第48卷 第2期
Vol.48 No.2
计 算 机 工 程
Computer Engineering
2022年2月
February 2022
·网络空间安全·
文章编号：1000-3428(2022)02-0147-09 文献标志码：A 中图分类号：TP309

基于历史查询概率的K-匿名哑元位置选取算法

杨 洋，胡晓辉，杜永文
（兰州交通大学 电子与信息工程学院，兰州 730070）

摘 要：基于历史查询概率的哑元位置隐私保护机制存在匿名度低、隐匿区域小和位置分布不均匀的问题。提出K-匿名哑元位置选取（K-DLS）算法用于位置隐私保护。通过综合考虑匿名集的位置离散度和零查询用户，增强哑元匿名集的隐私性。利用熵度量选择哑元位置，使得哑元匿名集的熵值最优，并根据位置偏移距离优化匿名结果，增加匿名集的位置离散度。仿真结果表明，K-DLS算法的哑元匿名集离散度优于DLS、DLP、Enhanced_DLP等算法，能够有效提高用户位置的隐私保护效果。

关键词：基于位置的服务；位置隐私；哑元位置选取；零查询用户；K-匿名；地理位置分布

开放科学（资源服务）标志码（OSID）：

中文引用格式：杨洋，胡晓辉，杜永文.基于历史查询概率的K-匿名哑元位置选取算法[J].计算机工程，2022，48(2)：147-155.

英文引用格式：YANG Y，HU X H，DU Y W.The K-anonymous dummy location selection algorithm based on historical query probability[J].Computer Engineering，2022，48(2)：147-155.

The K-Anonymous Dummy Location Selection Algorithm Based on Historical Query Probability

YANG Yang，HU Xiaohui，DU Yongwen
（School of Electronics and Information Engineering，Lanzhou Jiaotong University，Lanzhou 730070，China）

【Abstract】The dummy-based location privacy mechanism using historical query probability suffers from low anonymity，small coverage area and imbalanced location distribution. To address the problem，a K-anonymous dummy-based locationselection algorithm is proposed for position privacy protection.The privacy of dummy anonymous set is enhanced by comprehensively considering the location dispersion of anonymous set and zero-query users.The algorithm selects the location of dummy through entropymeasure to make the entropy of the anonymous dummy set optimal.Then the anonymous result is optimized based on the offset distance of the location，and the location dispersion of the constructed anonymous set is improved.The simulation results show that the proposed algorithm displays a higher location dispersion degree of the dummy-based anonymous set than DLS，DLP，Enhanced_DLP and other algorithms.It significantly improves the performance of location privacy protection for users.

【Key words】Location Based Service（LBS）；location privacy；dummy location selection；zero-query users；K-anonymity；geographic distribution of locations

DOI：10. 19678/j. issn. 1000-3428. 0060417

0 概述

随着物联网和移动定位技术的发展，基于位置的服务（Location Based Service，LBS）被广泛应用于导航、查找附近的服务、接收基于位置的广告等领域[1-2]。用户使用基于位置的应用向不可信的LBS服务器发送包含用户的地理位置信息、兴趣点等内容的服务请求，LBS服务器根据用户发送的服务请求为用户提供基于位置的服务[3]。用户提供的位置越精确，获得的服务质量越高；位置越模糊，服务质量越低[5]。如果不可信的服务提供商掌握用户的真实位置信息，与用户相关联的个人信息会进一步泄露，如家庭住址、社会关系等。

针对位置隐私保护中存在的问题，研究人员提出多种解决方案[6]，其中，最常用的位置隐私保护技术是基于历史查询概率的K-匿名技术[8-9]。K-匿名技术最早出现在数据库领域，对于准标识符属性，任意一条记录无法与至少$k-1$条记录区分[11]。哑元位

基金项目：国家自然科学基金（11461038，61163009）；甘肃省高等学校创新基金（2020A-033）；甘肃省科技支撑计划项目（144NKCA040）。
作者简介：杨 洋（1994—），女，硕士研究生，主研方向为位置隐私保护；胡晓辉，教授、博士；杜永文，副教授、博士。
收稿日期：2020-12-28 **修回日期：**2021-02-25 **E-mail：**1291936444@qq.com

置是一种与用户位置极相似的虚假查询位置，普遍采用K-匿名集的方式。K-匿名集技术的原理是将用户的真实位置与$K-1$个哑元位置组合形成包含k个位置的匿名集[12]。用户使用匿名集代替真实位置发起查询，LBS提供商响应查询并返回每个查询位置需要的服务列表，用户则根据其真实位置筛选查询结果，获取属于自己的服务[14]。

哑元位置作为位置泛化的重要方法，具有部署简单且不影响服务质量的优点[15]。现有的哑元选取方案都聚焦在如何能合理有效地选取哑元位置以防止攻击者从匿名集中获取用户的真实位置[16]。攻击者如果获取了用户位置相关的边信息并过滤一些不合理哑元，例如位于湖泊、海洋、沙漠等特殊的地理位置的哑元，则能够大幅增加攻击者获取用户真实位置的概率，降低匿名集的隐私级别。针对此类问题，文献[16]提出虚假位置选择(Dummy Location Selection，DLS)算法，该算法通过将地图上位置单元的历史查询概率作为一种边信息划分用户的位置隐私等级，并使用位置熵选择合适的哑元位置构造K-匿名集。但是DLS算法需要消耗大量的算力来选取具有最大熵值的匿名集，算法时间复杂度高，不适用于资源受限的物联网设备，同时算法也未考虑哑元位置的离散度问题。文献[17]通过分析DLS算法存在的问题，设计了DLS攻击(Attack for Dummy Location Selection，ADLS)算法，并验证ADLS攻击算法的有效性，进而提出一种新的哑元选取DLP算法，该算法具有较好的隐私保护效果，却忽略了历史查询概率为零的特殊位置情况。文献[18]根据DLP算法提出改进的Enhanced-DLP算法，引入增强型贪心算法，使其具有较高的匿名集生成效率和较强的隐私保护效果。文献[19]提出基于虚拟网格的GridDummy算法和基于虚拟圆的CircleDummy算法。这两种算法将K-匿名需要的K个位置全部用哑元位置代替，即提交的所有位置都不包含用户的真实位置，虽然提高了攻击者推测用户真实位置的难度，但是降低了服务质量，同时也未考虑到哑元匿名集的位置离散度问题，无法有效保证哑元位置在地理上均匀离散分布。

本文提出基于历史查询概率的K-匿名哑元位置选取算法，考虑用户历史查询概率为零的特殊情况，在保证用户位置信息不被泄露的同时，从地理分散度和零查询用户两个维度增强匿名集的隐私性，从而提高位置隐私保护的安全性。

1 相关理论

1.1 基本概念

本节主要介绍文中使用的相关概念。

1)位置查询概率

位置查询概率(Location Query Probability，LQP)将地图划分成$N\times N$的网格，每个网格代表一个位置，单元，对于网格地图上的位置单元i均有一个对应的历史查询次数n_i，地图上所有位置单元的查询次数总和为$\sum_{i=1}^{k}n_i$，则i位置单元的历史查询概率如式(1)所示：

$$P_{i_{\text{que}}}=\frac{n_i}{\sum_{i=1}^{k}n_i},\ P_{i_{\text{que}}}\in[0,1] \tag{1}$$

2)零查询用户

在网格化的地图位置单元中，历史查询概率为零的位置称为零查询位置。在零查询位置发起服务请求的用户称为零查询用户(Zero-Query Users，ZQU)。

3)用户最大偏移距离

用户u选取其位置附近的某个位置单元l'代替其真实位置l，l'称为l的偏移位置。位置偏移会造成用户服务质量的损失，用户服务质量由偏移位置与用户真实位置的数学期望表示。在用户可接受的最大服务质量损失时，用户与偏移位置的欧氏距离为用户最大位置偏移距离(User Maximum Offset Distance，UMOD)，如式(2)、式(3)所示：

$$Q_{\text{loss}}=\sum_{l'}^{l}p_l f(l'|l)D_{\text{Dis}}(l,l') \tag{2}$$

$$D_{\text{Dis}}(l,l')=\max\sqrt{(x-x')^2+(y-y')^2},\ Q_{\text{loss}}<Q_{\text{loss}}^{\max} \tag{3}$$

其中：p_l为用户真实位置的历史查询概率；$f(l'/l)$为在用户位置l的情况下获得偏移位置l'的概率，位置隐私的保护算法需要保证$Q_{\text{loss}}<Q_{\text{loss}}^{\max}$，因为超过该值时，所得到的位置服务请求结果就无法满足用户最低服务质量的要求。

1.2 LBS架构

LBS架构如图1所示。

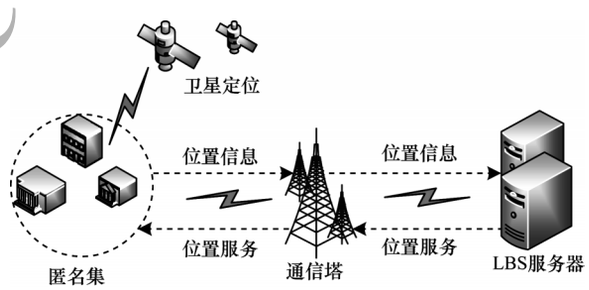

图1 LBS架构

Fig.1 Framework of LBS

LBS架构主要包括GPS卫星、具有定位功能和处理能力的终端、通信基站和位置服务提供商4个部分。

1)GPS卫星为移动终端提供当前的地理位置信息，位置隐私保护算法认为此过程是可靠的，默认在位置获取的过程中是安全的。

2)用户通过具有定位功能和处理能力的终端对请求服务的位置进行匿名保护，将用户的真实位置隐藏在哑元匿名集中发送至通信基站。

3）通信基站将接收到的服务请求发送至相应的LBS提供商的服务器，通信基站只对用户的服务请求进行转发，不对服务请求进行修改。通信基站在收到服务器响应后将其转发至用户终端，不对响应信息进行修改。但在通信过程中，相关的信息可能会被恶意攻击者获取，故本阶段是不安全的。

4）服务位置服务器在收到用户的服务请求后，解析请求内容，产生匿名集所有位置的查询结果，然后通过通信基站将结果返回给用户。由于恶意的服务提供商可能会通过泄露用户的隐私信息来获取利益，攻击者也可以通过攻击服务器获取用户相关的隐私信息，故本阶段也是不安全的。

1.3　隐私度量

隐私度量主要有基于历史查询概率K-匿名的隐私度量、基于信息熵的隐私度量、位置离散度。

1）基于历史查询概率K-匿名的隐私度量

根据哑元选取方案得到的哑元匿名集为C，$C=\{l_1,l_2,\cdots,l_{k-1},l_k\}$，即$|C|=k$。攻击者从匿名集中获取用户的真实位置概率$Q$，如式（4）所示：

$$Q=\frac{1}{|C|}=\frac{1}{k} \tag{4}$$

2）基于信息熵的隐私度量

位置查询概率经过地图网格化处理以后，生成$N\times N$个单元，从中选取包含用户真实位置在内的k个位置，根据位置单 元的历史查询概率p_i确定匿名集k中每个位置单元归一化后的查询概率q_i如式（5）所示，熵值如式（6）所示：

$$q_i=\frac{p_i}{\sum_{j=1}^{k}p_j} \tag{5}$$

$$H=-\sum_{i=1}^{k}q_i\times \mathrm{lb}\,q_i \tag{6}$$

隐私度量使用信息熵来度量匿名集的隐私程度[19]。熵值越大，表示集合中元素的区分度越低，攻击者从匿名集中识别用户的真实位置概率就越小，匿名集的隐私级别就越高。匿名集中位置单元的历史查询概率越相似，熵值越大。当历史查询概率相等时，熵值取得最大值。

3）位置离散度

哑元匿名集的位置离散度（P_{PD}）通过计算匿名集中所有位置单元之间的距离来获得，如式（7）所示：

$$P_{\mathrm{PD}}=\sum_{i\neq j}^{k}D_{\mathrm{Dis}}(l_i,l_j) \tag{7}$$

P_{PD}值越大，表示对于同一个位置数据集选取的哑元位置地理分布越离散，匿名集隐私效果越好；相反，哑元位置越聚集，越容易被攻击者缩小隐私区域，哑元匿名集隐私保护效果越差。

1.4　哑元位置选取

哑元位置选取方法的主要目的是使匿名用户选择合适的哑元位置构建K-匿名集，使得用户真实位置和其他$K-1$个哑元位置无法区分，将地图网格化为5×5的位置单元，根据每个网格的历史查询次数，计算各个位置单元的查询概率，用不同灰度代表位置单元查询概率的大小，颜色越深，历史查询概率越大，颜色越浅，历史查询概率越低，白色代表该位置单元的历史查询概率为零。

1.4.1　未充分考虑特殊查询概率位置单元

哑元K-匿名集经过地图网格化处理后存在零查询的位置单元。如果直接对零查询用户的位置单元进行哑元选取，则可能会生成含有多个零查询位置单元的匿名集。零查询位置单元在地理语义上可能是一些无人区、河流、沙漠等特殊位置。如果哑元匿名集含有这些特殊位置，则攻击者结合地理信息分析过滤这些特殊位置单元，从而降低哑元K-匿名集的规模。

5-匿名哑元位置选择如图2所示，从图2可以看出，右侧的数值代表对应位置的历史查询概率，位于新开发区的用户u_1通过熵度量的算法取得匿名度K=5的最佳隐私保护效果，需要选择4个与其能形成最大熵值的位置单元$\{u_2,u_3,u_4,u_5\}$。获取该匿名集的攻击者通过分析相关位置信息可过滤$\{u_3,u_4,u_5\}$，用户的隐私泄露概率由1/5增加到1/2，增加了隐私泄露的风险。

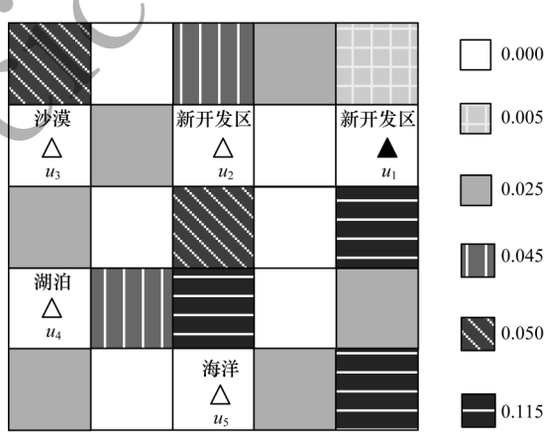

图2　5-匿名哑元位置选择

Fig.2　Location selection of five-anonymous dummy

1.4.2　未充分考虑哑元位置离散度

由于一些基于历史查询概率进行哑元选取的算法没有考虑位置单元的地理分布情况。对于一个给定k匿名规模的哑元匿名集，构造哑元匿名集的位置离散度应该尽可能大。

3-匿名哑元位置选择如图3所示。右侧数值代表对应位置的历史查询概率，位于新开发区的用户u_1通过熵度量的算法取得匿名度K=3的最佳隐私保护效果，

从与用户具有相同历史查询概率的$\{u_2,u_3,u_4,u_5,u_6\}$中随机选择2个位置单元，即$p_{u_1}=p_{u_2}=p_{u_3}$，构造位置匿名集$C_1=\{u_1,u_2,u_3\}$。在地理上用户$\{u_1,u_2,u_3\}$不是均匀离散分布，攻击者可以将匿名集划分成不同部分进行攻击，最坏的情况是攻击者第一次攻击就选择了用户所在的部分，识别用户的概率则由1/3增加到1/2，增加了隐私泄露的风险。

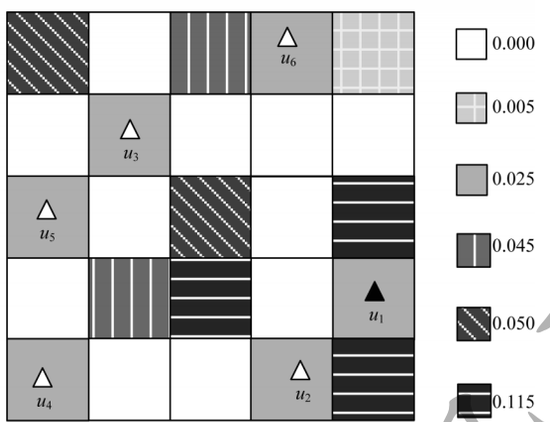

图3 3-匿名哑元位置选择

Fig.3 Location selection of three-anonymous dummy

如果构造的匿名集过于聚集，且位于某个特定的区域，攻击者可以通过地图信息获取与用户相关的隐私信息。同一区域哑元位置如图4所示，右侧数值代表对应位置的历史查询概率，根据3个位置都处在大学可以推测出用户的身份信息和相关学校，进一步推测更多的隐私信息。

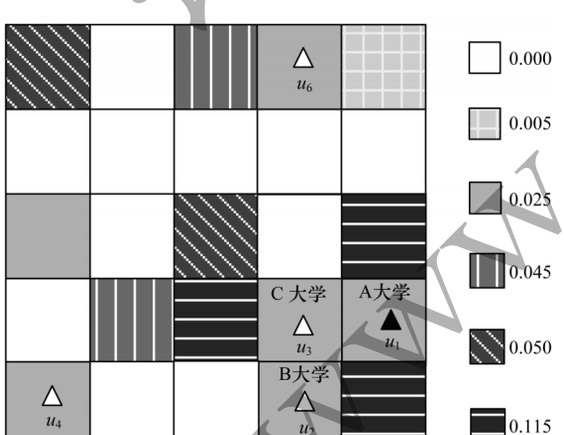

图4 同一区域哑元位置选择

Fig.4 Dummy location selection in the same region

因此，单一考虑位置单元的历史查询概率通过熵度量的方式构造哑元匿名集，无法抵御攻击者针对零查询用户的特殊地理信息攻击，同时无法避免由于匿名集中的位置分布不均造成离散的哑元位置易被过滤的问题。为了保证匿名度K，本文在基于熵度量的哑元选取原则上，加入了对零查询概率以及位置离散度两个方面的考量，设计一种改进算法。

2 攻防模型

位置隐私保护算法的设计也需要加入对常用攻击模型的考量，基于相应的攻击方式设计保护算法的防御思路。

2.1 攻击方式

攻击方式主要有边信息攻击和位置同质攻击。

1)边信息攻击

在位置隐私领域，边信息与真实位置不直接相关联，但是攻击者能利用它们筛选出更有价值的位置服务数据，达到辅助攻击的效果[21]。边信息攻击[22]是指攻击者根据已知边信息筛除一些不合理位置，增加获取用户真实位置信息的概率。假设位置匿名度为K，当选择的$K-1$个哑元位置中有多个哑元位置被攻击者过滤，则不满足K-匿名要求，降低隐私保护水平。

2)位置同质攻击

位置同质攻击[22]是指攻击者通过分析匿名集中的多个位置信息来缩小匿名区域。如果位置间非常接近，即使可以达到K匿名度，但是隐匿区域太小，攻击者就能获取相关的隐私信息；其次攻击者可以通过位置聚类的方法进行分类推理来增加推测用户真实位置的概率。

2.2 位置概率模型

对于含攻击者的位置隐私保护模型，隐私源为网格化地图的概率分布l，随机变量l的取值表示用户u的真实位置是网格化地图中某个哑元位置l_i，$\{l_1,l_2,\cdots,l_m\}$表示网格化的地图位置集。假设每个哑元位置对应的概率为p_i，则l概率模型如式(8)所示：

$$\begin{pmatrix} l \\ P \end{pmatrix}=\begin{pmatrix} l_1,l_2,\cdots,l_i,\cdots,l_k,\cdots,l_m \\ p_1,p_2,\cdots,p_i,\cdots,p_k,\cdots,p_m \end{pmatrix}$$

$$0\leqslant p_i\leqslant 1,\ \sum_{i=1}^{m}p_i=1 \tag{8}$$

用户的真实位置分布信息是隐私信息，为了防止攻击者直接获取用户的真实位置信息，需要对用户的位置进行匿名处理，将用户的位置泛化成可被攻击者直接观察到的位置分布l'，则l'的概率模型如式(9)所示：

$$\begin{pmatrix} l' \\ P \end{pmatrix}=\begin{pmatrix} l'_1,l'_2,\cdots,l'_i,\cdots,l'_k \\ p'_1,p'_2,\cdots,p'_i,\cdots,p'_k \end{pmatrix},\ 0\leqslant p'_i\leqslant 1,\sum_{i=1}^{k}p'_i=1 \tag{9}$$

攻击者获得用户的可观察位置分布l'后，结合相关的边信息对用户u进行位置攻击，攻击者得到用户的推断位置分布$\hat{l}$，设$\hat{l}=\{\hat{l}_1,\hat{l}_2,\cdots,\hat{l}_k\}$，$p_{\hat{l}_i}$表示攻击者推测的位置，$\hat{l}_i$为用户真实位置的概率，$\hat{l}$对应的概率模型如式(10)所示：

$$\begin{pmatrix}\hat{l}\\P\end{pmatrix}=\begin{pmatrix}\hat{l}_1,\ \hat{l}_2,\cdots,\hat{l}_i,\cdots,\hat{l}_k\\p_{\hat{l}_1},p_{\hat{l}_2},\cdots,p_{\hat{l}_i},\cdots,p_{\hat{l}_k}\end{pmatrix},\ 0\leqslant p_{\hat{l}_i}\leqslant 1,\ \sum_{i=1}^{k}p_{\hat{l}_i}=1 \tag{10}$$

具有边信息攻击者的攻防模型结构如图5所示。

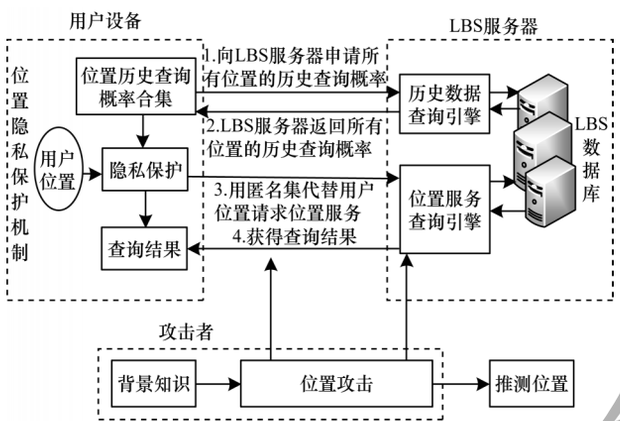

图5　攻防模型结构

Fig.5　Structure of attack and defense model

3　本文算法设计

本文算法将位置隐私保护的用户分为两种类型:1)用户发起位置服务请求所在位置单元的历史查询概率为零的零查询用户,零查询用户在进行哑元位置匿名集构造之间,根据偏移算法进行位置偏移,选择合适的偏移位置;2)用户发起位置服务请求所在位置单元的历史查询概率不为零,通过此位置单元的历史查询概率构造匿名集,无论是零查询用户还是非零查询用户,最后的匿名集要保证哑元位置地理上尽可能的均匀离散分布,构造哑元匿名集形成匿名区域要尽可能大。

3.1　最优哑元匿名候选集算法

哑元选取算法的目标是保护用户的位置信息不被泄露。攻击者通过重构匿名集与用户匿名集最相似的位置单元作为用户的真实位置,从而获取用户隐私信息。

最优的哑元匿名候选集算法是基于熵度量考虑攻击者可能掌握边信息的情况下,根据输入网格化的数据集S,用户的真实位置l,以及隐私匿名度K,利用增强型贪心算法高效地选取哑元位置,最后构造具有最大熵值哑元匿名集。

算法1　最优哑元匿名候选集算法

输入　地图数据集的网格化位置历史查询概率集合S,用户的位置l,隐私匿名度K

输出　最优的哑元匿名候选集C_h

1. 将集合S按照历史查询概率降序排序
2. K=K×2
3. K_b ← 从集合S中选择和用户位置的历史查询概率相同的位置
4. if size(K_b) ⩾ K, then
5. C_h ← 随机从K_b中选择K−1个位置作为哑元位置
6. 将用户的位置随机插入到集合C_h中
7. end if
8. if 1 < size(K_b) < k, then
9. $C_h \leftarrow C_h \cup K_b$, $S \leftarrow S/K_b$
10. $C_h \leftarrow C_h \cup l$, $S \leftarrow S/l$
11. D ← 从S集合中选择K−1个历史查询概率小于用户真实位置l的位置以及K−1个历史查询概率大于用户真实位置l的位置
12. for i = 1, i < K−2, i++, do
13. $P_{max} \leftarrow \max(C_h)$
14. $P_{min} \leftarrow \min(C_h)$
15. 从候选集D中找出一个位置,其历史查询概率是小于P_{min}的所有位置中的最大历史查询概率,用$P_{min-max}$表示
16. 从候选集D中找出一个位置,其历史查询概率是大于P_{max}的所有位置中最小历史查询概率,用$P_{max-min}$表示
17. if $H(C_h, P_{max-min}) > H(C_h, P_{min-max})$, then
18. $C_h \leftarrow C_h \cup P_{max-min}$, $D \leftarrow D/P_{min-max}$
19. else
20. $C_h \leftarrow C_h \cup P_{min-max}$, $D \leftarrow D/P_{min-max}$
21. end if
22. end for
23. else
24. S_b ← 从S集合中选择4K−ω−ε个候选位置,其中2K−ω个位置的历史查询概率小于用户位置的历史查询概率,2K−ε个位置的历史查询概率大于用户位置的历史查询概率
25. 从候选集S_b中随机选择一个位置i
26. $C \leftarrow H \cup i$, $S_b \leftarrow S_b/i$
27. for j = 1, j ⩽ k−2, j++, do
28. 从S_b集合中选择一个位置h,使得H(C, h)具有最大的值
29. $C_h \leftarrow C_h \cup h$, $S_b \leftarrow S_b/h$
30. end for
31. end if

最优哑元匿名候选集C_h。

3.2　零查询用户位置偏移算法

零查询用户位置偏移算法主要针对零查询用户发起服务请求时,根据用户给定的位置偏移距离$D_{\mathrm{Dis}}(l,l')$,选择合适的偏移位置l'代替用户的真实位置l。用户在进行请求时,通过最大服务质量损失$Q_{\mathrm{loss}}^{\max}$,给定算法最大的位置偏移距离$D_{\mathrm{Dis}}(l,l')$。首先算法根据用户给定的最大位置偏移距离,构造符合条件的偏移位置候选区C′;其次通过候选区构造候选偏移位置集,筛除候选集中历史查询概率为零的位置;最后随机在候选集中选择一个位置l'作为用户的替代位置,增加偏移位置的随机性。

算法2 零查询用户位置偏移算法

输入 地图数据集S,用户的位置l,用户最大偏移距离$D_{\mathrm{Dis}}(l,l')$

输出 偏移位置l'

1. 将地图数据集划分成的网格

2. 计算各个网格位置的查询次数

3. 计算各个网格位置的历史查询概率

4. 偏移位置候选集合C′←选择与l为中心,$\mathrm{D}_{\mathrm{Dis}}(\mathrm{l/l'})$为半径的区域中除用户外的所有网格位置形成偏移位置候选集C′

5. 偏移位置集合C′←将C′中历史查询概率为零的网格位置剔除

6. l′←在集合C′中随机选择一个位置

7. l′

3.3 K-匿名哑元位置选取算法

算法1构造的匿名集虽具有最大的熵值,但其哑元位置的地理离散度没有得到保障。攻击者即使没有获取用户的真实位置,也能根据该区域相关信息推测用户的隐私信息。除此之外,哑元位置的选取在地理上如果出现明显的分类,也不利于隐私保护。

离散度最大哑元匿名集构造算法的目的是尽可能选取距离较远的位置单元,保证哑元匿名集的位置离散度尽可能高。在选取哑元位置的过程中,计算两个位置之间的距离实现位置离散分布和,但这种方法在某些情况下无法获得最优的离散度。离散哑元位置选取示意图如图6所示。

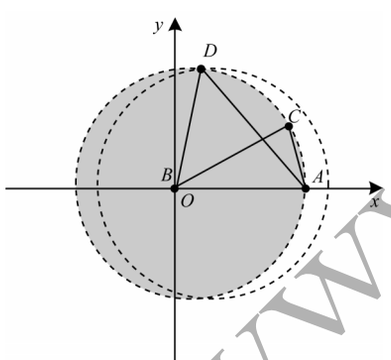

图6 离散哑元位置选取示意图

Fig.6 Schematic diagram of discrete dummy location selection

用户的位置为点B,当已经选择A作为哑元位置以后,选择C点和D点与A点、B点构成的哑元匿名集的熵值相同,如式(11)所示:

$$D_{\mathrm{Dis}}(C,A)+D_{\mathrm{Dis}}(C,B)=D_{\mathrm{Dis}}(D,A)+D_{\mathrm{Dis}}(D,B) \tag{11}$$

哑元位置的选取考虑到地理分布对于隐私的影响,选择D点优于选择C点加入哑元匿名集,如式(12)所示:

$$D_{\mathrm{Dis}}(C,A)\times D_{\mathrm{Dis}}(C,B)<D_{\mathrm{Dis}}(D,A)\times D_{\mathrm{Dis}}(D,B) \tag{12}$$

所以相对于选择两个位置的距离和,本文算法选择距离积作为构造匿名集的哑元选取方法。

本文构造离散度最大哑元匿名集$C_{\mathrm{list}}=\{l_1,l_2,\cdots,l_k\}$,描述为一个多目标优化问题,如式(13)所示:

$$\max\left\{-\sum_{i=1}^{k}P_i\times\mathrm{lb}\,P_i,\prod_{i\neq j}D_{\mathrm{Dis}}(l_i,l_j)\right\} \tag{13}$$

其中:P_i、P_j分别为l_i、l_j对应的历史查询概率。由于哑元位置的选取同时满足上述两个目标非常困难,本文将上述问题分解成两个部分,首先根据熵度量构造具有最大熵值的最优的$2K$哑元匿名候选集C_h,如式(14)所示:

$$C_h=\arg\max\left\{-\sum_{i=1}^{2K}P_i\times\mathrm{lb}\,P_i\right\} \tag{14}$$

基于熵度量构造具有最大熵值的最优的$2K$哑元匿名候选集C_h,通过位置距离乘积选择K个最离散的哑元位置匿名集C_{list},如式(15)所示:

$$C_{\mathrm{list}}=\arg\max\left\{\prod_{i\neq j}D_{\mathrm{Dis}}(l_i,l_j)\right\} \tag{15}$$

基于以上分析,本文首先根据用户的真实位置,综合使用算法1和算法2。通过算法1判断是否进行位置偏移,然后通过算法2选择$2K$个具有最大熵值的匿名候选集,最后选择K个具有最大离散度的哑元匿名集,从而构建基于历史查询概率的哑元选取算法。

本文进行K个最大离散度哑元匿名集的构造时,首先将用户的真实位置或者偏移位置加入匿名集C_{list},然后通过轮选择$K-1$个哑元位置加入匿名集C_{list}。

针对每一轮,本文对于候选集C_h中所有的位置单元都要计算其到C_{list}所有位置的乘积,然后计算每个位置单元占候选集C_h中所有的位置单元到C_{list}所有位置乘积之和的比率$\varpi_i=\frac{\prod_{i\in C_{\mathrm{list}},j\in C_h}D_{\mathrm{Dis}}(l_i,l_j)}{\sum_{i\in C_{\mathrm{list}}}\prod_{j\in C_h}D_{\mathrm{Dis}}(l_i,l_j)}$,最后构造候选集对应的候选概率矩阵,根据概率矩阵中的概率选择一个哑元位置加入匿名集。

本文每选择一个哑元位置都要对应计算一个概率矩阵,经过$K-1$轮形成包含K个用户的匿名集。在第一轮时,概率矩阵是通过候选集所有位置到用户位置的距离计算形成,其他轮则都是通过计算距离乘积。

算法3 基于历史查询概率的K-匿名哑元选取算法

输入 网格化地图数据集,用户的位置l,隐私匿名度K

输出 离散度最大哑元匿名集C_{list}

1. if $P_l = 0$, then
2. $l' \leftarrow$ 算法2选择零查询用户的偏移位置
3. $C_{list} \leftarrow l'$
4. else
5. $C_{list} \leftarrow l$
6. end if
7. $C_h \leftarrow$ 算法1选取出2K个候选位置
8. for i = 0, i < K−1, i++, do
9. 通过 $\bar{\omega}_i = \frac{\prod_{i \in C_{list}, j \in C_h} Dis(l_i, l_j)}{\sum_{i \in C_{list}} \prod_{j \in C_h} Dis(l_i, l_j)}$ 计算对应的候选概率矩阵，根据候选概率选择l_i作为候选位置单元
10. $C_{list} \leftarrow C_h \cup l_i$, $C_h \leftarrow C_h / l_i$
11. end for

4　安全性分析

假设匿名集C_{list}中的匿名度为K，因此攻击者从匿名集中推测出用户偏移位置的概率为$p_{ex} = 1/K$。设用户历史查询概率分布中的位置总数为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$o(n)$使得攻击者确定用户的真实位置的概率为p_{real}，满足$p_{real} \leqslant o(n)$。攻击者在获取用户的匿名集合C_{list}的情况下获得确切的用户位置信息的概率$p = p_{ex} p_{real} \leqslant o(n) \times (1/K)$。

攻击者根据已获得的地图信息和历史查询概率数据来推测哑元匿名集中用户的真实位置信息，本文算法能够有效应对此类进行攻击。在进行哑元位置的选取过程中，本文算法每次都选择与用户真实位置历史查询概率最相似或相等的位置单元作为哑元，即满足$p_i \approx p_j$（其中i表示用户，j表示哑元，p_i表示用户的历史查询概率，p_j表示匿名集中任一哑元的历史查询概率），根据式(6)可知，哑元位置的历史查询概率越接近，熵值越大，匿名集的信息混乱程度越高。本文算法在选择哑元时，将考虑熵值和地理离散度较高的位置选为哑元位置。因此，本文算法生成的哑元匿名集能抵抗攻击者基于历史查询概率的位置同质攻击和边信息攻击，有效控制攻击者推测出用户真实位置的概率约为$1/K$。

5　仿真实验

为验证本文算法的有效性，本文设计一系列仿真实验。实验采用Python3.8软件在PC机（Windows10操作系统，2.40 GHz Intel i7 CPU，12 GB内存）上进行模拟仿真。数据集为北美路网NA数据集，该数据集包含175 812个真实的位置记录，NA数据集分布如图7所示。实验将数据集网格化为100×100的网格地图，用户的最大偏移距离均设置为2。

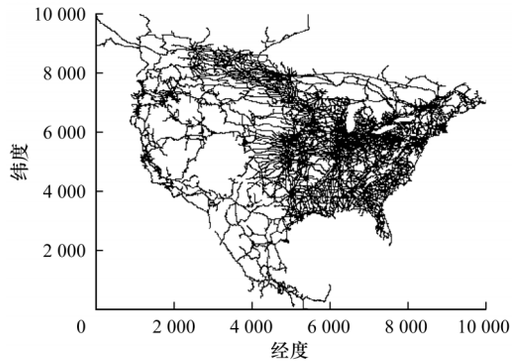

图7　NA数据集分布

Fig.7　Distribution of NA data set

5.1　离散度对比

离散度越大，哑元位置越分散。实验通过统计不同隐私度K下哑元匿名集位置单元的距离总和来衡量哑元匿名集的离散度。本文K-匿名哑元选取算法K-DLS、DLS算法、DLP算法、Enhanced_DLP算法、Optimal算法、CircleDummy算法、GridDummy算法在不同隐私度K下哑元匿名集的离散度对比如图8所示。

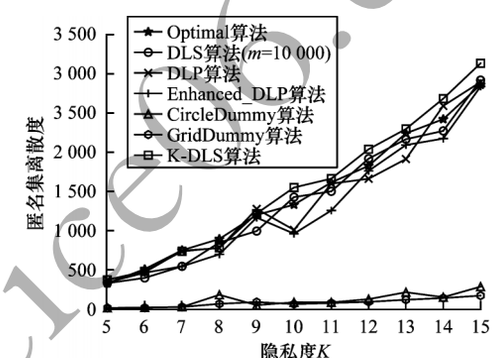

图8　不同算法的匿名集离散度对比

Fig.8　Anonymity set dispersion comparison among different algorithms

从图8可以看出，所有算法生成哑元匿名集的离散度都呈现上升趋势，在相同隐私度K下K-DLS算法构造的哑元匿名集离散度大于其他两种算法。随着K值增大，本文算法的匿名集离散度均高于其他算法。因为本文算法在选择哑元位置时，引入候选位置集与哑元位置集所有位置元素距离乘积作为指标，通过位置概率候选矩阵选择哑元位置，每次都选择了能构成最大离散度的位置作为哑元，保障生成的哑元匿名集具有最大的离散度。

5.2　哑元匿名集生成效率对比

哑元匿名集的生成效率是通过CPU的运行时间表示。为验证本文算法的哑元匿名集生成效率，K-DLS、DLP和Enhanced_DLP算法在不同隐私度K下生成时间对比如图9所示。

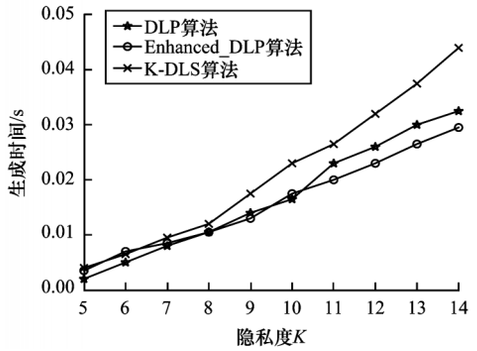

图9 隐私度K与匿名集生成时间

Fig.9 Privacy K and generation time of anonymity set

从图9可以看出,本文算法在构造哑元匿名集时,需要验证用户的历史查询概率,同时进行离散选择,故其运行时间比DLP和Enhanced_DLP算法的运行时间略高。本文算法在保证取得最大熵值的情况下,考虑了哑元位置的地理离散性,哑元匿名集的生成效果整体较好。

5.3 隐私保护熵值对比

在LBS的位置隐私保护方案中,位置熵可以衡量用户位置信息的不确定性,熵值越大表示哑元匿名集的隐私度越高。图10表示本文K-匿名哑元选取算法K-DLS、DLS算法、DLP算法、enhanced_DLP算法、Optimal算法、CircleDummy算法、GridDummy算法在不同隐私度K下生成哑元匿名集的熵值对比。

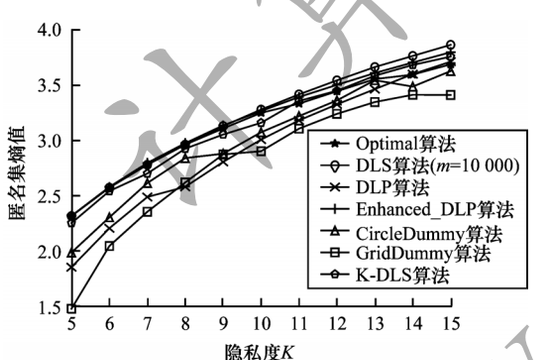

图10 不同隐私度K对匿名集熵值影响

Fig.10 The entropy of anonymity set versus different privacy K

从图10可以看出,随着K值不断增大,7种算法生成哑元匿名集的熵值整体呈现上升趋势。但是DLS、DLP、Enhanced_DLP、Optimal和K-DLS算法生成的熵值明显大于CircleDummy、GridDummy算法。K-DLS算法在保证哑元匿名集离散度最大的情况下,匿名集熵值也取得了最优的效果,能够有效保证用户的隐私。

5.4 攻击算法识别概率对比

攻击者获得用户的哑元匿名集C后,结合相关的边信息对用户进行位置攻击,攻击者得到用户的推断位置分布,根据分布概率确定用户的真实位置。K-DLS、DLS、DLP和Enhanced_DLP算法在不同隐私度K下生成哑元匿名集的位置识别率对比如图11所示,其中每个算法重复实验1 000次。

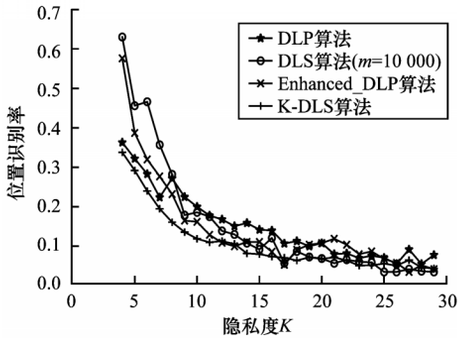

图11 不同隐私度K对位置识别率的影响

Fig.11 Location recognition probability versus different privacy K

从图11可以看出,K-DLS算法的位置识别率略低于其他算法,因为用户在进行位置隐私保护时,考察了历史查询概率为零的用户,对其进行位置偏移,随机在候选集中选择用户位置的偏移位置,增加匿名集的随机性;其次算法考量了位置离散度,生成哑元匿名集在地图上尽可能的均匀离散分布,攻击者很难通过位置聚类等特殊方法增加推测概率,能够有效提升用户位置的隐私保护效果。

6 结束语

本文分析现有位置隐私保护算法存在的不足,提出基于历史查询概率的K-匿名哑元位置选取算法。利用熵度量选择哑元位置,使得哑元匿名集熵值最大,并根据距离对匿名结果进行优化。仿真结果表明,与DLP、DLS等算法相比,K-DLS算法在保证匿名集离散度最大的情况下,匿名集熵值为最优,能够提高位置隐私保护的安全性。本文算法主要针对基于快照查询的LBS服务位置隐私保护,下一步将对用户连续查询移动轨迹的位置隐私保护进行研究。

参考文献

[1] CHEN L,THOMBRE S,JÄRVINEN K,et al. Robustness,security and privacy in location-based services for future IoT:a survey[J]. IEEE Access,2017,5:8956-8977.

[2] PARMAR D,RAO U P. Towards privacy-preserving dummy generation in location-based services[J]. Procedia Computer Science,2020,171:1323-1326.

[3] 张学军,桂小林,伍忠东. 位置服务隐私保护研究综述[J]. 软件学报,2015,26(9):2373-2395.
ZHANG X J,GUI X L,WU Z D. Privacy preservation for location-based services:a survey[J]. Journal of Software,2015,26(9):2373-2395. (in Chinese)

[4] PENG T,LIU Q,MEGN D C,et al. Collaborative trajectory privacy preserving scheme in location-based services[J]. Information Sciences,2017,384:165-179.

[5] 裴卓雄,李兴华,刘海,等. LBS隐私保护中基于查询范围的匿名区构造方案[J]. 通信学报,2017,38(9):125-132.
PEI Z X,LI X H,LIU H,et al. Anonymizing region construction scheme based on query range in location-based

service privacy protection[J]. Journal on Communications, 2017, 38(9): 125-132. (in Chinese)
[6] 裴媛媛，石润华，仲红，等. 面向位置服务的用户隐私保护[J]. 计算机工程，2015，41(10)：20-25.
PEI Y Y, SHI R H, ZHONG H, et al. User privacy protection for location-based service [J]. Computer Engineering, 2015, 41(10): 20-25. (in Chinese)
[7] 伍旭，罗敏. 稀疏环境下基于位置服务的专有资源方法[J]. 计算机工程，2017，43(5)：108-114.
WU X, LUO M. Privacy-preserving method of location based service in sparse environment [J]. Computer Engineering, 2017, 43(5): 108-114. (in Chinese)
[8] 王璐，孟小峰. 位置大数据隐私保护研究综述[J]. 软件学报，2014，25(4)：693-712.
WANG L, MENG X F. Location privacy preservation in big data era: a survey[J]. Journal of Software, 2014, 25(4): 693-712. (in Chinese)
[9] 马春光，周长利，杨松涛，等. 基于Voronoi图预划分的LBS位置隐私保护方法[J]. 通信学报，2015，36(5)：5-16.
MA C G, ZHOU C L, YANG S T, et al. Location privacy-preserving method in LBS based on Voronoi division[J]. Journal on Communications, 2015, 36(5): 5-16. (in Chinese)
[10] SWEENEY L. K-anonymity: a model for protecting privacy [J]. International Journal of Uncertainty Fuzziness & Knowledge Based Systems, 2002, 10: 557-570.
[11] ZHAO P, LIU W, ZHANG G, et al. Preserving privacy in WiFi localization with plausible dummy locations[J]. IEEE Transactions on Vehicular Technology, 2020, 69(10): 11909-11925.
[12] SUN G, CAI S, YU H, et al. Location privacy preservation for mobile users in location-based services [J]. IEEE Access, 2019, 7: 87425-87438.
[13] 王洁，王春茹，马建峰，等. 基于位置语义和查询概率的假位置选择算法[J]. 通信学报，2020，41(3)：53-61.
WANG J, WANG C R, MA J F, et al. Dummy location selection algorithm based on location semantics and query probability[J]. Journal on Communications, 2020, 41(3): 53-61. (in Chinese)
[14] ZHANG Y B, ZHANG Q Y, LI Z Y, et al. A k-anonymous location privacy protection method of dummy based on geographical semantics [J]. International Journal of Network Security, 2019, 21(6): 937-946.
[15] 刘海，李兴华，王二蒙，等. 连续服务请求下基于假位置的用户隐私增强方法[J]. 通信学报，2016，37(7)：140-150.
LIU H, LI X H, WANG E M, et al. Privacy enhancing method for dummy-based privacy protection with continuous location-based service queries[J]. Journal on Communications, 2016, 37(7): 140-150. (in Chinese)
[16] NIU B, LI Q, ZHU X, et al. Achieving k-anonymity in privacy-aware location-based services[C]//Proceedings of IEEE Conference on Computer Communications. Washington D. C., USA: IEEE Press, 2014: 754-762.
[17] SUN G, CHANG V, RAMACHANDRAN M, et al. Efficient location privacy algorithm for Internet of Things (IoT) services and applications[J]. Journal of Network & Computer Applications, 2017, 89: 3-13.
[18] DU Y W, CAI G, ZHANG X J, et al. An efficient dummy-based location privacy-preserving scheme for internet of things services[J]. Information (Switzerland), 2019, 10(9): 278-289.
[19] LU H, JENSEN C S, YIU M L. PAD: privacy-area aware, dummy based location privacy in mobile services[C]// Proceedings of the 7th ACM International Workshop on Data Engineering for Wireless and Mobile Access. New York, USA: ACM Press, 2008: 16-23.
[20] 邓密文. 融合边信息的双重匿名位置隐私保护方案[J]. 信息安全研究，2020，6(5)：421-426.
DENG M W. Double anonymous location privacy protection scheme with edge information [J]. Information Security Research, 2020, 6(5): 421-426. (in Chinese)
[21] 万盛，李凤华，牛犇，等. 位置隐私保护技术研究进展[J]. 通信学报，2016，37(12)：124-141.
WAN S, LI F H, NIU B, et al. Research progress on location privacy-preserving techniques[J]. Journal on Communications, 2016, 37(12): 124-141. (in Chinese)

编辑　薛晋栋