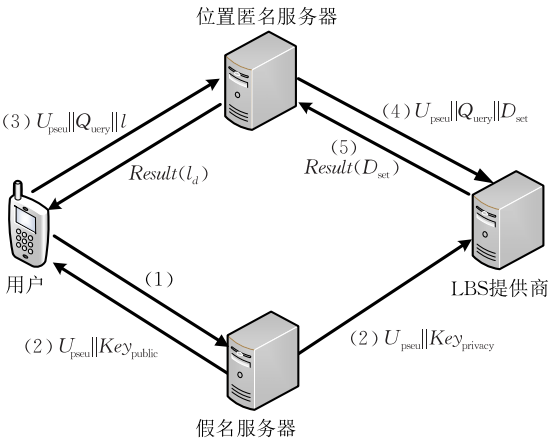


图 1 基于半可信第三方服务的系统结构

3.2 系统匿名过程

STTP 的匿名过程如下(图 1 所示):

(1)移动用户 User 在发起位置服务请求前,首先向假名服务器 PS 请求假名 U_{pseu} 和加密密钥 Key_{public};在限定时间 $t_{session}$ 内,对于同一位置的多次服务请求,User 只申请一次假名和密钥;当时间超过 $t_{session}$,或者位置 l 发生改变,User 会重新申请新的假名和密钥.

(2)假名服务器 PS 为用户生成对应的假名 U_{pseu} 和 RSA 密钥对〈Key_{public},$Key_{privacy}$〉,并将 U_{pseu} 和 Key_{public} 返回给用户,将 U_{pseu} 和 $Key_{privacy}$ 发送给 LBS 提供商(这里作为举例,只是使用经典的 RSA 算法进行加密,STTP 可以根据实际要求更换其他的加密算法).需要注意的是,STTP 要求 PS 只作为假名和密钥的提供者,因此 PS 在本地并不会对相关的假名和密钥进行保存.

(3)移动用户 User 通过密钥 Key_{public} 对自己的查询内容进行加密,并将自己当前的假名 U_{pseu}、加密后的查询内容 Q_{uery} 以及当前的精确位置 l 一起发送给位置匿名器 LAS.

(4)位置匿名器 LAS 在收到信息后,根据对应的位置匿名算法对用户的真实位置 l 进行隐匿,生成位置匿名集 D_{set},并将用户的假名 U_{pseu}、加密后的查询内容 Q_{uery} 和匿名集 D_{set} 发送给 LBS 提供商(对于连续的查询,同一 U_{pseu},使用相同的匿名集 D_{set}).

(5)LBS 提供商从匿名器 LAS 接收到位置服务请求后,会根据用户的假名 U_{pseu} 搜索其对应的解密密钥 $Key_{privacy}$,对用户的查询内容进行解密,并根据查询内容及匿名集 D_{set} 提供相应的服务结果,返回给 LAS.

(6)位置匿名器 LAS 在接收到相应的服务结果后,根据用户的假名 U_{pseu} 识别出其对应的位置 l_d,从服务结果中筛选出查询结果,并返回给移动用户 User.

3.3　系统结构分析

半可信第三方服务系统结构的优势在于用户的个人信息分别由 PS、LAS 和 LBSP 三方保存，具体用户信息的分布如表 1 所示. 攻击者不容易同时成功攻破三方的服务器，因此这种系统结构可以有效地保证用户隐私不被泄露.

表 1　用户的主要信息分布

组成	用户身份	假名	精确位置	匿名位置	密钥	查询内容
User	✓		✓			✓
PS	✓	✓			✓	
LAS		✓	✓	✓		
LBSP		✓		✓		✓

如表 1 所示，用户隐私信息主要包括用户身份、用户所处的精确位置以及查询内容. 在半可信系统结构中，PS 短暂了解用户的身份，LAS 保存用户的精确位置，而 LBSP 保存用户的查询信息，即使攻击者得到任意一方保存的用户数据，也不能精准定位用户，得到用户完整的隐私信息.

基于上述半可信第三方服务系统结构，本文提出了一种基于假位置和位置偏移的位置匿名算法，并基于 Stackelberg 博弈模型对位置匿名结果进行了优化.

4　基于假位置和位置偏移的位置匿名算法

基于假位置的位置匿名算法具有计算开销较小、不需要匿名合作、匿名成功率较高等优点，但相应地也存在着假位置容易被攻击者进行过滤的问题. 本节结合了假位置和位置偏移的特点，提出了一种基于用户查询历史的位置隐私匿名算法，最大限度避免攻击者的单点攻击，保护用户的位置隐私.

4.1　攻击者单点攻击模型

从时间维度上看，攻击者依靠截获的单次位置服务请求来推断用户的隐私信息，称为单点攻击模型[35]. 模型中，攻击者主要的攻击方式包括边信息攻击和位置同质攻击等.

边信息[24]是指攻击者所使用的用来过滤假位置、辅助缩小匿名度的信息. 例如，对于随机生成的假位置匿名集，某些位置可能处于河中或者无人区中，攻击者根据地图信息很容易将这些位置过滤掉. 假设位置匿名度要求为 k，当选择的 $k-1$ 个假位置中有 k' 个假位置被攻击者根据边信息过滤掉，则不满足 k 匿名要求，致使隐私保护水平下降.

位置同质攻击[35]是指攻击者分析匿名集中的多个位置，如果位置距离非常接近(如位于同一个广场内等)，虽然达到 k 匿名的要求，但由于隐匿空间太小，用户的位置隐私仍然无法得到很好的保护. 另外，攻击者也可以通过位置聚类的方式进行过滤，以此增加了推测出用户真实位置的概率.

综上，攻击者的单点攻击模型可以形式化表述为

$$h=\arg\max_{h} k'|h \tag{1}$$

其中 h 为攻击者所采取的操作(边信息过滤、位置聚类等)，k' 是攻击者通过 h 过滤掉的假位置个数，所以单点攻击的目的是在给定 h 的条件下最大化 k'.

4.2　基于查询历史的假位置选择算法

针对攻击者的单点攻击，本文提出了一种基于历史查询数据的假位置选择算法 HCLS(History based Cloaking Location Selection algorithm). 下面介绍算法的具体流程：

(1) 计算历史服务请求的位置概率分布

HCLS 将整个位置空间网格化，每个网格对应一个位置，空间划分的网格总数为 n.

由于位置匿名器 LAS 可以获得所有用户的位置服务请求，根据这些历史查询数据，HCLS 可以计算出每个位置 i 发起查询请求的概率 p_i，其满足如下条件：

$$\sum_{i}^{n} p_i=1 \tag{2}$$

即整个网格空间中所有位置发起请求的概率之和为 1.

图 2 为用户历史查询请求的位置概率分布示意图，图中颜色的深度表示服务查询概率的大小(颜色越深，概率越大)，白色区域表示该位置从未有过位置服务请求，所以这些空白位置可能是河流、荒山等容易被过滤的位置.

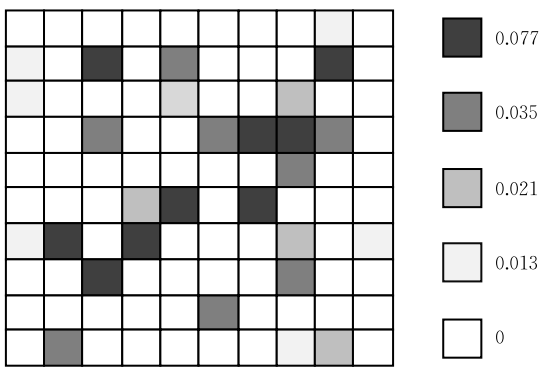


图 2　位置请求概率分布示意图

《计算机学报》

(2)选择偏移位置 l_d

位置偏移(Location Obfuscation)指的是通过加入噪声来降低位置精度,在此阶段,HCLS 通过选出一个偏移位置替代用户的真实位置放入匿名集中,以此来达到保护用户隐私的目的.

HCLS 在得到历史查询数据的位置概率分布后,在用户真实位置 l 的周围选择出距离最近的 $m-1$ 个 $p_i>0$ 的位置,与 l 一起组成偏移位置候选集合 $M_{\text{set}}=\{d_1,d_2,\cdots,d_{m-1},l\}$(这里,将 l 加入到 M_{set} 的原因是防止 l 周围没有合适候选位置的情况出现).从 M_{set} 中随机选择一个偏移位置 l_d 作为 l 的替代位置,具体流程如图 3 所示.

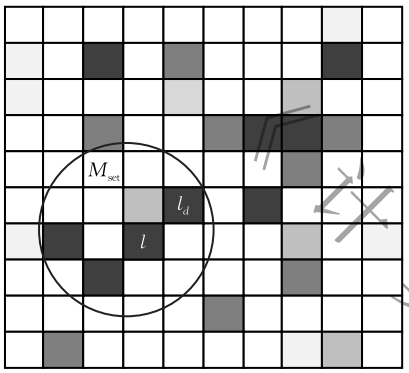


图 3 偏移位置选择示意图

(3)选择假位置集合

在整个网格空间中,选择出 $2k$ 个 $p_i>0$ 的位置(k 为匿名度),组成假位置候选集合 D_{Lset}.

为了解决匿名结果中位置过于集中的问题,HCLS 要对 D_{Lset} 进一步进行优化. HCLS 从 D_{Lset} 中选择出包含 l_d 在内的两两位置之和最大的 k 个位置,如式(3)所示,以此提高假位置的分散度(这里,利用 l_d 来生成匿名集的主要原因是为了进一步减小攻击者推测用户真实位置 l 的概率,增强算法的隐私保护效果).

$$D_{\text{set}}=\arg\max_{D_{\text{set}}}\left\{\sum_{i\neq j}^{k}Dis(d_i,d_j)\right\} \tag{3}$$

其中,$Dis(d_i,d_j)$ 表示任意两个假位置之间的距离,选出的 k 个假位置距离和越大,说明假位置之间越分散.

图 4 描述了假位置集合的选择过程,l_d 为偏移位置.为满足距离之和最大,选择 A、C 作为假位置组成匿名集.

(4)获得匿名集 D_{set}

通过步骤(3),HCLS 选择出 $k-1$ 个假位置,与偏移位置 l_d 一起,组成最终的匿名结果集

$$D_{\text{set}}=\{d_1,d_1,\cdots,d_{k-1},l_d\} \tag{4}$$

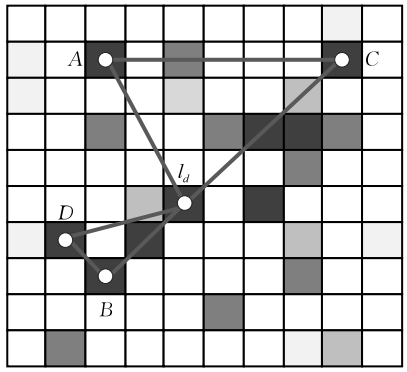


图 4 假位置选择示意图

HCLS 具体的算法流程,如算法 1 所示.利用用户的历史位置请求数据,HCLS 可以选择有效的假位置来实现对用户真实位置的隐匿,以此解决假位置容易被过滤或过于集中的问题,避免攻击者的单点攻击.

算法 1. 基于历史位置的假位置选择算法.

输入:用户位置 l、匿名度 k、历史位置查询数据 data

输出:假位置集合 D_{set}

1. 将空间划分为网格
2. 计算各网格位置查询分布
3. 偏移位置集合 M_{set}←选择与 l 最近的 $m-1$ 个 $p_i>0$ 的位置与 l 一起组成 M_{set}
4. 偏移位置 l_d←在集合 M_{set} 中随机选择一个位置
5. 假位置候选集合 D_{Lset}←在查询概率 $p_i>0$ 的位置中随机选择 $2k$ 个位置
6. 初始化含有 k 个假位置集合 D_{set}
7. WHILE $i<$迭代次数
8. D_{Lset}←从 D_{Lset} 中选择 $k-1$ 个假位置
9. D'_{set}←$D'_{\text{Lset}}+l_d$
10. IF D'_{set} 中的两两位置距离之和$>D_{\text{set}}$
11. $D_{\text{set}}=D'_{\text{set}}$
12. END IF
13. $i++$
14. END WHILE
15. RETURN D_{set}

5 基于 Stackelberg 博弈的位置匿名优化方法

为了阻止攻击者对用户位置隐私的推测攻击,本文引入 Stackelberg 博弈模型,通过将用户位置隐私的保护模型和攻击者的位置推测模型进行博弈,优化 HCLS 位置匿名过程,进一步地提高位置隐私保护水平.

5.1 攻击者推测攻击模型

在推测攻击模型中,攻击者是掌握一定的背景

《计算机学报》

知识的，此类背景知识包括攻击者收集到的用户的服务请求历史记录、攻击者了解到的位置隐私保护机制等.

利用收集到用户的服务请求历史记录(该记录不一定是由单一数据源中获取的，也可能是由多种数据源综合总结出的)，攻击者可以计算出该用户的查询概率分布 $\varphi(l)$. 当用户再次发送查询请求时，若匿名集内位置概率分布不均匀，则攻击者可以推断出用户很可能位于概率较高的位置上.

利用了解到的位置隐私保护机制，攻击者可以对截获的位置请求进行分析，结合匿名算法，推测匿名集中的每个位置是用户真实位置的概率，以此使推测攻击更为精确. 这里，我们将攻击者了解的隐私保护匿名机制定义为 $f(D_{set}|l)$.

根据已知知识(用户历史位置分布、隐私保护匿名机制以及 D_{set})，攻击者可以推测出的用户真实位置 l 的后验概率如式(5)表示：

$$\Pr(l|D_{set})=\frac{\Pr(l,D_{set})}{\Pr(D_{set})}=\frac{f(D_{set}|l)\varphi(l)}{\sum_{l}f(D_{set}|l)\varphi(l)} \tag{5}$$

根据推测出的所有可能的真实位置的分布，攻击者希望找到一个使用户位置隐私度最小的位置 l'，从而最大化自己的收益，式(6)为计算用户所有可能的位置与攻击者推测出的位置距离的期望.

$$\sum_{l}\Pr(l|D_{set})Dis(l',l) \tag{6}$$

其中，$Dis(l',l)$表示攻击者推测出的用户位置 l' 与用户真实位置 l 间的距离.

综上，攻击者可以通过式(7)来选择出最优的 l'，使用户的位置隐私度最低：

$$l'=\arg\min_{l'}\sum_{l}\Pr(l|D_{set})Dis(l',l) \tag{7}$$

当有多个位置同时满足条件时，攻击者随机从中选择一个作为推测结果.

5.2　Stackelberg 博弈优化过程

在这里，我们假定攻击者具有一定的背景知识，它会根据用户的历史位置请求分布 $\varphi(l)$ 以及位置匿名服务器 LAS 所使用的位置隐私保护算法 $f(D_{set}|l)$，结合匿名结果 D_{set}，尽可能推测用户的真实位置 l.

相对的，我们可以假设 LAS 知道的所有背景知识都会被攻击者利用，所以 LAS 可以将攻击者的最优攻击策略作为参数，优化位置匿名集 D_{set} 的生成过程，使攻击者推测位置与用户真实位置的距离尽可能大，最大化算法的效果.

对于上述过程，我们使用 Stackelberg 博弈进行建模. 在 Stackelberg 博弈模型中，我们将 LAS 作为领导者，攻击者作为跟随者. 博弈要求领导者先行做出决策，跟随者在观察到领导者的决策后再做选择，因此 LAS 需要先行给出位置匿名结果 D_{set}，而攻击者则根据背景知识以及 D_{set} 优化自己的攻击策略.

我们使用攻击者推测位置 l' 与用户真实位置 l 的距离来衡量博弈中参与者的收益：距离越大，LAS 收益越大，说明匿名算法隐私保护水平越好；相反，距离越小，攻击者收益越大，说明攻击者攻击策略越有效.

假设本博弈模型为零和博弈，攻击者推测出的用户位置用 l' 表示. l' 与用户真实位置 l 的距离即为 LAS 的收益. 距离越大，LAS 收益越大，说明匿名算法隐私保护水平越高；相反，距离越大，攻击者收益越小，说明其越不容易窥探用户的位置隐私.

Stackelberg 博弈优化的目的是找出强 Stackelberg 博弈纳什均衡[36](Strong Stackelberg Equilibrium，SSE)，让攻击者无法通过优化攻击策略获得更多的收益(即无法对用户的精确位置进行更准确的推测). 在 Stackelberg 博弈中，SSE 强制跟随者在多个策略收益相同时，总是选择对领导者最有利的策略，因此能够保证均衡的存在性[37]；另外一方面，领导者由于先手优势，能够通过选择无限接近于均衡解的策略使得跟随者倾向于选择对它有利的策略，从而达到更想要的强均衡，这也保证了 SSE 在本算法中的可用性[38]. 本文将经过 Stackelberg 博弈优化过的 HCLS 表示为 HCLS_SG.

需要注意的是，D_{set} 是 l_d 进一步匿名得到的结果，因此 $f(D_{set}|l)$ 可以用式(8)进一步表示：

$$f(D_{set}|l)=f(l_d|l)\cdot f(D_{set}|l_d) \tag{8}$$

在某些情况下，$f(D_{set}|l)$ 与 $f(l_d|l)$ 是相等的，因为攻击者在这些情况下可以将 l_d 除外的假位置过滤掉.

根据 HCLS，D_{set} 的生成概率可以表示为 $\Pr(D_{set})$，计算过程如式(9)所示.

$$\Pr(D_{set})=\sum_{l}\varphi(l)f(D_{set}|l) \tag{9}$$

结合攻击者推测攻击模型的定义(如式(7))，我们可以计算攻击者采取最优攻击策略下 HCLS_SG 的收益期望，具体过程如公式所示.

$$\sum_{D_{set}}\Pr(D_{set})\min\sum_{l}\Pr(l|D_{set})Dis(l',l)=$$
$$\sum_{D_{set}}\min\sum_{l}\varphi(l)f(D_{set}|l)Dis(l',l) \tag{10}$$

为了方便表述，用 x 对式(10)进行相应的简化，x 的定义如式(11)所示：

计算机学报

$$x=\min\sum_{l}\varphi(l)f(D_{\text{set}}|l)Dis(l',l) \tag{11}$$

因此,HCLS_SG 寻找最优 D_{set} 的过程,如式(12)所示.

$$D_{\text{set}}=\arg\max_{D_{\text{set}}}\sum_{D_{\text{set}}}x \tag{12}$$

该公式的含义可以理解为,在所有令攻击者收益期望最大化的策略中,找到一个 D_{set},让 HCLS_SG 的收益最大化.

为了计算方便,这里可以将式(11)变形为一个线性的条件约束,约束条件如式(13)定义:

$$x\leq\sum_{l}\varphi(l)f(l_d|l)f(D_{\text{set}}|l_d)Dis(l',l),\forall l' \tag{13}$$

另外,HCLS_SG 需要以保证用户服务质量为前提对用户的真实位置进行隐匿.为了保证服务质量,我们可以通过设置服务质量阈值 $Q_{\text{loss}}^{\max}$ 来限制服务信息质量损失的最大值,具体过程如式(14)所示.

$$\sum_{l,l_d}\varphi(l)f(l_d|l)Dis(l_d,l)\leq Q_{\text{loss}}^{\max} \tag{14}$$

综上,HCLS_SG 可以利用线性规划来进行求解,线性规划最后的定义如式(15)所示.

$$\begin{aligned}&\text{Maximize}\ \sum_{D_{\text{set}}}x\\&\text{Subject to}\\&x\leq\sum_{l}\varphi(l)f(l_d|l)f(D_{\text{set}}|l_d)Dis(l',l)\\&\sum_{l}\varphi(l)\sum_{l_d}f(l_d|l)Dis(l_d,l)\leq Q_{\text{loss}}^{\max}\\&\sum_{D_{\text{set}}}f(D_{\text{set}}|l)=1\\&f(D_{\text{set}}|l)\geq 0,\forall l,D_{\text{set}}\end{aligned} \tag{15}$$

其中,条件 1 用来让攻击者的收益最大化;条件 2 用来约束匿名集生成过程中信息质量损失;条件 3 表示匿名集的生成概率之和必须为 1;条件 4 表示每个备选匿名集的概率都大于 0.

HCLS_SG 在式(15)中的约束条件下对目标函数进行求解,得到最优的匿名集选择结果,在满足服务质量的条件下让位置隐私度最大化.

6 性能分析

由于 HCLS 和 HCLS_SG 是基于 STTP 的,此部分将结合 STTP,讨论本文整体方案的隐私保护性能.

6.1 隐私保护度量指标

定义 1. 位置暴露概率 *PERL*(Probability of Exposing Real Location),本文使用用户真实位置的暴露概率来衡量假位置的隐私度保护水平,其计算方式如式(16)所示.

$$PERL=\frac{1}{k-k'} \tag{16}$$

其中,k 为匿名度,k' 表示攻击者可以过滤掉的位置数量.

假位置被攻击者过滤掉的数量越多,*PERL* 越大,则暴露用户真实位置的概率就越大;*PERL* 越小,则位置隐私度越高.

定义 2. 位置分散度 *PD*(Product of Distances),匿名集假位置的分散程度可以通过计算匿名集中两两位置距离之和来获得,具体过程如式(17)所示.

$$PD=\sum_{i\neq j}^{k}Dis(d_i,d_j) \tag{17}$$

一般而言,位置分散度越高,匿名集的覆盖范围越大,因此会有更好的匿名效果.

定义 3. 隐私保护度 *PL*(Privacy Level),*PL* 可以用攻击者推测的用户位置 l' 与用户位置 l 间距离的数学期望来表示.

我们将攻击者产生 l' 的过程用 $h(l'|D_{\text{set}})$ 表示,则隐私保护度的计算过程如式(18)所示.

$$PL=\sum_{l,D_{\text{set}},l'}\varphi(l)f(D_{\text{set}}|l)h(l'|D_{\text{set}})Dis(l',l) \tag{18}$$

隐私保护度 *PL* 直接反应攻击者对用户真实位置的推测情况,*PL* 越大,攻击者推测结果越不准确,因此算法的位置隐私保护效果越好.

定义 4. 服务质量的损失 Qos_{loss},Qos_{loss} 可以用匿名集 D_{set} 中的偏移位置 l_d 与用户真实位置 l 距离的数学期望表示,具体如式(19)所示:

$$Qos_{\text{loss}}=\sum_{l,l_d}\varphi(l)f(l_d|l)Dis(l_d,l) \tag{19}$$

我们将用户可接受的最大服务质量损失用 $Q_{\text{loss}}^{\max}$ 表示,一般来说位置隐私保护算法需要保证 $Qos_{\text{loss}}\leq Q_{\text{loss}}^{\max}$,因为超过该值则得到的位置服务请求结果无法满足用户的需求.

6.2 安全性分析

STTP 方案设定 PS、LAS 和 LBSP 都是“负责且诚实”的,而且它们不会同时被一个攻击者控制.在以上设定的前提下,对几个定理进行证明,以说明整体方案的安全性.

引理 1. 在获得匿名集 D_{set} 的情况下,攻击者通过单点攻击和推测攻击获得用户真实位置 l 的概率是极小的.

证明. HCLS 在设计时考虑了 *PERL* 和 *PD* 两个度量指标,因此能够抵抗攻击者的单点攻击;而 HCLS_SG 考虑了 *PL* 和 Qos_{loss} 两个度量指标,因此

《计算机学报》

能够抵抗攻击者的推测攻击(此部分内容将在 7.2 节和 7.3 节通过实验进一步进行验证).

假设匿名集 D_{set} 中的假位置个数为 k,因此攻击者从匿名集 D_{set} 中推测出偏移位置的概率为 $p_1=1/k$.设用户历史分布中的位置个数为 n,攻击者通过偏移位置确定用户真实位置是可能发生的事件,但存在小概率函数 $negl_1(n)$ 使攻击者确定用户真实位置的概率 p_2 满足 $p_2\leq negl_1(n)$.那么在已知 D_{set} 的情况下获得确切的用户位置信息的概率 $p=p_1p_2\leq negl_1(n)/k\leq negl(n)$. 证毕.

定理 1. 当攻击者攻陷 PS 时,能够获取用户身份、假名和密钥的概率是极小的.

证明. 由 STTP 步骤(2)可知,STTP 要求 PS 只作为假名和密钥的提供者,不允许其在本地保存用户的假名和密钥;另一方面,PS 的假名分配是动态的,同一用户的不同请求会对应不同的假名;最后,STTP 设定 PS 负责且诚实,能够按照协议负责地完成自己的工作.因此,攻击者通过攻陷 PS 来获得用户的假名和密钥,进而定位用户身份是个小概率事件 $p\leq negl(s)$. 证毕.

定理 2. 当攻击者控制 LAS 时,能够同时获取用户身份、真实位置和查询内容的概率极小.

证明. LAS 与用户通信获得的请求信息主要包括$\langle U_{pseu},Q_{uery},l\rangle$.虽然攻陷 LAS,攻击者获取位置 l 的概率 $p_1=1$,但由定理 1 可知,攻击者获取用户身份和对应 U_{pseu} 的概率 $p_2\leq negl(s)$.因此,LAS 获取用户身份和真实位置的概率 $p_3=p_1p_2\leq negl(s)$.另外,由于 Q_{uery} 是由 RSA 加密过的,在没有私钥的情况下破译 Q_{uery} 是个小概率事件 $p_4\leq negl(s)$,因此同时获取用户身份、真实位置和查询内容的概率是极小的. 证毕.

定理 3. 当攻击者控制 LBSP 时,能够同时获取用户身份位置、真实位置和查询内容的概率极小.

证明. LBSP 获得的信息包括$\langle U_{pseu},Q_{uery},D_{set}\rangle$.由引理 1 可知,攻击者攻陷 LBSP,分析出位置 l 是个小概率事件 $p_1\leq negl(n)$;由定理 1 可知,攻击者获取用户身份与 U_{pseu} 对应关系也是个小概率事件 $p_2\leq negl(s)$;由于 LBSP 保存了用户的假名和私钥的对应信息,因此攻击者获取查询信息的概率 $p_3=1$,但攻击者同时获取用户身份、位置及查询信息的概率仍是极小的 $p=p_1p_2p_3\leq negl(n)$. 证毕.

定理 4. 攻击者相继(非同时)攻陷任意两方或三方服务器时,能够获取用户身份及关键隐私信息的概率是极小的.

证明. 按照 PS 的协议要求,假名的分配是动态且不被保存的.当攻击者相继攻陷 PS 和 LAS 时,即使攻击者从 PS 中获取了一些假名信息,但由于时间原因,这些假名都已失效,攻击者仍没办法定位用户;当攻击者相继攻陷 PS 和 LBSP 时,同样的原因,获取的假名也已失效,攻击者没有办法将查询内容和用户身份相匹配;最后,当 LAS 和 LBSP 两方,或 PS、LAS 和 LBSP 三方被相继攻陷,由于假名原因,攻击者没有办法定位用户,获取用户的完整隐私信息. 证毕.

这里需要说明的是,当攻击者同时攻陷任意两方服务器时,本方案确实会导致部分用户隐私的泄露,但是,在这种进攻成功率下(需要同时成功攻陷两个服务器,概率很低),传统的可信第三方系统结构 TTP(Trusted Third Party)服务器均已被攻陷,所有的用户信息也已泄露,与这种情况相比,本文的方案还可以保证用户部分信息的安全性.

综上,在保证三方服务器尽职且不共谋、服务器不被同时攻陷的情况下,整体方案不但能够达成所设定的安全目标,而且要比目前主流使用的 TTP 方案更加安全.

6.3 可用性分析

这部分,我们将从计算复杂度和通信消耗两方面来验证本文方案的可用性.

(1)计算时间

在 User 端,STTP 主要完成查询内容的加密操作,因此主要计算时间取决于加密算法的计算复杂度,一般加密算法的计算复杂度为 $O(C)$,C 为常数项.

在 LAS 端,STTP 主要完成位置匿名操作,因此 LAS 的计算时间主要取决于 HCLS 以及 HCLS_SG 算法的计算复杂度,这两个算法的计算复杂度均为 $O(k^2)$,k 为位置匿名度.

在 PS 端,STTP 主要是需要 PS 提供用户的假名和密钥,此类操作需要的时间为常数级,用 $O(C)$ 表示.

在 LBSP 端,STTP 主要完成查询内容解密、搜索 POI 信息操作,因此主要计算时间取决于解密算法的计算复杂度以及 POI 的搜索速度.一般解密算法的计算复杂度为 $O(C)$,而 POI 的搜索速度取决于搜索的位置个数 k,用 $O(k)$ 表示.

综上,STTP 的整体计算时间不会超过多项式复杂度(最高为 $O(k^2)$),因此有较高的可用性.

(2)通信消耗

PS 的通信发生在 PS 和 User 间以及 PS 和 LBSP 间.由于这些通信只是简单的请求以及传输少量的

《计算机学报》

数据，因此通信消耗可以定义为常数级，记为 $O(C)$.

LAS 的通信发生在 LAS 和 User 间以及 LAS 和 LBSP 间. LAS 和 User 间通信消耗为常数级 $O(C)$，而 LAS 和 LBSP 间主要传输匿名集以及匿名集的查询结果，因此通信消耗取决于匿名集的匿名度 k 以及 POI 的个数 m，记为 $O(km)$.

最后，我们考虑 LBSP 的通信消耗，LBSP 和 LAS 的通信消耗为 $O(km)$，LBSP 和 PS 的通信消耗为 $O(C)$.

综上，STTP 的整体通信时间不会超过多项式复杂度（最高为 $O(km)$），因此有较高的可用性.

7 仿真实验

7.1 实验数据集

本实验使用滴滴打车的订单数据①（由滴滴大数据比赛发布）进行算法验证. 数据集中给出了 M 市 2016 年连续三周的 900 万条数据信息：订单数据包括订单 ID、区域 ID、司机 ID、乘客 ID、价格以及具体的订单时间；附属表给出了天气、区域 POI 数量及拥堵情况等信息.

本文需要基于用户发起请求的位置概率分布来进行实验，所以对数据集进行了预处理. 我们将位置空间根据区域 ID 分区为 13×13 的网格，计算不同区域中的订单总数，然后通过式(20)计算每个分区中用户发起位置请求的概率 p_i^{num}：

$$p_i^{num}=\frac{Onum_i}{Tnum},\ 0<i<13\times13 \tag{20}$$

其中，$Onum_i$ 表示一个网格分区中的订单数量，$Tnum$ 表示订单总数.

7.2 HCLS 算法性能评价

图 5 展示了某时间段内服务请求的位置概率分布情况. 在图 5 中，不同的灰度表示用户位置服务请求的不同概率，后续实验都是基于这样的概率分布图来完成的.

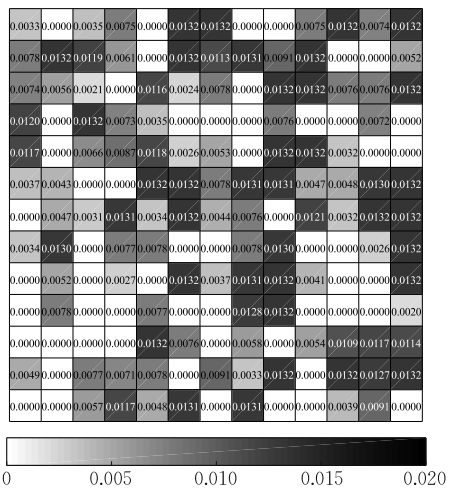
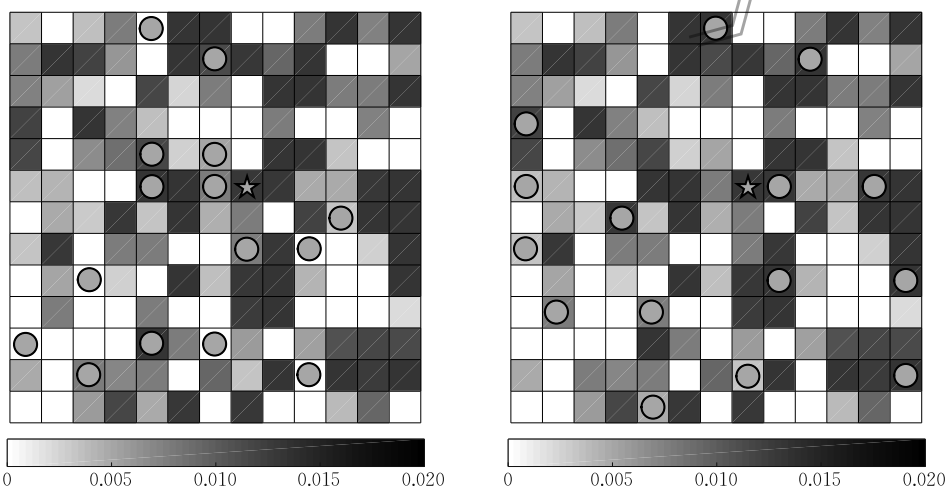


图 5 位置请求概率分布

为了验证算法的有效性，本文首先将 HCLS 与随机选择的方法（Random）进行了对比，具体匿名集合的位置分布情况如图 6 所示.



(a) Random假位置选择结果 (b) HCLS假位置选择结果

图 6 具体匿名集合的位置分布

图 6 中，标注星号的网格是用户的真实位置，而标注圆圈的网格即为选中的假位置. 图 6(a)是随机方式选择出的假位置的分布情况，图 6(b)是 HCLS 的假位置分布情况. 由图中可以看出，HCLS 选择出

① 滴滴打车数据集：http://research.xiaojukeji.com/competition/detail.action?competitionId=DiTech2016

《计算机学报》

的假位置结果更为分散，而 Random 选出的部分假位置出现在概率是 0 的位置上（这些位置容易被攻击者利用边信息进行过滤，因此会降低用户的隐私保护度，本文将这些位置定义为无效位置）.

本文采用改进的基于虚拟圆的 CirDummy 算法[24]以及基于虚拟网格的 GridDummy 算法[24]进行实验对比，以此验证 HCLS 算法的有效性. 需要注意的是，这三种算法提交的匿名集均不包含用户的真实位置. CirDummy 和 GridDummy 算法是将与用户真实位置最近的假位置（这里称之为“最近假位置”）的请求结果作为提交给用户的服务结果；而 HCLS 是通过偏移位置 l_d 来返回服务结果的.

(1) *PERL* 的比较

PERL 的定义如式(16)所示，三种算法的 *PERL* 对比结果见图 7. 由图中可以看出，随着 k 值的增加 *PERL* 呈下降趋势，这表示设置的 k 值越大，算法隐私保护效果越好，攻击者更加难以推测用户的真实位置.

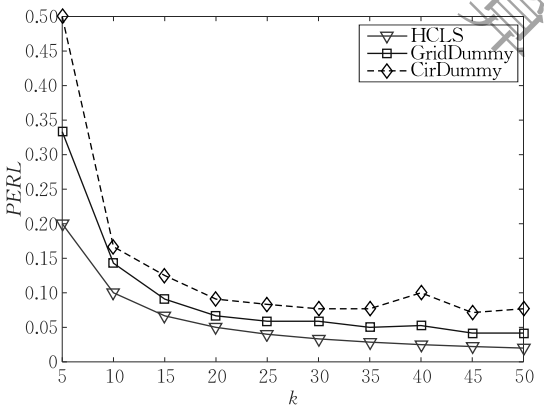


图 7　*PERL* 对比

通过图 7 可以发现，HCLS 算法泄露用户真实位置的概率更低，其它两种算法都会存在选择到无效位置的情况，无法保证 k 匿名. 所以，从实验结果可以看出，HCLS 可以提供更好的隐私度.

(2) *PD* 的比较

PD 的定义如式(17)所示，分散度的对比结果见图 8. 由图 8 可知，随着 k 值的增加，三种算法的 *PD* 总体呈上升趋势. 相对于 CirDummy 和 GridDummy 算法，HCLS 算法增加更多，且随着 k 值的增大，差距变得更为明显. 通过本实验可以看出，本文 HCLS 可以选择出更为分散的假位置对用户的真实位置进行隐匿，提供更好的混淆效果.

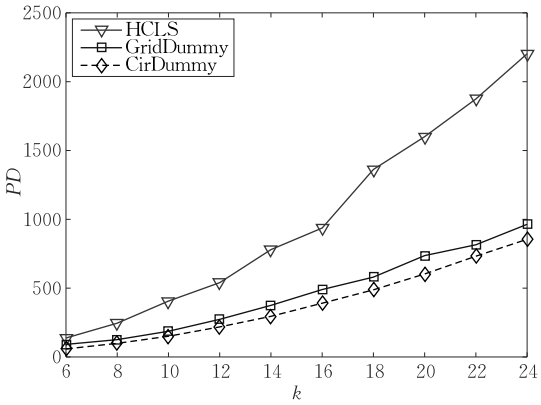


图 8　*PD* 对比

(3) 匿名域的比较

匿名域（Anonymity Area）面积是衡量隐私保护算法性能的重要指标，匿名域面积越大，攻击者定位用户越困难，因此隐私保护效果越好. 由于 CirDummy 和 GridDummy 算法需要提前指定请求区域限制，为了进行更为全面的比较，本实验在限制请求区域的情况下对三种算法的匿名域进行了对比. 这里，实验要求算法在限制的请求区域内进行假位置的选择，并将所生成的假位置的凸包作为最终的匿名域.

图 9 为不同请求区域限制下的匿名域面积对比，其中匿名度 k 设置为 12. 从图中可以看出，随着请求区域限制面积的增加，匿名域面积变化曲线呈上升趋势.

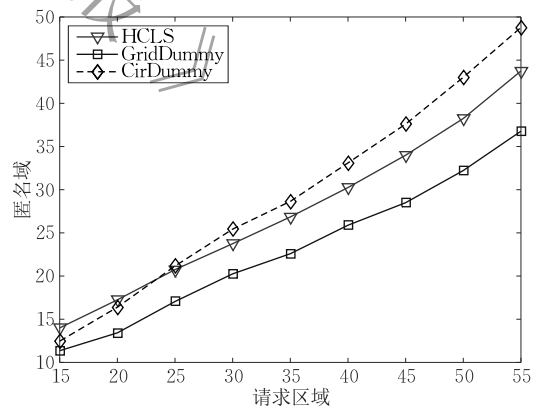


图 9　不同请求区域限制下的匿名域面积对比

由图 9 中可以看出，相较于 GridDummy 算法，HCLS 有明显的优势. 与 CirDummy 相比，初始时 HCLS 的匿名域略大于 CirDummy 算法，但随着请求限制面积的增加 HCLS 增长变得缓慢. 这是由于 HCLS 选择假位置的过程有一定的随机性，当限制

面积较小时选择的假位置更为容易覆盖更大范围；而随着限制面积变大，发送过请求的网格数量随之增多，HCLS 由于随机性的原因所覆盖的匿名域不稳定，多次实验取平均值得到的结果与 CirDummy 算法相比效果较差. 此处需要注意的是，CirDummy 无法抵御边信息攻击，同时 HCLS 的算法设计初衷是没有请求域限制要求的.

(4) 服务质量损失的比较

服务质量是衡量位置隐私保护算法的一个重要标准，服务质量即为用户发起位置服务请求后，得到的请求结果的准确程度，用户提交的位置越精确服务质量就越好. 由于三种算法都不包含真实位置，本文通过返回服务的假位置 l_{dummy} 与用户真实位置 l 的距离来评价服务质量的好坏，即

$$Qos_{\text{loss}}=Dis(l,l_{\text{dummy}}) \tag{21}$$

其中，Dis 指两个位置之间的距离，Qos_{loss} 越小，用户位置服务质量越好. 在 HCLS 中，l_{dummy} 即为偏移位置 l_d；在 CirDummy 和 GridDummy 中，l_{dummy} 为匿名集中与 l 距离最近的假位置.

服务质量损失的对比实验结果如图 10 所示，其中横坐标为指定的请求区域面积限制，纵坐标为服务质量损失，从图中可以看出 CirDummy 和 GridDummy 的服务质量损失随面积限制的增大总体呈上升趋势.

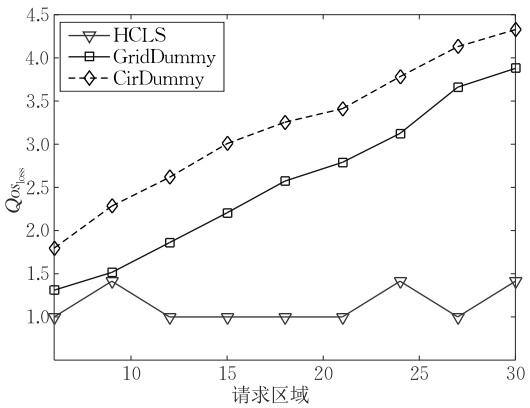


图 10 服务质量损失对比

由图 10 可以看出，HCLS 的 Qos_{loss} 在一定的范围内上下波动，但总体不受指定匿名域面积限制的影响. 这是由于 HCLS 中偏移位置 l_d 是提前选择出来的，无论匿名域面积限制增加还是降低，偏移位置 l_d 是不变的，因此 Qos_{loss} 就不会有明显的改变. GridDummy 和 CirDummy 两种算法随着指定请求区域面积的增加服务质量损失不断增大. 这是由于 CirDummy 与 GridDummy 选择的假位置会随着匿名域面积限制的增加而变得更为分散，因此“最近假位置”与真实位置的距离也会随之增大. 所以，HCLS 算法在服务质量损失上表现更优，能够在保护用户位置隐私的情况下为用户提供更好的服务.

上述实验从多个角度对算法性能进行了验证，实验结果表明 HCLS 在位置暴露率、假位置分散度和服务质量等方面有着很好的表现，产生的匿名集合可以对用户的真实位置进行隐匿，增加了攻击者推测用户真实位置的难度，同时也能将服务质量损失限制在可控范围内.

7.3 HCLS_SG 算法性能评价

本实验首先对攻击者的攻击模型进行定义：

攻击者所拥有的背景知识包括 HCLS 的匿名机制、用户的历史位置分布 $\varphi(l)$ 以及获取到的匿名集合 D_{set}. 通过上述知识，攻击者可以实施推测攻击，式(22)定义了这一攻击策略，其目的是根据已有知识选择位置 l' 去最小化位置隐私.

$$l'=\arg\min_{l'} PL \tag{22}$$

结合式(18)、式(22)可以用线性规划对其进行求解，求解过程如式(23)所示：

$$\begin{aligned}&\text{Min}\sum_{l,D_{\text{set}},l'}\varphi(l)f(D_{\text{set}}|l)h(l'|D_{\text{set}})Dis(l',l)\\&\text{Subject to}\\&\sum_{l'}h(l'|D_{\text{set}})=1,\forall D_{\text{set}}\\&h(l'|D_{\text{set}})\geqslant 0,\forall l',D_{\text{set}}\end{aligned} \tag{23}$$

实验利用式(23)定义的模型进行推测攻击，从隐私度 PL 和服务质量损失 Qos_{loss} 等方面验证 HCLS_SG 的有效性.

(1) PL 的对比

此部分，隐私度 PL 的定义如式(18)所示，PL 的值越大表示位置隐私保护机制的隐私效果越好.

如式(15)所示，预先设定的服务质量损失限制 $Q_{\text{loss}}^{\max}$ 和匿名度 k 对算法的隐私度有较大的影响. 所以，实验利用这两个因素将 HCLS_SG、HCLS 以及标准的 Stackelberg 博弈隐私保护算法 SG[30] 进行对比，验证算法的有效性. 因为算法在假位置选择过程中存在随机性，本文采用多次实验取平均值的方式进行验证，保证效果的一致性.

在 $Q_{\text{loss}}^{\max}$ 不同的情况下，HCLS 算法利用博弈模型优化的前后对比如图 11 所示. 在图 11 中，横坐标表示限定的最大服务质量损失 $Q_{\text{loss}}^{\max}$，纵坐标为隐私度 PL. 从图中可以看出，HCLS_SG 和 SG 明显优于 HCLS 算法，这是由于 HCLS 没有考虑攻击者的

推测攻击策略，而 HCLS_SG 优于 SG 的原因是 HCLS_SG 使用了假位置匿名，增加了攻击者的推测难度. 随着 $Q_{\text{loss}}^{\max}$ 的增加，三个算法的 PL 都有所增长，在 $Q_{\text{loss}}^{\max}$ 达到一定值后，算法的隐私度增长趋势变缓，这与用户位置分布 $\varphi(l)$ 有关，说明 $Q_{\text{loss}}^{\max}$ 对 PL 的影响存在一定的极限.

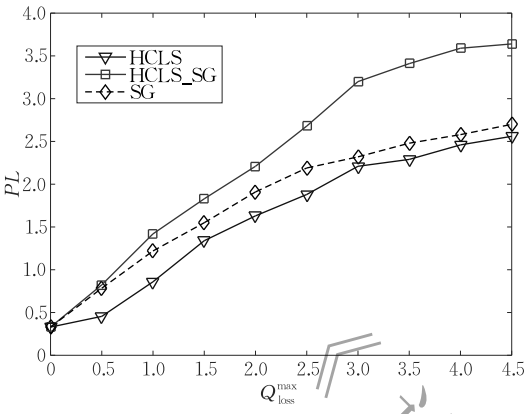


图 11　在不同服务质量损失限制下隐私度对比

在匿名度 k 不同的情况下，HCLS 算法利用博弈模型优化的前后对比如图 12 所示.

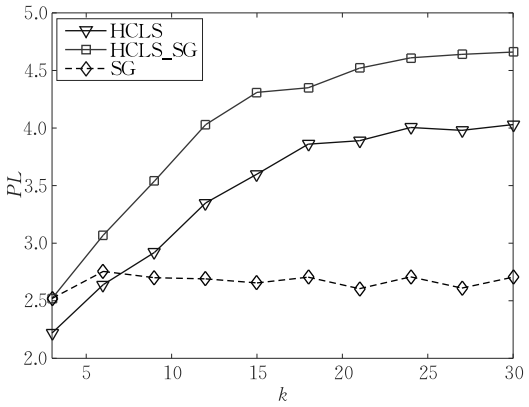


图 12　不同匿名度 k 下隐私度对比

在图 12 中，横坐标为匿名度 k，纵坐标为隐私度 PL. 随着 k 值的变大，HCLS_SG 和 HCLS 的隐私度有明显的提升，HCLS_SG 的表现要优于 HCLS；SG 的隐私度不会随匿名度 k 的增大而有明显改变，原因是 SG 本身只提供偏移位置，未考虑假位置匿名. 在分析 HCLS_SG 的偏移位置 l_d 时发现，l_d 的选择会随着 k 的变大慢慢接近服务质量损失限定的最大范围，也体现了博弈优化的可行性.

(2) $PERL$ 的比较

在本实验中，我们比较 HCLS 算法及 HCLS_SG 在攻击者推测攻击下的位置暴露率，实验结果如图 13 所示.

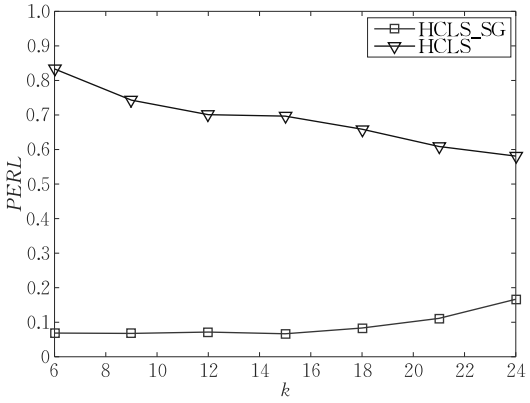


图 13　在不同匿名度 k 下的 $PERL$ 对比

在图 13 中，横坐标为匿名度 k，纵坐标为位置暴露率. 由图中可以看出，HCLS 在推测攻击下大部分假位置被过滤掉，随着匿名度 k 的增大，$PERL$ 有所减小；而 HCLS_SG 能够保证绝大部分假位置不被过滤，随着匿名度 k 的增大，HCLS_SG 的 $PERL$ 有所增长，这与用户位置分布 $\varphi(l)$ 的位置个数有关(HCLS_SG 有较高概率选取 $\varphi(l)$ 中的位置作为假位置，当 k 超过 $\varphi(l)$ 中的位置个数时，其他的假位置容易被攻击者所过滤).

(3) Qos_{loss} 的比较

服务质量损失 Qos_{loss} 与位置隐私度 PL 是存在密切关系的，在某些情况下，用户允许损失一定的服务质量来换取更高的隐私保护度，Qos_{loss} 的定义如式(19)所示. 本实验中，将用户允许的最大服务质量损失分别设定为 $Q_{\text{loss}}^{\max}=\{0,0.5,1.0,1.5,2.0,2.5,3.0,3.5,4.0,4.5\}$，以此为条件，分析服务质量损失与隐私度之间的关系，实验结果如图 14 所示. 图中横坐标为最大限定的服务质量损失 $Q_{\text{loss}}^{\max}$，左纵坐标为限定条件下的具体服务质量损失 Q_{loss}，右纵坐标为用户位置隐私度 PL.

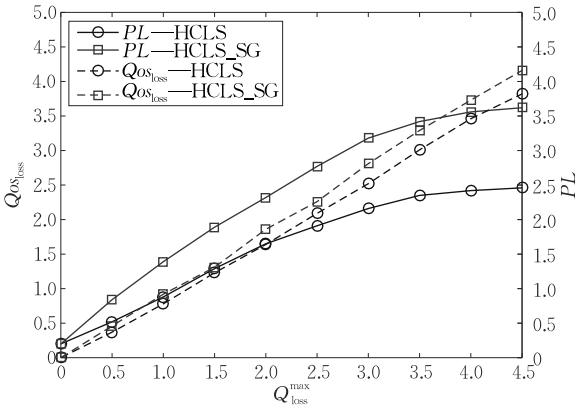


图 14　在不同约束下隐私度和服务质量损失的对比

计算机学报

由图 14 可以看出，随着限定服务质量损失的增加，具体服务质量损失和隐私度都随之增加. 另外，同等的 $Q_{\text{loss}}^{\max}$ 限制下，HCLS_SG 的位置隐私度要明显好于 HCLS，但服务质量损失在一定程度上也高于 HCLS，主要原因在于 HCLS_SG 会充分利用限定的最大服务质量损失去优化假位置选择结果，在确保不超过服务质量约束的条件下尽可能多地提高用户的位置隐私度.

上述对比实验说明，通过 Stackelberg 博弈模型优化后的 HCLS_SG 能够在一定程度上抵抗攻击者的推测攻击，使攻击者无法对用户真实位置进行更准确地推测.

8 结 论

为了解决位置隐私保护问题，本文对现有的位置隐私保护系统结构和各种位置隐私保护技术进行了研究，分析了它们的优势和缺陷，提出了基于半可信第三方服务的位置隐私保护系统结构. 基于这一系统结构，本文设计了一种基于假位置和 Stackelberg 博弈的位置匿名算法. 该算法利用用户历史请求位置分布选择假位置，提高用户位置隐私保护水平，并基于 Stackelberg 博弈对匿名结果进行优化，以此对抗攻击者的推测攻击，尽可能地保护用户的位置隐私. 本文利用真实数据集进行了实验，验证了算法的有效性.

参 考 文 献

[1] Wang S, Hu Q, Sun Y, et al. Privacy preservation in location-based services. IEEE Communications Magazine, 2018, 56(3): 134-140

[2] Zhou J, Cao Z, Dong X, et al. Security and privacy for cloud-based IoT: Challenges. IEEE Communications Magazine, 2017, 55(1): 26-33

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

[4] Pan X, Hao X, Meng X. Privacy preserving towards continuous query in location-based services. Journal of Computer Research & Development, 2010, 47(1): 121-129

[5] Zheng J, Tan X, Zou C, et al. A cloaking-based approach to protect location privacy in location-based services//Proceedings of the 33rd Chinese Control Conference(CCC 2014). Nanjing, China, 2014: 5459-5464

[6] Gambs S, Killijian M O, del Prado Cortez M N. Show me how you move and I will tell you who you are//Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Security and Privacy in GIS and LBS. San Jose, USA, 2010: 34-41

[7] Memon I, Ali Q, Zubedi A, et al. DPMM: Dynamic pseudonym-based multiple mix-zones generation for mobile traveler. Multimedia Tools and Applications, 2017, 76(22): 24359-24388

[8] Lu R, Lin X, Luan T H, et al. Pseudonym changing at social spots: An effective strategy for location privacy in VANETs. IEEE Transactions on Vehicular Technology, 2012, 61(1): 86-96

[9] Wiedersheim B, Ma Z, Kargl F, et al. Privacy in inter-vehicular networks: Why simple pseudonym change is not enough//Proceedings of the 7th International Conference on Wireless On-Demand Network Systems and Services (WONS 2010). Kranjska Gora, Slovenia, 2010: 176-183

[10] Ullah I, Wahid A, Shah M A, et al. VBPC: Velocity based pseudonym changing strategy to protect location privacy of vehicles in VANET//Proceedings of the 2017 International Conference on Communication Technologies (ComTech 2017). Rawalpindi, Pakistan, 2017: 132-137

[11] Mouratidis K, Yiu M L. Shortest path computation with no information leakage. Proceedings of the VLDB Endowment, 2012, 5(8): 692-703

[12] Chen X, Pang J. Measuring query privacy in location-based services//Proceedings of the 2nd ACM Conference on Data and Application Security and Privacy. San Antonio, USA, 2012: 49-60

[13] Etemad M, Küpçü A, Papamanthou C, et al. Efficient dynamic searchable encryption with forward privacy. Proceedings on Privacy Enhancing Technologies, 2018, 2018(1): 5-20

[14] Li J, Yan H, Liu Z, et al. Location-sharing systems with enhanced privacy in mobile online social networks. IEEE Systems Journal, 2017, 11(2): 439-448

[15] Ardagna C A, Cremonini M, di Vimercati S D C, et al. An obfuscation-based approach for protecting location privacy. IEEE Transactions on Dependable and Secure Computing, 2011, 8(1): 13-27

[16] Enhanced location privacy preserving scheme in location-based services. IEEE Systems Journal, 2017, 11(1): 219-230

[17] Andrés M E, Bordenabe N E, Chatzikokolakis K, et al. Geo-indistinguishability: Differential privacy for location-based systems//Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security. Berlin, Germany, 2013: 901-914

《计算机学报》

[18] Gedik B, Liu L. Protecting location privacy with personalized k-anonymity: Architecture and algorithms. IEEE Transactions on Mobile Computing, 2008, 7(1): 1-18

[19] Yigitoglu E, Damiani M L, Abul O, et al. Privacy-preserving sharing of sensitive semantic locations under road-network constraints//Proceedings of the 13th International Conference on Mobile Data Management. Bengaluru, India, 2012: 186-195

[20] Ye A, Li Y, Xu L. A novel location privacy-preserving scheme based on l-queries for continuous LBS. Computer Communications, 2017, 98: 1-10

[21] Kido H, Yanagisawa Y, Satoh T. An anonymous communication technique using dummies for location-based services//Proceedings of the 2005 International Conference on Pervasive Services (ICPS'05). Santorini, Greece, 2005: 88-97

[22] Lu H, Jensen C S, Man L Y. PAD: Privacy-area aware, dummy-based location privacy in mobile services//Proceedings of the 2008 ACM International Workshop on Data Engineering for Wireless and Mobile Access. Vancouver, Canada, 2008: 16-23

[23] Zha H, Xiaoling Y I, Wan J L. Privacy-area aware all-dummy-based location privacy algorithms for location-based services//Proceedings of the 2016 International Conference on Network and Communication Security. Beijing, China, 2016: 1-10

[24] Niu B, Li Q, Zhu X, et al. Achieving k-anonymity in privacy-aware location-based services//Proceedings of the 2014 IEEE Conference on Computer Communications (IEEE INFOCOM 2014). Toronto, Canada, 2014: 754-762

[25] Niu B, Li Q, Zhu X, et al. Enhancing privacy through caching in location-based services//Proceedings of the 2015 IEEE Conference on Computer Communications (IEEE INFOCOM 2015). Hong Kong, China, 2015: 1017-1025

[26] Pingley A, Zhang N, Fu X, et al. Protection of query privacy for continuous location based services//Proceedings of the 2011 IEEE Conference on Computer Communications (IEEE INFOCOM 2011). Shanghai, China, 2011: 1710-1718

[27] Wang Lu, Wang Xiaofeng. Location privacy preservation in big data era: A survey. Journal of Software, 2014, 25(4): 693-712

[28] Arain Q A, Memon I, Deng Z, et al. Location monitoring approach: Multiple mix-zones with location privacy protection based on traffic flow over road networks. Multimedia Tools and Applications, 2018, 77(5): 5563-5607

[29] Zheng X, Cai Z, Li J, et al. Location-privacy-aware review publication mechanism for local business service systems//Proceedings of the 2017 IEEE Conference on Computer Communications (IEEE INFOCOM 2017). Atlanta, USA, 2017: 1-9

[30] Shokri R, Theodorakopoulos G, Troncoso C, et al. Protecting location privacy: Optimal strategy against localization attacks//Proceedings of the 2012 ACM Conference on Computer and Communications Security. Raleigh, USA, 2012: 617-627

[31] Bordenabe N E, Chatzikokolakis K, Palamidessi C. Optimal geo-indistinguishable mechanisms for location privacy//Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security. Scottsdale, USA, 2014: 251-262

[32] Shokri R. Privacy games: Optimal user-centric data obfuscation. Proceedings on Privacy Enhancing Technologies, 2015, 2015(2): 299-315

[33] Chen J, He K, Yuan Q, et al. Blind filtering at third parties: An efficient privacy-preserving framework for location-based services. IEEE Transactions on Mobile Computing, 2018, 17(11): 2524-2535

[34] Shokri R, Theodorakopoulos G, Papadimitratos P, et al. Hiding in the mobile crowd: Location privacy through collaboration. IEEE Transactions on Dependable & Secure Computing, 2014, 11(3): 266-279

[35] Wan Sheng, Li Feng-Hua, Niu Ben, et al. Research progress on location privacy-preserving techniques. Journal on Communications, 2016, 37(12): 124-141(in Chinese)
(万盛，李凤华，牛犇等. 位置隐私保护技术研究进展. 通信学报，2016，37(12)：124-141)

[36] Leitmann G. On generalized Stackelberg strategies. Journal of Optimization Theory and Applications, 1978, 26(4): 637-643

[37] Breton, M, Alj A. Sequential Stackelberg equilibria in two-person games. Journal of Optimization Theory and Applications, 1988, 59(1): 71-97

[38] Von S, B, Zamir S. Leadership games with convex strategy sets. Games and Economic Behavior. 2010, 69(2):446-457

XIA Xing-You, Ph. D. candidate. His research interests include crowdsensing systems, location privacy protection methods, mobile big data analysis, etc.

BAI Zhi-Hong, M. S. His research interests include crowd sensing systems, location privacy protection methods, etc.

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

YU Rui-Yun, Ph. D., professor, Ph. D. supervisor. His research interests include crowdsensing systems, mobile big data analysis, virtual reality technologies, etc.

计算机学报

Background

Location privacy has been a hot topic in social media for the past decade. Users are increasingly aware of location privacy threats, and these threats range from an adversary's ability to localize an individual to profiling and identifying him based on the places he visits. A large body of researchs have focused on developing location privacy protection mechanisms (LPPMs), which can limit the amount of disclosed sensitive information when users use location based services (LBSs). These protection mechanisms often offset the real locations of a user, or hide the real location in some fake locations, in order to increase the difficulty of the attacker guessing the user's real location.

In this paper, a semi-trusted location privacy protection architecture is proposed. The architecture keeps the user's privacy information (the identifier, the location and the query) in the pseudonym server, the location anonymous server, and the LBS provider respectively. Therefore, even though the attackers get some data from a part of servers, they are unable to confirm the identity of the user or the related data. What's more, a location cloaking algorithm based on the dummy and the Stackelberg game is designed to improve the level of location privacy protection. The algorithm adopts the users' historical location requests to generate an anonymity set with k dummies, which will be sent to the location based services provider instead of the user's real location. Furthermore, the adversary might have some knowledge about the proposed location privacy protection algorithms as well as all the history data of a particular user. A Stackelberg game model is used to optimize the obfuscation location selection in the anonymity set when considering the above situation. Lastly, the real dataset is used to evaluate the proposed algorithm, and extensive results are provided to demonstrate the efficiency of the algorithm.

This work is supported by the National Natural Science Foundation of China (61672148, 61502092), the Fundamental Research Funds for the Central Universities (N171702001), the Program for Liaoning Innovative Research Term in University (LT2016007), the Liaoning BaiQianWan Talents Program (201514), and the Ministry of Education—China Mobile Research Fund (MCM20160201). An important task of these projects is to protect mobile users' location privacy under the environment of big data. This paper addresses the problem of location privacy protection, and proposes the location cloaking algorithm to limit adversaries speculating the users' sensitive locations through users' historical location requests. The algorithm can be integrated in many kinds of applications and systems easily.

《计算机学报》