

# 移动环境下的位置隐私

潘晓 肖珍 孟小峰

随着计算技术和无线通讯技术的发展与结合,一种全新的计算模式——移动计算模式逐渐产生并得到了迅速发展,给信息产业带来了一场深刻的变革,也给广大普通百姓带来了新鲜的感受。在未来的信息社会中,用户的移动性必将成为一大特性,移动对象/用户将成为移动计算环境下的运行主体。另一方面,传感器和手机,掌上电脑(Personal Digital Assistant,简称PDA)等手持无线设备的功能逐步变得强大,价格却在不断降低,它们具有体积小易于携带和通讯便捷的优点,在移动对象/用户中越来越普及。移动计算技术和无线设备的结合使得随时随地获得个人精确位置成为可能,也促进了新一类应用程序——基于位置的服务(Location Based Service, LBS)的产生和发展。简单而言,LBS是由服务内容提供商提供的基于用户位置的增值服务[1]。比如,基于位置的紧急救援服务(如查询“离我最近医院”等)、基于位置的信息娱乐服务(如查询“距离我五米内最近的饭店/电影院”等)和基于位置的广告服务(如“向所有在我咖啡店10m范围内的客人发送优惠券”等)等等。基于位置的服务有着极为广泛的用途。2003年,CSTB(Computer Science and Telecommunications Board)在“IT Roadmap to a Geospatial Future”中指出,基于位置的服务将会成为未来计算环境中非常重要的一部分,并渗入到未来生活的各个方面。市场研究公司ABI Research预测在2011年,全球享受位置服务的人数将由2006年的1.2千万增长到31.5千万。

## 1. 什么是移动环境下的位置隐私?

在人们享受各种位置服务的同时,移动对象个人信息泄露的隐私威胁也渐渐成为一个严重的问题,引起了研究者的注意。[2][3]报道了某人利用GPS跟踪前女友、[4]公司利用带有GPS的手机追踪监视本公司雇员行踪等等的案例。越来越多的事实说明了移动对象在移动环境下使用位置服务可能导致自己随时随地被人跟踪,被人获知曾经去过哪里、做过什么或者即将去哪里、正在做什么,换句话说,人们的隐私和安全受到了威胁。位置隐私[6]是一种特殊的信息隐私。**信息隐私**[7]是由个人、组织或机构定义的何时、何地、用何种方式与他人共享信息,以及共享信息的内容。而**位置隐私**则指的是防止其他人以任何方式获知对象过去、现在的位置。在基于位置的服务中,**敏感数据**[8]可以是有关用户的时空信息、可以是查询请求内容中涉及医疗或金融的信息、可以是推断出的用户的运动模式(如经常走的道路以及经过频率)、用户的兴趣爱好(如喜欢去哪个商店、哪种俱乐部、哪个诊所等等)等的个人隐私信息。而**位置隐私威胁**[9]是指攻击者在未经授权的情况下,通过定位位置传输设备、窃听位置信息传输通道等方式访问到原始的位置数据,并计算推理获取的与位置信息相关的个人隐私信息。比如,通过获取的位置信息可以向用户散播恶意广告,获知用户的医疗条件、生活方式或是政治观点。也可以通过用户访问过的地点推知用户去过哪所医院看病、在哪个娱乐中心消遣等等。位置隐私泄露的途径有三种[9]:第一,直接交流(Direct Communication),指攻击者从位置设备或者从位置服务器中直接获取用户的位置信息;第二,观察(Observation),指攻击者通过观察被攻击者行为直接获取位置信息;第三,连接泄露(Link Attack),指攻击者可以通过“位置”连接外部的数据源(或者背景知识)从而确定在该位置或者发送该消息的用户。

## 2. 位置隐私保护面临的挑战

在移动环境中,由于位置信息的特殊性及其对高质量的位置服务的需求,位置隐私保护技术面临的主要挑战为:

第一,保护位置隐私与享受服务是一对矛盾。移动环境下用户使用基于位置的服务时,需要发送自己的当前位置信息,位置信息越精确,服务质量越高,隐私度却越低,位置隐私和服务质量之间的平衡是一个难处理却又必须考虑的问题。这里考虑的服务质量包含响应时间,通讯代价等等,与具体的环境有关。

第二,位置信息的多维性特点。在移动环境下,移动对象的位置信息是多维的,每一维之间互相影响,无

法单独处理。这时采用的隐私保护技术，必须把位置信息看作一个整体，在一个多维的空间中，处理每一个位置信息。其中的处理包括存储，索引，查询处理等技术。

第三，位置匿名的即时性特点。在移动环境下，通常处理器面临着大量移动对象连续的服务请求以及连续改变的位置信息，使得匿名处理的数据量巨大而且频繁的变化。在这种在线（Online）的环境下，处理器的性能即匿名处理的效率是一个重要的影响因素，响应时间也是用户的满意度的一个重要衡量标准。其次，位置隐私还要考虑对用户的连续位置保护的问题，或者说对用户的轨迹提供保护，而不仅仅处理当前的单一位置信息。因为攻击者有可能积累用户的历史信息来分析用户的隐私。

第四，基于位置匿名的查询处理。在移动环境中，用户提出基于位置的服务请求。每一个移动对象不但关注个人位置隐私是否受到保护，同时还关心服务请求的查询响应质量。服务提供商根据用户提供的位置信息进行查询处理并把结果返回给用户。经过匿名处理的位置信息，通常是对精确的位置点进行模糊化处理后的位置区域。这样的位置信息传送给服务提供商进行查询处理时，得到的查询结果跟精确的位置点的查询结果是不一样的。如何找到合适的查询结果集，使得真实的查询结果被包含在里面，同时也没有浪费通讯代价和计算代价，是匿名成功之后需要处理的主要问题。

第五，位置隐私需求个性化。隐私保护的程度问题并不是一个技术问题，而属于个人事件。不同的用户具有不同的隐私需求，即使相同的用户在不同的时间和地点隐私需求也不同。例如用户在休闲娱乐时（比如逛街）隐私度要求比较低，但是在看病或参与政治金融相关的活动时隐私度比较高。所以，技术不能迫使社会大众共同接受一个最小的隐私标准。

### 3. 位置隐私保护技术

在位置隐私保护中主要有两方面的工作：第一，位置匿名（Location Anonymization）。匿名[6]指的是一种状态，这种状态下很多对象组成一个集合，从集合外向集合里看，组成集合的各个对象无法区别，这个集合称为匿名集。位置匿名是指[9]系统能够保证无法将某一个位置信息通过推理攻击的方式与确切的个人/组织/机构相匹配。在 LBS 中的位置匿名处理要求经过某种手段处理用户的位置，使得个体位置无法识别从而起到保护用户位置的目的。第二，查询处理。感知位置隐私的 LBS 系统中，位置信息经过匿名处理不再是用户的真实位置，可能是多个位置的集合也可能是一个模糊化（Obfuscation）的位置。所以，在位置服务器端，查询处理器的处理无法继续采用传统移动对象数据库中的查询处理方式，因为后者的技术均以确切的位置信息为基础。可以在原有技术的基础上进行改进和修改，从而使其适应新的查询处理要求。

#### 3.1 系统结构

在对移动对象的基于位置的服务请求进行响应时，必须首先确定所采用的系统结构。位置匿名系统的结构有三种：独立结构（Non-cooperative Architecture）、中心服务器结构（Centralized Architecture）和分布式的点对点结构（Peer-to-peer Architecture）。独立结构中用户仅利用自己的知识、由客户端自身完成位置匿名的工作，从而达到保护位置隐私的目的；中心服务器结构在独立结构的基础上，增加了一个可信第三方中间件，由可信的中间件负责收集位置信息、对位置更新做出响应、并负责为每个用户提供位置匿名保护；分布式点对点系统结构是移动用户与位置服务器的两端结构，移动用户之间需要相互信任协作从而寻找合适的匿名空间。现在大部分的工作集中在中心服务器结构和分布式点对点结构。

##### 1) 独立结构

独立结构[11]是仅有客户端（或者移动用户）与位置数据库服务器的 C/S 结构。该系统结构假设移动用户拥有能够自定位并具有强大的计算能力和存储能力的设备（比如 PDA）。移动用户根据自身的隐私需求，利用自己的位置完成位置匿名。

在此结构中一个查询请求的处理流程是：将匿名后的位置连带查询一起发送给位置数据库服务器；位置服务器根据匿名的位置，进行查询处理给出候选结果集返回给用户；用户知道自身的真实位置，所以可以根据真实位置挑选出真正的结果，换句话说，由用户自身完成查询结果的求精。总之，客户端需要自己完成位置匿名和查询结果求精的工作。

独立结构的优点是简单，容易与其他技术结合。但是它的缺点是对客户端的要求比较高。并且，它只利用自身的知识进行匿名，无法利用周边环境其他用户的位置等信息，所以比较容易受到攻击者的攻击。例如，

[1]中客户端降低空间粒度，生成了一个满足用户需求的匿名框，但是不幸的是如果在此匿名框中只有移动用户自身，那么任何从此匿名框处提出的查询都可以推断是由此移动用户提出的，查询内容与用户标识容易实现匹配，查询隐私泄露。

## 2) 中心服务器结构

中心服务器结构[12,13,22,25]除包含用户、基于位置的数据库服务器外，在二者之间加入了第三方可信中间件，称之为位置匿名服务器，其作用是：

接受位置信息：收集移动用户确切的位置信息，并响应每一个移动用户的位置更新。

匿名处理：将确切的位置信息转换为匿名区域。

查询结果求精：从位置数据库服务器返回的候选结果中，选择正确的查询结果返回给相应的移动用户。

之所以在用户与位置服务器之间加入可信的中间件，是因为我们无法确定位置数据库服务器是可信的，所以我们可以称其为半可信的[5]。不可信是因为会有一些不负责任的服务提供商出于商业目的将他所收集的位置记录卖给第三方。这样，攻击者可以锁定一些攻击对象，通过买来的数据获取这些对象历史所到之处，并推断未来的位置。而半可信是指，位置服务器会按照匿名框或者用户的真实位置确切无误的计算出查询结果。

在中心服务器结构中一个查询请求的处理过程如下：

1. 发送请求：用户发送包含精确位置的查询请求给位置匿名服务器。

2. 匿名：匿名服务器使用某种匿名算法完成位置匿名后，将匿名后的请求发送给提供位置服务的数据库服务器。

3. 查询：基于位置的数据库服务器根据匿名区域进行查询处理，并将查询结果的候选集返回给位置匿名服务器。

4. 求精：位置匿名服务器从候选结果集中挑出真正的结果返回给移动用户。

中心服务器结构的优点在于降低了客户端的负担，在保证高质量服务的情况下提供符合用户隐私需求的匿名服务。但是其缺点也很明显：

第一，位置匿名服务器是系统的处理瓶颈。移动用户位置频繁的变化，位置匿名服务器需要负责所有用户的位置收集，匿名处理以及查询结果求精。所以它的处理速度将直接影响到整个系统。如果位置匿名服务器出了什么问题，则将会导致整个系统瘫痪。

第二，当位置匿名服务器也变得不再可信的时候，如受到攻击者的攻击，由于它掌握了移动用户的所有知识，所以将会导致极其严重的隐私泄露。

## 3) 分布式点对点结构

分布式点对点系统结构由两部分组成：移动用户和位置数据库服务器。每个移动用户都具有计算能力和存储能力，它们之间相互信任合作。位置数据库服务器与其他两种系统结构中的作用一样，都是提供基于位置的服务。

分布式点对点结构与中心服务器结构的区别在于中心服务器结构中的第三方可信中间件需要负责位置匿名和查询结果求精等工作，而分布式点对点结构中每个节点都可以完成该工作，节点之间具有平等性。所以这将避免中心服务器结构中位置匿名服务器是处理瓶颈和易受攻击等缺点。与独立结构相比，表面上看两者都是两端结构，但是不同点在于独立结构中，移动用户仅利用自己的位置做匿名，并不考虑其他移动用户的信息。在分布式结构中，移动用户根据匿名算法找到其他一些移动用户组成一个匿名组（Group），利用组中的成员位置进行位置匿名。匿名处理过程可以由提出查询的用户本身完成也可以由从组中选出的头结点完成。查询结果返回给头结点，头结点可以选择出真实结果发送给提出查询的用户，也可以将查询结果的候选集发送给用户，由用户自己挑选出真实的结果。所以在分布式点对点结构中，除与其他两种结构相同的位置匿名处理和查询处理任务外，另一个重要任务就是选择头结点（Head），平衡网络负载。具有代表性的工作有 Group Formation[14]和 PRIVÉ[15]等。

## 3.2 位置隐私保护模型

在所有的系统结构下，位置隐私保护技术都需要定义一个合适的位置匿名模型，使得该模型既能够保证用户的隐私需求，又能够最好的响应用户的服务请求。

迄今为止，在位置匿名处理中，使用最多的模型是位置 k-匿名模型（Location K-Anonymity Model）。k-匿名模型[19]由美国 Carnegie Mellon 大学的 Latanya Sweeney 提出，最早使用在关系数据库的数据发布隐私保

护中[20,21],它指一条数据表示的个人信息和至少其他  $k-1$  条数据不能区分。其主要目的是为了解决如何在保证数据可用的前提下,发布带有隐私信息的数据,使得每一条记录无法与确定的个人匹配。

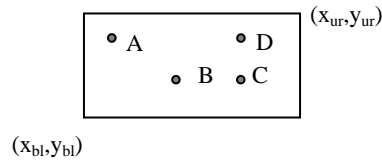


图 1 位置 4-匿名

Marco Gruteser[12]最先将k-匿名的概念应用到位置隐私上来,提出位置k-匿名 (Location K-Anonymity): 当一个移动用户的位置无法与其他  $(k-1)$  个用户的位置相区别时,称此位置满足位置k-匿名。通常采用的技术是把用户的真实位置点扩大为一个模糊的位置范围,使得该范围覆盖了k个用户的位置,从而隐藏了真实用户的位置。形式化来说,每一个用户的位置以一个三元组表示 $([x1, x2], [y1, y2], [t1, t2])$ ,其中 $([x1, x2], [y1, y2])$ 描述了对象所在的二维空间区域, $[t1, t2]$ 表示一个时间段。 $([x1, x2], [y1, y2], [t1, t2])$ 表示用户在这个时间段的某一个时间点出现在 $([x1, x2], [y1, y2])$ 所表示的二维空间中的某一点。除此用户外,还有其他至少  $(k-1)$  个用户也在此时间段内的某个时间出现在 $([x1, x2], [y1, y2])$ 的二维空间的某一点,这样的用户集合满足位置k-匿名。如图 1 是一个  $k=4$  的位置k-匿名的例子(为了叙述的方便,这里省掉了时间域)。A、B、C和D在经过位置匿名后,均用 $([x_{bl}, x_{ur}], [y_{bl}, y_{ur}])$ 表示,如表 1,其中 $(x_{bl}, y_{bl})$ 是匿名矩形框的左下角坐标, $(x_{ur}, y_{ur})$ 是匿名矩形框的右上角坐标。这样,攻击者只知道在此区域中有四个用户,具体哪个用户在哪个位置他无法确定,因为用户在匿名框中任何一个位置出现的概率相同,所以在位置k-匿名模型中,匿名集由在同一个匿名框中出现的所有用户组成,所以图 1 的匿名集为{A、B、C、D}。一般情况下,k值越大,匿名度越高。所以,以匿名集的大小表示匿名度。

表 1 位置 4-匿名

用户	真实位置	匿名后的位置
A	$(x_A, y_A)$	$([x_{bl}, x_{ur}], [y_{bl}, y_{ur}])$
B	$(x_B, y_B)$	$([x_{bl}, x_{ur}], [y_{bl}, y_{ur}])$
C	$(x_C, y_C)$	$([x_{bl}, x_{ur}], [y_{bl}, y_{ur}])$
D	$(x_D, y_D)$	$([x_{bl}, x_{ur}], [y_{bl}, y_{ur}])$

一般情况下,k值越大,匿名框也越大,但是这也与用户提出服务的所在位置的周围环境有关。假设提出查询请求的用户要求  $k=100$  的匿名度,如果此时用户正在一个招聘会上,一个很小的空间即可满足用户的需求,但如果用户此时在沙漠中,则返回的匿名空间可能非常的大。如何证明位置匿名模型的正确性。

这里的  $K$  和匿名框的大小都是衡量隐私保护性能的参数,也是用户用于表达自己对隐私保护和服务质量的要求。通常,移动对象的位置隐私需求可以用四个参数来表示:

$k$ : 即 k-匿名,用户要求返回的匿名集中至少包含的用户数。

$A_{min}$ : 匿名空间的最小值,即返回的匿名空间必须要超过此值,可以是面积或半径等。 $A_{min}$ 作用是为了防止在用户密集区,很小的空间区域即可满足用户k值的需求。极端情况下,在一个位置L上有k个用户,虽然满足k值的需求,但是位置还是暴露了。

$A_{max}$ : 匿名空间的最大值,即返回的匿名空间必须不能超过此值,也可以是面积或半径等。

$T_{max}$ : 可容忍的最长匿名延迟时间。即从用户提出请求的时刻起需要在 $T_{max}$ 的时间范围内完成用户的匿名。

$k$ 和 $A_{min}$ 是用户的位置匿名限制 (Location Anonymization Constraints),反映的是匿名质量的最小值; $A_{max}$ 和 $T_{max}$ 是位置服务质量限制 (Location Service Quality Constraints),反映的是最差服务质量。

### 3.3 位置匿名技术

#### 3.3.1 位置匿名算法

在位置隐私保护模型下,需要找到一个高效的位置匿名算法,使得既满足用户隐私需求又保证服务质量。首先,位置服务中的查询请求可以形式化为:  $(id, loc, query)$

其中,  $id$  表示提出位置服务请求的用户标识,  $loc$  表示提出位置服务时用户所在的位置坐标 $(x,y)$ ,  $query$  表示查询内容。举例而言,张某利用自己带有 GPS 的手机提出“寻找距离我现在所在位置最近的中国银行”,

则  $id=$  “张某”， $loc=$  “某医院地址”， $query=$  “距离我最近的中国银行”。

位置隐私保护主要目的[5]是防止/减少在服务提供系统中位置信息的可识别性。最早的方法是使用假名，即将此查询先提交给一个匿名服务器，将真实的唯一标识用户的  $id$  隐藏，换成假名  $id'$ 。这样攻击者即无法知道在此位置上的用户是谁，此查询是由谁提出的。此时查询三元组变为： $(id', loc, query)$ ，其中， $id'$  是用户的假名。

然而，不幸的是即使使用假名技术，位置信息  $loc$  也有可能引起位置隐私泄露。众所周知，Web 服务器会记录请求服务的 URL 和提出请求的 IP 地址。与 Web 服务器类似，位置服务器也以日志的形式记录自己收集到的所有服务请求。所以，在日志中包含的位置信息为攻击者提供了一扇方便之门。文献[12]中将位置作为媒介实现消息内容与用户匹配的隐私威胁分为两类：受限空间识别（Restricted Space Identification）和观察识别（Observation Identification）。例如，一个对象发送消息  $M$ ，其中包含了位置  $L$ 。攻击者  $A$  得到了此条消息，则他可以通过位置信息  $L$  确定消息  $M$  的发送者。受限空间识别是指如果攻击者  $A$  知道地点  $L$  是专属于用户  $S$  的，则任何从  $L$  发送的查询一定是由  $S$  发出的。比如，某别墅的主人在其家中发送了某条消息。可以通过消息中确切的位置  $(x, y)$  利用外部知识从而确定此别墅的主人。这样，攻击者即可确定这个用户发送了哪些查询。观察识别是通过一些外部观察知识实现用户标识和查询内容的匹配。如攻击者  $A$  之前被告知（或通过观察获知）时刻  $t$  对象  $S$  在位置  $L$  上，而攻击者又发现在时刻  $t$  从位置  $L$  发出的查询都来自同一人，则可以认为任何从  $L$  发送的消息  $M$  都是由  $A$  发出的。例如，一个对象在上一个消息中揭示了其标识与位置，那么在同一个位置上即使匿名了后面的消息，攻击者仍然可以通过消息中的位置识别出后来消息的来源。

由此可见，仅仅隐藏用户标识是不够的，需要将用户的位置也作一定的匿名处理，从而保护位置隐私。这正是近年来位置匿名研究的焦点，出现了一系列具有代表性的方法[9,10, 11, 13,16, 17,18]。迄今为止，广泛使用的位置匿名基本思想有三种：

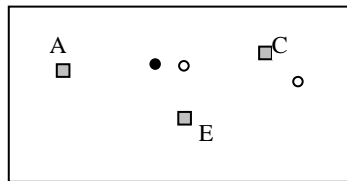


图 2 假数据示意图

第一，发布假位置，即不发布真实服务请求的位置，而是发布假位置，即哑元(Dummy)。如图 2 所示，圆点是查询点，方块是被查询对象。其中黑色的点是真实的位置点，为了保护用户的位置，发送给位置数据库服务器的是白色的假位置。由此可见，位置隐私就通过报告假位置而获得了保护，攻击者并不知道用户的真实位置。隐私保护程度和服务质量与假位置和真实位置的距离有关。假位置距离真实位置越远，服务的质量越差，但隐私保护程度越高；相反地，距离越近，服务的质量就比较好，但是隐私保护程度则比较低。

第二，空间匿名 (Spatial Cloaking)。本质上是降低对象的空间粒度，即用一个空间区域来表示用户的真实的精确位置。区域的形状不限，可以是任意形状的凸多边形，现在普遍使用的是圆和矩形，称这个匿名的区域为匿名框。如图 3 所示

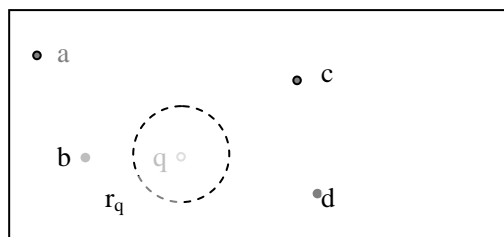


图 3 空间匿名示意图

用户  $q$  的真实位置点的坐标是  $(x, y)$ ，空间匿名的思想是将此点扩充为一个区域如图 3 中的虚线圆  $r_q$ ，即用这个区域表示一个位置，并且用户在此区域内每一个位置出现的概率相同。这样攻击者仅能知道用户在这个空间区域内，但是却无法确定是在整个区域内的哪个具体位置。

第三，时空匿名 (Spatio-Temporal Cloaking)。在空间匿名的基础上，增加一个时间轴。在扩大位置区域的同时，延迟响应时间，如图 4 所示。

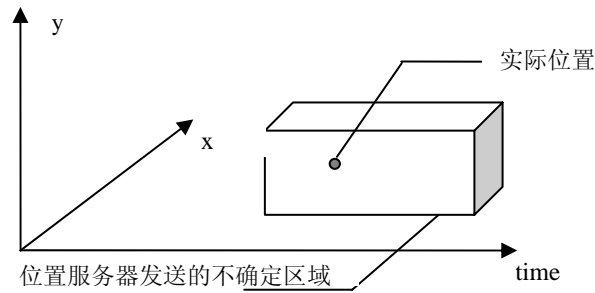


图 4 时空匿名示意图

延迟响应时间，可以在这段时间中出现更多的用户、提出更多的查询，隐私匿名度更高。与空间匿名相同的是，在时空匿名区域中，对象在任何位置出现的概率相同。

注意，无论是空间匿名还是时空匿名，匿名框的大小从一个侧面表示了匿名程度。匿名框越大可能覆盖的用户数就越多，匿名的效果可能越好，但是查询处理代价增加即服务质量降低；相反的，匿名框越小，匿名的程度就可能越低，服务质量就比较高，极端情况下匿名框缩小为一个确切的点，位置隐私泄露。以空间匿名为例。如图 3，用户查询“距离我最近的点”。传统的最近邻查询使用真实的位置点 $q$ ，返回给用户真实的查询结果 $b$ 。但是，在匿名的情况下，位置服务器只能返回距离此查询区域 $r_q$ 最近的对象集合 $\{b, c, d\}$ 。此集合是查询结果的候选集，也就是说，位置服务器在不知道用户真实位置的情况下，此集合中的任何一个对象都有可能成为真实的查询结果，它们是距离此匿名区域中某一个点最近的对象。所以，此后需要根据用户的真实位置对候选结果集求精，这个工作可以由用户完成，也可以由匿名服务器完成，这取决于系统结构。但是可以确定的是，匿名区域越大，候选集就越大，求精处理和传输代价就越高。所以，匿名区域的建立需要在隐私保护与服务质量之间寻求一个平衡点。所以空间/时空匿名算法最大的挑战[5]就是在满足用户隐私需求的前提下，如何高效的寻找最优的空间/时空匿名框。

### 3.3.2 算法评价标准

评价匿名算法的好坏可以至少从以下四个方面进行考虑[5,9, 18, 24, 25]:

#### (1) 匿名成功率 (Success Rate)

匿名成功率是评价一个位置匿名算法有效性的重要指标之一。匿名成功率越高，匿名算法就越好。匿名成功率可以定义为：成功匿名的消息数在所有移动用户提出的匿名请求的消息数中所占的比例，形式化的表示为

$$SR = \frac{|S'|}{S}$$

其中  $S$  是提出匿名请求的所有消息的集合， $S'$  是  $S$  的子集，是成功匿名的消息集合。

#### (2) 相对匿名度 (Relative Anonymity Level)

相对匿名度是指任意一条消息 $m_s$ ，它被匿名处理后所在的匿名集合的秩和 $m_s$ 所要求的匿名度 $k$ 的比值。相对匿名度的值一定大于等于 1。一般情况下，我们认为较高的相对匿名度意味着更好的匿名效果，即消息 $m_s$ 与更多的消息匿名在一起。但实际情况并不一定总是如此，因为由于匿名与服务质量之间的平衡问题，较高的相对匿名度可能意味着较高的查询代价。

#### (3) 相对空间粒度 (Relative Spatial Resolution)

相对空间粒度是表示匿名算法获得的匿名空间粒度的一个参数，其定义为匿名请求所定义的可容忍的最大匿名空间与匿名算法所获得的匿名空间的比。相对空间粒度越大，说明在满足隐私需求的前提下，匿名空间越小，更接近最优解。所以，相对空间粒度越大越好。

#### (4) 相对时间粒度 (Relative Temporal Resolution)

相对时间粒度与相对空间粒度类似，它表示的是匿名算法时间粒度的一个参数。其定义为匿名请求所定义的可容忍的最大匿名延迟时间与匿名框中时间维长度的比。相对时间粒度越大，说明在满足隐私需求的前提下，较短的时间范围内完成了匿名，在时间轴上更接近最优解。所以，相对时间粒度越大越好。相对时间粒度和相对空间粒度均大于等于 1。

相对匿名度、相对空间粒度和相对时间粒度反映的是服务质量。

#### (5) 消息处理时间 (Message Processing Time)

消息处理时间反映的是匿名算法的运行效率，它指的是一定规模移动用户的所有查询请求在多长时间可

以得到匿名处理。这是反映匿名算法好坏的重要指标之一。当然，处理时间越短越好，说明了匿名算法的高效性。

### 3.4 查询处理

在传统的移动对象数据库中，基于位置的查询目标分为两种：一种是移动对象（如汽车、移动用户等），一种是静态空间对象（如旅馆、医院等）。对移动对象的查询处理已经有大量的研究成果出现，但由于传统查询处理中查询对象都是一个位置点，而经过匿名处理之后的查询对象变成了一个匿名区域，所以我们需要改造已有的方法或者开发出新的查询处理方法，并解决几种典型查询类型。

带有隐私保护的查询处理是在基于位置的数据库服务器端完成的，查询处理技术与传统移动对象数据库中查询处理技术既有相同点又相互区别。根据匿名技术的不同，查询处理方式也分为两种：

第一，如果用户的位置匿名技术采用的是假数据，则移动对象数据库中的查询处理器无需作任何修改。因为发送给基于位置数据库服务器的是一些精确的位置点，可以直接利用已有的移动对象数据库中的查询处理技术完成各种查询。只是返回的结果不一定是真实结果。最坏情况下，查询结果误差为  $2d$ ，平均误差为  $d$ ，其中  $d$  是假数据与真实数据之间的距离[26]。

第二，如果位置匿名技术采用的是空间匿名或时空匿名，嵌入在基于位置的数据库服务器端的查询处理器需要做一定的修改。因为此时发送给位置服务器的位置不再是精确的位置点，而是一个匿名框，一个范围。匿名框内的用户在框内的每一个位置出现的概率相同，所以查询处理器无法获知移动用户在此范围内的确切位置。基于位置数据库服务器中的数据可以分为两种：公开数据（Public Data）和隐私数据（Private Data）。相应的，根据查询点和被查询点是否是隐私数据，可以将查询分为四种：基于公开数据的公开查询(Public Query over Public Data)、基于公开数据的隐私查询(Private Query over Public Data)、基于隐私数据的公开查询(Public Data over Private Data)和基于隐私数据的隐私查询(Private Query over Private Data)[22]。

基于公开数据的隐私查询是指查询点是隐私的，而被查询点是公开的。例如查询“离我最近的医院”、查询“在我前方 100 米的中国银行”等。查询点是一些带有隐私需求的用户，他们的位置不是一个精确的位置点，而是一个匿名框。被查询点是一些公开的公共信息，如加油站、医院、银行等公共设施的地址位置。

基于隐私数据的公开查询是指查询点的位置是公开的，而被查询点的位置是隐私的。这样的查询如“在中关村海龙 200 米内有多少车辆”、“距离我最近的请求急救的用户”等。在这类查询中，查询点是警察、医生等公众已知或用户愿意公开的位置，而被查询点是一些带有隐私保护需求的用户点。在这种情况下，查询处理器知道查询点的确切位置，而不知道被查询点的位置。

基于隐私数据的隐私查询是指查询点和被查询点的数据都带有隐私保护的需求。如查询“在我所在位置 100 米内，离我最近的朋友”。查询点“我”和被查询点“朋友”都带有一定的隐私需求，所以二者的位置都是一个匿名框，不是一个精确的位置点。所以，查询处理器不知道二者的确切位置。

基于公开位置的公开查询是指查询点和被查询点都是精确的位置信息，这是在移动对象数据库中处理的传统的查询。

无论哪一种查询都可以按照传统移动对象数据库中的查询分为：范围查询（Range Query）、最近邻查询（Nearest-Neighbor Queries）、聚集查询（Aggregate Queries）和连续查询（Continuous Query）等。

由于查询点和被查询点的位置不精确性，所以查询结果的表示有三种形式：

第一，在候选结果集中随机挑选一个对象作为结果返回给用户。很明显，返回的结果不一定是真实的。所以，真实结果与返回结果之间的距离表示了结果的精确度。匿名框越大，二者之间的距离就越大，所以随机挑选一个结果返回给用户，服务质量也随之减低。

第二，返回整个候选集。这种处理方式的缺点是：（1）匿名框越大，候选结果集就越大，传输与处理代价升高。（2）需要在客户端或者是第三方可信中间处根据真实位置或者应用语义进行结果求精，增加了移动用户或第三方中间件的负担。

第三，以概率查询处理技术[27]处理查询结果，也就是说，每一个候选结果都赋有一个表示是真实结果的概率。可以看出，这种方式是前两种方式的折中，为选取真实结果提供一定的依据，平衡降低的服务质量。

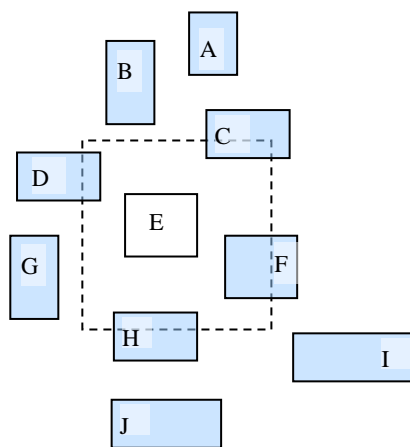


图 5 范围查询示意图

以一个基于隐私位置的公开范围查询为例，要查询在某一范围内出现的所有车辆。提出这个查询的是交通管理部门，在查询区域中的每一个感兴趣的对象（即车辆）涉及位置隐私，都不是一个精确的位置，而是一个空间匿名区域，如图 5 所示。A~J 是空间中的移动用户，虚线框是查询的范围。

最简单的方法，将所有与查询范围相交的匿名区域都作为候选集。如图 5 所有与虚线矩形相交的匿名框均为查询结果，即{C, D, E, F, H}。在匿名空间中对象均匀分布的前提下，可以根据匿名框与矩形框的重叠区域面积大小表示查询结果是真正结果的概率，查询结果可以表示为{(C, 0.3), (D, 0.2), (E, 1), (F, 0.6), (H, 0.4)}，这种查询结果可以进一步返回概率大于某一阈值的所有对象。例如，在上面的例子中，仅返回概率大于 50% 的查询结果，则结果集为{E, F}。当然，当匿名空间中，对象不再均匀分布，计算结果集的概率则比较麻烦。如果知道不同匿名空间的点在此空间的分布概率密度函数（Probabilistic Dense Function, PDF），则可以利用积分的方式计算概率。

## 4. 未来研究展望

位置隐私保护还是一个新兴的研究领域，很多具有挑战性的问题有待进一步解决：

### (1) 新的匿名模型

在位置匿名算法中，目前最广泛使用的模型是位置 k-匿名模型。位置 k-匿名模型是否够用或者说是否满足了位置 k-匿名模型就可以一劳永逸呢？回答是否定的。另一方面，现有的工作把位置隐私与查询隐私合二为一，不相互区分。即以前的工作都集中在三元组(id, loc, query)的 loc 位置匿名。如果满足了位置隐私，就认为查询隐私也满足了。现实的情况是否确实如此呢？答案也是否定的。

传统位置 k-匿名模型保证匿名框至少覆盖 k 个用户，即攻击者无法获知 k 个用户的真实位置。但是极端情况下，k 个用户位置都罗列在一个位置点上，或者都在同一个敏感区域（如医院），则位置隐私泄露。Ling Liu 在[24]中借鉴数据发布隐私处理中的 l-差异性（l-diversity）[28]的思想，提出需要保证位置 l-差异性的特性，l 表示的是查询位置的差异性。位置 l-差异性即要保证在一个匿名框中的用户除保证 k-匿名外，仍需要保证匿名框中包含 l 个不同的物理/实际位置。但是我们认为，除了位置 l-差异性外，也存在查询 l-差异性的情况。也就是说，即使用户的匿名框满足了位置 k-匿名，位置 l-差异性，但是 k 个用户提出的查询内容都是有关医院的信息，则攻击者可以推断出用户有访问医院的记录和患病的可能性。所以，我们提出查询 l-差异性的概念，即保证在匿名框中，需要保证有 l 种不同查询敏感度的查询。而如何有效的保证位置 l-差异性、查询 l-差异性至今仍是一个具有挑战性的开放性问题。另外，借鉴和参考数据发布中已有的其他隐私模型如 m-invariant 原则，结合位置隐私保护中的特点如位置连续更新，提出新的符合位置隐私保护下的匿名模型也是未来的一个研究方向。

### (2) 攻击者背景知识挖掘

现在的匿名算法中都是在各自的研究问题上拥有各自的假设，即不同的背景知识。设计的模型、算法也是在此背景知识的基础上设计出来的。所以，一旦用户拥有新的背景知识，即给出新的假设，则可能存在隐私泄露的现象，即被攻破。例如，已知数据分布情况、用户最大运动速度和存在连续查询，则存在异常点攻击模型、

最大速度运动速度攻击模型和连续查询攻击模型[22]。

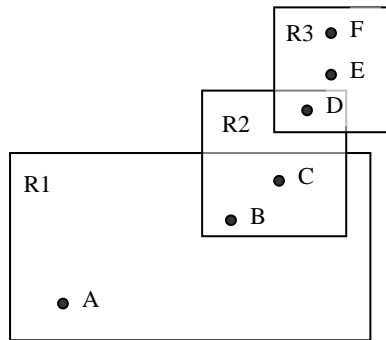


图 6 异常点攻击模型

异常点 (Outlier) 攻击模型是由于存在分布稀疏的数据点。实际生活中用户并非均匀分布，例如，大型购物中心是用户密集区，而在购物中心附近的小咖啡馆中却有寥寥可数的几个人，这是一个用户稀疏区。数据分布的不均匀性造成稀疏区域用户的匿名空间比密集区的匿名空间明显大很多，并且可能造成异常点隐私泄露。例如，图 6 中所示，A~F 是六个不同的查询用户。在  $k$ -匿名模型下 ( $k=3$ )，每个用户具有不同的匿名框。在用户 A 进入系统前，B、C 的匿名框均是 R2，覆盖 {B, C, D}；D、E、F 的匿名框均是 R3，覆盖 {D, E, F}。当用户 A 进入系统后，在  $k=3$  的情况下，用户 A 的匿名框将会是图中包含 B、C 的大矩形 R1。在最坏情况下，用户的位置已知，任何查询如果是从大矩形框中提出的，则一定是由用户 A 提出的，泄露查询隐私。

最大运动速度攻击模型存在的根本原因是攻击者可能知道移动用户的最大速度。现实生活中，这是可能的，比如所使用的车型或者所在公路的最大时速限制等。假设 (1) 考虑对象的连续位置更新；(2) 同一对象在不同时刻的不同位置发送请求的假名相同，即已知哪些请求是来自同一个用户，则如图 7 所示  $t_i$  时刻对象 A 的匿名框是  $R_{t_i}$ ，在  $t_{i+1}$  时刻的匿名框是  $R_{t_{i+1}}$ 。由于运动对象的最大速度已知，在可以计算出该对象在  $(t_{i+1}-t_i)$  这段时间可能运动到的范围 (Maximum Movement Boundary, MMB)，以后简称 MMB。在图 7 的例子中，A 的 MMB 是虚线圆。可以看到在 MMB 与  $R_{t_{i+1}}$  之间由一个交集，如图 7 的阴影所示，则攻击者可以知道对象在  $t_{i+1}$  时刻只可能存在于此。实际上，在  $t_{i+1}$  时刻的匿名空间变小了，可能已不满足用户的隐私需求，极端情况下相交区域缩小为一个点，则位置隐私泄露。

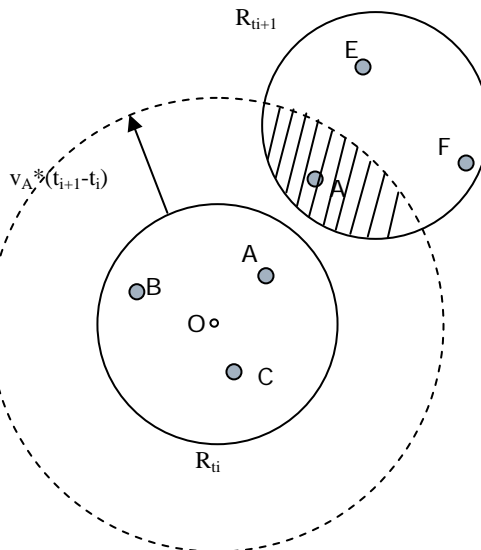


图 7 最大运动速度攻击模型

如果攻击者已知系统中存在一个连续查询，则攻击者可以通过结合各个不同时刻匿名框所包含的用户推知是哪个用户提出了这个连续查询，这种攻击模型可以称为连续查询攻击模型。例如，在图 8 中有 A~K 11 个用户。在时刻  $t_i$ ，用户 A 提出了一个连续查询，要求  $k$ -匿名模型  $k=5$ ，则匿名算法将 A~E 匿名到一个空间中，如图 8 (a)。此时攻击者只能推测从该区域提出的连续查询可能是 A~E 中的某一个人提出的，但是不能确定是哪一个人。时刻  $t_{i+1}$ ，用户均进行了位置更新，此时匿名算法为 A 生成的匿名框是包含的 A、B、F、G、H 的匿名空间，如图 8 (b)。将  $t_i$  和  $t_{i+1}$  时刻的用户集合取交，得 {A, B}，所以连续查询可能是由 A 提出的，也可能是由 B 请求的。在时刻  $t_{i+2}$ ，A、J、I、K、H 组成一个匿名框，则与 {A, B} 取交后，得 {A}，所以可知用户 A 提出了此连

续查询。

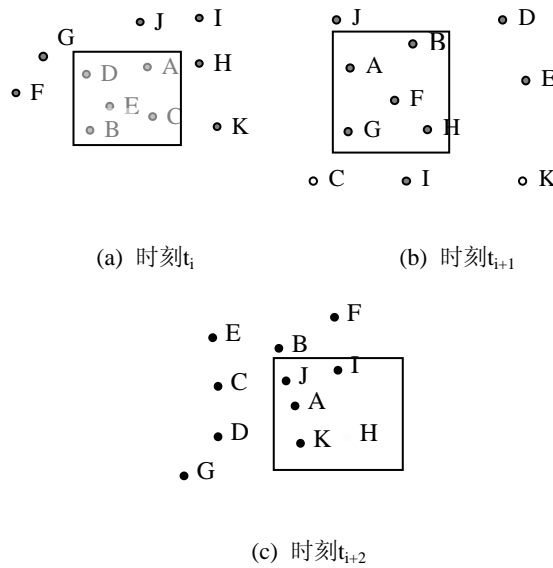


图8 连续查询攻击模型

可以看出，特定的匿名模型只在特定的背景知识下成立，每个匿名模型在有限的背景知识下成立。当攻击者具有新的背景知识的情况下，隐私可能泄露。所以，亟待设计一套可以量化背景知识的体系，量化的评价各匿名算法。这里涉及第一，考虑攻击者可能拥有的所有背景知识，这一点是非常难。第二，寻找某种量化工具量化背景知识，从而计算在可变背景知识下的某种隐私保护模型的隐私泄露率。

### (3) 最优解问题

目前已有的匿名工作中，都是以给出一种解为目的，并不关心解空间的最优性。

实质上，匿名空间解的最优性本质上是匿名质量与服务质量的一个平衡问题。从用户的隐私需求来看，反映匿名质量的是 $k$ 和 $A_{\min}$ ，反映服务质量的是 $A_{\max}$ 和 $T_{\max}$ 。现有的匿名空间解是用户定义三类参数中的某一中间值，即保证最低匿名质量的同时，不低于最差的服务质量。匿名空间中最优解可以在固定一个变量的同时，最大化另一个变量。例如，在固定匿名质量的情况下，获取最大的服务质量。反映在用户隐私需求上，即固定 $k$ -匿名模型中的 $k$ 值，获取最小的匿名空间（可以是周长也可以是面积）；或者在固定服务质量的同时，寻求好的匿名效果。反映在用户的隐私需求参数中，即固定匿名空间的面积或者周长，保证在匿名空间中覆盖最多的用户即最高的相对匿名度。解是否最优与定义的评价解得度量有关。

在位置匿名算法中求最优解问题是一个 NP 问题，所以寻找位置匿名的最优解将会是一个非常具有挑战性的工作。由于寻找最优解很难，在未来的研究中可以考虑寻找近似最优解，但是具体如何寻找和评价近似最优解仍是一个开放性的课题。

### (4) 查询处理

至今为止，大部分的隐私保护研究都集中在匿名模型、匿名算法的研究上。对于感知隐私的查询处理技术的研究工作相对较少。位置隐私保护中的查询处理在未来的研究中可以从查询类型、查询处理技术和查询结果的求精三个方面考虑。

现有的工作主要集中在最近邻(Nearest Neighbor, NN)查询、聚集查询(Aggregate Query)和范围查询(Range Query)上，在未来的研究工作中可以借鉴在移动对象数据库中已有查询处理技术，根据查询点与被查询点是否是隐私数据，分为四种情况(类似第六节所述)地进一步考虑解决更多的查询类型，如逆最近邻查询(Reverse Nearest Neighbor, RNN)、连续查询(Continuous Query)、空间匹配查询(Spatial Match Query)和 Skyline 查询(Skyline Query)等等。

在位置匿名中，通过发布某种形状的匿名框即用模糊的位置代替精确位置后，达到了隐私保护的目。这种位置的不确定性具有不确定数据(Uncertain Data)的特点，所以在未来的研究中可以借鉴在不确定数据上的概率查询中的模型、技术等解决在隐私保护中的查询处理问题。

在前面的讨论中，反复阐述过位置服务器返回的查询结果是候选结果集。在结果集中，并不是所有的对象都是真实的查询结果。现在普遍采用的方式是将候选结果集统一发送给客户端或位置匿名服务器。但是若以概率表示候选结果作为真实结果的可能性，则某候选对象的概率如果是 0.0005%，则该对象成为真正查询结果的

可能性非常的小。所以需要一定的方式从候选结果集中将其去除，从而求精。如何根据已知的条件，评价、求精候选集，以及查询结果的表示等都尚是一个开放的问题。

## 参考文献

- [1] Mokbel M F. Privacy in location-based services: start-of-the-art and research directions[C]// Proceeding of 8th International Conference on Mobile Data Management (MDM'07), Mannheim, Germany, May 2007
- [2] Man accused of stalking ex-girlfriend with GPS[OL]. Fox news.September,2004.  
<http://www.foxnews.com/story/0,2933,131487,00.html>.
- [3] Authorities: GPS system used to stalk woman[OL].USA Today.Dec.,2002.  
[http://www.usatoday.com/tech/news/2002-12-30-gps-stalker\\_x.htm](http://www.usatoday.com/tech/news/2002-12-30-gps-stalker_x.htm)
- [4] Michael Sciannamea. Companies increasingly use GPS-enable cell phones to track employees[IO]. Weblogsinc.September,2004.<http://wifi.weblogsinc.com/2004/09/24/companies-increasingly-use-gps-enabled-cell-phones-to-track/>
- [5] Gedik B, Liu L. Protecting location privacy with personalized k-anonymity: architecture and algorithms[J]. IEEE Transactions on Mobile Computing. (accepted, to appear).
- [6] Beresford A R , Stajano F. Location privacy in pervasive computing[J]. IEEE Pervasive Computing, 2003, 2(1):46–55.
- [7] Hong J I, Landay J A. An architecture for privacy-sensitive ubiquitous computing[C]// Proceedings of The International Conference on Mobile Systems, Applications, and Services(MobiSys'04), Boston, Massachusetts, USA, 2004:177–189.
- [8] Bettini C, Wang X S, Jajodia S. Protecting privacy against location-based personal identification[C]// Proceeding of the VLDB Workshop on Secure Data Management(SDM'05), 2005:185–199.
- [9] Liu L. From data privacy to location privacy: models and algorithms[C]//Proceeding of The 33rd International Conference on Very Large Data Bases(VLDB'07), Vienna, Austria, 2007:1429-1430.
- [10] Du J, Xu J, Hu H, et al. iPDA: supporting privacy-preserving location-based mobile services[C]// Proceeding of the 8th International Conference on Mobile Data Management (MDM'07), Mannheim, Germany, May 2007.
- [11] Cheng R, Zhang Y, Bertino E, et al. Preserving user location privacy in mobile data management infrastructures[C]//Proceedings of Privacy Enhancing Technology Workshop(PET'06), Cambridge, United Kingdom, 2006.
- [12] Gruteser M, Grunwald D. Anonymous usage of location-based services through spatial and temporal cloaking.[C]// Proceedings of the International Conference on Mobile Systems, Applications, and Services(MobiSys'03), San Francisco, USA, 2003:163–168.
- [13] Xiao Z, Meng X, Xu J. Quality-aware privacy protection for location-based services[C]//Proceedings of the International Conference on Database Systems for Advanced Applications(DASFAA'07), Bangkok, Thailand, April 2007.
- [14] Chow C, Mokbel M F, Liu X. A peer-to-peer spatial cloaking algorithm for anonymous location-based services[C]// Proceedings of the ACM Symposium on Advances in Geographic Information Systems(ACM GIS'06), Arlington, VA, November 2006:171-178
- [15] Ghinita G, Kalnis P, Skiadopoulos S. PRIVE: anonymous location based queries in distributed mobile systems[C]//Proceedings of International Conference on World Wide Web(WWW'07), Banff, Alberta, Canada, 2007:1–10.
- [16] Kido H, Yanagisawa Y, Satoh T. An anonymous communication technique using dummies for location-based services[C]// Proceedings of IEEE International Conference on Pervasive Services(ICPS'05), Santorini, Greece, 2005:88–97.
- [17] Ghinita G, Kalnis P, Skiadopoulos S. MOBIHIDE: A mobile peer-to-peer system for anonymous location-based queries[C]//Proceedings of the International Symposium on Advances in Spatial and Temporal Databases(SSTD'07), Boston, MA, USA, 2007.
- [18] Kalnis P, Ghinita G, Mouratidis K, et al. Preserving anonymity in location based services[R]. Technical Report TRB6/06, Department of Computer Science, National University of Singapore, 2006.
- [19] Sweeney L. K-anonymity: a model for protecting privacy[J]. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 2002,10(5):557-570.
- [20] Samarati P, Sweeney L. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression [J]. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 2002,10(5):571-588.
- [21] Samarati P. Protecting respondent's privacy in microdata release[J]. IEEE Transactions on Knowledge and Data Engineering, 2001, 13(6): 1010–1027.
- [22] Mokbel M F, Chow C, Aref W G. The New Casper: query processing for location services without compromising

- privacy[C]//Proceedings of the International Conference on Very Large Data Bases(VLDB'06), Seoul, Korea, September 2006: 763–774.
- [23] Mokbel M F, Chow C, Aref W G. The New Casper: a privacy-aware location-based database server[C]//Proceedings of the International Conference on Data Engineering (ICDE'07), Istanbul, Turkey, April 2007.
- [24] Bamba B, Liu L. PrivacyGrid: supporting anonymous location queries in mobile environments[R]. Technical Report, Georgia Institute of Technology, May 2007.
- [25] Gedik B, Liu L. Location privacy in mobile systems: a personalized anonymization model[C] // Proceeding of the International Conference on Distributed Computing Systems(ICDCS'05), Columbus, OH, USA, 2005: 620–629
- [26] Atallah M J, Frikken K B. Privacy-preserving location-dependent query processing[C]//Proceeding of the IEEE/ACS International Conference on Pervasive Services (ICPS'04), Beirut, Lebanon, July 2004:9–17.
- [27] Cheng R, Kalashnikov D V, Prabhakar S. Evaluating probabilistic queries over imprecise data[C]//Proceeding of ACM International Conference on Management of Data(SIGMOD'03), San Diego, California, USA, 2003: 551–562.
- [28] Machanavajjhala A, Gehrke J, et al. L-diversity: privacy beyond k-anonymity[C]// In Proceedings of 22nd International Conference on Data Engineering(ICDE'06), Atlanta, Georgia, USA, 2006:24-36.