doi:10.3969/j.issn.1000-436x.2014.11.021

KAP：一种面向定位服务的位置隐私保护方法

王宇航，张宏莉，余翔湛
（哈尔滨工业大学 计算机网络与信息安全技术研究中心，黑龙江 哈尔滨 150001）

摘 要：在移动互联网中，位置隐私保护一直是有待解决的重要问题。针对定位服务中的位置隐私问题，提出了一种位置隐私保护方法 KAP。KAP 适用于 WLAN 定位技术。首先，提出了一种区域热点拓扑模型，将 WLAN 热点的地理分布用带权无向图描述，该模型能够在不使用 WLAN 热点坐标的前提下，反映热点之间的位置关系。然后，基于拓扑模型并结合 *k*-匿名思想，提出了 3 种位置隐私算法，保证了位置不被攻击者准确获得。最后，通过仿真实验，验证了方法的正确性。

关键词：位置隐私；移动互联网；*k*-匿名；定位服务
中图分类号：TP393 **文献标识码**：A **文章编号**：1000-436X(2014)11-0182-09

KAP: location privacy-preserving approach in location services

WANG Yu-hang, ZHANG Hong-li, YU Xiang-zhan
(Research Center of Computer Network and Information Security Technology, Harbin Institute of Technology, Harbin 150001, China)

Abstract: Preserving location privacy is an essential requirement in mobile internet. A location privacy protection approach named KAP was proposed which aimed at the privacy issue of location service under the mobile Internet. Through the analysis on locating technology, a weighted adjacent graph-based topology model was given in order to describe the positional relationship between hot spots. Meanwhile, with the help of the model, combining the concept of *k*-anonymity, three privacy algorithms was shown to make sure the location can not be obtained precisely by attacker. The simulation results verified the correctness and performance of the approach.

Key words: location privacy; mobile internet; *k*-anonymity; location service

1 引言

随着智能移动终端的普及和移动互联网络的快速发展，定位服务正成为移动互联网领域的发展焦点之一。用户的位置是如今大量信息服务的驱动性数据要素，人们在使用这些服务时需要频繁地进行定位。通过向第三方提交周边用以定位的接入点数据（通信基站、Wi-Fi 热点等），人们即可享受便捷的定位服务，快速、准确地获取位置。

在定位服务中，用户不再像传统 GPS 定位技术那样依靠设备自身即可获得位置，而是转由向定位服务提供商“索要”位置，在定位服务流程中，用户的位置首先会在提供商处生成，随后再通过网络发送给用户。在系统设计中应重点解决好用户担心的隐私保护问题，即“如何安全地获得位置”。由此，本文针对定位服务中的位置隐私问题，设计了一种位置隐私保护方法。方法适用于三角定位技术，利用“*k*-匿名”思想，能够在定位服务提供商无法获知用户确切位置的前提下提供定位服务，且方法不影响定位精度，在不降低定位服务质量的前提下，实现了 *k*-匿名定位（KAP, *k*-anonymous positioning），有效保障了位置隐私。

收稿日期：2014-08-19；**修回日期**：2014-11-05
基金项目：国家重点基础研究发展计划(“973”计划)基金资助项目(2011CB302605)；国家自然科学基金资助项目(61173144, 61073194, 61202457)
Foundation Items: The National Basic Research Program of China(973 Program) (2011CB302605); The National Natural Science Foundation of China (61173144, 61073194, 61202457)

2　背景知识

2.1　定位服务和相关技术

定位服务是一种基于网络的实时位置提供服务，是新兴的重要移动互联网定位手段，发展迅猛，谷歌、Skyhook、百度等 IT 巨头均推出了自己的定位服务；IOS、Android 等主流操作系统也对定位服务提供支持。与 GPS 定位相比，定位服务具有能耗低，场景适用性强等优点，其定位精度也可以满足民用市场。其一般原理[1]如图 1 所示。

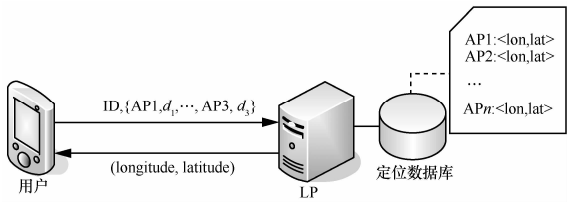


图 1　定位服务

定位服务提供商（LP, location provider）存储了大量通信接入点（WLAN 热点、手机基站等）的相关数据，这些数据被组织成定位数据库，当移动设备将感知到的周边接入点发送给 LP 后，LP 即可根据具体定位算法，生成用户的位置并返回给用户。

不同定位服务的区别主要表现在具体利用何种接入点数据上，现阶段，大多数定位服务采用的是基于 WLAN 热点的定位技术[2]，也有将 WLAN 定位与 GPS、手机基站等定位技术相结合的复合型定位技术[3]，但复合型定位技术中仍主要依赖 WLAN 定位技术。

三角定位法是一种常见的 WLAN 定位技术，三角定位法利用接入点（AP, access point）的地理位置和“信号—距离模型”相结合，进行三角测距定位，过程先后分为测距阶段和定位阶段：首先将 AP 的信号强度用传输损耗模型转换为几何长度[4]；随后结合 AP 坐标，使用三角测距算法计算用户坐标。图 2 是三角定位算法的一种基本情况。

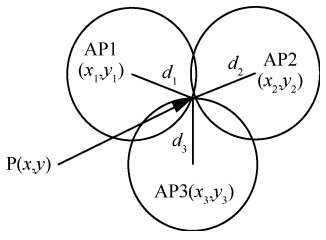


图 2　三角定位法示意

设接入点 AP1、AP2、AP3 的坐标分别为(x_1, y_1)、(x_2, y_2)、(x_3, y_3)，用户与三者距离分别为 d_1、d_2、d_3。则图 2 中用户 P 的坐标(x, y)可由方程组

$$\begin{cases}(x-x_1)^2+(y-y_1)^2=d_1^2\\(x-x_2)^2+(y-y_2)^2=d_2^2\\(x-x_3)^2+(y-y_3)^2=d_3^2\end{cases}$$

求得。

2.2　相关工作

移动互联网下的位置隐私保护相关研究已经取得了一定成果，早期工作大多专注于研究对已生成位置的隐私保护，主要通过匿名、模糊化等技术降低位置隐私泄漏的风险[5]，其中，基于 *k*-匿名（*k*-anonymity）的位置隐私保护技术取得了显著成果。*k*-匿名是信息安全领域的重要概念，指数据发布时，真实数据应首先“去标识符”化，并与其他若干（至少为 *k*−1）个数据同时发布[6]。*k*-匿名使攻击者无法一次准确辨别出真实数据及其所属个体。通常，*k* 称为匿名度，真实数据连同$(k-1)$个其他数据组成的集合成为匿名集。

k-匿名最早由 GRUTESER M 等[7]引入位置隐私保护领域并大量应用。*k*-匿名位置隐私保护技术通常会使用一个可信第三方充当匿名服务器，将用户的真实位置与一定区域内的其他$(k-1)$个用户的位置混合成含 *k* 个位置的匿名集，使攻击者无法准确定位用户[8~10]。

对定位服务中的位置隐私保护研究起步相对较晚，2013 年，DAMIANI M L 等[1]指出该问题，并提出一种基于隐私政策的保护方法，该方法利用可控的坐标粒度控制机制来限制 LP 能够获得的位置精度，从而实现保护位置隐私的目的；2014 年，PENG Z T 等[11]针对定位服务过程中的位置欺诈攻击(location-spoofing attack)，提出了一种位置隐私攻击判定方法，该方法的核心思想是当 LP 收到一个定位请求后首先判断接入点数据的真实性，主要依据 RSS 信号强度和距离之间的概率关系以及用户的历史位置数据，当一次定位服务请求的接入点数据判定为假时则拒绝服务。该方法有效抵御了位置欺诈，但位置欺诈攻击本身并没有将 LP 视为威胁来源。

3　分析和假设

3.1　分析

3.1.1　问题表述

KAP 的目的是实现定位服务的 *k*-匿名化。达成该目的需要解决的问题是：如何在一次定位中，使

LP 生成包含真实位置在内的 k 个位置且无法分辨其中真实位置。

3.1.2 分析

根据 2.1 节介绍的定位原理，LP 需要一个可定位的 AP 集来计算位置，因此，解决办法是向 LP 发送包含真实 AP 集在内的多个可定位 AP 集，并保证计算出的位置彼此不同。该问题可以分解为以下 2 个子问题。

Q1：现实中移动设备仅能感知到其周边 AP，如何获取其感知区域以外的 AP。

Q2：在设备感知区域以外的 AP 中，如何找寻一个可定位 AP 集，进行正确的定位计算。

3.1.3 结论和对应策略

利用移动设备本身获取其感知区域外的 AP 代价较大，KAP 采用一个第三方服务器来代替移动设备去获取区域外 AP。该服务器称为匿名器。匿名器的责任包括如下。

1) 维护区域内的大量 AP 数据，用户在定位服务之前，首先向该匿名器提出 k-匿名请求。

2) 对区域内的 AP 数据进行组织，使其能够反映 AP 之间的位置关系，以便能够在其中查找可定位的 AP 集。

3) k-匿名成功后，对 LP 返回的多个位置进行筛选，将真实 AP 集对应的位置返回给用户。

至此，匿名器可以解决 Q1，同时为 Q2 的解决提供了基础。

3.2 假设

3.2.1 威胁假设

首先，移动设备是可信的，其中包括定位服务请求的可信性，以及提供 AP 的可信性。

其次，不可信的 LP 是定位服务中的威胁来源。虽然在移动互联网通信场景下可能存在多种威胁来源，例如，对 LP 的攻击行为，或对通信信道的窃听行为，但 LP 是最具有威胁性的一方，因此假设 LP 是威胁来源有助于问题理解和威胁模型简化。

最后，3.1 节提出的匿名器是可信的，但仍对其通过技术手段加以必要防范。KAP 利用限制匿名器拥有的知识，来减少潜在威胁的可能性，具体将在第 4 节介绍。

3.2.2 空间模型和 RSS 模型

KAP 基于二维欧式空间模型，在建立坐标参考系后，用户和 AP 的位置均表示为形如(x, y)的二元组坐标。

KAP 的目的旨在保护位置隐私，因此抽象 RSS 信号强度和信号传播模型等底层问题。方法假设 AP 的可感知范围为圆形，并用圆面积表征 AP 的覆盖区域大小。图 3 展示了 AP 以及 AP 覆盖区域的空间表示。同时，将 RSS 生成的实际距离等同为覆盖半径。

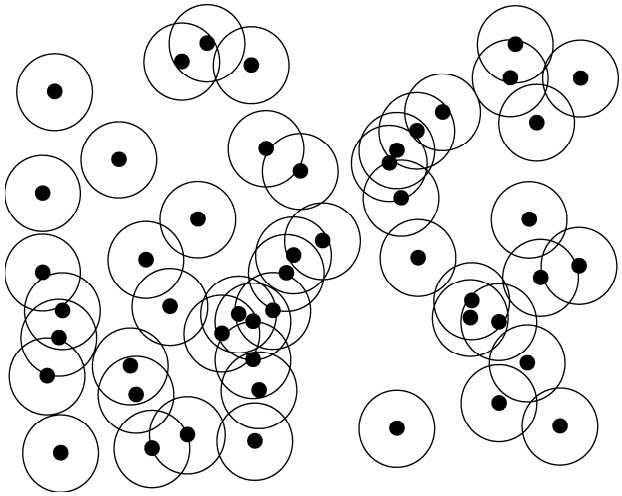


图 3 AP 的空间示意

4 KAP 方法设计

4.1 架构模型和服务流程

图 4 是 KAP 使用的中心型架构模型，包括三类角色：移动设备、匿名器和 LP。

移动设备为请求定位服务的一方，负责感知其周边的 AP 信号，并与匿名器进行直接通信。

匿名器作为用户与 LP 之间的第三方，负责对用户的定位请求实现 k-匿名，同时还负责对 LP 的返回结果进行筛选。

LP 是定位的最终实现者。LP 从匿名器处接收经 k-匿名处理后的定位请求数据，利用定位数据库查找用户位置，并将结果返回给匿名器。

利用该模型，KAP 的定位服务流程如下。

1) 用户向匿名器发起定位服务请求，发送的信息包括：身份标识、周边 AP 集合、指定匿名度 k。

2) 匿名器为每个定位请求实现 k-匿名，包括去标识符，以及根据指定的匿名度 k，添加$(k-1)$个冗余的 AP 集合，并保证这些冗余 AP 集合能够正确定位。最后向 LP 发起定位请求。

3) LP 使用匿名后的若干 AP 集合，生成包含真实位置在内的位置集合，并返回给匿名器。

4) 匿名器根据用户的真实 AP，从位置集合中筛选出真实位置，返回给用户。

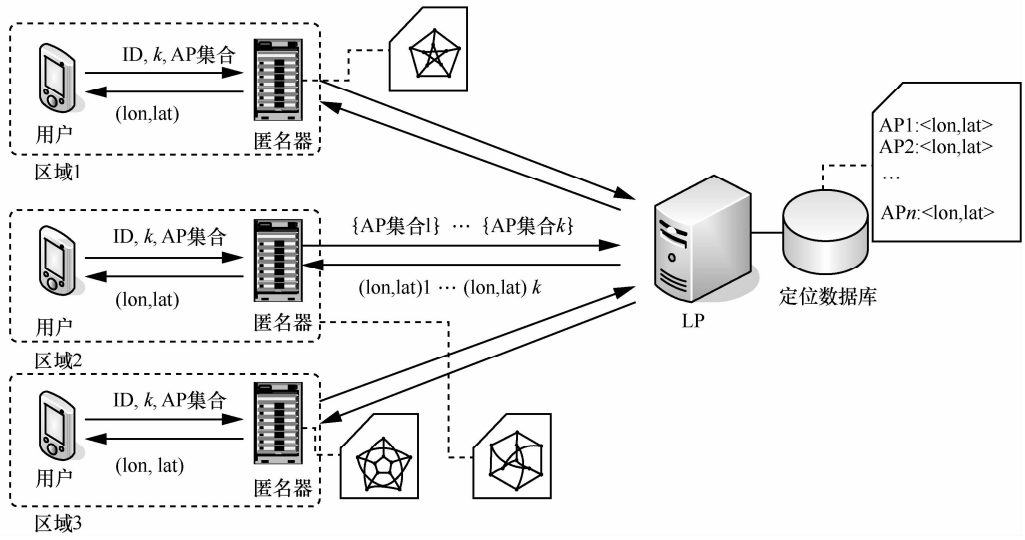


图 4　架构模型

该模型下，原始的移动设备定位服务请求方式和 LP 定位算法均没有变动。略有所不同的是，设备在请求定位服务时应指定匿名度 k；设备与 AP 间的距离简化为 AP 的覆盖半径；同时，LP 在返回 k 个位置时，应对每个位置标注其对应的 AP 集合。

4.2　AP 拓扑模型

4.2.1　AP 拓扑模型构建

KAP 使用无向图作为 AP 拓扑模型。用 $G(V,E)$ 表示区域内的 AP 拓扑模型，其中，点集 V 表示全部 AP 的集合，边集 E 表征 AP 覆盖区域之间彼此交叠的关系，即 $\exists e(v_x,v_y)\in E \Leftrightarrow C_x \cap C_y \neq \varnothing$，其中 C_x 和 C_y 分别为 v_x 和 v_y 的覆盖区域。

基于该方式，可以将区域内的 AP 位置及其覆盖交叠关系，转换为一张无向图。例如，图 5 即是对图 4 中 AP 的拓扑建模。

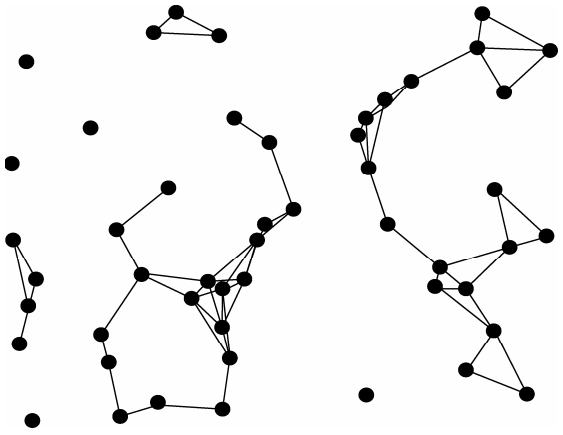


图 5　图 4 对应的拓扑模型

4.2.2　可定位 AP 集的拓扑表示

根据定位原理，若用户发送了若干 AP，即意味着该用户位于这些 AP 覆盖区域的重叠区域中，因此这些 AP 覆盖区域的交集必不为空，继而，将这些 AP 映射于拓扑模型中的点后，其对应的子图必能构成一个完全图。

然而，反之，若拓扑中存在某完全子图，将其点集还原在平面空间下后，却不一定能代表一个覆盖区域全部重叠的 AP 集。图 6 展示了 2 种此类特殊情况。

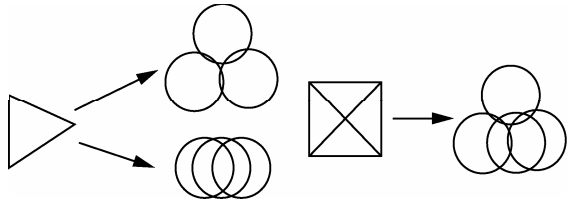


图 6　拓扑对应平面模型时的特殊情况

但幸运的是，“AP 集的覆盖区域全部重叠”也并非三角定位法能够定位的必要条件。对于类似于图 6 展示的一类情况，三角算法仍可以计算位置[12]。因此，在本文提出的拓扑模型下，可以利用拓扑中（含至少 3 个点）的完全子图来表征一个可定位 AP 集。

基于该拓扑模型，对于每一个 AP，匿名器仅保留其唯一标识符，而不需存储其地理坐标。因为通过上述分析可以发现，在该拓扑下，对“可定位 AP 集”的查找，就是对完全子图的查找，这可以不依赖 AP 坐标而实现。同时，这样做的优点在于：

1)节省匿名器内存空间，也简化后续 *k*-匿名计算的代价；2)符合威胁假设，使匿名器不具有其职能所需以外的不必要信息。

4.2.3 拓扑模型更新

实际条件下区域中 AP 会发生增减和变化，拓扑模型应随方法的不断执行而更新，表 1 为涉及更新情况对应的更新操作。

表 1 拓扑更新

操作	情况	拓扑更新说明
插入点	用户发送 AP 集中出现拓扑中没有的 AP	拓扑 $G(V, E)$, 新 AP 集{AP1,…, APn}; V= V∪ {AP1,…, APn}; E = E∪E_{new}，其中 E_{new} 是以{AP1,…, APn}为点集生成完全图的边集；
插入边	出现拓扑中现有 AP 的新定位集	拓扑 $G(V, E)$; 新定位集{AP1,…, APn}; E= E∪E_{new}，其中 E_{new} 是以{AP1, …, APn}为点集生成的完全图的边集；
删除点	LP 中没有对应 AP 的 ID	删除 G 中点 AP 及其关联的所有边
删除边	LP 根据对应 AP 集无法定位	删除 G 中所有以该 AP 集构成完全图中包含的边

对于插入点算法需要特殊说明的是：随着点和边的不断插入，拓扑中可能生成非先验的（即由拓扑建模过程生成的，或由用户发送的）完全子图。基于与 4.2.1 节相同的论证，仍可断定这些完全图满足可定位条件。因此，插入点算法不会破坏原有拓扑的特性。

4.3 *k*-匿名的实现

在 4.2 节拓扑模型基础上，将 *k*-匿名的实现转化为在拓扑中寻找完全子图问题。本节提出 3 种定位 *k*-匿名算法。表 2 是文中的缩写以及符号的含义。

表 2 相关符号含义

符号	含义	符号	含义
G	整体拓扑	AP_E	可定位 AP 集
V_G	整体拓扑点集	K_x	包含 x 个点的完全子图
E_G	整体拓扑边集	Nei_v	点 v 的邻居组成的点集
AP_T	用户的真实 AP 集	E_v	以 v 为端点的所有边集合

4.3.1 随机实现 *k*-匿名

从 G 中随机地选取(k−1)个 AP_E，连同 AP_T 形成匿名集。在随机前提下，算法尽量规避可能由随机性带来的 2 个问题，一是生成的 AP_E 与 AP_T 在 G 中距离过近，导致位置离散性降低，可能无法抵御位置同质性攻击[13]；二是 AP_E 的点个数与 AP_T 的点个数差异性过大。对于前者，算法将尽力找寻与 AP_T 保持一定（拓扑）距离的点充当 AP_E，从而确保 AP_E 与 AP_T 在空间位置上的离散性；对于后者，算法采用在有限次循环下，提供与 AP_T 相似度尽可能高的 AP_E 迭代策略。随机 *k*-匿名算法核心步骤如下。

算法 1 random_based_KAP

输入：G, AP_T, k //拓扑，真实点集和匿名度 k

输出：AP_{E1},…, $AP_{E(k-1)}$ //k−1 个可定位 AP 集

```
DEFINE MAX//最大跳数
Vertex v_ori = random_Pick_Vertex( APT );
//随机在 APT 中选取一点
size = sizeof( APT ); //标记 APT 大小
for (i from 2 to k){ //开始寻找完全图
  while( APEi == null){
  Vertex v_ano = jump_To(G, v_ori, MAX);
//从随机原始点出发，在 G 中随机跳不大于 MAX 次，至目标点
   APEi =find_K (v_ano, G, size);
//找寻 G 中一个包含目标点在内的完全子图
  }//endwhile
}//endfor
return APE1 to APE(k-1);
```

算法首先随机选取 AP_T 中的一点 *v_ori* 作为起始点（第 1 行），再以 *v_ori* 为基础，执行 jump_To 函数，选取一个目标随机点 *v_ano*（第 5 行），jump_To 的具体过程是：首先考虑从 *v_ori* 出发，沿 E_v 中的随机某边，“跳”随机次，到达 *v_ano*；但若 G 中存在不连通子图且 *v_ori* 所在的子图过小时，则利用从其他子图中随机选点的办法确定 *v_ano*。随后，基于 *v_ano* 寻找可能存在的完全子图（第 6 行），函数 find_K 从 K_{size} 开始寻找，若 *v_ano* 处没有 K_{size} 则寻找 $K_{(size-1)}$，直至找到一个完全子图，或 *size*<3 为止。若 *v_ano* 处没有任何点数大于 2 的完全子图，重新选择 *v_ano*，循环直至找到一个完全子图（第 4 至 7 行）。算法找到(k−1)个完全子图（第 3 至 8 行）后，返回这些完全子图的点集（第 9 行）。

G 中至少存在一个满足 find_K 函数要求的完全子图，就是 AP_T 映射的子图本身，因此，最坏的情况即是返回了(k−1)个与 AP_T 完全相同的 AP_E，此时算法仍然是有穷的。

4.3.2 基于查找集的 *k*-匿名算法

根据 4.2.2 节的论证，G 中本身含有大量 AP_E 的拓扑表示，即完全子图。因此可以在用户请求之前，预先从 G 中查找出这些完全子图，作为 *k*-匿名算法的查找基础。相较于随机法，基于查找

集的算法能降低 k-匿名实现的复杂度，提高匿名器工作效率。算法先后分为训练阶段和 k-匿名阶段。训练阶段将 G 中所有 AP_E 找出作为查找集，通过遍历 G 中的子图，判定某子图的边数 E 与点数 V 是否满足关系

$$E=\frac{V(V-1)}{2}$$

来判定其是否为完全子图。由于当 V_G 规模庞大时，遍历代价将过大，因此应对训练算法优化。考虑到若 G 中已没有 K_x，则必不再含有 $K_{(x+1)}$，因此算法采取递增循环方式，从 K_3 开始查找，直至 G 中不再有 $K_{(x+1)}$ 为止。此时用所有完全子图作为 G 的 AP_E 查找集。

基于查找集的算法的目标在于快速实现 k-匿名，因此在 k-匿名阶段不借助 AP_T，而是直接从查找集中随机选择(k−1)个 AP_E 作为算法的输出。

4.3.3　基于贪心策略的 k-匿名

为了降低在 G 中实现 k-匿名的计算代价，同时又不借助庞大的查找集，本节给出一种贪心算法，把整体问题从“在 G 中寻找 k 个 AP_E 问题（全局最优解）”，转换为“根据拓扑特性抓取点（局部最优策略）问题”。算法通过观察 G 中点的分布疏密特性，从中直接抓取若干个点作为一个点集，并尽最大努力使抓取的点集能够构成 AP_E。

难点在于：从 G 中何处进行抓取更可能获得 AP_E。经分析发现，G 中任意点 v_i 附近存在完全子图的概率，可以用该点的聚集系数表征。聚集系数（clustering coefficient）用于描述图中与同一点相邻的点间也互为相邻关系的程度。用 x_i 表示点 v_i 所连接的邻居个数，e_i 表示 x_i 个邻居之间实际存在的无向边数，则节点 v_i 的聚集系数可表示为

$$C_i=\frac{2e_i}{x_i(x_i-1)}$$

显然，$C_i\in[0,1]$，当 $C_i=1$ 时，Nei_{vi} 能构成 K_{xi}；$v_i\bigcup Nei_{vi}$ 则能构成 $K_{(xi+1)}$。且 C_i 越大，Nei_{vi} 中存在完全子图的概率越大。

同时，Nei_{vi} 中可能存在的完全子图的点个数上限，由边数 e_i 决定，上限 $x_{\max}$ 与 e_i 的关系为

$$x_{\max}=\left\lfloor\frac{1+\sqrt{1+8ei}}{2}\right\rfloor$$

结合上述分析，本文给出点的“邻接完全图概率”定义。

定义 1　（邻接完全图概率）：G 中某点 v_i 的邻居 Nei_{vi} 中蕴含有 K_x 的概率，称为点 v_i 的“x-邻接完全图概率”，记作 $P_{v_i}(x)$

$$P_{v_i}(x)=\begin{cases}0, & x>x_{\max}\text{或}C_i=0\\ \dfrac{C_{x_i}^{x}\times C_{E_{x_i}-E_x}^{e_i-E_x}}{C_{E_{x_i}}^{e_i}}, & 1<x\leqslant x_{\max}\\ 1, & x=1\text{或}C_i=1\end{cases}$$

其中，E_{x_i} 和 E_x 分别表示 K_{x_i} 和 K_x 的边数。

综合上述，贪心算法基于聚集系数及定义 1 实现。首先计算 G 中各点的聚集系数和其邻接完全图概率，采集聚集系数较大的点作为索引。索引形式如图 7 所示。

贪心算法首先利用索引找到 G 中目标点，再从其附近抓取若干点组成点集作为输出。具体步骤如下。

算法 2　greedy_based_KAP

输入：索引、AP_T、匿名度 k、概率阈值 p

输出：(k−1)个点集

step1　确定 AP_T 中的点个数 x。

点	邻居个数	C	$P_v(x)$	…	$P_v(n)$	指针
$V1$	3	1				
$V2$	4	1				
$V3$	2	1				
$V4$	3	0.91	0.84			
$V5$	4	0.87	0.72	0.6		
$V6$	4	0.8	0.61	0.52		

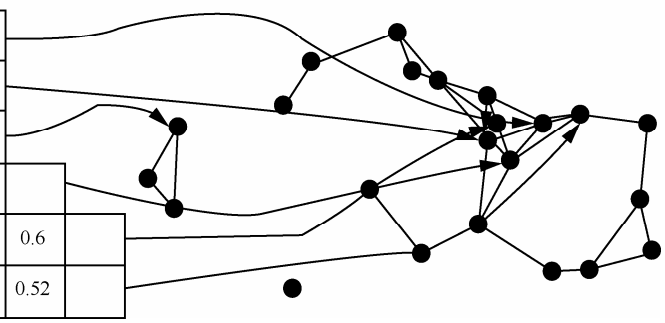


图 7　索引示意

step2 从索引中随机选择邻居个数 n 大于或等于$(x-1)$的点 v，并保证 v 不在 AP_T 中。根据 v 的聚集系数，若 $C_v=1$，转至 step3；否则跳至 step4。

step3 返回 $v \cup Nei_v$ 中任意 x 个点组成的点集。

step4 判定点 v 的各个邻接完全图概率与 p 的大小关系：若 $P_v(x) \geqslant p$，从 Nei_v 中随机选择$(x-1)$个点，返回 v 和这些点的点集；否则，找出索引中满足 $P_v(n) \geqslant p$ 的最大点个数 n，从 Nei_v 中随机选择$(n-1)$个点，返回 v 和这些点的点集。

step5 循环 step2 至 step4，至返回$(k-1)$个点集。

4.3.4 安全性分析

3 种算法的核心是实现对 AP_T 的 k-匿名，因此，算法的安全性符合一般 k-匿名数据的安全性定义：对每一个敏感数据（定位服务请求），都将其敏感属性（用户 ID）掩盖，同时将非敏感属性（AP 集）泛化，并保证每条数据都至少和其他$(k-1)$条数据无法区分开来。算法杜绝了攻击者（LP）将每个 AP 集与用户 ID 进行准确匹配的可能。因此保护了定位服务的隐私。同时，由于实现方式各异，3 种算法在对同一 AP_T 多次执行后生成的匿名集在空间上会产生一定的聚集特性。

此外，受益于本文提出的拓扑模型，k-匿名算法所需的 AP_E，并不依赖 AP 的具体位置坐标而生成。这意味着算法在 TTP 处执行时，其输入为用户发送的真实 AP_T，输出则是 k 个在 LP 处可以实现定位的 AP_E，而这些 AP_E 不包含任何具体位置坐标信息。因此，在 4.1 节指出的 1)~3)阶段，TTP 处并没有用户的位置信息。综上，算法是安全的。

5 仿真实验和对比

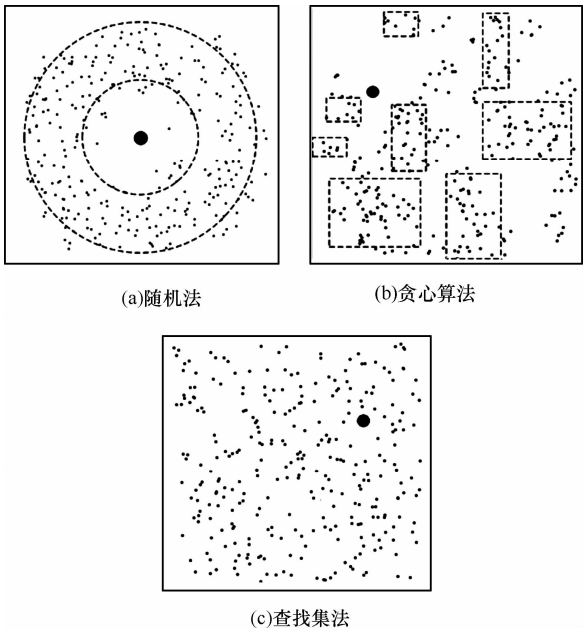
本节对 KAP 方法进行仿真实验，建立了一个简单的模拟用户请求发生器，用以向 k-匿名算法提出定位匿名化请求；以及一个模拟定位器，用以对 k-匿名算法输出的点集进行合理性验证和定位坐标分析。在此基础上，实验主要对 3 种 k-匿名算法进行验证和比较，其中考察的方面包括：1) 位置离散性，在一次定位 k-匿名中，利用生成的 k 个点集所计算出的实际位置坐标的空间关系；2) 匿名成功率，即考察 k-匿名算法的正确性，k-匿名成功率即算法输出的点集中，可定位的点集数与全部点集数的比值；3) 算法实际运行时间。表 3 是本次实验的环境配置。

表 3 实验环境配置

CPU	内存	操作系统	开发平台	使用语言
Intel 双核 E4500 2.2GHz	DDR2 667 4G	Windows 7	Visual Studio 2008	C++

5.1 离散性

实验在随机数据集上进行，随机数据集采用在一定大小区域内随机布点的方式，结合统一半径生成无向网模型，共设有 2 000 个 AP 点和 3 792 条边，对同一 AP_T 执行 3 种 3-匿名算法 100 次并将输出点集转换为空间坐标，随后观察这些生成位置的空间关系。图 8 展示了空间上这些算法生成位置的分布情况。



(a)随机法 (b)贪心算法

(c)查找集法

图 8 算法生成的位置离散性

实验结果表明，随机法由于采用从 AP_T 出发“跳”至目标 AP_E 的方式，因此保证了生成位置与真实位置具有一定的距离，也就是离散性，而随着算法执行次数的增加，生成位置在统计上会具有以真实点为中心圆的范围特征，但对于单次 k-匿名而言，真实位置仍是无法分辨的；贪心算法生成的位置离散性来源于对索引项的随机选择和在 G 中对目标 AP 的随机抓取，因此多次执行下，不具有随机法中的真实位置趋于中心特性，而是统计上向索引项中的 AP 点坐标聚集趋势；基于查找集的算法则在满足随机离散性的基础上，在多次执行后其生成位置分布最为平均。

5.2 成功率

随机算法和基于查找集的算法实现设计都具

有完备性，实验主要对贪心算法的成功率进行测试。同时，为了能够反映贪心算法的成功率与拓扑规模的关系，实验基于多组随机数据，点数从 500 开始，以 500 的增量递增至 2 500。图 9 给出了在各个数据集上当 k=3、p=0.9 时，对随机输入的 AP_T 执行贪心算法 1 000 次后的成功率。实验显示，对于不同的拓扑规模，贪心算法总能保证接近 95%的成功率，且随着拓扑规模的增大，算法成功率基本保持稳定且有所提升。实验证明了贪心算法的高成功率和稳定性。

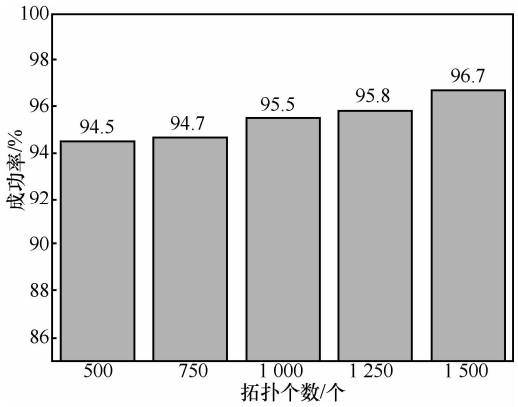


图 9　贪心算法的成功率

5.3　算法实际运行时间

为了提高对算法运行时间测试的真实性，实验使用真实数据集评估 3 种算法的处理时间，真实数据集拓扑使用美国纽约市的真实 WLAN 热点及其地理坐标[14]生成，选取 1 250 点。在随机输入 AP_T 前提下，将匿名度 k 从 5 递增至 25。

如图 10 所示，随着匿名度的增大，3 种算法运行时间均有增加，其中基于查找集算法的时间最短且比较稳定，这是由于从查找集中随机选取 AP_E 并不依赖 AP_T 完成，而是直接从查找集中随机选取。该算法的快速运行时间依赖于预先训练的查找集之上；随机法随着匿名度的增大，运行时间最长，这主要由于随机法在一次匿名中，需要在 G 上执行 $(k-1)$次完全子图的查找；贪心算法的运行时间较为折中。

结合实验，可以看出，3 种算法各有其优点和适用性。

其中，随机法由于采用从拓扑中即时计算完全子图的策略，因此其时间代价较高，优点是算法无需任何数据预处理过程，也不占用额外的内存和磁盘空间，随机法的运行时间随着匿名度的提高而增长显著，因此适用于拓扑规模适中且匿名度较低的场景；基于查找集的算法本质上是用大量预处理工作换取了算法快速的运行时间，随着匿名度 k 的增长，其运行时间几乎没有增加，适用于匿名度较大的条件下。最后，贪心算法虽然是以一定概率给出正确的 AP_E，但其预处理代价和运算时间均较为合理，此外，根据成功率实验，其实际成功率也能维持在较高水平。

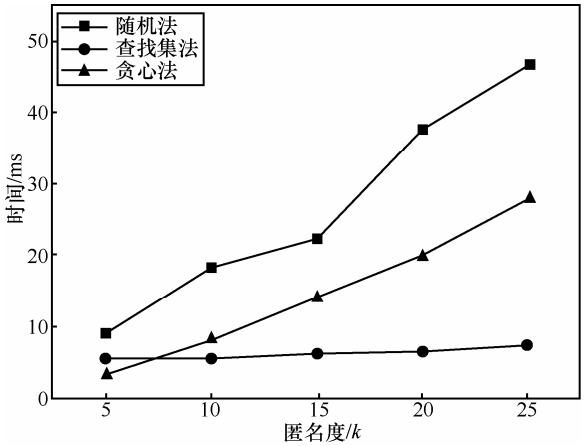


图 10　算法实际运行时间对比

对于位置离散性，随机法能够在单次查询中给出具有较高离散性的点集，，且对同一 AP_T 多次执行的条件下则会呈现出比较明显的规律性；基于查找集的算法在多次执行条件不呈现聚集规律；贪心算法也会出现一定的聚集规律，相对于随机法，其生成位置的空间分布更为离散。

6　结束语

本文针对定位服务中的隐私泄漏问题，提出了一种位置隐私保护方法：KAP。方法基于 k-匿名策略，采用以匿名器为中心的中心型架构模型实现，设计了一种空间 AP 的拓扑建模方法，该方法可以在不依赖 AP 坐标的前提下，反映 AP 之间的可定位关系。然后，基于该拓扑，设计了 3 种 k-匿名算法：随机法、基于查找集的算法和贪心算法。最后的仿真实验证证明了 3 种 k-匿名算法的正确性以及各自特性，实验也表明了贪心算法具有较高的适应性和可靠性。

对于定位服务中的位置隐私保护，未来仍有若干问题需要继续研究。定位服务中的轨迹隐私问题，以及更有威胁性的背景知识攻击问题，将是下一阶段工作的研究重点。

参考文献：

[1] DAMIANI M L, CUIJPERS C. Privacy challenges in third-party location services[A]. IEEE 14th International Conference on Mobile Data Management(MDM 2013)[C]. Milan, Italy, 2013.

[2] LI B, SALTER J, DEMPSTER A G, *et al*. Indoor positioning techniques based on wireless LAN[A]. First IEEE International Conference on Wireless Broadband and Ultra Wideband Communication(2007)[C]. 2008.13-16.

[3] FICCO M, PALMIERI F, CASTIGLIONE A. Hybrid indoor and outdoor location services for new generation mobile terminals[J]. Personal and Ubiquitous Computing, 2014, 18(2): 271-285.

[4] GEZICI S. A survey on wireless position estimation[J]. Journal of Wireless Personal Communications, 2008, 44(3):63-282.

[5] DUCKHAM M, KULIK L. Location privacy and location-aware computing[M]. FL: CRC Press, 2006.

[6] WERNKE M, SKVORTSOV P, DURR F, *et al*. A classification of location privacy attacks and approaches[J]. Personal and Ubiquitous Computing, 2012, 18(1): 163-175.

[7] GRUTESER M, GRUNWALD D. Anonymous usage of location-based services through spatial and temporal cloaking[A]. Proceedings of the 1st International Conference on Mobile Systems, Applications and Services (MOBISYS 2003)[C]. San Francisco, California, 2003. 31-42.

[8] ZHANG C Y, HUANG Y. Cloaking locations for anonymous location based services: a hybrid approach[J]. Geoinformatica, 2009, 13(2): 159–182.

[9] DIVANIS A G, KALNIS P, VERYKIOS V S. Providing *k*-anonymity in location based services[J]. SIGKDD Explorations Newsletter, 2010, 12(1): 3-10.

[10] CHOW C Y, MOKBEL M F. Trajectory privacy in location-based services and data publication[J]. SIGKDD Explorations, 2011, 13(1): 19-29.

[11] PENG Z T, KAJI K, KAWAGUCHI N. Privacy protection in Wi-Fi based location estimation[A]. Seventh International Conference on Mobile Computing and Ubiquitous Networking(ICMU 2014)[C]. Singapore, 2014. 62-67.

[12] CAVALIERI S. WLAN-based outdoor localization using pattern matching algorithm[J]. International Journal of Wireless Information Network, 2007, 14(4): 265-279.

[13] MACHANAVAJJHALA A, KIFER D, GEHRKE J, *et al*. L-diversity: privacy beyond *k*-anonymity[J]. ACM Transactions on Knowledge Discovery from Data, 2007, 1(1): 3.

[14] Location of Wi-Fi hotspots in the city with basic descriptive information[EB/OL]. https://data.cityofnewyork.us/.

作者简介：

王宇航（1987-），男，黑龙江哈尔滨人，哈尔滨工业大学博士生，主要研究方向为移动互联网和信息安全。

张宏莉（1973-），女，吉林榆树人，哈尔滨工业大学教授、博士生导师，主要研究方向为网络与信息安全、网络测量与建模、网络计算、并行处理等。

余翔湛（1973-），男，黑龙江哈尔滨人，哈尔滨工业大学教授，主要研究方向为网络容灾和信息安全。