计算机科学
COMPUTER SCIENCE

抵御背景信息推理攻击的假位置生成算法

张学军, 杨依行, 李佳乐, 田丰, 黄海燕, 黄山

引用本文

张学军, 杨依行, 李佳乐, 田丰, 黄海燕, 黄山. 抵御背景信息推理攻击的假位置生成算法[J]. 计算机科学, 2023, 50(11A): 221000036-9.

ZHANG Xuejun, YANG Yixing, LI Jiale, TIAN Feng, HUANG Haiyan, HUANG Shan. Dummy Location Generation Algorithm Against Side Information Inference Attack [J]. Computer Science, 2023, 50(11A): 221000036-9.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.221000036

抵御背景信息推理攻击的假位置生成算法

张学军[1]　杨依行[1]　李佳乐[1]　田　丰[2]　黄海燕[1]　黄　山[3]
1 兰州交通大学电子与信息工程学院　兰州 730070
2 陕西师范大学计算机科学学院　西安 710062
3 兰州交通大学土木工程学院　兰州 730070

摘　要　针对已有的假位置生成算法,设计了一种多次查询请求攻击算法(Multiple Query Request Attack algorithm,MQRA)来测试其安全性。为有效保护用户的位置隐私,提出了一种抵御背景信息推理攻击的假位置生成算法(Dummy Location Generation Algorithm against Side Information Inference Attack,DLG_SIA),该算法综合考虑了查询概率、时间分布、位置语义和物理分散度等背景信息来生成有效的假位置集以抵御概率分布攻击、位置语义攻击和位置同质攻击,避免攻击者结合背景信息过滤掉假位置。用户首次请求时,DLG_SIA 算法先利用位置熵和时间熵选取当前请求时间下查询概率相似的位置点来生成假位置集,并通过调整的余弦相似度生成满足语义差异性的位置点;然后通过距离熵保证选取的位置点间具有更大的匿名范围,并将当前请求位置的最佳假位置集进行缓存。安全性分析和仿真实验结果表明:MQRA 算法能以很高的概率识别出假位置集中用户的真实位置;与已有的假位置生成算法相比,DLG_SIA 算法能有效抵御背景信息推理攻击,保护用户的位置隐私。

关键词:基于位置的服务;查询概率;位置语义;时间分布;物理分散度
中图法分类号　TP309

Dummy Location Generation Algorithm Against Side Information Inference Attack

ZHANG Xuejun[1],YANG Yixing[1],LI Jiale[1],TIAN Feng[2],HUANG Haiyan[1] and HUANG Shan[3]
1 School of Electronics and Information Engineering,Lanzhou Jiaotong University,Lanzhou 730070,China
2 School of Computer Science,Shaanxi Normal University,Xi'an 710062,China
3 College of Civil Engineering,Lanzhou Jiaotong University,Lanzhou 730070,China

Abstract　Aiming at the existing dummy location generation algorithm,a multiple query request attack algorithm (MQRA) is designed to test its security. In order to effectively protect user's location privacy,a dummy location generation algorithm against side information inference attack(DLG_SIA) is proposed. It comprehensively considers the side information such as query probability,time distribution,location semantics and physical dispersion to generate an effective dummy location set to resist probability distribution attacks,location semantics attacks and location homogeneity attacks,and avoid attackers filtering dummy locations with side information. When the user requests for the first time,the DLG_SIA algorithm first uses the location entropy and time entropy to select the location points with similar query probability at the current request time to generate a dummy location set,and then uses the adjusted cosine similarity to generate the location points that meet the semantic differences. Next,distance entropy is used to ensure that the selected location points have a larger anonymous range,and the best dummy location set of the current request location is cached. Security analysis and simulation results show that MQRA algorithm can identify the real location of users in the dummy location set with high probability. Compared with the existing dummy location generation algorithm,DLG_SIA algorithm can effectively resist the side information inference attack and protect the user's location privacy.

Keywords　Location-based service,Query probability,Location semantics,Time distribution,Physical dispersion

1　引言

随着 5G 时代的到来和移动定位技术的空前繁荣,基于位置的服务(Location-Based Service,LBS)作为一种新型网络服务模式,已成为移动互联网的基础服务和标准配置。虽然 LBS 为用户提供了多样化的服务,但也引发了严重的隐私泄露问题。为了获取相关服务,用户需要向不可信或恶意的位置服务提供商(本文定义为攻击者)提交包含位置信息及查询

基金项目:国家自然科学基金(61762058,61901201);甘肃省自然科学基金(21JR7RA282);兰州交通大学百人青年人才培养计划基金;甘肃省教育厅产业支撑计划项目(2022CYZC-38);中央高校基本科研业务费(GK202103090);陕西省自然科学基础研究计划项目(2022JM-329)
This work was supported by the National Natural Science Foundation of China(61762058,61901201),Natural Science Foundation of Gansu Province(21JR7RA282),Foundation of A Hundred Youth Talents Training Program of Lanzhou Jiaotong University,the Education Department of Gansu Province:Industrial Support Plan Project(2022CYZC-38),Fundamental Research Program of the Central Universities (GK202103090) and Natural Science Basic Research Project of Shaanxi Province(2022JM-329).
通信作者:张学军(xuejunzhang@lzjtu.edu.cn)

属性在内的服务请求。然而，附加在服务请求上的上下文信息会披露用户的生活习惯、兴趣爱好及健康状况等敏感信息[1]。如何在享受高质量 LBS 服务的同时对用户的隐私信息进行保护受到了学术界和工业界的广泛关注。

针对用户位置隐私保护问题，国内外研究学者提出了许多隐私保护策略[2-7]，并有效保护了用户的位置隐私。结合 k 匿名技术的位置隐私保护方法是其中最常用的技术之一，其核心思想是由用户的真实位置与其他 $k-1$ 个无法区分的位置来构建假位置集[8-12]，这样不可信的位置服务提供商很难从假位置集中区分出用户的真实位置。但是，位置 k 匿名隐私保护方法[13-15]在面对攻击者掌握的背景信息时，如用户的历史查询概率、地图信息等[2]，存在严重的隐私泄露风险。为此，Niu 等[16]提出了基于熵度量的假位置选择(Dummy Location Selection，DLS)算法，通过用户位置历史查询概率背景信息，构建具有相同查询概率的假位置集来迷惑攻击者，最终有效地保护了用户的位置隐私。但是，DLS 算法的计算开销很高，而且攻击者可以利用 ADLS[17]算法以很高的概率从 DLS 生成的假位置集中识别出用户的真实位置。Sun 等[17]针对 ADLS 攻击算法，提出了新的假位置隐私保护(Dummy Location Privacy-preserving，DLP)算法，该算法在熵度量的基础上利用贪心策略搜索一个最优的假位置集，在有效抵御 ADLS 攻击的同时减少了计算开销。Du 等[18]在 DLP 算法的基础上，引入了增强型贪心算法，设计出 Enhanced-DLP 算法，该算法较 DLP 算法在响应时间上有较大的提升。Xia 等[19]设计了一种基于假位置和 Stackelberg 博弈的位置匿名算法，该算法将用户位置隐私保护模型和攻击者位置推测模型进行博弈，从而得到了更为优化的匿名结果。然而当假位置集中用户真实位置与 $k-1$ 个假位置为同一语义类型时，即使攻击者不知道用户的真实位置，其仍会暴露用户的其他隐私信息。为使生成的假位置集满足语义差异性，Chen 等[9]和 Wang 等[20]利用语义距离来刻画假位置集的语义差异性，并采用树形组织结构计算位置节点间的跳数来获取语义距离。但是该算法需要提前构建位置语义树，适用于兴趣点类型多的地区，在有些地区可能会匿名失效。Shi 等[21]提出了一个基于位置语义量化的假位置生成算法，通过不同时间段内某个位置区域内访问的用户数量来构建位置语义向量，并使用余弦相似度来度量两个位置语义的相似性，但是当假位置集中包含沙漠、湖泊等查询概率为零的位置时，攻击者很容易将它们过滤掉。另外，如果用户在同一位置发起多次 LBS 查询请求，攻击者能够利用生成的多个假位置集以很高的概率识别出用户的真实位置。

本文首先分析了针对假位置生成算法的多次查询请求攻击算法，通过实验验证了 MQRA 算法识别用户真实位置的有效性；然后充分考虑了攻击者可能拥有的关于用户 LBS 查询的相关背景信息以及 MQRA 推理攻击，提出了一种新的假位置生成算法。该算法在首次请求 LBS 查询后会缓存当前位置的最佳假位置集，在下次请求时直接以缓存的假位置集发起查询请求，生成的假位置集满足：1)没有查询概率为零的假位置；2)各位置点在当前请求时间下查询概率相似；3)具有语义差异性；4)各位置点间的地理位置尽可能分散。所提算法解决了攻击者通过背景信息和推理攻击过滤部分假位置的问题，能有效保护用户的位置隐私。仿真实验从位置熵、时间熵、语义差异性、物理分散度、平均匿名时间、攻击者的推理能力和 MQRA 算法的成功识别率 7 个方面将所提算法与已有的假位置生成算法进行对比，验证了所提算法抵御背景信息推理攻击的有效性。

2 预备知识

2.1 系统架构

随着无线通信技术的发展，政府为公众提供了无处不在的 Wi-Fi，因其具有较强的计算能力和存储能力，已在文献[9,16,20]等 LBS 保护方法中广泛应用。本文的系统架构主要由卫星定位系统、Wi-Fi 接入点(Access Point，AP)、移动终端和 LBS 服务器组成。

在本文架构中，移动终端被认为是可信的，通过卫星定位系统确定当前位置信息；由于位置语义信息和历史查询概率之类的背景信息不经常发生变化，因此采用 Wi-Fi AP 收集其覆盖范围内的地图信息、位置语义信息和历史查询概率；LBS 服务器为用户提供服务，通常被认为是不可信的。

具体过程如图 1 所示，用户在发起请求前根据定位系统获取自身位置，从 Wi-Fi AP 获取当前覆盖范围内的地图信息、位置语义信息和历史查询概率；并根据相应的位置隐私保护算法生成一个满足自身隐私需求的假位置集发送给 LBS 服务器进行服务请求；LBS 服务器解析请求内容，查询出相应的结果集返回给移动终端，经移动终端处理后用户获取到当前请求的最终结果。

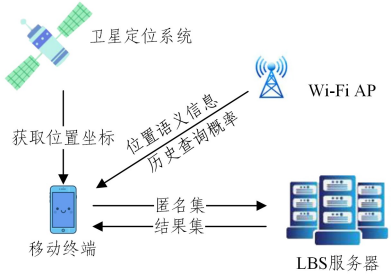


图 1 系统架构

Fig. 1 System structure

2.2 假位置集生成原则

定义 1(查询概率) 将实验区域作为样本空间划分为 $n\times n$ 的网格地图，每个网格确定一个位置单元，图 2 是划分的一个 5×5 的样本空间地图。

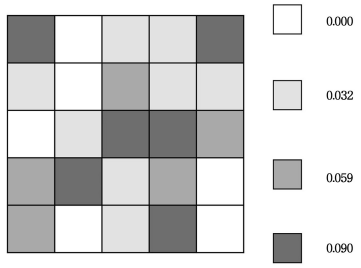


图 2 样本空间划分

Fig. 2 Sample space division

使用不同灰度表示不同的查询概率，颜色越深表示历史查询概率越大。位置单元 i 的历史查询概率 p_{L_i} 的计算如式(1)所示：

$$p_{L_i}=\frac{num(i)}{\sum_{i=1}^{n^2}num(i)} \tag{1}$$

其中,$num(i)$表示在过去一段时间内位置单元 i 被查询的次数,$\sum_{i=1}^{n^2}num(i)$表示所有位置单元在过去一段时间内被查询的总次数,且$\sum_{i=1}^{n^2}p_{L_i}=1$,即所有位置单元的历史查询概率之和为 1。

定义 2(零查询位置) 划分的样本空间中查询概率为零的位置单元在本文中定义为零查询位置(如图 2 中的白色单元格)。假设攻击者可以过滤零查询位置,用户在零查询位置进行服务请求,根据现有的匿名方法生成的假位置也是零查询位置,因此攻击者的成功识别率会提高。

定义 3(时间分布) 通常,不同时间段用户对同一单元格的访问概率不同。为避免攻击者通过时间分布识别出假位置集中的真实位置,本文将查询概率按照时间进行划分,其定义如式(2)所示:

$$p_{L_{it}}=\{p_{L_{i1}},p_{L_{i2}},\cdots,p_{L_{i24}}\} \tag{2}$$

其中,$p_{L_{it}}$表示位置单元格 i 在 t 时刻的查询概率。

定义 4(位置语义) 位置语义指当前位置聚集的用户所具有的相似行为或属性(如医生聚集在医院)。本文将位置语义划分为 4 种类型,分别为餐饮娱乐、教育科学、行政办公和医疗救护。每个位置单元格中 4 种类型的语义位置占比不同,假设用户所在单元格医疗救护的占比最高,生成的假位置也是医疗救护的占比最高,攻击者可以根据位置语义类型推断出用户的隐私信息。为避免此类情况发生,本文采用调整的余弦相似度来构建具有语义差异性的假位置集,如式(3)所示:

$$Sem(u,d)=\frac{\sum_{c\in K}(p_{uc}-\overline{p_u})(p_{dc}-\overline{p_d})}{\sqrt{\sum_{c\in K}(p_{uc}-\overline{p_u})^2}\sqrt{\sum_{c\in K}(p_{dc}-\overline{p_d})^2}} \tag{3}$$

其中,$Sem(u,d)$表示单元格 u 和 d 的语义相似度,K 为包含 4 种类型语义位置的集合,c 表示 K 中的某种语义类型,p_{uc}和 p_{dc}分别表示语义类型为 c 时位置单元 u 和 d 的访问概率,$\overline{p_u}$和$\overline{p_d}$分别表示 K 集合中所有语义类型在位置单元格 u 和 d 处访问概率的均值。

定义 5(距离熵) 距离熵用来度量物理分散度,熵值越小,假位置集中各位置点间的距离越分散。因此,在生成假位置集时选取熵值小的候选集能提高隐私保护度。距离熵的计算公式为:

$$HD(C)=-\sum_{i=1}^{k}lbd(C_i,C_{\text{center}})\times d(C_i,C_{\text{center}}) \tag{4}$$

其中,k 为位置点的个数,lb 为以 2 为底的对数,C_{center}表示假位置集 C 的最小覆盖圆的圆心,$d(C_i,C_{\text{center}})$表示位置点 C_i 到 C_{center}的欧氏距离。

2.3 攻击模型

(1)隐私威胁

由于 LBS 服务器可能会被恶意攻击者俘获或者出于利益考虑去分析挖掘用户的敏感信息,所以本文假设移动终端是可信的,而 LBS 服务器是不可信(忠实但好奇的)的攻击者,其目标是依赖地图信息、历史查询概率及位置语义等背景信息从用户提交的位置查询请求中推理出用户的隐私信息。隐私威胁主要来自 3 个方面:1)概率分布攻击[22],攻击者具有当前地图的历史查询概率,针对假位置集的查询概率进行分析,认为查询概率较高的位置大概率为用户真实位置;2)位置语义攻击[22],攻击者根据位置语义信息,结合假位置集的语义类型,当语义类型一致时可以推断出用户真实位置的语义信息,从而造成用户隐私泄露;3)位置同质攻击[23],攻击者分析匿名区域内的多个位置点,若距离非常接近,即使假位置集满足 k 匿名的要求,用户的位置隐私仍然无法得到有效保护。

(2)攻击者的推理能力

针对不同的位置匿名隐私保护算法,攻击者从生成的假位置集中推测出用户真实位置的概率是不同的。假设事件 X 表示攻击者从假位置集中识别出用户真实位置,用 $P(X=0)$表示攻击者识别的失败概率,$P(X=1)$表示攻击者识别的成功概率。理想情况下,$P(X=0)$的最大概率为$(k-1)/k$,$P(X=1)$的最大概率为 $1/k$,则攻击者的推理能力可以表示为事件 X 的期望 $E(X)$,如式(5)所示:

$$E(X)=\frac{1}{k} \tag{5}$$

3 MQRA 算法

假位置生成算法通常是用户在当前位置生成不可区分的 $k-1$ 个假位置,并和用户真实位置组成 k 个假位置集 C 进行位置服务请求。MQRA 的主要目标是通过获取用户在同一位置发起的多次查询请求服务,从假位置集 C 中识别出用户的真实位置,即攻击者从假位置集 C 中成功识别出用户真实位置的概率大于 $1/k$。

MQRA 算法获取用户在同一位置发起多次查询请求时生成的 k 个假位置集 $C_i(1\leqslant i\leqslant n)$,最终得到集合 $C_{\text{list}}=\{C_1,C_2,\cdots,C_n\}$,其中 n 为请求次数。由于每次请求生成的假位置集 C_i 互有差异,且用户真实位置必定存在于 C_i 中,因此攻击者对 n 个假位置集 C_i 求交集得到 $C_{\text{update}}=C_1\cap C_2\cap\cdots\cap C_n$,且 $1\leqslant|C_{\text{update}}|\leqslant k$,其中$|\cdot|$表示集合大小,即攻击者识别出用户真实位置的概率可能会大于 $1/k$。最坏情况为 C_{list}中的所有假位置集都一样,即更新得到的 C_{update}与 C_{list}中的任意一个假位置集都相等,则攻击算法失效,实际中这种情况出现的概率极小。

以文献[16]提出的 DLS 算法为例说明 MQRA 算法的有效性。DLS 算法通过从排序的 $2k$ 个查询概率相同的位置点中随机选择 k 个位置点构成假位置集,并选取熵值最大的一组作为最终生成的假位置集。由于假位置集是随机选取的,经过 MQRA 算法后,C_{list}中至少存在一个假位置集 $C_i(1\leqslant i\leqslant n)$与其他假位置集中的位置点间存在差异,则更新假位置集 C_{update}的大小小于隐私保护度 k,攻击者识别出用户真实位置的概率大于 $1/k$。MQRA 算法伪代码如算法 1 所示。

算法 1 MQRA 算法

输入:DLS 算法,运行次数 n

输出:更新假位置集 C_{update}

1. $C_{update}\leftarrow$Run(DLS);
2. for(i=2;i<=n;i++)do;
3. $C_i\leftarrow$Run(DLS);
4. $C_{update}=C_{update}\cap C_i$;
5. end for;
6. return C_{update};

图 3 给出了利用 MQRA 算法攻击 DLS 算法的实验结果。本文将实验区域划分成 100×100 的大小相等的位置单元,假设每个单元格已经存在历史查询概率。k 设置为 2~20,每次实验运行 DLS 算法 10 次,重复实验 1 000 次取平均值。

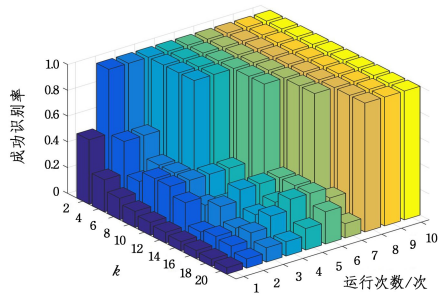


图 3 MQRA 成功识别率

Fig. 3 MQRA success recognition rate

由图 3 可以看出，当 k 值不变时，随着运行次数的增加，MQRA 算法的成功识别率也在增加；当运行次数不变时，随着 k 值的增大，MQRA 算法的成功识别率逐渐降低，但是在运行两次时 MQRA 算法已达到攻击目的，即成功识别率大于 $1/k$。因此，本文所提 MQRA 算法能够有效降低 DLS 算法的隐私保护强度。

4 DLG_SIA 算法

为解决现有假位置生成算法在面对多种位置推理攻击时存在的隐私泄露问题，本文通过引入缓存机制，提出了一种抵御背景信息推理攻击的假位置生成算法 DLG_SIA，该算法先判断当前请求是否命中缓存，若命中则直接使用缓存的假位置集发起请求，否则采用边生成边判断的策略，综合零查询特殊位置、查询概率、时间分布、位置语义及物理分散度等因素输出一个最佳假位置集并进行缓存。为抵御拥有背景信息的攻击者，需确保假位置集的位置熵和时间熵最大、距离熵最小且满足语义差异性。DLG_SIA 算法的基本思想如下：

1)在发起请求前，移动终端先判断当前请求位置是否命中缓存，若是则以缓存的假位置集发起请求；否则判断是否为定义 2 定义的零查询位置，若是则选择距离请求位置最近且查询概率不为零的位置单元作为用户真实位置，然后生成假位置。

2)确定用户真实位置后，选取 m 组 $6k$ 大小的候选集 D_6，以 D_6 的位置熵为 key，候选集 D_6 为 value 添加到 map 集合中，从 map 中选取位置熵最大的 value 作为候选集 C_6；生成 m 组 $4k$ 大小的候选集 D_4，要求当前请求时间 t 下候选集中的位置单元与用户真实位置的查询概率相似，并选取时间熵最大的一组集合作为候选集 C_4。上述过程保证了候选集 C_4 能够抵御概率分布攻击。

3)移动终端通过定义 4 判断 C_4 中任意两个位置单元之间是否满足用户设置的语义阈值，选取 $4k$ 个满足语义阈值 θ 的位置点构成集合 C_{sem} 并升序排列，取前 $2k$ 个位置点构成候选集 C_2。该过程确保了生成的候选集 C_2 在抵御概率分布攻击的基础上能够抵御位置语义攻击。

4)在候选集 C_2 的基础上，移动终端通过 C_2 中各位置点构成的最小覆盖圆的圆心到各位置点间的距离计算定义 5 提出的距离熵，并选取熵值最小的集合与用户真实位置构成假位置集 C。该过程确保了生成的假位置集 C 还可以抵御位置同质攻击。

(5)最终输出能够抵御背景信息推理攻击的最佳假位置集 C，将该假位置集进行缓存。

DLG_SIA 算法伪代码如算法 2 所示。

算法 2 DLG_SIA 算法

输入：用户所在位置单元 L_u，请求时间 t，隐私匿名度 k，排序查询概率集 p_L，候选集个数 m，语义阈值 θ，时间分布 p_{L_t}

输出：最佳 k 假位置集 C

```
1. if(Cached.contains(Lu));
2.     return Cached.get(Lu);
3. end if;
4. if(pLu==0);
5.     Lu=Lu', (pLu'>0) and min(dis(Lu,Lu'))/*对零查询位置发起请求的位置点进行转移处理*/;
6. end if;
7. D8←从 pLu 左右两侧选取 8k 个位置点;
8. for(i=1;i<=m;i++) do;
9.     D6i←从 D8 中随机选取 6k 个位置点;
10.    map.put(HP(D6i),D6i)/* HP(D6i)表示根据式(12)计算的候选集 D6i 的位置熵*/;
11. end for;
12. C6←map.get(max(HP(D6i)));
13. for(i=1;i<=m;i++) do;
14.    D4i←从 C6 中选取 4k 大小的位置点，要求 pLit≈pLut;
15.    map.put(HT(D4i),D4i)/* HT(D4i)表示根据式(13)计算的候选集 D4i 的时间熵*/;
16. end for;
17. C4←map.get(max(HT(D4i)));
18. for(i=1;i<=|C4|;i++) do;
19.    if(Sem(Lu,C4.get(i))<θ);
20.        Csem←C4.get(i)/*将满足语义阈值 θ 的位置单元添加到集合 Csem 中*/;
21.    end if;
22. end for;
23. C2←ASC(Csem),取前 2k 个位置点/*ASC(Csem)表示将 Csem 按升序排列*/;
24. for(i=1;i<=m;i++) do;
25.    Di←从 C2 随机选取 k-1 个位置点;
26.    计算 Di 的最小覆盖圆的圆心 Ccenter;
27.    map.put(HD(Di),Di)/*HD(Di)表示根据式(4)计算的候选集 Di 的距离熵*/;
28. end for;
29. C←map.get(min(HD(Di)))∪Lu;
30. Cached.put(Lu,C);
31. return C;
```

该算法第 1－3 行判断当前请求是否命中缓存，是则以缓存的假位置集发起请求；算法第 4－6 行零查询特殊位置进行转移处理，选取查询概率大于零且最近的位置点发起请求；算法第 7－12 行生成位置熵最大的候选集 C_6，其时间复杂度为 $O(6k\times m)$；算法第 13－17 行生成时间熵最大的候选集 C_4，其时间复杂度为 $O(4k\times m)$；算法第 18－23 行考虑位置语义，候选集 C_2 由 $2k$ 个满足位置语义的位置点构成，其时间复杂度为 $O(4k)$；算法第 24－29 行生成距离熵最小的候选集与用户真实位置构成最佳假位置集 C，其时间复杂度为 $O(2k\times m)$；最后将最佳假位置集 C 进行缓存然后返回。

通过分析可知，DLG_SIA 算法的时间复杂度最大为 $O(n^2)$ 且能抵御背景信息推理攻击，说明 DLG_SIA 算法既能保证服务质量又能满足用户的位置隐私需求。

5 安全性分析

本文未采用集中式架构，这样做不仅降低了攻击者通过链路攻击获取用户位置隐私的风险，而且不存在第三方匿名服务器被攻破的威胁。DLG_SIA 算法综合考虑零查询特殊位置、历史查询概率、时间分布、位置语义及物理分散度，能够有效抵御拥有背景信息的攻击者和 MQRA 推理攻击。

定理 1 DLG_SIA 算法可以抵御 MQRA 攻击。

证明：LBS 服务器作为攻击者获取到用户在同一位置多次请求的假位置集 $C_{\text{list}}=\{C_1,C_2,\cdots,C_n\}$，根据 MQRA 算法得到 $C_{\text{update}}=C_1\cap C_2\cap\cdots\cap C_n$。假设攻击者推理成功，即 $1\leqslant|C_{\text{update}}|<k$，则 C_{list} 中至少存在两个集合满足式(6)：

$$L_m\in C_i\wedge L_m\notin C_j,\exists(1\leqslant m\leqslant k,1\leqslant i\neq j\leqslant n) \tag{6}$$

由算法 2 第 1－3 行可知，当用户在同一位置第二次请求时直接以缓存假位置集发起请求，即攻击者获取到的 $C_{\text{list}}=\{C_1,C_2,\cdots,C_n\}$ 满足式(7)：

$$C_i=C_j,\forall(1\leqslant i\neq j\leqslant n) \tag{7}$$

式(6)和式(7)互相矛盾，则假设不成立，即 DLG_SIA 算法可以抵御 MQRA 攻击。

定理 2 DLG_SIA 算法可以抵御概率分布攻击。

证明：通过 DLG_SIA 算法生成的位置点 L_i 为用户真实位置的概率如式(8)所示：

$$P\{L_i\in L_u\mid L_u\in C\}=\frac{P\{L_i\in L_u\cap L_u\in C\}}{P\{L_u\in C\}}=\frac{p_{L_i}}{P\{L_u\in C\}} \tag{8}$$

同理可得，假位置集中的位置点 L_j 为用户真实位置的概率如式(9)所示：

$$P\{L_j\in L_u\mid L_u\in C\}=\frac{p_{L_j}}{P\{L_u\in C\}} \tag{9}$$

若攻击者需要成功推理出用户的真实位置，则位置点 L_i 和 L_j 必须满足式(10)：

$$p_{L_i}\neq p_{L_j},\forall(1\leqslant i\neq j\leqslant k) \tag{10}$$

由算法 2 第 7－12 行和第 13－17 行可知，算法通过位置熵和时间熵生成位置点，即使在攻击者获知了请求时间的情况下，假位置集中任意两个位置点仍不满足式(10)，故 DLG_SIA 算法可以有效抵御概率分布攻击。

定理 3 DLG_SIA 算法可以抵御位置语义攻击。

证明：假设攻击者试图通过假位置集 $C=\{L_1,L_2,\cdots,L_k\}$ 中各位置点的语义信息来推理用户真实位置的语义信息，则需要判断假位置集中任意两个位置点 L_i 和 L_j 是否满足式(11)：

$$Sem(L_i,L_j)<\theta,\forall(1\leqslant i\neq j\leqslant k) \tag{11}$$

若满足则推理失败，否则推理成功。由算法 2 第 18－23 行可知，算法通过调整余弦相似度选取满足式(11)的 $4k$ 个位置点加入候选集，并选取语义相似度最小的前 $2k$ 个位置点进行下一步筛选。由此可知，攻击者推理失败，即 DLG_SIA 算法可以有效抵御位置语义攻击。

定理 4 DLG_SIA 算法可以抵御位置同质攻击。

证明：假设攻击者试图通过假位置集 $C=\{L_1,L_2,\cdots,L_k\}$ 构成的匿名区域 CR 来判断用户真实位置所在区域，若 CR 过小则推理成功，否则推理失败；由算法 2 第 24－29 行可知，算法通过定义 5 来选取距离熵最小的 $k-1$ 个位置点与用户真实位置构成假位置集 C 输出。由式(4)可知，距离熵越小，则 $d(C_i,C_{center})$ 越大，假位置集构成的匿名区域 CR 就越大，因此攻击者推理失败，表明 DLG_SIA 算法可以抵御位置同质攻击。

非理想情况下，拥有背景信息攻击者的推理能力增强为 $E(X)=1/(k-k')$，其中 k' 为攻击者通过背景信息识别出的位置点个数。通过上述分析可知，本文算法综合多方面因素，k' 将会小于其他方法，即降低了攻击者的推理能力，因此本文方法具有更好的隐私保护效果。

6 实验与分析

6.1 实验设置

为验证本文算法的有效性，实验采用 Geolife[24] 真实数据集，该数据集是微软研究院记录的轨迹数据，包含 182 名用户 5 年内的 17 621 条移动轨迹。实验将样本空间划分为 100×100 的矩形网格，每个网格用其中心点的坐标表示。实验将位置语义划分为 4 类：餐饮娱乐、教育科学、行政办公和医疗救护。用户隐私偏好 k 设为 2～20，语义阈值 θ 设为 0，对每一个 k 值重复实验 1 000 次取其平均值。

实验以 IDEA 作为开发平台，采用 JAVA 语言编程实现。硬件环境为：Intel(R) Core(TM) i5-7300HQ CPU @ 2.50 GHz，内存为 8.00 GB，操作系统为 Windows 10。

6.2 评价指标

1)位置熵

通常在位置隐私保护方法中，位置熵用于度量用户真实位置的不确定性，熵值越大表明假位置集 C 中各位置单元格的历史查询概率越相似，攻击者识别出用户真实位置的概率就越小，即假位置集的隐私保护程度越高。位置熵的计算如式(12)所示：

$$HP(C)=-\sum_{i=1}^{k}lbp_{L_i}\times p_{L_i} \tag{12}$$

其中，p_{L_i} 为位置单元 i 的查询概率，lb 为以 2 为底的对数。

2)时间熵

时间熵用于衡量在当前请求时间 t 下，假位置集中各位置点的访问概率是否相似，时间熵越大表明当前时刻攻击者通过请求时间识别出用户真实位置的可能性就越小。时间熵的计算如式(13)所示：

$$HT(C)=-\sum_{i=1}^{k}lbp_{L_{it}}\times p_{L_{it}} \tag{13}$$

其中，$p_{L_{it}}$ 表示位置单元 i 在 t 时刻的查询概率，lb 为以 2 为底的对数。

3)语义差异性

最终生成的假位置集中各位置点之间需要满足语义差异性，该值用来度量本文算法抵御位置语义攻击的能力，语义差异性越大表明抵御位置语义攻击的效果越好。语义差异性的计算如式(14)所示：

$$SD=1-\frac{|Sem>\theta|}{C_k^2} \tag{14}$$

其中，$|Sem>\theta|$ 表示 k 假位置集中任意两个位置点之间不满足用户语义偏好的个数，C_k^2 为组合运算公式。

4)物理分散度

物理分散度用于衡量假位置集中位置点之间的离散程度，本文采用距离熵来度量假位置集中各位置点的物理分散

度。距离熵越小表明假位置集中各位置点之间越离散，构成的匿名区域越大。

5)平均匿名时间

通常，为提高隐私保护效果会牺牲一定程度的服务质量，但服务质量又是影响用户体验的直观因素。本文采用平均匿名时间来度量用户在使用 LBS 时的服务质量，在保证服务质量和隐私保护效果的前提下，平均匿名时间越短越好。

6)攻击者的推理能力

由于攻击者掌握着一定的背景信息，因此本文采用非理想情况下的 $E(X)$ 来刻画攻击者的推理能力，$E(X)$ 值越大，表明攻击者的推理能力越强，即成功识别用户真实位置的概率越大。

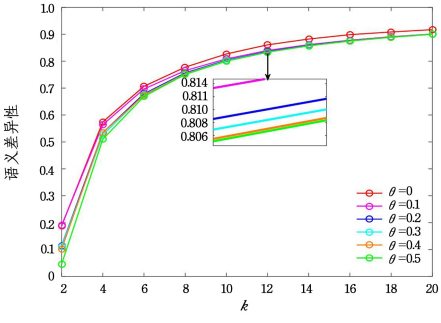
6.3 实验结果与分析

实验从假位置集的位置熵、时间熵、语义差异性、物理分散度、平均匿名时间及攻击者的推理能力 6 个方面验证 DLG_SIA 算法的有效性，并将其与同类假位置生成算法 DLS 算法[16]、enhanced-DLS 算法[16]、DLP 算法[17]、VLBS 算法[21]及 MMDS[20]算法进行对比，最后验证了本文算法抵御 MQRA 算法的有效性。

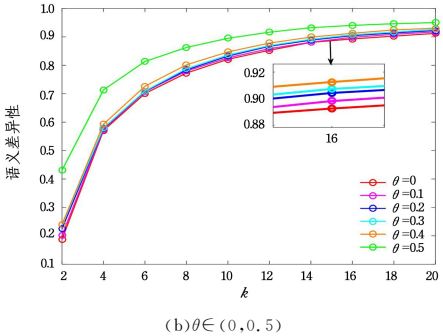
1)语义阈值 θ 对语义差异性的影响

用户可以自定义语义阈值 θ，该值是判断语义差异性的重要参数，语义差异性直接反映了生成的假位置集抵御位置语义攻击的效果，为确定一个最佳语义阈值，本实验针对不同语义阈值对语义差异性的影响进行分析。

用户自定义的语义阈值 θ 理论上的取值范围为[−1,1]，本文分析 θ 在[−0.5,0.5]时对语义差异性的影响。从图 4 中可以看出，本文算法在 k 值不变的情况下，随着语义阈值 θ 的减小，语义差异性逐渐减小。当语义阈值 θ 设置过小时可能会无法抵御位置语义攻击，语义阈值 θ 设置过大时可能无法找到相应的位置点。为在二者之间取得一个权衡，本文实验设置语义阈值 θ 为 0。



(a)$\theta\in(-0.5,0)$



(b)$\theta\in(0,0.5)$

图 4 不同语义阈值对语义差异性的影响

Fig. 4 Semantic differences at different semantic thresholds

2)位置熵

图 5 表示本文假位置生成算法 DLG_SIA 和 DLS 算法、DLP 算法、enhanced-DLS 算法、VLBS 算法、MMDS 算法及 random 算法在不同 k 值下生成的假位置集位置熵对比结果。

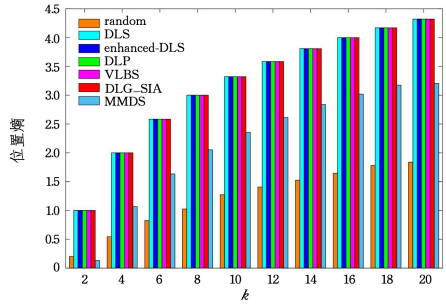


图 5 不同 k 值下的位置熵

Fig. 5 Location entropy at different k values

通过实验对比可以看出，本文 DLG_SIA 算法和 DLS 算法、DLP 算法、enhanced-DLS 算法及 VLBS 算法均首先考虑历史查询概率，因此各算法位置熵在不同 k 值下基本接近理想情况 lbk；随机算法因未考虑历史查询概率故位置熵在不同 k 值下都是最小值；MMDS 算法的位置熵低于其他算法是因为该算法首先筛除了语义相近的位置点，然后才通过查询概率进行选取。本文 DLG_SIA 算法生成假位置集的位置熵取得了最优结果，从历史查询概率的角度来看，本文算法能够很好地抵御概率分布攻击。

3)时间熵

在本文中，时间熵用来度量用户在不同时间下的历史查询概率，熵值越大表示假位置集的隐私保护度越高。图 6 为不同 k 值时各个算法的时间熵，假位置集的时间熵最佳可以达到 lbk。

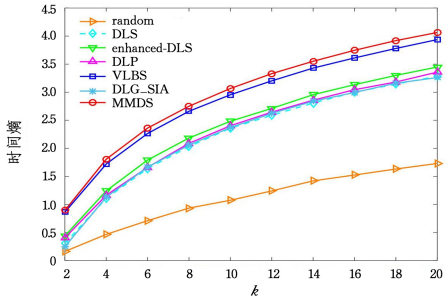


图 6 不同 k 值下的时间熵

Fig. 6 Time entropy at different k values

从图 6 可以看出，随着 k 值的增大，7 种算法的时间熵都呈上升趋势，相较于 VLBS 算法，本文算法在时间熵上平均提升了 3.46%且远大于其他算法。这是因为 DLG_SIA 算法和 VLBS 都以不同时间段用户对同一位置单元的访问量来选取假位置。VLBS 算法略低于本文算法是因为其先选取物理分散度大的位置点，然后才通过不同时间的访问量量化位置语义来间接选取位置点。由此可见，DLG_SIA 算法生成的假位置集不仅能够抵御概率分布攻击，而且综合了时间因素，隐私保护能力得到提高。

4)语义差异性

实验通过语义差异性来度量假位置集的语义丰富程度，语义差异性越大则攻击者越难通过位置语义信息来推理用户

所处位置的语义信息。

从图 7 中可以看出，MMDS 算法的语义差异性均在 0.9 之上，DLG_SIA 算法低于 MMDS 算法是因为 MMDS 算法首先保证选出的位置点满足语义差异性。相较于 VLBS 算法，DLG_SIA 算法在语义差异性上平均提高了 12.51%，这是因为 VLBS 算法只通过时间来量化位置语义，并未考虑位置语义类型对语义差异性的影响，其语义差异性均处于 0.8 以下且低于本文算法。DLS，DLP 及 enhanced-DLS 低于其他算法是因为三者都未涉及位置语义，但 enhanced-DLS 略高于 DLS 和 DLP 是因为后两者没有考虑距离因素，由此可以看出语义差异性和物理分散度呈正相关，也验证了针对位置语义攻击，本文算法具有良好的抵御效果。

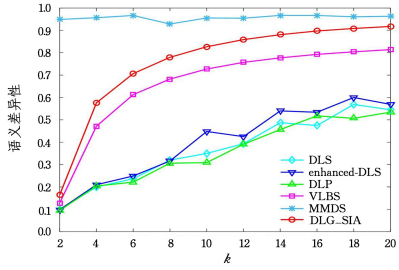


图 7　不同 k 值下的语义差异性

Fig. 7　Semantic differences at different k values

5）物理分散度

实验采用物理分散度来度量匿名面积的大小，物理分散度越大，匿名面积就越大。本文通过统计不同 k 值下的距离熵来度量物理分散度。图 8 是不同 k 值时各种算法的距离熵，图 9 是不同 k 值时各种算法生成假位置集的平均匿名范围。

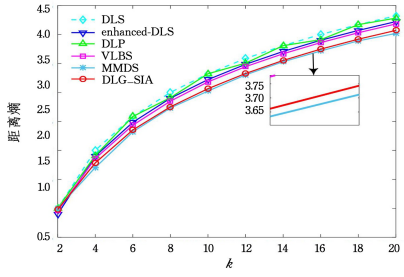


图 8　不同 k 值下的距离熵

Fig. 8 Distance entropy at different k values

由图 8 可知，随着 k 值的不断增大，距离熵整体呈上升趋势，MMDS 算法较本文算法略降低了 1.37%，但相较于其他算法，DLG_SIA 算法平均降低了 4.32%。理论上距离熵越小则假位置集的覆盖范围越大，通过图 9 可以看出理论分析与实验结果相互印证。由于 MMDS 算法首先保证位置点的语义差异性，即保证了位置点间的距离，因此其距离熵小于 DLG_SIA 算法。

由图 9 可以看出，MMDS 算法的匿名范围最大。此外，在 k 值相同的情况下，DLG_SIA 算法的平均匿名范围较 VLBS 算法和 enhanced-DLS 平均提升了 2.19% 和 7.07%，其原因是，DLG_SIA 算法在生成假位置时将距离熵作为选取标准，只选取距离熵最小的候选集作为最终假位置集，保证了最终假位置集具有最大的物理分散度。由此可见，DLG_SIA 算法在抵御位置同质攻击方面具有较强的能力。

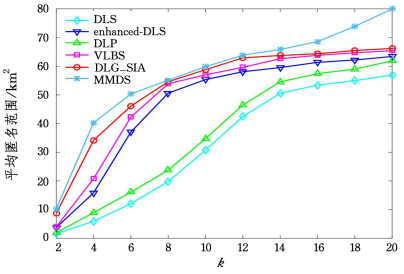


图 9　不同 k 值下的平均匿名范围

Fig. 9　Average anonymity range at different k values

6）平均匿名时间

本文采用平均匿名时间来度量算法性能，为验证 DLG_SIA 算法生成假位置集的效率，在不同 k 值下对几种算法进行实验对比，实验结果如图 10 所示。

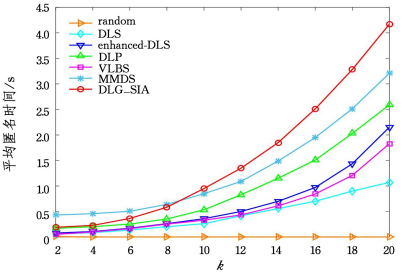


图 10　不同 k 值下的平均匿名时间

Fig. 10　Average anonymity time at different k values

从实验结果可知，随着 k 值的增大，各算法的平均匿名时间也逐渐增加。各算法的平均匿名时间几乎始终保持小于 DLG_SIA 的趋势。原因是，DLS 算法未考虑物理分散度；VLBS 算法在奇数轮和偶数轮根据不同权重选择假位置；enhanced-DLS 算法需在每一轮中计算距离乘积；DLP 算法针对 enhanced-DLS 算法的缺陷进行多条件判断，增加了算法的复杂性。相较于前几种算法，DLG_SIA 算法不仅考虑了零查询位置、时间分布、物理分散度，还探究了位置语义，故其平均匿名时间略高。MMDS 的平均匿名时间大于除 DLG_SIA 外的其他算法是因为其每次选取时都需要遍历整个数据集，筛除不满足语义阈值的位置点。虽然本文 DLG_SIA 算法的平均匿名时间大于其他算法，但其在用户可接受的范围内具有更高的位置隐私保护效果。

7）攻击者的推理能力

攻击者的推理能力越强，用户隐私泄露的概率就越大，本文通过不同 k 值下的 $E(X)$ 来度量攻击者的推理能力，实验结果如图 11 所示。

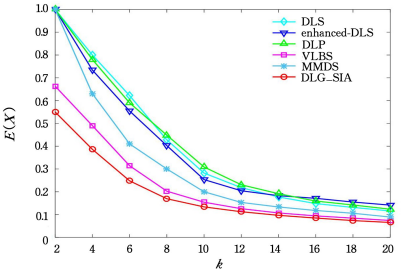


图 11　不同 k 值下的 $E(X)$

Fig. 11　$E(X)$ at different k values

通过实验对比可知，随着 k 值的不断增大，$E(X)$ 逐渐

减小，这是因为 k 值的增加使得 $P(X=1)=\frac{1}{k-k'}$ 逐渐减小，即攻击者的推理能力下降，这符合客观分析。针对不同的保护算法，在相同 k 值下，其他算法的攻击者推理能力较高是因为：DLS 算法只考虑了概率分布攻击；enhanced-DLS 算法未考虑位置语义攻击；DLP 算法仅比 DLS 算法有更多的条件判断。相较于 VLBS 算法，本文算法平均降低 12.41%是因为 VLBS 算法未结合实际语义类型；相较于 MMDS 算法，本文算法平均降低 39.27%是因为 MMDS 算法未考虑时间因素。由此可见，本文算法针对概率分布攻击、位置语义攻击以及位置同质攻击具有更好的隐私保护效果。

8）MQRA 识别率

前面以 DLS 算法为例说明本文所对比的算法不能抵御 MQRA 算法攻击。为验证 DLG_SIA 算法抵御 MQRA 攻击算法的有效性，本文在不同 k 值下进行实验对比，实验结果如图 12 所示。

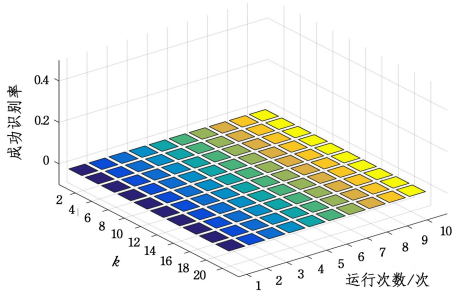


图 12 MQRA 成功识别率

Fig. 12 MQRA success recognition rate

由图 12 可以看出，在不同 k 值时，不论用户在同一位置请求几次服务，即使攻击者获取到每次请求生成的假位置集后也无法直接识别出用户的真实位置。这是因为 DLG_SIA 算法将用户首次请求进行缓存，同一位置的下次请求直接命中缓存，攻击者通过获取的假位置集识别出用户真实位置的概率始终为 $1/k$，即 DLG_SIA 算法针对此攻击具有很好的隐私保护效果。

结束语 针对已有的假位置生成算法，本文提出了一种攻击算法使得攻击者猜测用户真实位置的概率大于 $1/k$。针对攻击者掌握历史查询概率、位置语义及地图信息等背景信息时存在的隐私泄露问题，本文综合查询概率、时间分布、位置语义和物理分散度，提出了一种抵御背景信息推理攻击的假位置生成算法。该算法在抵御所提攻击算法的基础上不仅能够生成满足用户位置隐私需求的假位置集，而且能够有效抵御概率分布攻击、位置语义攻击以及位置同质攻击。同时，安全性分析和仿真实验结果进一步证明了所提算法的有效性和可行性。所提算法主要是针对快照查询下的位置隐私保护，下一步将研究实际路网环境中的位置隐私保护方案。

参 考 文 献

[1] ZHANG X J, GUI X L, WU Z D. Overview of Research on Privacy Protection of Location Services[J]. Journal of Software, 2015, 26(9): 2373-2395.

[2] YAN G H, LIU T, ZHANG X J, et al. Service Similarity Location k Anonymous Privacy Protection Scheme against Background Knowledge Inference Attack[J]. Journal of Xi'an Jiaotong University, 2020, 54(1): 8-18.

[3] 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.

[4] ZHANG X J, GUI X L, JIANG J H. User Centered Privacy Protection Method for Differential Disturbance Location[J]. Journal of Xi'an Jiaotong University, 2016, 50(12): 79-86.

[5] ZHANG X J, YANG H Y, LI Z, et al. Differentially Private Location Privacy-preserving Scheme with Semantic Location[J]. Journal of Computer Science, 2022, 48(2): 147-155.

[6] HUANG G, DENG K, XIE Z, et al. Intelligent Pseudo-location Recommendation for Protecting Personal Location Privacy[J]. Concurrency and Computation: Practice and Experience, 2020, 32(2): 5435-5446.

[7] ZHU X Y, CHI H T, NIU B, et al. MobiCache: When k-anonymity meets cache[C]// 2013 IEEE Global Communications Conference(GLOBECOM). IEEE, 2013: 820-825.

[8] 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.

[9] CHEN S, SHEN H. Semantic-Aware Dummy Selection for Location Privacy Preservation[C]// 2016 IEEE Trust-com/BigDataSE/ISPA. IEEE, 2016: 752-759.

[10] 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.

[11] DEWRI R, THURIMELLA R. Exploiting Service Similarity for Privacy in Location-Based Search Queries[J]. IEEE Transactions on Parallel and Distributed Systems, 2014, 25(2): 374-383.

[12] ZHENG L J, YUE H H, LI Z X, et al. k-Anonymity Location Privacy Algorithm Based on Clustering[J]. IEEE Access, 2018, 6: 28328-28338.

[13] ZHAO Z M, HU H D, ZHANG F, et al. K-anonymous Location Privacy Protection Method Based on Circular Region Partition [J]. Journal of Beijing Jiaotong University, 2013, 37(5): 13-18.

[14] WANG J, LI Y, YANG D, et al. Achieving Effective k-Anonymity for Query Privacy in Location-Based Services[J]. IEEE Access, 2017, 5: 24580-24592.

[15] GEDIK B, LIU L. Protecting Location Privacy with Personalized k-Anonymity: Architecture and Algorithms[J]. IEEE Transactions on Mobile Computing, 2008, 7(1): 1-18.

[16] NIU B, LI Q, ZHU X, et al. Achieving k-anonymity in Privacy-aware Location-based Services[C]// IEEE INFOCOM 2014-IEEE Conference on Computer Communications. NJ: IEEE, 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, 2016, 89(7): 3-13.

[18] DU Y, CAI G, ZHANG X, et al. An Efficient Dummy-Based Location Privacy-Preserving Scheme for Internet of Things Ser-

vices[J]. Information(Switzerland),2019,10(9):1-15.

[19] XIA X Y,BAI Z H,LI J,et al. A Location Cloaking Algorithm Based on Dummy and Stackelberg Game[J]. Journal of Computer Science,2019,442(10):92-108.

[20] WANG J,WANG C R,MA J F,et al. Dummy Location Selection Algorithm Based on Location Semantics and Query Probability[J]. Journal of Communication,2020,41(3):53-61.

[21] SHI X J,ZHANG J R,GONG Y. A Dummy Location Generation Algorithm Based on the Semantic Quantification of Location [C]// 2021 IEEE International Conference on Artificial Intelligence and Computer Applications(ICAICA),NJ:IEEE,2021:172-176.

[22] WANG S,LI F H,NIU B,et al. Research Progress on Location Privacy-preserving Techniques[J]. Journal of Communication,2016,37(12):124-141.

[23] 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):3-5.

[24] ZHENG Y,ZHANG R,XIE X,et al. GeoLife:Managing and Understanding Your Past Life over Maps[C]// The Ninth International Conference on Mobile Data Management. IEEE,2008:211-212.

ZHANG Xuejun, born in 1977, Ph. D, professor, is a senior member of China Computer Federation. His main research interests include edge computing, differential privacy, network security, data privacy and machine learning, etc.