软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2014,25(4):693−712 [doi: 10.13328/j.cnki.jos.004551]

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

位置大数据隐私保护研究综述*

王 璐, 孟小峰

(中国人民大学 信息学院,北京 100872)

通讯作者: 王璐, E-mail: luwang@ruc.edu.cn

摘 要: 大数据时代移动通信和传感设备等位置感知技术的发展形成了位置大数据,为人们的生活、商业运作方法以及科学研究带来了巨大收益.由于位置大数据用途多样,内容交叉冗余,经典的基于“知情与同意”以及匿名的隐私保护方法不能全面地保护用户隐私.位置大数据的隐私保护技术度量用户的位置隐私,在信息论意义上保护用户的敏感信息.介绍了位置大数据的概念以及位置大数据的隐私威胁,总结了针对位置大数据隐私的统一的基于度量的攻击模型,对目前位置大数据隐私保护领域已有的研究成果进行了归纳.根据位置隐私的保护程度,可以把现有方法总结为基于启发式隐私度量、概率推测和隐私信息检索的位置大数据隐私保护技术.对各类位置隐私保护技术的基本原理、特点进行了阐述,并重点介绍了当前该领域的前沿问题:基于隐私信息检索的位置隐私保护技术.在对已有技术深入分析对比的基础上,指出了未来在位置大数据与非位置大数据相结合、用户背景知识不确定等情况下保护用户位置隐私的发展方向.

关键词: 大数据;位置大数据;位置隐私保护技术

中图法分类号: TP311 **文献标识码**: A

中文引用格式: 王璐,孟小峰.位置大数据隐私保护研究综述.软件学报,2014,25(4):693−712. http://www.jos.org.cn/1000-9825/4551.htm

英文引用格式: Wang L, Meng XF. Location privacy preservation in big data era: A survey. Ruan Jian Xue Bao/Journal of Software, 2014,25(4):693−712 (in Chinese). http://www.jos.org.cn/1000-9825/4551.htm

Location Privacy Preservation in Big Data Era: A Survey

WANG Lu, MENG Xiao-Feng

(School of Information, Renmin University of China, Beijing 100872, China)

Corresponding author: WANG Lu, E-mail: luwang@ruc.edu.cn

Abstract: Development of mobile communication and sensing technologies forms location based big data, bringing revolution to human’s living, business pattern and scientific research. Diversity of usage patterns and redundancy among various sources of location based big data make it impossible for classical location preservation methods to protect privacy systemically. Privacy preservation for location based big data measures user’s location privacy in all possible aspects and therefore protects user’s privacy in information theory semantic. Starting with an introduction to the concept of location based big data, its associated privacy threats and a universal measurement-based attack model, this paper surveys the state of the art of privacy preservation techniques for location based big data. Based on different privacy protecting strength, various big privacy preservation techniques can be categorized into heuristic privacy measurement, probability deduction and private information retrieval based technologies. The principles, mechanisms and characteristics of various techniques are described in detail, with special emphasis on a proceeding research topic: Private information retrieval based technology. Following a comprehensive analysis and comparison of existing techniques, privacy protecting for location based big data

* 基金项目: 国家自然科学基金(61379050, 91224008); 国家高技术研究发展计划(863)(2013AA013204); 高等学校博士学科点专项科研基金(20130004130001)

收稿时间: 2013-08-13; 定稿时间: 2013-12-05; jos 在线出版时间: 2014-01-13

CNKI 网络优先出版: 2014-01-13 14:11, http://www.cnki.net/kcms/doi/10.13328/j.cnki.jos.000012.html

under situations like combination of location information and non location information and attacker's arbitrary background knowledge is highlighted as future research directions.

Key words: big data; location based big data; location privacy-preserving

大数据时代,移动通信和传感设备等位置感知技术的发展将人和事物的地理位置数据化.移动对象中的传感芯片以直接或间接的方式收集移动对象的位置数据:一方面,内置在手机、车载导航等移动设备中的 GPS, WiFi 等定位设备可以直接获得移动对象任意时刻准确的位置信息,并经过各种途径发布这些采集的位置信息,比如,移动社交网络的一些新型应用可以发布任意时刻用户所处的位置信息[1];另一方面,近期得到广泛应用的可穿戴设备等传感设备采集到的加速度、光学影像等数据经过处理后也可以准确地确定使用者的位置信息[2−4].

传感器自动采集位置信息的速度和规模远远超过现有系统的处理能力.根据统计,每个移动物体平均 15s 提交一次当前位置,这样,全球上亿手机、车载导航设备等移动对象每秒钟提交的位置信息超过 1 亿条[5].未来,移动传感设备的进步和通信技术的提升会更频繁地产生位置信息.大数据时代,这样的产生速度和数据规模为人们的生活、企业的运作以及科学研究带来巨大的变革[6].我们称这类由于包含位置信息且具有规模大、产生速度快、蕴含价值高等满足被普遍认可的大数据的特点[7]的数据为位置大数据.

位置大数据在人们的生产与生活中有诸多运用:

- 一方面,从个人生活层面上说,通过推测一个人居住的地点和每天常去的地方,可以为他提供便捷的服务.例如:文献[8,9]利用人们大量的历史活动轨迹数据,为每个人的出行和旅游给出路线推荐;文献[10]根据当前的交通流量情况,为用户推荐可以乘坐的公共交通;总部位于亚特兰大的 AirSage 公司每天通过处理来自上百万手机用户的 150 亿条位置信息,为超过 100 个美国的城市提供实时交通信息[11].当前,这些基于位置大数据的新型服务逐渐形成了一个正在迅速增长的市场.一份来自 Pyramid Research 的调查报告显示,2010 年,诸如导航或移动社交网络等基于位置的服务已具有 28 亿美元的市场.到 2015 年,这一数字将达到 103 亿美元[12];
- 更重要的是,位置大数据改变了商业运作方式并为科学研究提供了新的方法.例如,传统的车险业通过考虑一个群体的平均风险确定车险定价,当保险公司通过获得的车辆出行时间、常见行驶地点和实际行驶过程等位置大数据后,转变为对每个用户个性化的分析定价,改变了车险业的运作方式[13].与此同时,联合包裹运输公司(united parcel service incorporation)收集自己旗下运输车辆的行驶信息,为它们提供最佳行车路线以减少燃油、故障成本,在商业模式上取得了巨大成功.仅 2011 年,UPS 公司旗下的车辆就节省了 4 828 万公里的路程、1 136 万升的燃料和 3 万吨二氧化碳的排放,同时减少了容易出事故的路线[14].此外,无线数据科技公司(Jana)使用来自 100 多个国家的、超过 200 个无线运营商提供的、覆盖了拉丁美洲、非洲、欧洲的大约 35 亿人口的手机数据,试图回答疾病如何传播以及城市如何繁荣这样重大的科学问题[15].

位置大数据在带给人们巨大收益的同时,也带来了泄露个人信息的危害.这是因为位置大数据既直接包含用户的隐私信息,又隐含了用户的个性习惯、健康状况、社会地位等其他敏感信息.位置大数据的不当使用,会给用户各方面的隐私带来严重威胁.已有的一些案例说明了隐私泄露的危害,例如:某知名移动应用由于不注意保护位置大数据,导致根据三角测量方法可以推断出用户的家庭住址等敏感位置,已引发多起犯罪案件[16].同时,某著名移动设备厂商在未获得用户允许的情况下大量收集用户的位置数据[17,18],攻击者可以通过这些位置数据推测用户的身体状况等个人敏感信息[19−21].而在为用户提供了合适的位置隐私保护后,更多的人愿意将自己的移动数据提交给智能交通、智能城市等分析系统,进而为人们的日常生活提供更多的便利.

经典的位置隐私保护技术经过较长时间的发展,从最早将位置数据视为一般数据使用"知情与同意"[22]等访问控制方法发展到针对单个位置数据的匿名化隐私保护方法,再进一步完善到对轨迹数据的匿名化隐私保护方法.但是,"知情与同意"以及匿名化等经典的位置隐私保护方法在大数据时代不能有效地保护用户隐私:

(1) 大数据尚未想到的用途无法提前告诉用户,企业也无法承担发现位置大数据的创新性用途后通知每

个用户并请求用户同意再进行使用的成本.因此,“知情与同意”等保护方法要么限制了对位置大数据价值的挖掘,要么无法保护个人隐私;

(2) 由于位置大数据来源众多,这些数据之间可以相互补充,最近的研究对精心匿名的位置数据进行了成功的反匿名化[23].

大数据时代,经典位置隐私保护方法不能解决的主要问题是:攻击者可以从多种途径获得各个角度关于用户的位置数据或非位置数据,这些数据可以直接或者间接地重构出用户希望保护的位置隐私.比如:

(1) 单纯针对位置数据.用户在服务 A 中保护起来的数据可能在服务 B 中被泄露,如果攻击者同时获得服务 A 和服务 B 中的数据,就可以重构出用户的准确数据;

(2) 考虑位置与非位置数据相结合的情况.位置数据与非位置数据由于是同一用户产生,因此,用户的某些个性就成为了位置数据与非位置数据之间的联系.攻击者根据这些个性可以区分不同用户的位置数据,进而对用户的身份等敏感信息进行推测.

位置大数据隐私保护技术面对这两种威胁全面地控制用户位置信息的泄露.位置大数据隐私保护技术可以保证针对用户不同的隐私需求进行信息论意义上的全面保护.因此,位置大数据隐私保护技术需要考虑以下 3 个具有挑战性的问题:

(1) 如何度量用户的敏感信息的泄露程度;

(2) 如何实现对位置大数据隐私全面的保护;

(3) 如何兼顾隐私保护的程度和基于位置服务的可用性.

传统的位置隐私保护方法(如基于加密的方法等)没有考虑对用户敏感信息泄露的度量问题,也不能实现对位置隐私的全面保护,因此无法在隐私的保护程度和可用性之间做出权衡.当前,位置大数据隐私保护技术主要针对不同程度的隐私需求,权衡隐私保护效果和服务可用性.

本文综述位置大数据隐私保护技术的最新进展:一方面,介绍位置大数据的基本概念以及总结出针对位置大数据隐私的统一的基于度量的攻击模型等研究背景;另一方面,以统一的攻击模型为依据,根据不同隐私保护技术在隐私保护程度和服务可用性之间的权衡情况,分类阐述位置大数据的隐私保护技术,分析不同技术的优缺点、适用场景等.其中,重点介绍当前该领域的前沿问题、基于隐私信息检索的隐私保护技术.本文在对位置大数据的隐私保护技术进行综合对比和分析后,探讨了位置大数据未来的研究方向.

考虑到大数据时代的攻击者可以获得和位置数据相关的非位置数据[24],可以从其他角度获得或者分析用户的历史位置数据得到有关用户的背景知识,本文认为,位置大数据与非位置大数据结合产生的隐私问题将是未来的研究热点之一.其中,作为位置数据与非位置数据结合的特例,移动社交网络由于天然地将位置数据与非位置数据结合起来,将成为未来的研究热点之一.本文在未来工作中介绍了一种通用的位置数据与非位置数据结合后的位置隐私保护方法,同时介绍了移动社交网络中的隐私保护方法.除此以外,由于位置大数据隐私的统一的基于度量的攻击模型与攻击者具有的背景知识有关,而差分隐私由于其对背景知识不敏感,在量化隐私时具有重要意义,本文认为,将差分隐私应用到位置大数据隐私保护中,也是未来的研究热点.

本文第 1 节介绍位置大数据的隐私保护的概念和统一的基于度量的攻击模型等背景知识.第 2 节对位置大数据的隐私保护技术的分类和评估方法进行介绍.第 3 节~第 5 节分别对 3 大类位置大数据的隐私保护技术:基于启发式隐私度量、基于概率推测和基于隐私信息检索的位置大数据隐私保护技术进行阐述.第 6 节~第 7 节对各类技术进行对比分析,并指出未来的研究方向.第 8 节为全文结语.

1 位置大数据隐私与隐私度量

1.1 位置的表示与定位技术

位置是指移动对象在某一时刻的经纬度,通常由三元组〈经度 x,纬度 y,时刻 t〉表示.文献[25]总结了目前对移动对象进行定位的 5 种常用方法以及攻击者获得移动对象位置信息常用的 3 种方法.其中,移动对象获得自己的位置常用以下 5 种方法:(1) 全球定位系统部署的卫星与移动设备经过通信,根据多个卫星与同一移动设备之

间通信时在时间上的延迟,使用三角测量方法得到目前最为精准的移动物体的经纬度[26],目前常见的 GPS 设备可以实现 5m 以下的精度;(2) 因为 WiFi 访问点与它们的准确位置之间的对应关系,可以从类似文献[27]的数据库中查找到,当移动物体连接到某个 WiFi 访问点时,用户的位置也可以较精确地对应到一个经纬度[28];(3) 当移动设备位于 3 个手机基站的信号范围内时,三角测量同样可以获得用户的经纬度[29],这种方法和方法(2)都避免了 GPS 系统无法在建筑物内进行定位的缺点;(4) 移动设备接入互联网时会被分配一个 IP 地址,IP 地址的分配是和地域有关的,利用已有的 IP 地址与地区之间的映射关系,可以将移动物体的位置定位到一个城市大小的地域[30];(5) 目前的很多研究显示,通过传感器捕获的加速度、光学影像等信息,可以用于识别用户的位置信息[2−4].

移动用户使用上述方法获得自己的位置信息以后,攻击者通常使用以下 3 种方法获得移动用户的位置信息:(1) 攻击者发布恶意的基于位置的应用,当移动用户使用这些应用请求基于位置的服务时,这些应用会向攻击者报告用户的当前位置,当前,手机应用中流行的移动社交网络的签到应用和导航应用等都有可能被攻击者利用;(2) 一些网站,例如文献[31],可以通过用户的 IP 地址获取用户的位置信息,当移动设备访问这些网络信息时,自己的位置就可能暴露给攻击者;(3) 用户在使用移动设备浏览网页时,会因为响应一些网页中的请求将当前的位置信息发送给网页指定的攻击者[32,33],这种方法与方法(2)类似,但攻击者不需要掌握从 IP 地址到位置的映射方法.

1.2 位置大数据的隐私威胁

类似一般的隐私定义[34],我们认为,位置大数据的隐私是移动对象对自己位置数据的控制.大数据时代,位置数据的来源极为广泛,位置大数据中包含的移动对象不同时刻的位置信息与背景知识结合,会泄露用户的健康状况、行为习惯、社会地位等敏感信息.比如:观察到用户出现在医院附近,可以推测出用户大致的健康状况;考虑用户轨迹开始和结束的地点,可以推测出用户的家庭住址等信息[20].此外,加速度传感器等收集到的只包含部分位置的信息,也可以让攻击者有效推测用户的行为模式[35].

攻击者利用类似上述各种数据推测用户某时刻的隐私,在传统的位置隐私保护工作中通常被称为观察攻击[36−38]或者关联攻击[39,40],但这些攻击模型不能概括大数据时代用户的位置隐私面对全方面推测的威胁.我们将使用各种类型的数据推测用户位置隐私的行为总结成位置大数据的隐私攻击模型:模型包括攻击者获得的数据 D 以及攻击者希望据此推测出的攻击对象 U 在 t'时刻处于位置 i 的概率 $P\{U_i^{t'} \mid D\}$.

由于前文提到“知情与同意”、匿名等经典的隐私保护策略在大数据时代均失效,如何防止攻击者利用收集到的各方面数据推测用户的隐私信息,成为大数据时代亟待解决的位置大数据的隐私保护问题.

位置大数据隐私保护技术研究的早期,并没有专门针对位置大数据的保护手段,研究者仅简单通过用户对数据进行分类,并提供访问控制列表或者数据使用列表等隐私控制策略,避免不可信对象对用户敏感位置数据的获得[41−44]以及数据的不正当应用[22].之后,针对位置大数据隐私保护的研究集中在如何避免向攻击者发布移动对象某一时刻的精确位置,同时获得基于位置大数据的服务,这类技术的典型方法包括位置 k 匿名等基于单点位置的启发式隐私度量的方法[45−49].随着位置大数据隐私保护技术的发展,人们开始注意到轨迹信息包含用户的移动在时间上的相关性[28,50−52],于是,保护用户的轨迹信息的方法受到重视.由于位置之间在时间上的相关性难以把握,一些基于轨迹的启发式的隐私度量方法(比如将位置数据随机化的方法、对空间数据的模糊化方法和对时间数据的模糊化方法)被提了出来.

但在大数据时代,提供可以量化的位置大数据的隐私保护效果是十分重要的[53],因此,基于概率推测的位置大数据隐私保护方法从信息论的角度给出位置隐私完整的度量方式,量化每个位置数据暴露的用户隐私.同时,基于隐私信息检索的位置大数据隐私保护技术提供了完美隐私(见下文定义 1 中相关内容).

1.3 位置大数据隐私的攻击模型

位置大数据隐私的攻击模型使用统一的度量方法描述对位置大数据隐私的攻击效果.不同攻击方法的效果由根据发布后的位置大数据能够提供给攻击者多少用户处于某敏感位置的信息增益来刻画.攻击者收集的用户的位置数据是包含了用户的时空数据的集合$\{r:r=\langle p,t\rangle\}$,其中,$r$ 代表攻击者收集到的一条位置数据,p 代表

用户的位置信息,t 代表收集到这条数据的时刻.那些可能暴露用户隐私的敏感位置也组成一个集合$\{s_1,\ldots,s_n\}$,其中,s_i代表将敏感位置编号后的第 i 个敏感位置.

根据收集到的位置数据,攻击者可以通过推测用户在 t 时刻处于某个敏感位置 s_i 的概率 P\{用户在时刻 t 的位置为 s_i|攻击者收集到的用户的历史位置序列\},从而推测用户的隐私信息.为了量化攻击效果,我们定义任意时刻用户发布的位置数据不泄露用户处于某一敏感位置的θ隐私.

定义 1(位置大数据的θ隐私). 在任意时刻 t',用 $P\{U_i^{t'}\}$ 表示用户在 t'时刻处于位置 s_i 的概率,用 L_t 表示攻击者收集到的用户在时刻 t 之前发布的位置数据,则,

$$P\{U_i^{t'} \mid L_t\} - P\{U_i^{t'}\} \leqslant \theta ,$$

其中,θ是用户给定的隐私需求,也是攻击者能够获得的最大攻击效果; $P\{U_i^{t'} \mid L_t\}$ 表示攻击者收集到用户 t 时刻之前的位置数据后推测用户在 t'时刻处于敏感位置 s_i 的后验概率;$P\{U_i\}$是攻击者推测用户处于敏感位置 s_i 的先验概率.定义 1 的含义是:攻击者收集到的用户的位置数据为攻击者推测用户的敏感位置带来的信息量不能超过θ,即,获得用户位置序列后推测用户某一时刻处于某个敏感位置的后验概率与先验概率之差小于θ.按照信息论的定义[42],在任何时刻 t',攻击者掌握的用户所处位置的先验信息量为 $-\sum_i P\{U_i^{t'}\}\log P\{U_i^{t'}\}$.用户的位置大数据暴露给攻击者用户所处位置的信息量可以如下计算:

$$\sum_i P\{U_i^{t'}\}\log P\{U_i^{t'}\} - \sum_i P\{U_i^{t'} \mid L_t\}\log P\{U_i^{t'} \mid L_t\}.$$

根据定义 1,满足θ隐私的保护方法可以保证对位置大数据隐私的攻击方法的效果 $n\theta$,其中,n 是敏感位置的个数.显然,如果用户需要保证发布位置数据不泄露用户处于任意敏感位置的θ隐私,只需在每个时刻都满足定义 1 中的θ/n 隐私.在给定时刻,具有强隐私需求的用户可以将θ设置为 0,这时称为完美隐私.

以上的位置大数据隐私攻击模型针对单一数据类型的位置数据,但在第 7 节对未来工作的展望中我们可以看到,面对位置大数据的隐私,未来将位置数据与非位置数据结合的研究方向,该攻击模型同样适用.

总的来说,位置大数据的攻击模型采用统一的基于度量的方法将位置数据对用户隐私的披露风险使用定义 1 进行量化.从定义 1 我们可以看出,用户在任意时刻 t'的位置隐私与攻击者收集的 t 时刻之前的位置数据集 L_t 有关,由于攻击者可以持续收集位置数据集 L_t 以至于 $t>>t'$,用户任意时刻的位置隐私实际上与 t'时刻前和 t'时刻后都有关.特别地,当定义 1 中的θ为 0 时,这样的隐私保护技术称为完美隐私技术.

2 位置大数据隐私保护技术分类及性能评估

2.1 位置大数据隐私保护技术的分类

不同的位置大数据隐私保护技术都以定义 1 为统一的攻击模型,但出于不同的隐私保护需求以及实现的原理不同,在实际应用中各有优缺点.本文将位置大数据隐私保护技术分为 3 类:

(1) 基于启发式隐私度量的位置大数据隐私保护技术.

在定义 1 中,任意时刻 t'的位置信息发布后,暴露的用户敏感信息与攻击者收集到的时刻 t'之前和之后的位置数据都有关,针对这些完整的数据攻击和保护用户的位置隐私代价很大.对于一些隐私保护需求不严格的用户,基于启发式隐私度量的位置大数据隐私保护技术假设用户在 t'时刻的位置信息只与当前时刻攻击者收集到的数据有关.相应的方法包括经典的基于单点或轨迹的位置隐私保护技术,如针对位置或轨迹的 k 匿名[54]、l 多样性[55]、t 紧密性[56]、p 敏感性[57]、m 不变性[58]或空间匿名框等方法.直接应用这些方法会遭受针对数据特征的攻击.比如:经过空间匿名框处理以后的数据,在考虑移动物体的移动速度以后,某时刻发布的匿名框的一部分由于移动物体从上一时刻的匿名框中无法到达从而导致匿名失败.为此,这类方法针对常见的攻击手段,如考虑匿名框的面积等技术,对发布的位置数据进行处理,以降低攻击者推测出用户敏感位置的可能性.

(2) 基于概率推测的位置大数据隐私保护技术.

这类方法严格按照定义 1 量化攻击模型的效果,并进而限制任意时刻 t'发布的位置数据包含的信息量.基于概率推测的隐私保护技术假设攻击者具有全部背景知识,并由此对每个发布的位置数据计算其披露风险,判断

发布当前的位置数据是否违反用户的隐私要求.因此,这种位置大数据的隐私保护技术可以在攻击者具有完全的背景知识的情况下,在统一的位置大数据攻击模型下,定量地保护用户的位置隐私.

(3) 基于隐私信息检索的位置大数据隐私保护技术.

当用户要求定义 1 中的完美隐私时,由于发布位置信息或多或少地会为攻击者带来一些信息,这时会导致没有数据可以发布,因而无法获得基于位置大数据的服务.基于隐私信息检索的位置大数据保护技术,可以在任何情况下完全地保护移动用户的隐私.基于隐私信息检索的方法与加密方法类似,都是为了完全保护用户的隐私.但在位置大数据上的应用服务中,由于用户查询本身包含位置信息,很长时间内都不存在可以在不解密用户查询的情况下回答复杂的基于位置的查询的加密算法.尽管最近的研究结果发现,基于同态映射的加密方法[59]可以在不暴露用户位置隐私的情况下返回正确的查询结果,但最新的结果显示,因为高效的数据访问方法暴露了数据之间的顺序,可以提供完美隐私的高效加密方法是不存在的[60].

这 3 类技术各有自己的优势和缺陷:基于启发式隐私度量的隐私保护技术效率通常比较高,但位置信息存在一定程度的不准确性,另外也易遭受研究者没有考虑到的针对数据特点的攻击;基于隐私信息检索的位置大数据隐私保护技术正相反,可以完全保证数据的准确性和安全性,但预计算开销和运行时计算开销比较大;而基于概率推测的位置大数据隐私保护技术则可以提供相对平衡的隐私保护程度和运行效率.

在第 3 节~第 5 节,本文将分别深入地阐述这 3 类位置大数据隐私保护技术.

2.2 位置大数据隐私保护技术的性能评估

位置大数据隐私保护技术需要在保护用户位置隐私的同时兼顾服务的可用性以及开销.本文从以下 3 个方面度量位置大数据隐私保护技术的性能:

(1) 隐私保护程度.这通常由隐私保护技术的披露风险来反映,定义 1 中的θ就是用户可以接受的最大披露风险.披露风险θ越小,隐私保护程度越高;

(2) 服务的可用性.这是指发布位置信息的准确度和及时性,它反映通过隐私保护技术处理后用户获得的基于位置数据的服务质量.通常,服务的可用性与隐私保护程度之间具有一个权衡,提高隐私保护程度有时会降低服务的可用性;

(3) 开销.位置大数据隐私保护技术的开销包括预计算和运行时发生的存储和计算代价.存储代价主要发生在预计算时.预计算的代价在现有技术中通常可以接受,并在选择隐私保护技术时被忽略.运行时的计算代价根据位置大数据隐私保护技术的特点一般利用 CPU 时间以及文件块访问次数的时间复杂度进行度量.比如,基于启发式隐私度量和基于概率推测的位置大数据隐私保护技术一般使用时间复杂度来度量开销,基于隐私信息检索的位置大数据隐私保护技术通常使用文件块的访问次数的时间复杂度来度量开销.

3 基于启发式隐私度量的位置大数据隐私保护技术

定义 1 中,t'时刻的位置数据发布后的后验概率与攻击者收集到的用户的全部历史数据有关,因此,计算后验概率是计算量很大的一项任务.文献[61]仅考虑连续提交的位置数据中包含的速度信息暴露给攻击者的信息量,但这样依然导致计算上过于复杂,最终将连续取值的位置数据离散化.即使不要求定义 1 中的完美隐私,一些为了避免攻击者考虑到用户移动速度造成的匿名失败的隐私保护方法也面临求解 NP 问题[62].因此,早期的位置大数据隐私保护技术通常保护启发式的隐私度量,假设用户在 t'时刻所处的敏感位置只与当前时刻发布的位置数据有关.这样,同时满足直觉上的隐私需求,提供较为高效的隐私保护算法和较快响应的基于位置的服务.直到人们开始深入地研究人的移动模式,高效地满足定义 1 中信息论意义上完整的隐私条件的隐私保护技术才相继出现(见第 5 节).

启发式的隐私度量主要包括让用户提交不真实的位置数据来避免攻击者获得用户的真实位置数据,采取的主要技术包括随机化、空间模糊化和时间模糊化技术.这些技术一般假设在移动用户和服务器之间存在一个可信任的第三方服务器来将用户的位置数据转换成不真实的位置数据,以及将对模糊数据的查询结果转化成

用户需要的结果.

3.1 随机化

随机化是在原始位置数据中加入随机噪声.可信任的第三方服务器在接收到用户的准确位置以后,将噪音和准确位置都发送给服务提供商,并根据用户的真实数据过滤服务提供商返回的查询结果,并将过滤后的结果返回给用户.文献[63]首先提出了随机化方法,在每一时刻,都根据上一时刻的位置按照随机的速度和随机的方向进行移动,并将获得的随机位置点加入到原始数据中进行发布.然而,这些位置点组成的历史数据中的移动特征与真实的移动对象的特征具有很大差别,甚至提交的位置可能是一些实际上不可达的位置,因此很容易被攻击者区分.为此,文献[64]在产生随机位置数据的时候加入了路网、移动速度等对移动的约束条件.文献[63,64]都假设物体在不停地移动;文献[65]考虑到移动对象不会不停地移动,根据移动对象的周围环境等因素让移动对象随机地产生停顿,以进一步防止攻击者区分这些噪声.

3.2 空间模糊化

空间模糊化在一定程度上通过降低发布的位置数据的精度以满足用户的隐私需求,同时,没有妨碍地获得服务.空间模糊化的隐私保护技术通常将用户提交的位置精度从一个点模糊到一个区域,直至任一用户提交的位置数据都包含其他若干用户使得攻击者无法获得某个用户清晰的位置.图 1(a)是某一时刻 A 到 E 这 6 个移动用户在空间中的位置.空间被 Quad-tree 等技术划分成若干区域,用户希望每次发布的位置数据不要准确到区域中只有自己一个用户.于是,用户 $A(B,C)$在发布自己的位置时,可以发布左下角阴影区域作为自己的位置,用户 $D(E,F)$可以发布右上角阴影区域作为自己的位置.每个用户的近邻查询会向服务提供商发送自己所在的阴影区域,服务提供商需要根据计算包含阴影区域中任何一点的最近邻的区域,返回其中包含的全部用户.对于包含右上角阴影区域的近邻查询,服务器只需要返回阴影区域和黑色区域包含的全部用户,发起查询的用户就可以自行计算出自己的近邻.

实际上,用户 $D(E,F)$发出的近邻查询不需要返回整个右上角 4 个区域中的全部用户,只需要返回与右上角的浅色区域相近的若干用户即可,这样就显著减少了需要传输的数据量,提升了服务的可用性.利用三角形两边之和大于第三边的性质,可以求出与右上角浅色区域相近用户距离浅色区域的边界的最大值.这样,传输更小的区域中包含的全部用户给发起查询的用户也是正确的[47].文献[66]考虑到隐私需求应当保证两个连续提交的区域间在速度上是可达的,并根据用户的实际速度修正前面方法提交的区域.这种隐私需求在文献[48,49]中得到扩展,允许每个用户具有不同的模糊要求.比如,A 可以要求提交的区域中至少具有两个不同对象,B 可以要求提交的区域中至少具有 3 个不同对象.这种方法用图 1(b)表示每一时刻各个用户之间的关系,其中,考虑到速度因素后依然可以保护隐私的用户间有边相连.为了应对个性化的隐私需求,每个用户都维护图中包含自己的最大团.如果团的大小大于区域中用户要求的最少用户个数,那么包含这个团的区域可以直接作为用户的提交区域.由于在图中寻找最大团是一个 NP-hard 问题,文献[49]增量地维护每个用户的最大团,缩短了用户请求的响应时间.

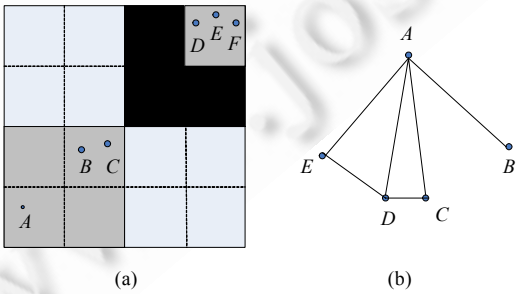
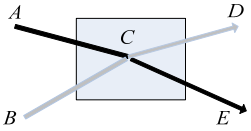


Fig.1 Illustration for space vagueness

图 1 空间模糊化示意图

3.3 时间模糊化

时间模糊化通过增加位置数据的时间域的不确定性,以减少位置数据的精度.一个简单的时间模糊的例子如图 2 所示:图 2(a)是两个移动物体在路网上移动的示意图,物体 1 的移动轨迹用黑色表示,物体 2 的移动轨迹用灰色表示.没有经过时间模糊时,他们要提交的位置信息如图 2(b)前两行所示.假设用户希望 t_2 时刻在黑色框内的用户不唯一,经过时间模糊后的位置数据如图 2(b)后两行所示.在 t_2 时刻,黑色框内物体 1 和物体 2 同时在 C 出现,达到了用户的隐私要求.



(a)

物体 1	$\langle A,t_1\rangle, \langle C,t_2\rangle, \langle E,t_3\rangle$
物体 2	$\langle B,t_0\rangle, \langle C,t_1\rangle, \langle D,t_2\rangle$
物体 1	$\langle A,[t_1,t_2]\rangle, \langle C,[t_2,t_3]\rangle, \langle E,[t_3,+\infty]\rangle$
物体 2	$\langle B,[t_0,t_2]\rangle, \langle C,[t_1,t_3]\rangle, \langle E,[t_3,+\infty]\rangle$

(b)

Fig.2 Illustration for time vagueness

图 2 时间模糊化示意图

由于时间模糊易于操作且实际应用通常不需要很大程度的模糊,它被广泛应用在隐私保护中.考虑到移动对象在某些敏感位置在时间上具有特征,比如在交通路口,当红色交通信号灯亮时,附近的移动对象会较长时间没有位置上的变化.对这些位置数据的时间域进行模糊,能够避免攻击者察觉到用户处于交通路口这一事件的发生[67].同时,有时满足用户隐私要求的区域不存在,这时,使用时间模糊化可以实现用户的隐私需求[48,49,66].

4 基于概率推测的位置大数据隐私保护技术

由于实现位置大数据的完美隐私代价较高,早期基于启发式隐私度量的方法只考虑当前时刻的位置信息是否会暴露用户的敏感位置信息.因此,用户的隐私信息可能会由于数据在时间和空间上的关系而泄露.

根据定义 1,隐私保护的目标是限制攻击者收集到用户的历史位置数据以后在某一时刻推测用户处于某敏感位置的概率的信息增益,即,知道用户的历史位置数据后计算出的用户某一时刻处于某敏感位置的后验概率与用户处于同一敏感位置的先验概率之差.计算某一时刻用户处于敏感位置的概率需要把握用户的位置数据在时间和空间上的关系.最新的研究成果显示:用户在未经保护的情况下,位置数据在时间和空间上的关系可以通过多种模型来刻画.当前主要使用隐马尔可夫模型[68]及其一般化模型图模型[69]刻画用户的位置数据在时间和空间上的关系.

满足隐私定义 1 的一种平凡的方法是抑制所有位置数据的发布,但是:(1) 这样做无法获得基于位置的服务;(2) 当用户放松隐私保护需求时,这样做在服务的性能上没有收益.另一方面,仅仅发布非敏感位置并不能有效地保护位置大数据的隐私,这是因为攻击者会根据自己掌握的模型推测当用户发布的数据为抑制时,用户处于一些特定的敏感位置.因此,可行的方法是为用户所处的每个可能的位置关联一个发布位置数据的概率,这些概率形成一个发布概率向量,用户在每个位置根据其关联的概率值对自己的位置进行发布或抑制发布.这样,攻击者无法区分敏感位置和非敏感位置,从而无法以较高的后验概率推测出用户处于哪个敏感位置.

4.1 基于隐马尔可夫模型进行概率计算

图 3 是将用户移动模式用隐马尔可夫模型建模的示意图,用户从每天的起始位置开始移动,以后的各个时刻会根据上一时刻用户位置的不同转移到各个其他位置,其中一些位置是敏感位置,在图中用浅色阴影表示.转移的概率由模型建立时形成的参数确定.攻击者推测用户处于敏感位置的先验概率是 0.2,假设无论用户处于敏感位置或非敏感位置都以 0.5 的概率发布位置,但攻击者接收到抑制发布的位置信息时,根据贝叶斯公式,其推测用户处于敏感位置的后验概率为

$$P(\text{用户处于敏感位置}|\text{发布位置信息})=\frac{P(\text{用户处于敏感位置且发布位置信息})}{P(\text{用户发布位置信息})}=\frac{0.2\times0.5}{0.2\times0.5+0.8\times0.5}=0.2.$$

若用户要求定义 1 中θ=0.1 的位置大数据隐私,则当前以向量(0.5,0.5)为发布概率向量就可以满足要求.

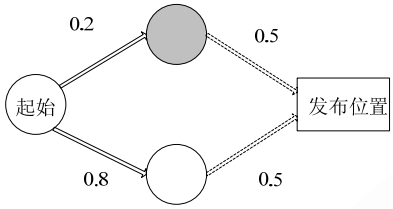


Fig.3 Illustration for hidden Markov model of user's moving pattern

图 3 用户移动的隐马尔可夫模型示意图

目前,由于隐马尔可夫模型具有模型简单且捕捉到的用户的位置数据在时间和空间上的关系较准确的特点[70],基于隐马尔可夫模型进行概率推测的位置大数据的隐私保护技术受到重视[71].

隐马尔可夫模型认为,用户发布的位置数据只与其当前所处的位置有关,因此,用户当前所处的位置对能否安全地发布位置数据具有很大的影响.比如:不敏感的当前位置对用户的位置隐私威胁很小因而容易被发布;同时,考虑到历史数据暗示了用户当前位置是否敏感,因此,用户在当前时刻前发布的历史数据也对当前位置是否能够安全地发布具有很大影响.总之,当前时刻所处的位置和当前时刻前发布的历史数据对是否发布位置数据直接影响了基于位置服务的可用性.为此,文献[71]提出了两种不同的抑制位置信息的方法,它们针对各自的用户群体具有各自适宜的服务可用性:

第 1 种方法首先针对用户的隐马尔可夫模型进行各个位置的发布概率向量 $p=(p_1,\ldots,p_n)$的预计算,其中,p_i表示将各个位置编号后,用户处于第 i 个位置的时候以 p_i 的概率发布当前位置.预计算分为两个阶段:第 1 个阶段生成一个候选的发布概率向量,第 2 个阶段判断这个发布概率向量是否能保护用户的位置隐私.这两个阶段都涉及到计算量大的问题.

在第 1 个阶段中,因为每个概率值是一个实数,枚举全部可能的概率向量是不可行的.注意到,发布概率的降低一定会保证更好的隐私以及发布概率的提高一定会保证更好的服务可用性,预计算从不发布全部位置信息开始,即,p=(0,…,0),利用贪心优化方法 MONDFRIAN[72]和 ALGPR[73]逐渐调整各个位置发布的概率,并检查在当前发布概率向量 p 是否能在任何情况下满足隐私保护的要求.

面对第 1 个阶段每次生成的一个候选的发布概率向量,第 2 个阶段需要考虑按照这个发布概率向量进行发布后,攻击者能够收集到的位置数据是否满足用户的位置隐私需求,即,没有泄露超过用户指定的信息量阈值.发布概率向量 p 确定以后,给定某个移动轨迹,用户在该轨迹下各个时刻发布的位置数据序列的概率很容易计算;反过来,利用贝叶斯公式可以根据用户发布的位置数据计算用户在某时刻处于某个位置的概率.利用这一点,根据攻击者可能获得的位置数据,计算用户属于敏感位置的概率(后验概率);同时,根据用户的移动模式已经建立起来的隐马尔可夫模型可以计算每个时刻用户处于某个敏感位置的概率(先验概率).根据后验概率与先验概率之差,判断 p 是否满足隐私需求.不断地修改 p 中各个元素的值,最终 p 会收敛.在实际运行中,系统按照收敛后的发布概率向量在各个位置发布用户的位置信息.

第 2 种方法省去了预计算过程,根据用户在当前时刻前的历史位置数据,计算用户当前位置处于各个敏感位置的先验概率;然后,考虑未来所有可能的位置发布,计算发布当前位置以后用户在各个时刻处于各个敏感位置的后验概率;根据这两个概率之差,判断发布当前位置是否破坏隐私需求,在线地决定是否发布当前位置信息.文献[74]将"考虑未来所有可能的位置发布"这样一项时间复杂度极高的任务,简化成多项式时间内可完成的任务.

4.2 基于图模型进行概率计算

隐马尔可夫模型在对人的移动模式进行建模时,假设人下一时刻移动到的位置只与当前位置有关,而与曾经的位置无关.这样的假设有利于模型的高效创建,但对用户在某时刻处于某位置的概率计算并不十分准确.图模型是对隐马尔可夫模型的一般化,它允许用户在某时刻处于某位置与历史位置数据有关.文献[75,76]使用图模型对用户的移动模式进行建模,获得了更为准确的位置数据在时间和空间上的关系,其先验概率和后验概率的计算与隐马尔可夫模型的计算是类似的,采用的均是教科书中标准的高效算法.

5 基于隐私信息检索的位置大数据隐私保护技术

上面的两类方法通过发布不精确的位置数据和抑制发布位置数据达到对位置大数据隐私保护的目的.然而,当用户的隐私需求较高或者需要完美隐私时,这两类方法都面临要么发布非常不精确的位置要么无法发布当前位置,因此,用户无法获得基于位置大数据的服务.基于隐私信息检索(PIR)的位置大数据隐私保护技术在确保服务可用的情况下不会泄露任何用户的位置信息,实现了完美隐私.

5.1 隐私信息检索方法简介

隐私信息检索理论最早被应用于访问网络中的外包数据,用户可以检索一个不可信的服务器上的任意数据项而不暴露用户检索的数据项信息[77].实现PIR的方法可以按照隐私保护的强弱分为基于信息论的PIR方法和基于计算能力的 PIR 方法.基于信息论的 PIR 方法保证攻击者无论拥有怎样的计算能力,都不能区分用户对不同数据项的访问;基于计算能力的 PIR 方法假设攻击者不具有计算求解某个难题的能力而保证攻击者不能区分用户对不同数据项的访问.

基于信息论的 PIR 方法有且只有一个平凡的解法:将全部信息都发送给客户端[78],这需要的传输代价是 $O(n)$,其中,n 是数据库的大小.因此,当前通常采用基于计算能力的PIR方法.目前常用的PIR方法有两种:一种基于一个难解的二次剩余假设问题[79],另一种基于伪随机函数的可实现性[80].这两种方法通常被硬件实现,安装在服务器端以消除查询中多个 PIR 访问间客户端与服务器之间的通信代价.

利用二次剩余假设问题的难解性实现 PIR 的原理如下:假设 m 是一个正整数,如果存在一个整数 x,满足 $x^2=a \bmod m$,则整数 a 被称作模 m 的“二次剩余(QR)”;否则,a 称为模 m 的“非二次剩余(QNR)”.二次剩余假设认为:在不给定 m 的因式分解的情况下,识别一个数 a 是否是模 m 的二次剩余,是一个 NP 问题.如图 4 所示(b_r 是非二次剩余,根据矩阵中每一列的值计算出的 P_i 如果是二次剩余,则说明元素 X_{ri} 为 0,否则说明元素 X_{ri} 为 1),基于计算能力的 PIR 方法将数据库看成一个矩阵$[x_{ij}]$,假设用户对 x_{ij} 感兴趣,用户预先计算好 $s-1$ 个模 m 的二次剩余 $a_1,\ldots,a_{i-1},a_{i+1},\ldots,a_s$ 和一个非二次剩余 b_i,并将向量$(a_1,\ldots,a_{i-1},b_i,a_{i+1},\ldots,a_s)$发往服务器端.服务器由于计算能力有限,无法区分哪个是二次剩余,哪个不是二次剩余,因此,服务器将这个向量看作$(u_1,\ldots,u_s)$,然后根据下述方法计算一个向量$(p_1,\ldots,p_s)$:

p_i 是根据矩阵的第 i 列计算出来的,它是那些值为 1 的 x_{ri} 对应的 u_r 的乘积.例如,如果第 i 列的 x 值都为 1,那么 $p_i=\prod_{j=1}^{s}u_j$.

如果 x_{ij} 为 0,那么 p_i 是那些二次剩余的乘积,因此 p_i 仍是一个二次剩余;反之,如果 x_{ij} 为 1,那么 p_i 的乘积中有一项非二次剩余,因此p_i是非二次剩余.客户端通过这个性质来知道x_{ij}的值,而不暴露自己想查哪个位置的值.

$$\begin{array}{cl} \begin{bmatrix} X_{11} & \cdots & X_{1s} \\ \vdots & \ddots & \vdots \\ X_{s1} & \cdots & X_{ss} \end{bmatrix} & \begin{array}{l} a_1(\mathrm{QR}) \\ b_r(\mathrm{QNR}) \\ a_s(\mathrm{QR}) \end{array} \\ \begin{array}{ccccc} P_1 & \cdots & P_i & \cdots & P_s \end{array} & \end{array}$$

Fig.4 Illustration for quadratic residue based PIR method

图 4 基于二次剩余方法的 PIR 示意图

利用伪随机函数实现 PIR 的基本方法如下:

首先,定义随机排列:

对于有 n 个元素的数据库 DB,其上的一个随机排列按照以下对应规则π将 DB 转换成 DB_π:$DB[i]=DB[\pi[i]]$.例如,对于数据库 $DB=\{o_1,o_2,o_3\}$,如果排列π是[2,3,1],则 $DB_\pi[1]=DB[\pi[1]]=DB[2]=o_2$,$DB_\pi[2]=DB[\pi[2]]=DB[3]=o_3$,$DB_\pi[3]=DB[\pi[3]]=DB[1]=o_1$.因此,$DB_\pi=\{o_2,o_3,o_1\}$.

当应用 PIR 方法时,首先对数据库中的数据块进行加密重排处理.之后,每当查询到来时,选取处理后的数据块返回客户端,并将获取的数据块再次与其他数据块混合重排.另外,控制每一个查询都访问相同次数的页面,从而保证服务器所有的查询都是不能区分的[81].

尽管不同的 PIR 方法具有不同的通信代价和计算量,但对基于隐私信息检索的位置大数据隐私保护方法来说,通过 PIR 访问服务器端数据的方法都是通过提供一个数据块编号的方法向服务器安全地获取这个数据块的内容.

5.2 基于隐私信息检索方法保护位置大数据隐私的框架

将隐私信息检索方法应用到位置大数据隐私保护上时,其流程如图 5 所示.用户在本地进行计算,计算所需的数据使用 PIR 方法向服务器端获取,于是,服务器不知道用户的位置数据就给出了用户查询结果,这是实现完美隐私的基础.但是仅仅保护每一次获取数据时的隐私,不能保证用户完全的隐私.用户的一个查询通常需要多次访问才能获得查询结果,比如导航应用中,不同的起始点和目的地需要的 PIR 访问次数也不相同.由于用户不同的位置会导致一个查询中不同的 PIR 访问次数,攻击者可以对用户的当前位置进行推测[82].因此,基于 PIR 的方法需要通过访问一些无用的数据块让用户在不同位置提交的查询都具有相同的 PIR 的访问次数,以避免攻击者的推测.然而,PIR 方法每次不泄露信息的访问带来性能上很大的开销,为了避免这些性能上的开销,有些工作通过对原始数据进行预处理来加速服务的访问.

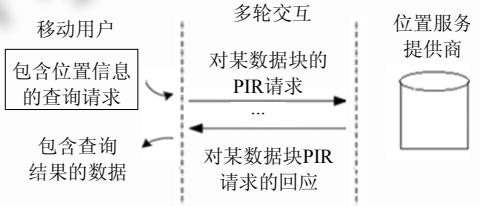


Fig.5 Illustration for workflow of PIR based privacy preservation of location based big data

图 5 使用 PIR 方法实现位置大数据的隐私保护流程示意图

根据用户不同的查询目的,当前,基于 PIR 的方法主要针对两类被广泛应用的查询:最短路径计算以及近邻查询.

5.3 最短路径计算中基于PIR的方法

最短路径的计算,是众多位置大数据上的应用服务需要使用的技术.在最短路径的计算中,用户需要提供自己当前的位置,这存在暴露位置隐私的风险.基于 PIR 的方法既可以保护用户的位置隐私,还可以保证得到正确的结果.

最短路径的计算基于 Dijkstra 算法[83]或 A*搜索算法[84],PIR 提供的访问接口可以看作实现了完美隐私的对传统存储介质(内存或磁盘)的访问.用户使用任何已有的最短路径计算方法将其中访问存储的接口改为使用 PIR 的访问接口,就可以获得正确的结果.但是,由于不同的起点和终点对(S,T)在计算最短路径时需要的访问次数不同,比如距离较近的起点终点对需要的 PIR 访问次数较少,而距离较远的需要访问的次数较多,为了避免不同起点和终点对之间 PIR 访问次数不同导致的位置隐私泄露,使用 PIR 方法的查询计划需要保证每次请求相同次数的 PIR 访问请求.

一种平凡的保护隐私的方法是:针对任何一个起点、终点对(S,T),服务器都预计算并保存 S 到 T 的最短距离,当用户请求 S 到 T 的最短距离时,可以只进行一轮访问一个存储位置 $S\times\mathrm{Max}+T$(其中,Max 是任何点编号中的最大值)的 PIR 请求,就可以完成最短路径的计算.但这种方法要求用户发起查询的位置是静态的,为了服务大数据时代任何位置的用户,文献[82]将平凡的保护方法中预计算的粒度由点扩大到区域.空间被划分成若干区域,如图 6 所示,路网被划分成了 8 个不同的区域.每个起点、终点对之间的最短距离由它们所在区域间所有最短路径经过的区域来表示.其中,区域 R_1 与区域 R_7 之间的最短路径经过的区域由 $S_{1,7}$ 表示.在图 6 中,R_1 和 R_7 之间只有两条最短路径,分别用实线折线和虚线折线表示,即 $S_{1,7}=\{R_3,R_4\}$.用户在请求 R_1 中的点与 R_7 中的点的最短路径时,使用 PIR 接口获取 R_3 和 R_4 的数据,结合同样使用 PIR 接口获得的 R_1 和 R_7 中包含的数据,在本地构建出包含所求最短路径的子图,通过运行已有的最短路径计算方法得到查询结果.由于不同的 $S_{i,j}$ 具有不同的集合大小,为了对任何查询都使用相同次数的 PIR 访问磁盘块,所有 $S_{i,j}$ 的大小都等于 $S_{i,j}$ 中最大的.

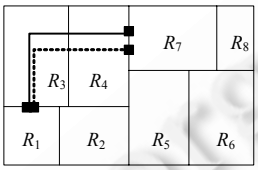


Fig.6 Illustration for PIR based shortest path calculation

图 6 利用 PIR 计算最短路径示意图

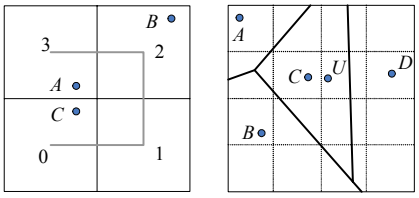
5.4 近邻查询中基于PIR的方法

用户除了查询自己与目的地之间的最短路径以外,还会查询自己周围有哪些物体,近邻查询同样具有最短路径计算中的隐私问题.现有的工作逐渐解决了保证完美隐私的最近邻查询问题[85]和 k 最近邻查询问题[81,86].

最近邻的查询问题分为近似最近邻的寻找和精确最近邻的寻找.近似最近邻用于那些需要实时查询自己周围有哪些物体的用户,但并不保证获取的位置一定是用户请求位置的最近邻.图 7(a)是近似最近邻查询的一个示意图,服务器为每个数据点用希尔伯特曲线进行降维,比如,A(B,C)的希尔伯特值分别为 3(2,0),在查询某个点的近似最近邻时,只需要向服务器查询与要查询点的希尔伯特值最接近的希尔伯特值对应的空间位置.比如,A 点的近似最近邻是 B 点,因为 B 的希尔伯特值为 2,与 A 的希尔伯特值 3 最接近.服务器将各空间点的希尔伯特值组织成一棵 B 树,由于 B 树是平衡的,查询与任何点最接近的希尔伯特值就可以通过相同次数的 PIR 访问来完成.因此,攻击者区分用户的不同查询,进而无法推测用户的位置隐私.

对于那些需要准确结果的用户,服务器将空间按照网格的方式划分成若干区域,用户保存这个划分结果,对于某个空间点的 k 近邻查询,用户首先在网格中定位该空间点,然后寻找近似最近邻,以这两个点之间的距离为半径、以待查询的空间点为圆心画圆,圆形与网格重叠的格子里一定包含最近邻[81].此外,为了进一步降低 PIR 的访问次数,服务器为空间中的点计算维诺图[87],每个位置点都存在于一个多边形区域中,每个网格区域与若干包含位置点的多边形区域重叠,如图 7(b)所示.根据维诺图的定义,每一个多边形区域只包含一个位置点,若某个要查询最近邻的位置被包含于某个多边形区域,这个点的最近邻一定是这个多边形区域对应的那个位置点.每一个空间划分后的网格区域都记录与其有重叠的多边形区域,其中每个网格区域最多与 $P_{\max}$ 个多边形区域产生重叠,$P_{\max}$ 在预计算过程中确定.比如在图 7(b)中,A 所在的网格区域只包含一个维诺区域,而图 7(b)中第 4 行第 3 列的网格区域则包含 3 个维诺区域,因此 $P_{\max}=3$.这样,当最近邻查询请求到来时,首先通过一次 PIR 访问获得查询点所在的区域,然后使用 $P_{\max}$ 次 PIR 访问获得所有可能的最近邻.这样,任何精确最近邻查询都使用 $P_{\max}+1$ 次 PIR 访问获得查询结果,没有泄露用户的任何隐私.

一般的 k 近邻查询也被众多服务请求,比如,根据位置进行附近商家的推荐.文献[81,86]分别通过寻找距离查询点最近的 k 个希尔伯特值对应的空间点,以这些空间点中与查询点的最远距离为半径、以查询点为圆心画圆,圆圈与网格区域相交的部分一定包含了查询点的 k 近邻.



(a) (b)

Fig.7 Illustration for PIR based k nearest neighbor calculation

图 7 利用 PIR 计算近邻示意图

从 PIR 方法的两个不同类型的应用可以看出,PIR 方法的关键在于查询计划如何使不同位置的查询不可区分.概括来说,现有方法使用空间划分和预计算的方式为不同的位置确定 PIR 接口需要访问次数的上限,对于无需访问这么多位置的查询,则提交相应数量的无用查询.

6 总 结

位置大数据隐私保护技术具有广泛的应用,是近年来学术界新兴的研究方向.本文对位置大数据隐私保护研究现状进行了综述.本文介绍了位置大数据的概念以及位置大数据的隐私威胁,总结了针对位置大数据隐私的统一的基于度量的攻击模型,并以此为依据,对目前位置大数据隐私保护领域已有的研究成果进行了归纳,介绍了基于启发式隐私度量、概率推测以及隐私信息检索(PIR)的 3 大类隐私保护技术,特别是对当前位置大数据隐私保护的研究前沿问题“基于隐私信息检索的位置大数据隐私保护技术”进行了比较详尽的阐述与分析.

容易看出,每类位置大数据隐私保护技术都有不同的特点,针对不同的应用需求,我们将各种隐私保护技术的分析比较结果列在表 1 中,从表 1 中可以看出,它们的适用范围、性能表现等不尽相同.当对位置数据的隐私程度要求较高且对计算开销要求较高时,基于概率推测的位置大数据隐私保护技术更适合;当关注位置信息的完美隐私保护时,则应考虑基于隐私信息检索的位置大数据隐私保护技术,这时,计算量以及响应时间上的代价较高.基于启发式隐私度量的位置大数据隐私保护技术能以较低的计算开销实现对一般隐私需求的保护,表 2 对位置大数据隐私保护技术作了进一步的对比分析.

Table 1 Performance evaluation for privacy preservation of location based big data

表 1 位置大数据的隐私保护技术性能评估

	隐私保护度	运行时开销	预计算开销	数据缺失
基于启发式隐私度量的位置大数据隐私保护技术	中	中	低	中
基于概率推测的位置大数据隐私保护技术	中高	低	高	中
基于隐私信息检索的位置大数据隐私保护技术	高	高	高	低

Table 2 Comparison and analysis for privacy preservation of location based big data

表 2 位置大数据的隐私保护技术对比分析

	主要优点	主要缺点	代表技术
基于启发式隐私度量的位置大数据隐私保护技术	计算开销中等,实现简单	位置解析度失真,会受到基于数据特征推测的攻击	k 匿名技术[62] 匿名框技术[47,48] 考虑数据特征的方法[49]
基于概率推测的位置大数据隐私保护技术	计算开销小,位置发布准确	数据缺失,服务可用性降低,预计算开销大,依赖于位置数据模型,不同的模型需设计不同的算法	基于隐马尔可夫模型[71,74] 基于图模型[75]
基于隐私信息检索的位置大数据隐私保护技术	完全隐私保护,服务可用性高	预计算代价大,运行时代价大,需要针对应用设计优化方法	针对最短路径[82] 针对最近邻[85,88] 针对 k 近邻[86,88]

7 未来展望

在大数据时代,获取信息的渠道越来越多,获取到的位置信息也越来越多样化,类型也不限于第 2 节中介绍的单一的位置数据.下面从位置数据与非位置数据相结合、移动社交网络以及背景知识攻击这 3 个方面介绍未来的位置大数据的隐私保护技术.

7.1 位置数据与非位置数据结合的位置大数据的隐私保护技术

大数据时代,攻击者可以从多种渠道获得用户和位置数据相关的其他类型数据,并结合位置数据共同推测用户的隐私信息.此时,位置数据与非位置数据之间使用用户的个性或者行为模式进行匹配,成为联系位置数据与非位置数据的通用方法.位置数据与非位置数据相结合的位置大数据隐私可以如下定义[20]:

设 $P=\{p_1,\ldots,p_n\}$为从某个角度对用户个性的度量,比如,P 可以是从对商品喜好角度对用户个性的度量,这时,p_i 表示一个用户对编号为 i 的物品的喜好程度.同时,$L=\{l_1,\ldots,l_m\}$为攻击者收集的用户的位置数据,考虑一个偏好函数 F 将某个用户的 L 映射为对用户个性的度量 P.如果攻击者可以收集到相同角度度量用户个性的数据 $X=\{x_1,\ldots,x_n\}$,其中,这个例子中 x_i 表示某个用户对编号为 i 的商品的喜好程度.这种攻击方式一样可以纳入第 1 节中介绍的位置大数据的统一攻击模型.用 $U_i^{t'}$ 表示某个用户 i 在 t'时刻的位置数据与 x_i 匹配成功,用 L_i^t 表示用户 i 到 t 时刻为止发布的位置数据,在任一时刻 t,为了避免攻击者获得用户位置信息后推测用户的敏感信息(比如身份),与第 1 节中针对单一位置数据类型的攻击模型一样,位置数据与非位置数据相结合的位置大数据隐私一样可以类似定义 1 来定义:

$$P\{U_i^{t'} \mid L_i^t\} - P\{U_i^{t'}\} \leqslant \theta,$$

其中,先验概率与后验概率的计算除了第 1 节中提到的与攻击者收集到的用户的历史数据有关外,还与由用户的位置信息映射成的个性向量与用户的位置信息构成的个性向量之间的匹配难度有关.根据文献[89]提出的基于信息论的匹配方法,匹配成功的难度取决于攻击者收集到不同用户的位置数据的差异性,可以由以下公式定义:

$$d_{\min}=\min p_i, p_j \in P|\{k|p_{i,k}!=p_{j,k}, k\in(1,|p_i|)\}|,$$

其中,$d_{\min}$表示任意不同用户的位置数据映射到用户个性向量p_i和p_j不同的元素个数的最小值,它代表了最接近的两个向量的差异程度.显然,$d_{\min}$ 较大时,不同的用户之间很好区分;而当 $d_{\min}$ 较少时,区分就会变得困难甚至不可能.

根据以上介绍,位置数据与非位置数据相结合后进行隐私保护的研究还有许多重要问题需要解决.

在大数据时代,含有多种类型的位置大数据包含了用户的行为模式信息,它们会导致用户位置大数据隐私的泄露.例如,超市的购物数据包含了用户对各种商品的喜好,攻击者可以根据这些喜好匹配一些明显的行为模式.文献[20]借鉴 Narayanan 等人将 NetFlix 发布的匿名数据集[90]通过与购物数据相关联进行反匿名化的方法,提出了将用户的位置数据集合$\{L_i:L_i=[l_1,\ldots,l_n]\}$,其中,$l_j$代表匿名用户 i 在时刻 j 的位置映射为该用户对各种商品的喜好程度 $P=\{p_1,\ldots,p_n\}$(其中,p_i表示该用户对编号为 i 的商品的喜好程度),并结合使用由商场购物信息得到的用户的偏好向量集合$\{X_i:X_i=[x_1,\ldots,x_n]\}$来确定每个匿名用户与商场购物信息中的用户的匹配程度,其中,x_{ij} 表示用户 i 对商品 j 的偏好程度.为了进行匹配,计算每个用户经过位置数据映射后的个性向量与攻击者收集到的不同用户的个性向量之间的差异度.差异度最小的匿名用户数据与商场购物数据匹配成功.如果若干匿名用户的个性向量之间差异不可区分,则匹配失败.

为了保护位置数据与非位置数据结合后位置大数据的隐私,我们认为,位置大数据隐私保护方法的设计者应研究用户位置与行为模式之间的映射关系,以设法降低攻击者根据从匿名用户位置推测出的个性向量与攻击者收集到的不同用户之间的个性向量的匹配程度.我们认为,这与以下两个因素有关:

(1) 匿名用户的位置数据增多后,其映射成的偏好程度向量会变得准确;反之,偏好程度向量会变得模糊;

(2) 匿名用户偏好程度的向量变得准确,有利于与攻击者收集到的个性向量进行匹配,且映射后不可区分的个性向量会变少;反之,不可区分的匿名数据数量会增多.

因此,我们认为保护用户的位置大数据隐私的关键是:在确保服务可用的前提下,尽量让映射后的个性向量变得模糊.这可以从两个方面入手:(1) 减少用户发布的位置数据的数量,这相当于减少了映射后的个性向量的信息量,因此映射出的个性向量应更模糊;(2) 降低用户的位置数据中的元素映射成有效个性向量中元素的能力.但是,针对不同应用的特点,如何在减少位置数据规模的同时保持基于位置服务的准确性,是一个具有挑战性的问题.

7.2 移动社交网络中位置大数据的隐私

第 7.1 节中介绍的领域相关的映射方法将位置数据与非位置数据关联,不同的领域需要相应的领域专家设计各自的映射方法.大数据时代的新应用移动社交网络,为位置数据与文本、图片和用户个人信息进行了自然的结合,为攻击者和隐私保护者提供了新问题.在移动社交网络中,用户的位置信息与用户的身份信息显式地结合,比如在移动社交网络的签到服务中,用户使用带有定位功能的移动设备将自己的位置和自己的标识发布在社交网络中.移动社交网络上,位置大数据的隐私保护问题是前面提到的位置数据与非位置数据结合的位置大数据隐私问题的特例,因此,其上的隐私同样可以使用统一的θ隐私来度量.当前,移动社交网络上位置大数据的隐私保护的研究还处于初级阶段,当前的保护手段主要是通过对数据进行隐私等级的分类,然后使用访问控制策略防止攻击者访问敏感位置数据[91].此外,最新的研究试图将单一位置大数据隐私保护方法应用到移动社交网络中,提出了 k 匿名的轨迹隐私保护方法[92]以及概率推测的方法[93].这些方法的提出顺序与单一位置数据进行保护的方法的发展过程类似,从使用访问控制策略的方法到对单点的位置保护,逐渐发展到对历史位置数据的保护.其采用的技术大多是已有方法的变种,但由于研究时间短,还有大量针对单一位置数据的位置大数据的隐私保护方法没有在移动社交网络中得到应用.我们认为,针对移动社交网络的位置大数据的隐私保护技术是未来位置大数据的隐私保护技术的重要组成部分,考虑移动社交网络中用户相关的位置数据与非位置数据之间的关系,并防止攻击者利用该关系推测用户的敏感信息,是未来位置大数据的隐私保护技术的研究方向.

7.3 针对背景知识的位置大数据隐私保护技术

现有的基于概率推测的位置大数据隐私保护技术假设攻击者对用户的背景知识的掌握是固定的,但在大数据时代,攻击者可以持续地收集用户的历史数据从而具备学习用户的背景知识的能力.这时,现有的基于概率推测的位置大数据隐私保护技术不能有效地保护用户的位置隐私[71].差分隐私是迄今为止针对攻击者的先验知识进行普遍保护的最有效的技术,尽管目前差分隐私主要针对离线数据,不能直接应用于位置大数据的隐私保护,但我们认为,将差分隐私保护技术应用到位置大数据隐私保护技术中,是未来有潜力的研究方向.差分隐私同样遵循定义 1 中位置大数据的攻击模型:

如果 L_t 是攻击者收集到 t 时刻之前的历史位置数据,$U_i^{t'}$ 表示用户在 t'时刻处于敏感位置 S_i,按照差分隐私的定义[94],如果 $P\{U_i^{t'}\}/P\{U_i^{t'}\mid L_t\}\leqslant e^{\varepsilon}$,那么当前位置数据满足差分隐私条件可以被发布;否则,不能直接发布或需要经过某种变换机制才能进行发布.注意到,只要令$\varepsilon=-\log(\theta)$,差分隐私即满足定义 1 中位置大数据的隐私中的$\theta$ 隐私.

当前,差分隐私的研究主要集中在如何添加假数据干扰分析结果,主要的技术包括针对数据类型是整数的 Laplace 机制[95]、针对数据类型是浮点数的 exponential 机制[96]以及它们在性能上的改进 geometric 机制[97].差分隐私在数据分析上的应用很多,但还没有针对位置大数据的研究.尽管目前为止差分隐私的研究已经涉及到与位置数据有关的特定多维索引四叉树,但我们认为,还没有提出能够满足位置大数据的隐私保护技术.由于差分隐私对于攻击者背景知识的假设十分保守,同时,大数据时代攻击者获取背景知识的渠道和途径十分广泛,我们认为,将差分隐私保护技术应用到位置大数据隐私保护中是具有广阔前景的研究方向.

8 结束语

大数据时代用户的位置数据会从不同的视角和渠道被收集.位置大数据在给人们的生活、商业运作方式以及科学研究带来巨大收益的同时,由于其中蕴含了人们的行为模式、习惯偏好和敏感信息等隐私信息,为人们

带来了严重的隐私威胁.由于经典的基于“知情与同意”以及匿名的位置隐私保护方法不能全面地保护用户的位置隐私,目前的研究者已在位置大数据的隐私保护方面做了大量的工作.本文对位置大数据隐私保护技术的研究成果进行了回顾和总结,综述了位置大数据隐私保护技术研究的现状,总结位置大数据隐私统一的基于度量的攻击模型,并以此为依据,分析和对比了基于启发式隐私度量、概率推测和隐私信息检索的位置大数据隐私保护方法,指出各种方法适宜的情况和特点.最后,对未来位置数据与非位置数据相结合等位置大数据的隐私保护研究方向进行了探讨.总之,位置大数据隐私保护技术属于大数据带来的新兴研究领域,仍然有大量关键问题需要深入而细致的研究.

致谢 在此,我们向对本文的工作给予支持和建议的同行表示感谢.

References:

[1] Jabeur N, Zeadally S, Sayed B. Mobile social networking applications. Communications of the ACM, 2013,56(3):71−79. [doi: 10.1145/2428556.2428573]

[2] Sousa M, Techmer A, Steinhage A, Lauterbach C, Lukowicz P. Human tracking and identification using a sensitive floor and wearable accelerometers. In: Proc. of the IEEE Int’l Conf. on Pervasive Computing and Communications (PerCom). San Diego, 2013. 166−171. [doi: 10.1109/PerCom.2013.6526728]

[3] Ugolotti R, Sassi F, Mordonini M, Cagnoni S. Multi-Sensor system for detection and classification of human activities. Journal of Ambient Intelligence and Humanized Computing, 2013,4(1):27−41. [doi: 10.1007/s12652-011-0065-z]

[4] Anguelov D, Dulong C, Filip D, Frueh C, Lafon S, Lyon R, Ogale A, Vincent L, Weaver J. Google street view: Capturing the world at street level. Computer, 2010,43(6):32−38. [doi: 10.1109/MC.2010.170]

[5] Civilis A, Jensen CS, Pakalnis S. Techniques for efficient road-network-based tracking of moving objects. IEEE Trans. on Knowledge and Data Engineering, 2005,17(5):698−712. [doi: 10.1109/TKDE.2005.80]

[6] Mayer-Schönberger V, Cukier K. Big Data: A Revolution that Will Transform how We Live, Work, and Think. Eamon Dolan/Houghton Mifflin Harcourt, 2013. 102−105.

[7] Dijcks JP. Oracle: Big Data for the Enterprise. White Paper. Oracle, 2012.

[8] Zheng K, Shang S, Yuan NJ, Yang Y. Towards efficient search for activity trajectories. In: Proc. of the 29th IEEE Int’l Conf. on Data Engineering (ICDE 2013). Brisbane, 2013. 230−241. [doi: 10.1109/ICDE.2013.6544828]

[9] Abowd G, Atkeson C, Hong J, Long S, Kooper R, Pinkerton M. CyberGuide: A mobile context-aware tour guide. Wireless Networks, 1997,3(5):421−43. [doi: 10.1023/A:1019194325861]

[10] NextBus Inc. 2004. http://www.nextbus.com/

[11] Smith CW, *et al*. System and method for providing traffic information using operational data of a wireless network. CI: G08G 1/01. US Pat 10/243, 589, 2002.

[12] Sythoff J, Morrison J. Location-Based services. 2011. http://www.pyramidresearch.com/store/Report-Location-Based-Services.htm

[13] Litman T. Distance-Based vehicle insurance: Feasibility, costs and benefits. Comprehensive Technical Report, Victoria Transport Policy Institute, 2011.

[14] Bacheldor B. UPS slashed the time it takes to determine the least-expensive route from months to days to hours and wants to make that information available in real time. 2004. http://www.informationweek.com/breakthrough/d/d-id/1023066

[15] Hill S, Banser A, Berhan G, Eagle N. Reality mining Africa. In: Proc. of the AAAI Spring Symp. on Artificial Intelligence for Development. 2010. http://ai-d.org/pdfs/Hill.pdf

[16] Chi HB. Three circle at three location: Weibo can locate. 2013. http://www.fawan.com.cn/html/2013-07/03/content_442649.htm

[17] Williams C. Apple under pressure over iphone location tracking. 2011. http://www.telegraph.co.uk/technology/apple/8466357/Apple-underpressureover-iPhone-location-tracking.html

[18] Cheng J. How apple tracks your location without your consent and why it matters. 2011. http://arstechnica.com/apple/news/2011/04/how-appletracks-your-location-without-your-consent-and-why-it-matters.ars

[19] Hansell S. AOL removes search data on vast group of Web users. 2006. http://query.nytimes.com/gst/fullpage.html?res=9504E5D81E3FF93BA3575BC0A9609C8B63

[20] Wicker SB. The loss of location privacy in the cellular age. Communications of the ACM, 2012,55(8):60−68. [doi: 10.1145/2240236.2240255]

[21] Narayanan A, Shmatikov V. Robust de-anonymization of large sparse datasets. In: Proc. of the IEEE Symp. on Security and Privacy. Oakland, 2008. 111−125. [doi: 10.1109/SP.2008.33]

[22] Beresford AR, Rice A, Skehin N, Sohan R. MockDroid: Trading privacy for application functionality on smartphones. In: Proc. of the 12th Workshop on Mobile Computing Systems and Applications. ACM Press, 2011. 49−54. [doi: 10.1145/2184489.2184500]

[23] Beresford AR, Stajano F. Location privacy in pervasive computing. Pervasive Computing, IEEE, 2003,2(1):46−55. [doi: 10.1109/MPRV.2003.1186725]

[24] Agrawal D, Bernstein P, Bertino E, Davidson S, Dayal U, Franklin M, Gehrke J, Haas L, Halevy A, Han J, Jagadish HV, Labrinidis A, Madden S, Papakonstantinou Y, Patel JM, Ramakrishnan R, Ross K, Shahabi C, Suciu D, Vaithyanathan S, Widom J. Challenges and opportunities with big data—A community white paper developed by leading researchers across the United States. 2012. http://cra.org/ccc/docs/init/bigdatawhitepaper.pdf

[25] Tsai J, Kelley P, Cranor L, Sadeh N. Location-Sharing technologies: Privacy risks and controls. 2009. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1997782

[26] Sadeh N. M-Commerce: Technologies, Services, and Business Model. Wiley, 2002.

[27] Glass J. Shyhood is location. 2014. http://www.skyhookwireless.com/

[28] Kim M, Fielding JJ, Kotz D. Risks of using AP locations discovered through war driving. Pervasive Computing, 2006,3968:67−82. [doi: 10.1007/11748625_5]

[29] Frommer D. Loopt location to update in the background on iphone. 2009. http://www.businessinsider.com/loopt-to-run-in-the-background-on-iphone-2009-6

[30] Roberts P, Challinor S. IP address management. BT Technology Journal, 2000,18(3):127−136. [doi: 10.1023/A:1026749131441]

[31] Loki. http://loki.com/

[32] FireEagle. http://info.yahoo.com/privacy/us/yahoo/fireeagle/

[33] Google latitude. http://www.google.com/latitude/apps/badge

[34] Zhou SG, Li F, Tao YF, Xiao XK. Privacy preservation in database applications: A survey. Chinese Journal of Computers, 2009,32(5):847−861 (in Chinese with English abstract). [doi: 10.3724/SP.J.1016.2009.00847]

[35] Fitzpatrick M. Mobile that allows bosses to snoop on staff developed. BBC News. 2010. http://news.bbc.co.uk/2/hi/technology/8559683.stm

[36] Decker M. Location privacy—An overview. In: Proc. of the 7th Int'l Conf. on Mobile Business. Barcelona, 2008. 221−230. [doi: 10.1109/ICMB.2008.14]

[37] Ngai ECH, Rodhe I. On providing location privacy for mobile sinks in wireless sensor networks. Wireless Networks, 2013,19(1):115−130. [doi: 10.1007/s11276-012-0454-z]

[38] Wernke M, Skvortsov P, Dürr F, Rothermel K. A classification of location privacy attacks and approaches. Personal and Ubiquitous Computing, 2014,18(1):163−175. [doi: 10.1007/s00779-012-0633-z]

[39] Bonchi F. Privacy preserving publication of moving object data. In: Bettini C, *et al*., eds. Proc. of the Privacy in Location-Based Applications. Berlin, Heidelberg: Springer-Verlag, 2009. 190−215. [doi: 10.1007/978-3-642-03511-1_9]

[40] Fung B, Cao M, Desai BC, Xu H. Privacy protection for RFID data. In: Proc. of the 2009 ACM Symp. on Applied Computing. Honolulu, 2009. 1528−1535. [doi: 10.1145/1529282.1529626]

[41] Bertino E, Catania B, Damiani ML, Perlasca P. Geo-RBAC: A spatially aware RBAC. In: Proc. of the 10th ACM Symp. on Access Control Models and Technologies. Stockholm, 2005. 29−37. [doi: 10.1145/1063979.1063985]

[42] Gunter CA, May MJ, Stubblebine SG. A formal privacy system and its application to location based services. In: Proc. of the 4th Int'l Workshop on Privacy Enhancing Technologies.Toronto, 2004. 256−282. [doi: 10.1007/11423409_17]

[43] Myles G, Friday A, Davies N. Preserving privacy in environments with location-based applications. Pervasive Computing, 2003, 2(1):56−64. [doi: 10.1109/MPRV.2003.1186726]

[44] Snekkenes E. Concepts for personal location privacy policies. In: Proc. of the 3rd ACM Conf. on Electronic Commerce. Tampa, 2001. 48−57. [doi: 10.1145/501158.501164]

[45] Yiu ML, Jensen CS, Møller J, Lu H. Design and analysis of a ranking approach to private location-based services. ACM Trans. on Database Systems (TODS), 2011,36(2):10. [doi: 10.1145/1966385.1966388]

[46] Gruteser M, Grunwald D. Anonymous usage of location-based services through spatial and temporal cloaking. In: Proc. of the 1st Int'l Conf. on Mobile System, Application, and Services. San Francisco, 2003. 31−42. [doi: 10.1145/1066116.1189037]

[47] Mokbel MF, Chow CY, Aref WG. The new Casper: Query processing for location services without compromising privacy. In: Dayal U, ed. Proc. of the 32nd Int'l Conf. on Very Large Data Bases. New York: ACM Press, 2006. 763−774.

[48] Gedik B, Liu L. Protecting location privacy with personalized *k*-anonymity: Architecture and algorithms. IEEE Trans. on Mobile Computing, 2008,7(1):1−18. [doi: 10.1109/TMC.2007.1062]

[49] Pan X, Xu J, Meng X. Protecting location privacy against location-dependent attacks in mobile services. IEEE Trans. on Knowledge and Data Engineering, 2012,24(8):1506−1519. [doi: 10.1109/TKDE.2011.105]

[50] Xu T, Cai Y. Exploring historical location data for anonymity preservation in location-based services. In: Proc. of the 27th IEEE Int'l Conf. on Computer Communications. Phoenix, 2008. 547−555. [doi: 10.1109/INFOCOM.2008.103]

[51] Chow CY, Mokbel MF. Trajectory privacy in location-based services and data publication. ACM SIGKDD Explorations Newsletter, 2011,13(1):19−29. [doi: 10.1145/2031331.2031335]

[52] Nergiz ME, Atzori M, Saygin Y. Towards trajectory anonymization: A generalization-based approach. In: Proc. of the SIGSPATIAL ACM GIS 2008 Int'l Workshop on Security and Privacy in GIS and LBS. Irvine, 2008. 52−61. [doi: 10.1145/1503402.1503413]

[53] Machanavajjhala A, Reiter JP. Big privacy: Protecting confidentiality in big data. XRDS: Crossroads. The ACM Magazine for Students, 2012,19(1):20−23. [doi: 10.1145/2331042.2331051]

[54] Huo Z, Meng XF. A survey of trajectory privacy-preserving techniques. Chinese Journal of Computers, 2011,34(10):1820−1830 (in Chinese with English abstract). [doi: 10.3724/SP.J.1016.2011.01820]

[55] Liu F, Hua KA, Cai Y. Query l-diversity in location-based services. In: Proc. of the 10th Int'l Conf. on Mobile Data Management. Taipei, 2009. 436−442. [doi: 10.1109/MDM.2009.72]

[56] Bamba B, Liu L, Pesti P, Wang T. Supporting anonymous location queries in mobile environments with privacy grid. In: Huai J, ed. Proc. of the 17th Int'l Conf. on World Wide Web. New York: ACM Press, 2008. 237−246.

[57] Chen J, Xu H, Zhu L. Internet of Things. Berlin, Heidelberg: Springer-Verlag, 2012. 157−165.

[58] Liu L. From data privacy to location privacy: Models and algorithms. In: Koch C, ed. Proc. of the 33rd Int'l Conf. on Very Large Data Bases. New York: ACM Press, 2007. 1429−1430.

[59] Gentry C. A fully homomorphic encryption scheme [Ph.D. Thesis]. Stanford: Stanford University, 2009.

[60] Li FF, Xiao XK. Secure nearest neighbor revisited. In: Proc. of the 29th IEEE Int'l Conf. on Data Engineering. Brisbane, 2013. 733−744. [doi: 10.1109/ICDE.2013.6544870]

[61] Xu J, Tang X, Hu H, Du J. Privacy-Conscious location-based queries in mobile environments. IEEE Trans. on Parallel and Distributed Systems, 2010,21(3):313−326. [doi: 10.1109/TPDS.2009.65]

[62] Gedik B, Liu L. A customizable *k*-anonymity model for protecting location privacy. Technical Report, Georgia Institute of Technology, 2004. https://smartech.gatech.edu/bitstream/handle/1853/100/git-cercs-04-15.pdf;jsessionid=FC0FA9422D52E947651B4E565E690BE1.smart1?sequence=1

[63] Kido H, Yanagisawa Y, Satoh T. Protection of location privacy using dummies for location-based services. In: Proc. of the 21st Int'l Conf. on Data Engineering. Tokyo, 2005. 1248−1248. [doi: 10.1109/ICDE.2005.269]

[64] Suzuki A, Iwata M, Arase Y, Hara T, Xie X, Nishio S. A user location anonymization method for location based services in a real environment. In: Proc. of the 18th ACM SIGSPATIAL Int'l Symp. on Advances in Geographic Information Systems. Sana Jose, 2010. 398−401. [doi: 10.1145/1869790.1869846]

[65] Kato R, Iwata M, Hara T, Suzuki A, Arase Y, Xie X, Nishio S. A dummy-based anonymization method based on user trajectory with pauses. In: Proc. of the 20th ACM SIGSPATIAL Int'l Conf. on Advances in Geographic Information Systems. Redondo, 2012. 249−258. [doi: 10.1145/2424321.2424354]

[66] Yigitoglu E, Damiani ML, Abul O, Silvestri C. Privacy-Preserving sharing of sensitive semantic locations under road-network constraints. In: Proc. of the 13th IEEE Int'l Conf. on Mobile Data Management (MDM). Bengaluru, 2012. 186−195. [doi: 10.1109/MDM.2012.48]

[67] Palanisamy B, Liu L. Mobimix: Protecting location privacy with mix-zones over road networks. In: Proc. of the 27th Int'l Conf. on Data Engineering (ICDE). Hannover, 2011. 494−505. [doi: 10.1109/ICDE.2011.5767898]

[68] Eddy SR. Hidden Markov models. Current Opinion in Structural Biology, 1996,6(3):361−365. [doi: 10.1016/S0959-440X(96)80056-X]

[69] Lafferty J, McCallum A, Pereira FCN. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Brodley CE, ed. Proc. of the 18th Int'l Conf. on Machine Learning. San Francisco: Morgan Kaufmann Publishers, 2001. 282−289.

[70] Kim E, Helal S, Cook D. Human activity recognition and pattern discovery. Pervasive Computing, 2010,9(1):48−53. [doi: 10.1109/MPRV.2010.7]

[71] Götz M, Nath S, Gehrke J. MaskIt: Privately releasing user context streams for personalized mobile applications. In: Proc. of the 2012 ACM SIGMOD Int'l Conf. on Management of Data. Scottsdale, 2012. 289−300. [doi: 10.1145/2213836.2213870]

[72] Mannini A, Sabatini AM. Accelerometry-Based classification of human activities using Markov modeling. Computational Intelligence and Neuroscience, 2011,15(11):1−10. [doi: 10.1155/2011/647858]

[73] Arasu A, Götz M, Kaushik R. On active learning of record matching packages. In: Proc. of the 2010 ACM SIGMOD Int'l Conf. on Management of Data. Indianapolis, 2010. 783−794. [doi: 10.1145/1807167.1807252]

[74] Goetz M, Nath S, Gehrke J. MaskIt: Privately releasing user context streams for personalized mobile applications. Technical Report, MSR-TR-2012-29, Microsoft Research, 2012. [doi: 10.1145/2213836.2213870]

[75] Parate A, Chiu MC, Ganesan D, Marlin BM. Leveraging graphical models to improve accuracy and reduce privacy risks of mobile sensing. In: Proc. of the 11th Annual Int'l Conf. on Mobile System, Applications, and Services. Taipei, 2013. 83−96. [doi: 10.1145/2462456.2464457]

[76] Kuenzer A, Schlick C, Ohmann F, Schmidt L, Luczak H. An empirical study of dynamic bayesian networks for user modeling. 2001. http://www.research.rutgers.edu/~sofmac/ml4um/mirrors/ml4um-2001/papers/AK.pdf

[77] Chor B, Goldreich O, Kushilevitz E, Sudan M. Private information retrieval. Journal of the ACM, 1998,45(6):965−981. [doi: 10.1145/293347.293350]

[78] Chor B, Goldreich O, Kushilevitz E, Sudan M. Private information retrieval. In: Proc. of the 36th Annual Symp. on Foundations of Computer Science. Piscataway: IEEE, 1995. 41−50.

[79] Kushilevitz E, Ostrovsky R. Replication is not needed: Single database, computationally-private information retrieval. In: Proc. of the 38th Annual Symp. on Foundations of Computer Science. Miami Beach, 1997. 364−373. [doi: 10.1109/SFCS.1997.646125]

[80] Goldreich O, Goldwasser S, Micali S. How to construct random functions. Journal of the ACM, 1986,33(4):792−807. [doi: 10.1145/6490.6503]

[81] Goldreich O, Ostrovsky R. Software protection and simulation on oblivious RAMs. Journal of the ACM, 1996,43(3):431−473. [doi: 10.1145/233551.233553]

[82] Mouratidis K, Yiu ML. Shortest path computation with no information leakage. Proc. of the VLDB Endowment, 2012,5(8):692−703.

[83] Dijkstra EW. A note on two problems in connexion with graphs. Numerische Mathematik, 1959,1(1):269−271. [doi: 10.1007/BF01386390]

[84] Hart PE, Nilsson NJ, Raphael B. A formal basis for the heuristic determination of minimum cost paths. IEEE Trans. on Systems Science and Cybernetics, 1968,4(2):100−107. [doi: 10.1109/TSSC.1968.300136]

[85] Ghinita G, Kalnis P, Khoshgozaran A, Shahabi C, Tan KL. Private queries in location based services: Anonymizers are not necessary. In: Proc. of the 2008 ACM SIGMOD Int'l Conf. on Management of Data. Vancouver, 2008. 121−132. [doi: 10.1145/1376616.1376631]

[86] Papadopoulos S, Bakiras S, Papadias D. Nearest neighbor search with strong location privacy. Proc. of the VLDB Endowment, 2010,3(1-2):619−629.

[87] Berg MD, Cheong O, Kreveld MV, Overmars M. Computational Geometry. Berlin, Heidelberg: Springer-Verlag, 2008. 147−170.

[88] Khoshgozaran A, Shahabi C, Shirani-Mehr H. Location privacy: Going beyond *K*-anonymity, cloaking and anonymizers. Knowledge and Information Systems, 2011,26(3):435−465. [doi: 10.1007/s10115-010-0286-z]

[89] Shannon CE. Communication theory of secrecy systems. Bell System Technical Journal, 1949,28(4):656−715. [doi: 10.1002/j.1538-7305.1949.tb00928.x]

[90] Netflix prize rules. http://www.netflixprize.com//rules

[91] Cheng Y, Park J, Sandhu R. Preserving user privacy from third-party applications in online social networks. In: Leslie C, *ed al*., eds. Proc. of the 22nd Int'l Conf. on World Wide Web Companion. New York: ACM Press, 2013. 723−728.

[92] Huo Z, Meng XF, Huang Y. PrivateCheckIn: Trajectory privacy-preserving check-in services in MSNS. Chinese Journal of Computers, 2013,36(4):716−726 (in Chinese with English abstract).

[93] Huo Z, Meng XF, Zhang R. Feel free to check-in: Privacy alert against hidden location inference attacks in GeoSNs. In: Proc. of the 18th Int'l Conf. on Database Systems for Advanced Applications. Wuhan, 2013. 377−391. [doi: 10.1007/978-3-642-37487-6_29]

[94] Yang Y, Zhang Z, Miklau G, Winslett M, Xiao XK. Differential privacy in data publication and analysis. In: Proc. of the 2012 ACM SIGMOD Int'l Conf. on Management of Data. Scottsdale, 2012. 601−606. [doi: 10.1145/2213836.2213910]

[95] Dwork C. Differential privacy. In: Bugliesi M, ed. Proc. of the Automata, Languages and Programming. Berlin, Heidelberg: Springer-Verlag, 2006. 1−12.

[96] McSherry F, Talwar K. Mechanism design via differential privacy. In: Proc. of the 48th Annual IEEE Symp. on Foundations of Computer Science. Providence, 2007. 94−103. [doi: 10.1109/FOCS.2007.41]

[97] Ghosh A, Roughgarden T, Sundararajan M. Universally utility-maximizing privacy mechanisms. SIAM Journal on Computing, 2012,41(6):1673−1693. [doi: 10.1137/09076828X]

附中文参考文献:

[34] 周水庚,李丰,陶宇飞,肖小奎.面向数据库应用的隐私保护研究综述.计算机学报,2009,32(5):847−861.

[54] 霍峥,孟小峰.轨迹隐私保护技术研究.计算机学报,2011,34(10):1820−1830.

[92] 霍峥,孟小峰,黄毅.PrivateCheckIn:一种移动社交网络中的轨迹隐私保护方法.计算机学报,2013,36(4):716−726.

王璐(1986－),女,河北邢台人,博士生,主要研究领域为位置隐私保护.
E-mail: luwang@ruc.edu.cn

孟小峰(1964－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为 Web 数据管理,移动数据管理,XML 数据管理,云数据管理.
E-mail: xfmeng@ruc.edu.cn